Web Topic 1.1
Animal Communication and Science Education

Introduction

The authors have been teaching a course in animal communication since 1970, either at Cornell University or the University of California, San Diego (UCSD). One comment frequently made in student evaluations is “I understood and learned more physics in this one term biology course than I learned in a year of regular college physics.” In the 1980s and 1990s, UCSD had over 4000 biology majors and only a tenth as many physics majors. Despite the fact that most of these biology majors had to take college physics, the standard courses largely focused on classical topics of greatest interest to physics majors (and physics faculty). Given the recurrent comments made by the 200–300 students taking animal communication each year, we asked several physics faculty whether they were interested in integrating more animal communication topics into their courses. For various reasons, they were not.

The last decade has seen an enormous effort worldwide to improve STEM (science, technology, engineering, and mathematics) education. One common theme is better integration of the different STEM disciplines in K–12 education. In the USA, central online clearinghouses have been created to promote and distribute innovative and proven STEM curricula and teaching plans. Examples include the websites of the National Science Digital Library (NSDL, http://nsdl.org/) and the National Science Teachers Association (NSTA, http://www.nsta.org/). The Biosciences Education Network (BEN, http://www.biosciednet.org/portal/) focuses on biology topics, but makes a major effort to integrate other science and math disciplines into its curricula. Despite this broader approach, few programs have sought to exploit animal behavior generally or animal communication specifically as entry points to other science disciplines. A notable exception is the website of the American Biology Teacher, of the National Association of Biology Teachers (NABT, http://www.nabt.org/websites/institution/index.php?p=30), which has published theme issues on animal behavior, including bioacoustics lessons suitable for high school and college courses. Some other sites that are likely to develop biology/physics interface modules include Merlot (http://biology.merlot.org/), Bioquest (http://bioquest.org/), Ecological Society of America (http://tiee.esa.org/vol/toc_all.html), and ABLE (http://www.ableweb.org/proceedings/index.php).

Animal behavior as an educational springboard

A major difference between plants and animals is that animals overtly behave. This behavior typically takes the form of movements or the emission of signals, or both. Few anatomical or biochemical adaptations in animals are effective without some coupled behavior that invokes their use. As a result, behavior is now recognized as a major factor in the biology of any animal, and, although the study of animal behavior is occasionally claimed as a subfield of ecology, psychology, neurobiology, or physiology, the study of animal behavior is now a separate discipline. It has its own highly subscribed journals, academic departments, and international societies, including the Animal Behavior Society (http://animalbehaviorsociety.org/), the Association for the Study of Animal Behavior (http://asab.nottingham.ac.uk/), and the International Society for Behavioral Ecology (http://www.behavecol.com/pages/society/welcome.html).

Animal behavior is intrinsically an interdisciplinary science. Many of the components studied singly by other disciplines come together in the study of behavior. For example, why different species adopt different behaviors is closely tied to their differing ecologies (Wilson 1975). The kinematics of movement (Alexander 2002), the design and mechanisms for producing and detecting signals (this book), and the energetics of behavior (McNab 2002) are just some of the many facets of animal behavior that are closely tied to basic principles of physics. The physiologies of muscles, brains, digestion, reproduction, immune systems, hormonal controls, and even aging are tightly linked to the behaviors animals perform (Alcock 2009). Animal behavior studies have proved to be superlative testing grounds for modern theories of economics and decision-making (Houston and McNamara 1999; Maynard Smith 1982), and animal models are widely used to help understand the origins of learning and culture in our own species (DeWaal and Tyack 2003; Dugatkin 2009; McGrew 2004). Finally, behavior is now seen as a critical component of any conservation or wildlife management...
The interdisciplinary nature of animal behavior makes the field superbly adapted for both novel science and for stimulating science education. The enormous diversity of behaviors performed by different species when faced with similar challenges allows scientists to examine the relative roles of ecological function, physical or chemical constraints, physiological mechanisms, economic optimality, and cultural adaptation. Scientists can also examine the conservation repercussions of behaviors by undertaking experimental manipulations or by comparing the different solutions achieved by different species, identifying correlates of either convergence or divergence, and then testing emergent hypotheses by examining additional species.

Educational opportunities arise because humans, especially children, are naturally drawn to the behavior of animals. Television programs on nature are extremely popular with both children and adults. Ecotourism is similarly popular, and often focuses on the behavior of charismatic fauna. The universal intrigue of behaving animals creates a potentially powerful entry point for teaching physics, chemistry, physiology, economics, culture, and conservation. However, several problems have hindered this approach. First, many current teachers were never exposed to courses in ecology or animal behavior in their own education. Many do not even know that animal behavior is a legitimate field of science. Thus, they do not have the background to bridge animal behavior to physics, chemistry, or even other areas of biology. Second, those that do have the background and inclination to use this approach are overwhelmed by the amount and diversity of rich media resources on the Internet. Where computer tools might help bridge the disciplines, which ones should they use? Finally, the current emphasis on standardized testing in the United States severely constrains which principles of physics, chemistry, or physiology are to be taught at each grade level. Many behaviors are interesting to students, but to bridge the disciplines effectively, a teacher must find an example that is both interesting to students but also one that leads to the teaching of a current standard. Given that few teachers have time to develop any new curricula on their own, these challenges can be insurmountable.

A major motivation for the establishment of NSDL, NSTA, NABT, and similar programs was to resolve the problems noted above. This meant bringing scientists and educators together to design new curricula that integrated disciplines, selected suitable rich media and experiential (e.g., lab and field) exercises, and aligned content to state standards. The resulting curricula are increasingly used nationwide and even worldwide. Many “hooks” for interesting younger students in physics, chemistry, and mathematics have been devised in these curricula. However, only a few have focused on animal behavior and animal communication. We describe two of these below. There thus remains much unexploited potential for this type of educational bridging.

**Innovative curricula starting with animal communication**

Several recent efforts have focused specifically on linking animal communication behavior with other disciplines, such as physics. We hope that these examples will inspire future curriculum efforts at all levels.

**The Macaulay Library Project**

The Macaulay Library (http://macaulaylibrary.org/index.do) at the Cornell Lab of Ornithology is the world’s largest archive of animal sounds, with a growing parallel library of videos. With funding from NSDL and the National Science Foundation (USA), the Macaulay Library undertook an ambitious project to integrate its rich media collection with students’ natural interests in animal behavior and communication and with the teaching of basic physics. Scientists trained in animal communication and K–12 science teachers worked together to identify topics required by state standards, identify the best animal examples in the vast library, and develop lessons and hands-on exercises that demonstrated the focal physics principles. The project was performed in collaboration with the Center for Nanoscale Systems Institute for Physics Teachers (http://www.cns.cornell.edu/cipt/) and the New York Wayne-Finger Lakes Board of Cooperative Educational Services (http://wflboces.org/). The former collaboration focused on modules for high school physics classes, whereas the latter examined opportunities at all ages in K–12 education. The resulting animal communication modules focus on the physics of sound (using animal acoustic signals) and the physics of light (using the generation of colors in bird plumages). Additional modules discuss aerodynamics (by examining bird flight) and the
Physics of forces (by examining bird beaks). These lessons and associated media are available at the Library’s “Physics of Animal Behavior” website (http://macaulaylibrary.org/physics). Further bioacoustics lessons for college and AP high school biology can be obtained from the Online Research in Biology project website (http://birds.cornell.edu/orb), an NSF-funded educational effort from the education program at the Cornell Lab of Ornithology and the Macaulay Library.

The Sea of Sound Project

A project from the Cornell Lab of Ornithology, funded by the National Science Foundation and the National Oceanographic Partnership Program (http://www.nopp.org/), uses sound in the oceans as the “hook” topic to teach students a variety of state-standard principles in physics and biology. It is aimed at grades 6–12. The sound signals used by whales and dolphins figure prominently, but the curriculum also includes sound production by marine invertebrates and discussion of anthropogenic noise and its effects on communication of marine animals. These curriculum resources feature high-definition video footage of marine organisms and other multimedia segments that explore underwater communication and how it is impacted by sound from human-created activities, such as shipping and oil exploration. They also highlight the right whale monitoring efforts led by scientists from the Bioacoustics Research Program at the Cornell Lab of Ornithology (http://listenforwhales.org). Activities include everything from role-playing debates on the use of sonar to examining whale sounds recorded by multiple underwater buoys to calculate the speed of sound in salt water. Tables in the educator materials provide an at-a-glance overview of alignment between specific elements of the curriculum and the National Science Education Standards for middle school and high school. These classroom activities can be found at the Sea of Sound website (http://seaofsound.org/).

Literature Cited


Web Topic 1.2
Information and Communication

Introduction
Throughout this book, the term information is used repeatedly. In recent years, a number of authors have questioned the utility and even the propriety of invoking information concepts in studies of animal communication. Here we summarize some of these concerns, explain why we feel they are unnecessary or unsupported by recent studies, and indicate the specific steps during communication where we feel information concepts play an essential role.

The concerns
Below we list the problems that are most often raised about the use of the term “information” in studies of animal communication, ranging from doubts that the term has been sufficiently defined and consistently applied to proposals for models of animal communication in which the provision of information is irrelevant.

The duality problem
Following a lead by Cherry (Cherry 1966), W. John Smith proposed a distinction between the message and the meaning of animal signals (Smith 1968, 1977). He defined the message as “each kind of information that a display makes available about a referent.” The provision of information was dependent on the assignment of specific signals to specific referents using a code that was shared by the sender and the receiver. He then proposed two aspects of meaning. One arose from the consequences to receivers of adjusting subsequent actions on the basis of received signals, and another focused on the consequences to senders of these changes in receiver actions. In this view, without a message, there could be no meaningful change in consequences for either party, and without meaning, there was no point for either party to communicate.

While this is intuitively appealing, given the parallels with human speech, the application of these two terms (message and meaning) has proved challenging. The codes mapping animal signals on referents are invariably imperfect (e.g., different referents may elicit the same signal, and different signals may be given for the same referent), which makes identification of the message by both receivers and researchers a quantitative rather than a qualitative task. Assuming receivers accurately identify an incoming signal, subsequent authors have claimed the meaning to be the inferred referent given the code, the appropriate receiver action when that referent is present, or the likely fitness effects of taking the appropriate action. There is thus considerable confusion in the literature as to whether one of these stages or some combination is the appropriate sense of receiver meaning. There is the further complication that different receivers often respond differently and experience different consequences for the same signal, and different senders of the same signal might experience different consequences of whatever actions receivers perform. The seemingly indeterminate links between signals and referents, and between messages and meanings have undermined support for the utility of these terms, at least as originally defined.

Rather than discard the entire duality, some authors have championed one of the original components while minimizing emphasis on the other. The development of information theory (Shannon and Weaver 1949) provided new tools for measuring the amount of information provided by a signal relative to that needed to completely resolve some prior uncertainty. These computations explicitly incorporated the coding system of the sender. The emphasis here was thus the message; consequences were not considered. For over a decade, these tools were enthusiastically applied to a wide variety of animal signal systems (Quastler 1958; Atteave 1959; Johnson 1970; Dingle 1972; Wilson 1975; Hailman 1977; Bell and Gorton 1978; Losey 1978). However, the same concerns that had arisen over the initial duality were raised again: if different receivers had reasons to invoke different prior uncertainties, each would receive different amounts of information for the same signal. And even if most receivers obtained the same amount of information from a signal, the fitness consequences could differ markedly between them. The indeterminate relationships between a given coding system and the amount of information (due to variable priors), and the fact that fitness consequences, not
information, are the ultimate focus of evolution, led other authors to adopt the opposite extreme and argue that information provision should not be a part of any definition of communication (Rendall et al. 2009; Owren et al. 2010). The philosopher Scott-Phillips (2008, 2010) even argued that the provision of information was at best “incidental” to the basic communication process.

Other philosophers and economists have taken a totally opposite view. D. K. Lewis (1969) used classical game theory to examine signaling games in which coding was totally conventional and both signaling and receiving were costless (often called “cheap talk” in the philosophical and economic literature). He found that perfect codes shared by senders and receivers were most likely to lead to a stable outcome. Concerns about how a population might arrive at such an equilibrium were resolved by invoking evolutionary game theory and adaptive dynamics models (Skyrms 1996, 2002; Huttegger 2007a). These analyses again found that perfect or, at worst, moderately imperfect signaling was the only likely ESS. The Lewis model has been extended to a wide variety of contexts, including coding with more than binary alternatives, finite populations, senders in one species signaling to receivers in another, presence or absence of mutation, and biologically relevant signaling (e.g., the Sir Phillip Sydney game) (Huttegger 2007a, b; Pawlowitsch 2007; Hofbauer and Huttegger 2008; Skyrms 2009; Huttegger et al. 2010; Huttegger and Zollman 2010; Skyrms 2010a, b). The outcome of these models is nearly always the same: reliable coding is not only essential to communication, it is the only stable outcome.

An alternative approach replaced the original notion of message with signal reliability (Maynard Smith and Harper 2003; Searcy and Nowicki 2005) and the original notion of meaning with the value of information (Gould 1974; Stephens 1989; Bradbury and Vehrencamp 2000; Koops 2004; Dall et al. 2005; McLinn and Stephens 2006, 2010). The value of information integrates signal reliability with receiver decoding, decision making, and fitness consequences into one number that is subject to selection. We outline an example of its use in more detail at the end of this document. This and related arguments have led to a chorus of support for the continued inclusion of information concepts in studies of animal communication (Hasson 2000; Dall et al. 2005; Stegmann 2005, 2009; Castellano 2009; Carazo and Font 2010; Font and Carazo 2010; Seyfarth et al. 2010). In fact, information sharing is now recognized as one of the key adaptations that has led to major evolutionary changes throughout organismal history (Maynard Smith and Szathmáry 1995; Maynard Smith 1999; Lachmann et al. 2000; Maynard Smith 2000; Jablonka 2002).

**The arms race problem**

Early ethologists largely ignored the possibility of deceit in animal communication either because they assumed that neither party was capable of actions outside the norm (e.g., they were locked into fixed action patterns) or because they assumed communication was a cooperative venture. These views were challenged by Krebs and Dawkins who argued that senders should, and usually do, try to manipulate receivers to the sender’s advantage, and receivers should, and usually do, try to “read the minds” of senders to the receiver’s advantage (Dawkins and Krebs 1978; Krebs and Dawkins 1984). There is now little disagreement that senders and receivers can have conflicts of interest: the optimal interaction for one party may not be identical with that of the other. The question is then where the subsequent evolutionary trajectories will lead. One possibility is that the arms race between the two parties is unending over evolutionary time—each adaptation that gives one party an advantage will eventually be overcome by a counter-adaptation in the other party. This does not seem to be the case. As outlined in the text, most (but not all) communication systems in animals appear to be at some sort of equilibrium. There are three possible types of equilibria: one in which the sender is able to keep the system at its optimum at a net cost to the receiver; one in which the receiver is at its optimum at the expense of the sender; and one (or more) in which both parties have a net benefit, but neither does as well as it would at its own optimum.

Some authors have asserted that arms races are only likely to end when the sender acquires a strategy that the receiver cannot counter (Rendall et al. 2009; Owren et al. 2010). This might occur if the sender can mimic some stimulus that the receiver attends to for other reasons (sensory exploitation); any attempt by the receiver to escape the exploitation would undermine some other necessary adaptation. The equilibrium in these authors’ view is thus a sender dominant one. Since receivers cannot help but respond, the issue of information provision is irrelevant and these authors have argued that term be dropped from discussions of signal evolution.
The weight of evidence does not support these propositions. A number of the studies originally proposed in support of persistent sensory exploitation now appears better explained by recurrent loss and recovery of an early adaptation that benefitted both parties (see Ron 2008 for túngara frog; see Chapter 10 for other examples). In addition, there are numerous evolutionary models identifying realizable conditions under which receivers can eventually escape persistent exploitation. These models usually lead to an intermediate equilibrium in which both parties obtain an average net benefit that is contingent upon the provision of minimally reliable information to the receiver by the sender. The last few decades have seen a major effort to test these models in a wide variety of taxa and signaling contexts. As reviewed in the text, the conditions favoring intermediate equilibria are much more often found to be present than absent, and where fitnesses can be measured, both parties gain an average net benefit by communicating. While sender exploitation of receiver sensory biases remains one of several likely starting points for signal evolution, it appears that most systems subsequently move on to intermediate equilibria.

The coding problem

Traditional models of animal communication assume that senders and receivers have at least reasonably concordant, if not identical, coding schemes. Signals given randomly cannot provide information to receivers. Signals given selectively can provide information, but senders must show some consistency in which signal they emit for a given referent, and receivers must have some previously acquired expectations about likely sender assignments.

Several authors have questioned the use of the coding concept in animal communication. Some argue that the terms “encoding” and “code” are never explicitly defined, and are indiscriminately applied to very different signaling phenomena (Rendall et al. 2009; Owren et al. 2010). Other authors challenge the utility of the coding concept because the contexts in which a given signal is emitted may completely change the assigned referent implied: examples include human sarcasm (Scott-Phillips 2010) and context-dependent signaling in birds (Smith 1977).

In response, it is instructive to note that a reliance on coding schemes is not unique to communication, but instead is a condition for most sensory processing. The primary function of sensory organs is to detect changes in ambient conditions. Animals must be able to categorize conditions or there would be no benefit to sampling the environment. This is true even for the simplest case in which an animal only monitors the presence or absence of a single stimulus; it is even more relevant to the many organisms that routinely discriminate between multiple stimuli. The sampling and sorting of ambient stimuli, particularly those emitted by other organisms, is ubiquitous in animals (Danchin et al. 2004; Wagner and Danchin 2010). Stimuli vary in how well their emission is correlated with specific and unique conditions. Few stimuli of interest to animals will be “natural” in the sense that they are guaranteed to be available if and only if a given condition is true (Scarantino 2010). Many will be “normative,” in that the correlation between a given stimulus and a given condition is sufficiently high that it is worth attending to them (Millikan 1989, 2004; Stegmann 2009). Stimuli that are moderately reliable indicators of the presence of particular conditions are called cues. Whether acquired through genetic inheritance, learning, or some combination of the two, nearly all animals rely on interpretive codes to make sense of the many cues that they might encounter. They can then generalize these codes as necessary to deal with novel stimuli (Ghirlanda and Enquist 2003). The reception and classification of a cue stimulus, when combined with the interpretive code, provides information that can then be used to influence decisions about whether and how to change current actions and physiological states.

Seen in this light, signal codes are just a special case of a more general sensory strategy used by all animals. In fact, many receivers appear to rely on a combination of cues and signals to make decisions. The relative weighting of signals and contextual cues can be quite variable: some signals elicit the same responses regardless of contexts, whereas others elicit responses that are very sensitive to ambient conditions (Smith 1977; Marler et al. 1992). Reliable signals that are more heavily weighted than ambient cues are often singled out as “referential.” However, there appears to be a continuum rather than a dichotomy in most taxa, and it is not surprising that relative weightings should vary with the relative reliabilities of the stimuli and the fitness consequences of alternative actions.

Secondly, the claim that coding schemes are never explicitly defined in animal signal literature (Rendall et al. 2009; Owren et al. 2010) is inaccurate. Explicit and quantitative definitions of animal signal coding schemes are provided in
the prior and current editions of this text and in countless other publications (including a whole book, Hailman 2008) ignored by these authors. The critics may not agree with these definitions, but claiming they don't exist is spurious. While the same authors who claim coding schemes are never defined acknowledge that receiver actions may be sensitive to both cues and signals, they see the observed variability in relative weightings as undermining any formal definition of signal coding schemes. We would argue that once you acknowledge the overlap of cue and signal coding schemes, variable weighting is an adaptation that natural selection is sure to favor.

Finally, it seems appropriate to ask just how reliable animal signal codes really are. If the codes do not enhance decision making above random choice, any invocation of coding is moot. As discussed in the text, signal reliability depends on the consistency with which senders assign signals to referents, the level of signal distortion during propagation, and the degree to which receivers share the sender's coding scheme and can correctly assign incoming stimuli to expected templates. Because the minimal reliability that justifies communication depends significantly on the fitness consequences of receiver actions (Bradbury and Vehrencamp 2000; Koops 2004), one might expect observed values to be highly variable among taxa, modalities, and contexts. In fact, measures of signal reliability obtained in recent decades find most animals using intermediate levels of reliability: signals provide much better information than relying on prior probabilities alone, but signal coding is almost never perfect (see Chapter 8; Maynard Smith and Harper 2003; Dall et al. 2005; Searcy and Nowicki 2005; Seyfarth et al. 2010). In exceptional cases where signal reliability is found to be surprisingly low, the contexts are such that the value of information for those signals remains positive for all parties (Gyger and Marler 1988; Møller 1988).

**The black box problem**

The use of metaphorical models for animal and human behaviors has been a long tradition in psychology. It was also an early tool of ethologists who tried to explain phenomena such as vacuum and displacement behaviors (see Chapter 10) by postulating hypothetical “drives” whose dynamics and interactions could be adjusted to replicate the observed patterns. Perhaps the most famous of these was the hydraulic model proposed by Lorenz and Leyhausen (1973). However, as neurobiology became more sophisticated, one after another of these hypothetical constructs was found wanting (see Web Topic 10.4; Berridge 2004). In parallel, the enthusiasm for applications of information theory to animal behavior in the 1970’s waned as it became clear that receiver actions after receipt of a signal were not a good guide to underlying decision processes: did a receiver fail to respond to a stimulus because it could not discriminate it from some alternative (an amount of information issue), or because it did not pay to change its current behavior (a fitness consequences issue)? As a result, most ethologists, and those in the descendant field of behavioral ecology, began to eschew speculations about brain mechanisms and instead focus on the economics of animal behavior: what ecological factors caused one species to be polygynous but a related species to be monogamous, what payoffs justified being territorial in a given habitat, and what were the costs to a male of directing carotenoids into coloration instead of into immune function? Reviewers often chastised authors who treated the brain as anything except a black box and suppressed any speculations about the mechanisms behind assessment and decision making.

Luckily, recent advances in cognitive science and neurobiology have changed the situation completely. Clever signal detection theory paradigms now allow one to measure the amount of information and the value of information separately and non-invasively (see Web Topic 8.10). A multitude of neurobiological efforts now focus explicitly on elucidating how the brains of animals and people accomplish the tasks associated with communication. On the sender’s side, the neurobiology of *Drosophila* displays, frog and cricket calls, and passerine song acquisition and production are largely worked out. On the receiver’s side, significant advances have been made in our understanding of sensory processing, stimulus categorization, encoding and decoding, the storage of perceived valence, and decision making in a wide variety of taxa. In growing numbers, the relevant genes have been identified.

It is thus no longer taboo to ask whether animal receivers can use receipt of a given signal to perform a Bayesian update on stored probability estimates or instead invoke some sort of heuristic shortcut. One can now identify specific parts of a vertebrate or invertebrate brain that carry out individual stages in decision making (see Web Topic 8.7). The many efforts to understand this type of process in humans, where self-reporting can be used to confirm neurobiological models, are now being applied and tested in animals (see Chapter 8). Results so far confirm the basic model of
receiver updating and decision making outlined in this book. The steps outlined in this and other models can increasingly be tested at both the proximal (mechanistic) and ultimate (fitness consequence) levels. So far, results are supportive. The initial success of the basic Bayesian design has spawned second-generation models that can explain hierarchical processing (e.g., Yang and Shadlen 2007; Tennenbaum et al. 2011). Clearly, the black box has come a long way since the early days of hydraulics.

The math problem

The most effective way to view the interaction between the amount of information in a signaling system and the fitness consequences involves algebraic formulations (see below). It is a curious fact that nearly all of the publications arguing against the incorporation of information in definitions of communication rely on entirely verbal arguments, excluding algebra. Most do not even cite the many models defining the quantitative conditions that favor stable signaling equilibria, and, even if these appear in the reference list, the models themselves receive no serious attention in the associated texts (Scott-Phillips 2008; Rendall et al. 2009; Owren et al. 2010; Scott-Phillips 2010). While some of these authors are philosophers, for whom a persuasive verbal argument is the gold standard, the avoidance of any mathematics seems odd when so much effort has gone into deriving rigorous evolutionary models for communication.

While algebraic formulation does not guarantee that all terms will be included or clearly defined, it often makes deletions conspicuous and badly defined terms clearly unmeasurable. In contrast, it is very easy, as can be seen in several of the cited papers, to construct a plausible verbal argument that hides the omission of contrary citations and data. Verbal arguments also make it easy to claim that a critical term (such as information) is “poorly defined” or to recast an opposing argument with such hyperbole that it becomes an easily disproved straw man.

Perhaps the most pernicious aspect of verbal arguments is the perceived need to partition quantitatively varying phenomena into discrete categories. Much of the dissent over definitions of communication arises from one group finding a case that cannot be assigned to available discrete categories or is inappropriately assigned by another’s definition. The problem is that many of the phenomena associated with animal communication do not fit into tidy, discrete categories. Behaviors performed during physical conflicts can both provide information to an opponent and set up a tactical advantage. The relative importance of the two can vary continuously between fights, and even shift during the same fight by the same two animals. Is such a behavior a signal or a fighting tactic? Discrete categories simply cannot handle these cases. Many biological parameters of interest vary continuously—forcing them into discrete categories, though intellectually convenient, is thus artificial. Many discrete definitions for biological phenomena end up having multiple criteria. What should one do with cases that meet all but one of these criteria? Such cases are often the most instructive, and ignoring them is foolish. Slavish obedience to discrete definitions is a recurrent problem; it is much better to accept the existence of continua and mixtures. Since this is often hard to do verbally, it is best left to algebraic expressions.

An integrated model of animal communication

Below, we briefly summarize two similar algebraic treatments of animal communication that explicitly integrate information provision with fitness consequences. Several more recent formulations are also available, but these two set the scene and will suffice to make our point. One model was published by Bradbury and Vehrencamp (1998, 2000) and the other by Koops (2004). The two models share the following assumptions and components:

- **Basic question**: Both models address the question of if and when a receiver should incorporate signals into a decision about subsequent actions.

- **Basic format**: Both models invoke the value of information as the relevant criterion subject to selection. This measure compares the average fitness of a receiver when it incorporates a given set of signals in its decisions against when it does not incorporate them. Signal usage will only be favored if the value of information is positive.

- **Alternative comparisons**: The value of information can compare the use of a given signal set to any of a variety of alternatives. For example, one could compare the value of information if the animal were to switch from using one set of signals to using an alternative set in the same contexts. However, to answer the basic question listed
above, the two cited models focus on the value of information for receivers using signals when compared to those using some non-signal default strategy.

- **The default strategy**: Some prior discussions of signal evolution have naïvely assumed that animals without access to signals would resort to random decisions. That is probably never the case: heritable biases and prior personal experience will allow most animals to estimate both the likelihoods of alternative conditions being true and the relative fitness consequences of alternative actions given each alternative condition. Both models thus assume that receivers without signals will combine some estimate of prior probabilities with potential payoffs to identify an optimal default action. In the absence of signals or new cues, the receiver will always perform this default action. Since this is only a best guess, sometimes the default action will be the right thing to do, but at other times it will be the wrong action. Despite some errors, the default action is the best choice on average.

- **The role of reliability**: Most animal signal schemes are also imperfect. Thus, a receiver relying on a given code and sender signals will sometimes make the right choice of action and sometimes the wrong one. The difference between relying on signals versus using the default strategy is a quantitative shift in the relative frequencies of correct and erroneous decisions.

- **The costs of communication**: Acquiring the coding scheme and investing time and sensory organs in attending to signals will impose costs on receivers that are not experienced by receivers adopting the default strategy.

The way in which these components can be combined into the value of information is most easily seen in the Bradbury and Vehrencamp model. Here is a brief outline of that approach:

- **Alternative conditions**: Suppose a receiver is concerned about which of two alternative conditions, \( C_1 \) or \( C_2 \), is currently true (e.g., predator present versus predator absent).

- **Prior probabilities**: Based on recent experience, the receiver can expect \( C_1 \) to occur a fraction \( p \) of the time, and \( C_2 \) to occur \((1-p)\) of the time. Note that both probabilities are nonnegative by definition.

- **Fitness consequences**: The receiver can perform either of two actions: \( A_1 \) and \( A_2 \). Suppose \( A_1 \) is the correct choice (higher payoff) when \( C_1 \) is true, and \( A_2 \) is the correct choice when \( C_2 \) is true. Let \( \Delta W_1 \) be the difference in fitness payoffs for making the correct versus the wrong choice of action when \( C_1 \) is true, and \( \Delta W_2 \) be the equivalent difference between right and wrong decisions when \( C_2 \) is true. Note that both \( \Delta W_1 \) and \( \Delta W_2 \) are nonnegative by definition.

- **Costs of communication**: Let the absolute value of the costs of acquiring the coding scheme and attending to signals be \( K \). This number is also nonnegative, but will be subtracted from the value of information because it is a cost.

- **Reliabilities**: Suppose that when the receiver only uses its default strategy, it correctly identifies that \( C_1 \) is true a fraction \( \phi_1 \) of the time and makes the wrong identification \((1 - \phi_1)\) of the time; similarly, it correctly identifies that \( C_2 \) is true a fraction \( \phi_2 \) of the time and makes the wrong identification \((1 - \phi_2)\) of the time. When the receiver relies on signals, it is now correct in identifying \( C_1 \) as the current condition a fraction \( \phi_1' \) of the time, and is incorrect \((1 - \phi_1')\) of the time; when \( C_2 \) is true, it is now correct a fraction \( \phi_2' \) of the time, and is incorrect \((1 - \phi_2')\) of the time. As with consequences, it is simpler to consider the changes when signaling is used compared to when it is not. Thus we define \( \Delta \phi_1 = \phi_1' - \phi_1 \) and \( \Delta \phi_2 = \phi_2' - \phi_2 \).

- **Value of information**: After Bradbury and Vehrencamp 1998, the value of information for this simple 2-condition/2-signal/2-action case can then be computed to be

\[
VI = p \Delta \phi_1 \Delta W_1 + (1 - p) \Delta \phi_2 \Delta W_2 - K.
\]

\( VI \) must be \( > 0 \) if signals are to be favored over the default strategy by selection. When is this the case? Note that:

- The values of \( p \), \((1 - p)\), \( \Delta W_1 \), and \( \Delta W_2 \) are all positive and are unchanged in value whether a receiver uses signals or not.
- The only variables that are sensitive to the use of signals are the values of \( \Delta \phi_1 \) and \( \Delta \phi_2 \) and the cost \(-K\).
Because a receiver using the default strategy is always correct in its choice of action when one of the conditions is true, and always wrong when the other condition is true, the switch to using signals has to decrease reliability for one condition and increase it for the other. This means that one of $\Delta \phi_1$ or $\Delta \phi_2$ has to be negative and the other has to be positive.

Assuming neither $\Delta W_i = 0$, this means that one of the $p_i \Delta \phi_i \Delta W_i$ terms in the value of information expression is negative, and the other one is positive. For the overall value of information to be positive, the positive term on the right side of the expression has to be large enough to more than cancel out the two other terms which are negative.

The magnitude of the positive term on the right hand side of the expression depends on the magnitudes of the corresponding $p_i$, $\Delta \phi_i$, and $\Delta W_i$. We thus see that both the amount of information, as measured by the reliability of the signals, $\phi_i^\prime$, and the fitness consequences of making correct instead of wrong decisions, $\Delta W_i$, are important in determining whether it is worth relying on signals or not. The amount of information, as measured by reliability, is not just an “incidental” property: it is an essential property determining the selection for or against communication.

Note that as the reliabilities of the signal set increase, the positive $p_i \Delta \phi_i \Delta W_i$ term gets larger and the negative $p_i \Delta \phi_i \Delta W_i$ term gets closer to zero. Both effects increase the value of information.

If the reliabilities are not sufficiently high, the value of information will not be positive and the use of signals will not be favored by selection. Both models show that this effect alone predicts a minimum reliability that must be present before it pays for receivers to shift from the default strategy to attending to signals.

Because increasing reliabilities usually increases the costs of participating in communication ($-K$), the optimal reliability for the receiver will depend in part on the shape of the cost function. If it accelerates with increasing reliability, or is linear but improvements in reliability are asymptotic, the optimum reliability for receivers is predicted to be intermediate between default values and that providing perfect information.

Note that a similar expression for the value of information can be derived for senders. The optimal value for senders will depend on their own set of fitness consequences ($\Delta W_i$) and their own set of costs ($-K_s$). The optima for the two parties need not be the same.

**Relevance to earlier concerns**

**What is information?**

Information in these models is the change in a receiver’s estimated probabilities that a given condition is currently true. It is not a substance so it cannot “flow” from sender to receiver. The $\Delta \phi_i$ can be used as measures of the amount of information provided by signals. Note that the change will depend on the prior probabilities: receipt of a signal can create big change if the initial expectation was chance, but a small change if the signal only confirms the receiver’s strong prior bias. An average for a signal set can be obtained by discounting each $\Delta \phi_i$ by the probability that it will be used,

$$\overline{\Delta \phi} = p \Delta \phi_1 + (1 - p) \Delta \phi_2.$$

This can be scaled in various ways to make it more useful. The typical approach is to scale it relative to the maximum reliability. Since ratios can get very small or very large, log scales (bits) are often used.

**Restoring the duality**

While none of the terms in the computation of $VI$ fit the original definitions of message and meaning, it should be clear from the algebraic model that an improvement in either signal reliability or payoffs can trigger a shift from a default strategy to reliance on signals. While it is true that it is the net fitness payoff (value of information) that is the focus of selection, this payoff is equally dependent on how much information is provided and how much getting it right versus getting it wrong affects fitness. Quantifying reliabilities and fitness payoffs are thus equally important tasks when
examining signal economics.

Where is the code?

Reliability is a measure of the probability that a receiver correctly identifies the current condition given available cues and signals. To do that, it must combine receipt of a particular cue or signal, consultation of the coding scheme, and its prior probability estimates to generate an updated estimate of the probability that a condition is true. The protocols by which the receiver processes and categorizes the signal, retrieves correlations from a stored coding scheme, generates an update, and makes a decision are all parts of increasingly well-understood brain functions. It no longer pays to ignore this formerly “black box.” Sensory processing and classification have been dissected in detail in many species. Many animals appear to use Bayesian updating or nearly Bayesian heuristics to generate updates. How this is achieved neurobiologically is currently a subject of intense research but considerable progress has already been made. We do not specify how a given reliability is achieved in the model presented above, but likely scenarios are discussed in Chapter 8 and its associated Web Topics.

Receiver variability

The problem that different receivers might invoke different priors and have different values of $\Delta W_i$ can be accommodated for by computing a different $V_I$ for each type of receiver. While critics might argue that this simply puts numbers on the problem of indeterminacy noted earlier, the fact is that variation among individuals in fitness consequences for a given strategy is a normal part of evolutionary dynamics. The selective advantage of using signals must be based not on a few individual cases but on the population-wide average value of information. One expects that this average will be positive in populations where the use of signals is the norm.

Arms races

The problem that senders and receivers might have different optima is best handled with one of the many models of signal evolution using evolutionary game theory. If the optima for both parties are intermediate, (due to accelerating costs or decelerating improvements in reliability), the interesting question is whether or not the equilibria predicted by game theory are above the minimum reliabilities for both parties. Most existing models do not predict a net fitness loss (negative value of information) for either party at an equilibrium (ESS).

Verbal versus algebraic descriptions

The problem that receivers may rely on variable weightings of cues and signals for decisions remains intractable given an insistence on discrete verbal classifications, but is easily accommodated by these algebraic models.

Conclusions

The amount of information provided by a set of signals, the differences in fitness payoffs for correct versus wrong decisions, and the costs of participating in communication all play parallel roles in determining whether selection will favor signaling over alternative strategies. Because cue monitoring grades into signaling, and most receivers base decisions on both cues and signals, discrete categories separating what is a signal and what is not can be very misleading. Conflicts of interest are common in signaling dyads, and these can interact in complicated ways to determine the equilibrium levels of reliability. However, most animal signals appear to have an intermediate level of reliability. This may reflect the opposing forces exerted by the two parties on the system, or it may be more a result of the fact that increased investments (costs) in communication likely result in asymptotic benefits for both parties, and thus the optima for both have intermediate values. Any conflict is then over which optimum is closer to the equilibrium.

Literature Cited


Huttegger, S. M., B. Skyrms, R. Smead and K. J. S. Zollman. 2010. Evolutionary dynamics of Lewis signaling games:


Web Topic 2.1  
Measuring Sound Pressure

Overview

Biologists measure sound pressures of animal signals for one or more of the following reasons:

- To record and characterize the temporal and spectral patterns in those signals
- To measure absolute signal intensities
- To use multiple microphone arrays to monitor locations of sound-producing animals
- To undertake playbacks to animals to determine the amount of information or value of information provided by acoustic signals

In all cases, a sensor called a microphone is used to capture the sound signals and convert pressure waveforms into electrical waveforms. The electrical waveforms are then fed into measuring devices or recorded for later analysis. We examine each of these steps and approaches briefly below.

Microphone types

Broadly defined, a microphone is any sensor that responds to ambient variations in sound pressure and converts these variations into an electrical analogue that can be measured or stored. An ideal microphone would respond equally to all frequencies over a broad range, to all amplitudes, and to even the most rapidly varying waveforms. No microphone meets all of these requirements perfectly, and none works equally well for air, water, or substrate recording. One thus needs to match the microphone to the context and the task. Some basic types of microphones currently used for monitoring animal sounds include:

- **Condenser microphones**: A condenser (or capacitor) microphone consists of two metal plates arrayed parallel to each other. The plate that faces the sound source must be very thin, and is often a sheet of plastic with an ultra-thin layer of metal coated onto one side. This is called the “diaphragm” of the microphone. The other plate can be thicker. An insulating ring filled with air separates the two plates and allows for slow equalizing with ambient pressure through a tiny hole. A voltage is then applied across the plates causing electrons to aggregate on one plate, and to become scarce (relative to the status when no voltage is present), on the other plate. The excess of electrons on one plate and shortage on the other depends on the voltage applied and the distance between the two plates. When a propagating sound arrives at the microphone, the pressure of the air inside the ring separating the plates stays relatively constant at ambient levels whereas the pressure outside the diaphragm rises and falls with the sound waves. When the exterior pressure is greater than inside, the diaphragm is bent into the ring cavity and closer to the second plate and when the exterior pressure drops below that inside the cavity, the diaphragm is bent outwards and away from the other plate. As the distance between the diaphragm and the second plate varies, the electrical field force between them varies, causing the appropriate numbers of electrons on each plate to change. Electrons thus move into or out of the plates, and this current can be detected as a varying voltage across a resistor in series with the voltage source. The result is an electrical replica of the sound pressure variation that can then be stored, measured, and analyzed. Condenser microphones tend to have very low electronic self-noise levels and respond similarly to all frequencies over a broad range including into the ultrasonic region. The use of very thin and small diameter diaphragms improves the response of the microphone to sounds with sudden transients and minimizes overshooting and ringing effects. These microphones do however require a carefully regulated source of polarizing voltage for the two plates, and may not function properly in high humidity. However, recent designs use the output of the condenser to modulate a radio-frequency (RF) carrier signal instead of just feeding the output into amplifiers. These RF designs result in extremely low-noise, flat frequency response, and relatively humidity-insensitive microphones that are now very popular among terrestrial field recordists. They do not work under water because the incompressibility of water...
limits the range of movement that would be induced in the diaphragm.

- Electret microphones: Electret microphones work on the same principle as a condenser microphone but differ in that the thin membrane responding to sounds consists of a material that is permanently polarized electrically (in a manner similar to a magnet that is permanently polarized magnetically). No additional voltage source is thus required to detect movements of the membrane in a sound field. Unfortunately, the greater thickness required for the diaphragm slows down its movements and limits its responsiveness to higher frequencies. Newer models place the polarized material on the back plate and use the same metalized diaphragm found in condenser microphones. This makes their frequency range as good as that for a condenser microphone. Electrets are cheap to produce, and can be made very small. Some electrets are sensitive to high humidity. Electret and condenser microphones are the two most widely used types of sensors for monitoring of terrestrial animal sounds.

- Piezoelectric microphones: This type of microphone consists of a crystal or similar material that generates a voltage when accelerated. Ceramic phonograph cartridges are a historical example. Modern uses include accelerometers (which measure the rate at which a solid substrate changes speed in a given direction when vibrating), and hydrophones (microphones used to monitor sounds under water). Piezoelectric microphones do not require a polarizing voltage, but do require amplification given the low electrical signals generated. A related device called a piezoresistive accelerometer changes its resistance as a result of pressure variation and, given an exterior source of voltage, can monitor very low frequencies propagating in solid substrates.

- Dynamic microphones: In a typical dynamic microphone, a wire coil is attached to the inside surface of a thin plastic diaphragm. When sound waves force the diaphragm to oscillate in and out of the microphone cavity, the coil passes back and forth over a magnet generating a varying electrical current in the coil that emulates the pressure variations. Alternatively, the magnet is attached to the diaphragm and sound waves move it into and out of a fixed coil. Many studio microphones have this dynamic design. A limitation is that attaching the diaphragm to a coil or magnet limits the diaphragm to slower movements and thus lower frequencies than is possible with condenser and electret microphones. In addition, the signal generated is usually very small, requiring significant amplification before usage. On the other hand, dynamic microphones do not require a polarizing voltage, and are completely immune to humidity problems.

- Ribbon microphones: Ribbon microphones, like dynamic microphones, produce an electric signal by letting sound waves vary the position of a metallic object in a magnetic field. However, they do this without a diaphragm by suspending a thin corrugated strip of metal between magnets. The microphone casing is open on both ends and the ribbon is set into motion by the difference in sound pressures at the two ends of the casing. As the ribbon moves in the magnetic field, it generates electrical currents that are analogues of the variations in pressure gradient at the two ends of the casing. Ribbon microphones tend to be very fragile because the supporting attachments of the ribbon must be small to maintain ribbon mobility. This limits their utility in field situations.

- Laser vibrometers: These devices bounce a laser beam off of a vibrating surface and compare the outgoing and reflected frequencies of light. When the surface is moving towards the laser source, the reflected light is Doppler-shifted (see text for definition) to a higher frequency than that emitted; when the surface is moving away from the source, the reflected light is Doppler-shifted to a lower frequency. The device monitors these light frequency deviations over time and thus can track the changing velocity of the surface as it moves repeatedly towards and then away from the laser vibrometer. This is largely a laboratory tool as the measured surface must be relatively immobile except for the vibrations. However, it has been used successfully to measure vibrations on the surfaces of frog vocal sacs and ears, and the substrate signals of leafhoppers on plant leaves and stems.

More information: More details on these and other kinds of microphones can be found at:

- General microphone intro on Wikipedia: http://en.wikipedia.org/wiki/Microphone
Microphone directionality

Microphones differ dramatically in their directionality. An omnidirectional microphone picks up all sounds from any direction. For measurements of ambient noise levels, say in a tropical forest, this is a desirable feature. However, when recording individual animals ambient noise can significantly reduce the utility of the resulting recordings. Field recordists thus favor more directional devices that maximize the signal being generated directly in front of the microphone and reduce the amplitudes of signals arriving from other directions. There are several options available to increase directionality:

- **Cardioid microphones**: These microphones have a sensitivity (polar) pattern that is heart-shaped (cardioid): sounds coming from in front of the microphone are strongly favored and sensitivity drops off steadily as the angle of the sound source moves to the sides of the microphone. Sensitivity is least for sounds behind the microphone. Most commercial microphones have a cardioid sensitivity pattern.

- **Ribbon microphone**: Because a ribbon microphone measures the pressure difference at the two ends of its casing, it is inherently bi-directional. Sounds that are located to the sides of the microphone axis tend to arrive at the two casing ends with the same amplitude and phase and thus generate no movement in the ribbon. Sounds originating at either end of the microphone produce the strongest response. While they can be highly directional, ribbon microphones are often too fragile to be of broad use in field recording.

- **Hyper-cardioid (shotgun) microphones**: These consist of long hollow tubes with slits cut at intervals along the length and a condenser or electret sensor at the end of the microphone furthest from the sound source. The microphone is most sensitive to sound sources located along the main axis of the tube and very insensitive to sound sources to either side. It has a second, but lower, peak of sensitivity for sound sources behind the microphone. These microphones work by canceling out sounds at the sensor that arrive both directly from the opposite end of the tube and from the slits in the side. There is no amplification of arriving signals, but only elimination of lateral noise. Shotgun mikes are widely used by field biologists to record animal sounds with moderate to high signal-to-noise ratios.

- **Parabola and microphone combinations**: Another way to obtain directionality is to aim a large metal or plastic parabolic reflector at the sound source and record the reflected and focused sound waves with a microphone placed at the parabola’s focal point. This combination has the advantage over shotgun microphones that it amplifies those sounds (without adding concomitant electronic noise) that originate along the main axis of the parabola and excludes sounds from other directions. It has the disadvantage that it can become highly frequency dependent when the wavelengths of the sounds of interest are as large or larger than the parabola. Given the weight and width of very large parabolas, this method has largely been used to record small to moderately-sized animals (which tend to produce higher frequency sounds than large animals).

- **Substrate recording**: A host of recent studies have examined how sounds produced by arthropods propagate inside plant leaves and stems. One difficulty with such recordings is that a single microphone placed on a plant part may not be at a location in which bending and other internal waves cause detectable surface vibrations. Recent work suggests that placing at least two microphones at right angles is more likely to capture any propagating waves. The article “A method for two-dimensional characterization of animal vibrational signals transmitted along plant stems” in the Journal of Comparative Physiology A has details (McNett, G. D, R. N. Miles, D. Homentcovschi, and R. B. Crocroft. 2006. 192: 1245–1251).

**Sound recorders**

Biologists often want to record the waveforms of animal sounds for later pattern analysis. This has usually meant
moving some magnetic medium over a metal recording head that converts the electrical signals from microphones or microphone amplifiers into a reasonably permanent magnetic record on the medium. Running the medium back over a playback head at a later time then reconverts the magnetic record back into electrical signals for visualization and analysis. The earliest magnetic recorders used spools of metal wire, but these were soon displaced by magnetic tape. The early 1950s saw the introduction of the first field portable tape recorders and this began a golden age of animal sound recording. Reel-to-reel tapes were replaced by cassette tapes in the 1970s, which enhanced portability even further. Up to this point, waveforms were recorded in analog fashion on the magnetic materials. However, in the mid-1980s, digital magnetic tapes (R-DATs) became available and quickly replaced the analog cassette as a preferred recording format. While digitizing throws away tiny segments of the original waveforms, the losses are usually undetectable by most human ears at commercial rates (44.1 kHz or 48 kHz), and digitizing at even higher than that can be used to eliminate detectable losses for species such as katydids and bats as well. The digital environment provides the opportunity to store a much broader range of amplitudes accurately than affordable analog systems. Digital recordings can also be copied and cloned without loss of information, a problem that plagued analog magnetic systems that often suffered medium decomposition over time. Digitization has also now solved one additional problem with field recording: as long as some medium has to be moved physically to make a recording, there is always a chance for dirt, humidity, or insects to gum up the working parts. Today, modern digital recorders can store sounds on a flash or other solid memory device without moving a single part. This is an enormous advance and modern field recorders are now remarkably robust. See below for links to a more detailed review of available recording devices.

**Sound level meters**

Biologists interested in the pattern structure of a sound are mainly interested in the relative variation in sound pressure. Put another way, they are interested in the shape of a recorded waveform, and not in the absolute amplitudes. However, there are times when measuring the absolute values of the sound pressures are essential to the study. Examples include comparisons of maximal sound amplitude among displaying males, attenuation of animal sounds in different habitats, measures of ambient noise that signaling animals must exceed to be heard, etc. Measuring absolute sound pressures requires a calibrated device called a sound level meter. Consider that arriving sound pressures are usually converted into physical movements by a microphone diaphragm. These movements are then converted into electrical signals that in turn are likely to be amplified before they can be read from the position of a needle on a meter. Each step of this process involves a conversion of one replica of the waveform into another. At best, each new replica is a proportional (linear) version of the prior one and all frequencies in the sound are treated equally. However, that is not always possible and most microphones have some frequency dependence. Even if successive versions are proportional copies, the proportionality constant is likely to be different for different steps. In short, the only way to determine what a given needle reading means is to calibrate the entire system against sounds of different frequencies and known pressure amplitudes.

Sound level meters record the amplitude of a signal in decibels relative to some standard. The scale is thus logarithmic. See [Web Topic 2.3](#) for definitions and alternative standards. Available models differ in how accurately they are calibrated and how often they must be recalibrated. The most accurate are Type I meters which tend to be quite expensive. Most environmental noise standards allow for measurements at the Type II level, a lower accuracy, and this is often sufficient for biological measurements in the field. While most sound level meters assume that one is making a measurement in the far field, where sound intensity can be measured completely by measuring pressure, some meters allow for measurements in the near field where both pressure and medium velocity contribute to overall sound intensity. These are naturally more expensive than the standard far field devices.

Most sound level meters have the microphone sensor located at the end of a pointed or tubular end of the meter. This is designed to minimize reflections from the body of the meter that might interfere with measurements at the sensor. Sound level meters that work with hydrophone sensors or that measure acceleration due to substrate vibrations may have the sensor located on an even longer probe. The electronics and meter needle in most meters can be adjusted to respond slowly (thus reducing the potentially over-riding effects of transients in the sound) or instead to record the peak value of an impulse sound by having the needle “stick” at the highest value detected. Most models also provide
both “flat” (unaltered) and alternative (A, B, and C) “weightings” that convert absolute pressure levels into values that more accurately match the sensitivity of the human auditory system. Meters may use either analog or digital processing of the incoming sounds. Analog meters tend to be easier to use, but digital meters allow for more accurate measurements over a wider range of amplitudes. Note that reported sound level measurements must specify the distance from the sound source at which the measurement was taken: because sound attenuates as it propagates, this distance will critically affect the values recorded.

Microphone array recording

A final type of sound recording task is to identify where a sound source is located. In open environments, this can sometimes be accomplished visually by looking for concurrent physical movements of the sound producing animal. However, if animals do not show visible movements when vocalizing, it can be challenging to determine who made which sound. While a human observer may be able to determine the direction and distance to a sound source acoustically, this is really only practical for high frequency sounds. Trying to decide by ear which elephant in a study site emitted a low frequency rumble is usually impossible. In water or forest, there are no visible cues about who made a sound at all. Assigning specific sounds to specific animals has become of great importance as researchers tackle networking: the interactions between local assemblages of animals. When songbirds countersing, which bird makes which song and do other neighbors enter into the exchange? Do singing humpbacked whales match each other’s song themes? Which birds contribute to a dawn chorus and is there any synchrony in the process? These and other problems require accurate assignment of sounds to specific individuals.

The current availability of laptop computers has solved this problem by allowing one to deploy an array of microphones at known locations and record from all microphones simultaneously. Microphones can be connected to the computer by cables or radio links. Alternatively, one can distribute multiple automatic recording units (ARUs) that record sounds simultaneously and use satellite signals to embed timing information for later synchronization. However recorded, the multichannel recordings are then analyzed with software that uses the arrival times of any given sound at the separate microphones to compute the location of its source. If the animals are sufficiently territorial, caller identities can then be determined from the locations. If they are not, concurrent videos of animal positions or other visual information can be used to identify which animal was in a given location at that time.

Some recent descriptions of this method and its use:

- **Recent publications on method**:

- **Active telemetry for network/array analysis**: Encounternet Project: [http://encounternet.net/aboutus.html](http://encounternet.net/aboutus.html)
Array recording analysis software:
- Bioacoustics Research Program’s XBAT units for Matlab: http://xbat.org/home.html

Chickadee and wren applications: http://www.uwindsor.ca/dmennill
Banded wren applications: http://www.birds.cornell.edu/brp/BandedWren.html
Blackbirds and sage grouse applications: http://www.eve.ucdavis.edu/gpatricelli/Patricelli_Research_Interests.html
Forest elephant applications: http://www.birds.cornell.edu/brp/elephant/sections/field/tools.html
Bowheaded whale applications: http://www.birds.cornell.edu/brp/research/acoustic-census-of-migrating-bowhead-whales

Equipment reviews

The technology for recording animal sounds keeps changing rapidly and there are many different suppliers and options available. A site that provides a free and neutral overview of relevant equipment is provided by the Macaulay Library at the Cornell University Laboratory of Ornithology (http://macaulaylibrary.org/inside/record/equipment/audio/index.do).

Web Topic 2.2
Visualizing Sound Waves

Overview

Thanks to recent efforts to promote science, technology, and mathematics education and to the rapid growth of the Internet, there are currently many informative websites devoted to visualizations of the principles of physics, and of particular interest here, the principles of acoustics. For our examples, we redirect you to “Acoustics and Vibration Animations” by Dr. Dan Russell of Kettering University at http://www.kettering.edu/~drussell/demos.html.

To get a second view, many of these phenomena are also animated at “Sound Waves” from the Institute of Sound and Vibration Research (ISVR) at the University of Southampton at http://www.isvr.soton.ac.uk/SPCG/Tutorial/Tutorial/StartCD.htm.

Molecular movements in different kinds of sound waves

At “Longitudinal and Transverse Waves” by Dr. Dan Russell of Kettering University (http://www.kettering.edu/~drussell/Demos/waves/wavemotion.html), the movements of individual molecules can be seen at the same time that a particular type of sound wave is propagating. In each example, fix your eye on one molecule and watch how it moves as it helps propagate the wave past its own immediate region. The examples cover:

- Longitudinal sound waves
- Transverse sound waves
- Water surface waves
- Ground surface (Rayleigh) waves

Patterns of sound wave interference

When two waves in air or water of different frequency or phase arrive at the same location, their effects are additive. If they are in phase, the sum can be greater than either alone; if they are out-of-phase, they might cancel each other. Two waves that are similar in amplitude and only slightly different in frequency create a sum with a regular variation in amplitude called beats. A useful website is “Beats” by Dr. Dan Russell of Kettering University, at http://www.kettering.edu/~drussell/Demos/superposition/superposition.html.

Sound at boundaries

Depending upon the relative acoustic impedances of two adjacent media, sound traveling in one may be reflected or refracted when it encounters the boundary between them. Reflected energy stays in the original medium but travels at a new angle; refracted energy passes into the second medium, and also changes its direction of travel. Go to “Reflection” (Dr. Dan Russell, Kettering University), at http://www.kettering.edu/~drussell/Demos/reflect/reflect.html, and “Refraction” at http://www.kettering.edu/~drussell/Demos/refract/refract.html.

Scattering and diffraction

When sound waves propagating in a medium encounter either a boundary containing a hole in it, or an object in the medium, their trajectories will be altered. The way in which the altered waves move depends largely on the relative size of the slit or object and the sound wavelengths. Go to “Diffraction of sound waves around objects” at the Salford Acoustics Pages at http://www.acoustics.salford.ac.uk/feschools/waves/diffract.htm#object.
Doppler shifts

Doppler shifts are changes in apparent frequency when either the sender or the receiver of a sound signal is moving rapidly with respect to the other. See the website “Doppler shifts” by Dan Russell (Kettering University), at http://www.kettering.edu/~drussell/Demos/doppler/doppler.html.

Sound fields

The distribution of sound pressures around a sound source differs depending upon the type of sound source and whether there are reflecting surfaces or objects nearby. A good website is “Examples of sound fields” (Dan Russell, Kettering University), at http://www.kettering.edu/~drussell/Demos/rad2/mdq.html. It covers monopoles, dipoles, and quadrupoles.

Resonance and filtering

Most structures and enclosed cavities tend to oscillate at particular frequencies determined by their shapes, dimensions, and acoustic impedances. Resonant frequencies are those that “fit” into the structure and can even build up in amplitudes over successive cycles; filtered frequencies are those that do not fit well into the structure or produce reflected versions that cancel out initial versions. Try the following site where you can adjust the frequency of a sound source that is inserted into an object of fixed size by varying its wavelength relative to the size of the object:

- University of Colorado at Boulder, Physics:
  http://www.colorado.edu/physics/2000/microwaves/standing_wave2.html

The resonant frequencies for a structure are also called its natural modes. Look at these examples of natural modes for:

- A rectangular membrane, Dr. Dan Russell, Kettering University:
  http://www.kettering.edu/~drussell/Demos/MembraneSquare/Square.html
- A circular (drum) membrane, Dr. Dan Russell, Kettering University:
  http://www.kettering.edu/~drussell/Demos/MembraneCircle/Circle.html
- An acoustic guitar body, Dr. Dan Russell, Kettering University:
  http://www.kettering.edu/~drussell/guitars/hummingbird.html
- An empty beer bottle, Dr. Dan Russsell, Kettering University:
  http://www.kettering.edu/~drussell/Demos/BeerBottle/beerbottle.html
- An open pipe, ISVR, University of Southampton:
  http://www.isvr.soton.ac.uk/SPCG/Tutorial/Tutorial/Tutorial_files/Web-standing-pipe.htm

Web Topic 2.3
Quantifying and Comparing Sound Amplitudes

Overview

Sound pressures can vary from levels that are little more than molecular noise to those generated by massive explosions. This poses a challenge to animals that may need to be able to detect the tiny sounds of an approaching predator as often as the roars of threatening competitors. A related problem applies to how animals compare sounds: it is often more useful to compare the ratio of two sounds than to compute the absolute difference in their amplitudes. Most animals solve these problems by using a logarithmic scale to measure sound amplitudes and use ratios to compare sounds (see Web Topic 8.6 on Weber’s Law for additional discussion of this scaling and its consequences). It was thus also natural for scientists to measure and compare sound amplitudes using the logarithm of the ratios of amplitudes. Here, we define this scale in more detail, show how it is typically invoked, and provide some reference values to keep in mind when using it.

Definition of the decibel

Microphones usually measure sound pressure variation or some time derivative of the pressure variation. However, pressure represents only part of the overall power carried by a sound wave; the rest of the power is carried by the velocity of medium particles. The power (also called intensity or energy flow) of a sound wave is the product of pressure and velocity. Because the ratio of pressure to velocity changes, even if the power were to stay constant, as a sound signal moves from the solid vibratory organs of a sender into the propagating medium and then back into the hearing organs of the receiver, scientists selected power as the appropriate basis for measuring and comparing sound amplitudes. In many cases, the power per unit of sampling area, called intensity, is the actual measure used.

Relative amplitudes of two sounds are thus measured by taking the logarithm (to the base 10) of the ratio between their powers or intensities. In the past, the resulting unit was called the bel. However, because this did not provide a fine enough scale, it has since been replaced by a second unit, the decibel (abbreviated dB), which equals one-tenth of a bel. For two sounds with intensities $I_1$ and $I_2$ respectively, their relative amplitudes in dB would be

$$\text{Relative amplitude (dB)} = 10 \log_{10} \left( \frac{I_1}{I_2} \right)$$

Whichever sound is used in the denominator of the ratio is considered the reference value. Thus, if the denominator sound had twice the intensity of the numerator sound, the ratio would be 0.5 and the numerator sound would be said to be –3 dB lower in intensity than the denominator sound. If the numerator sound had twice the intensity of the denominator, then one would say that the numerator was 3 dB higher in intensity than the denominator. Note that over much of the range of frequencies that humans can hear, 1 dB is very close to the minimal difference in amplitude that is necessary before a human will judge one sound to be louder than another. This amounts to about a 26% difference in intensities.

Since most microphones only measure pressure, how can we use decibels to compare sounds? If one is making the measurement sufficiently far away from the sound source to be in the far field (see text for details), the local velocity of the medium particles is proportional to the local pressure and the proportionality constant is the reciprocal of the acoustic impedance. Intensity ($I$) then depends on pressure ($p$) and acoustic impedance ($z$) as

$$I = p \times (p/z) = p^2/z.$$ 

If two sounds are measured in their respective far fields in the same medium, their relative amplitudes can then be written as
Thus, at least in the far field and in the same medium, we can use pressure measurements to compute relative amplitudes. The difference is that the coefficient of the log is 20 for pressure, but only 10 for intensity. Note that if the ratio of intensities of the two sounds is 0.5, the ratio of their pressures will be $\sqrt{0.5} = 0.707$, but the relative amplitude will still be $-3$ dB. Similarly, if the ratio of the two intensities is 2, the ratio of their pressures will be $\sqrt{2} = 1.41$, and their relative amplitude will again be 3 dB.

**Impulses and averages**

Most animal signal waveforms have quite variable amplitudes over time. What measure of pressure should we use to compare amplitudes? One obvious approach is to find the largest maximum or minimum in a sequence of waves and use the absolute value of this peak as the amplitude measure. In some cases, we may be interested in measuring such single spikes of pressure (called impulses): examples include the very short sonar calls of porpoises or the explosive calls of bellbirds. However, more often, a single spike in a waveform is not representative of the amplitude of the rest of the signal and one wants some sort of average pressure measurement. We cannot simply average the recorded sound pressures in a signal because this will only give us the ambient pressure around which a sound wave is oscillating! One solution is to measure the differences in pressure between successive maxima and minima (which should always yield positive numbers and a non-zero average). These are called peak-to-peak measures. However, even this approach could be misleading if peaks are few and far between. A common solution is to sample the waveform at many successive points, compute the difference between the sound pressure at each point and ambient levels, square this value (to eliminate negative differences), and add the squares together. The square root of this sum is called the root-mean-square (abbreviated rms) and is widely used in statistics and physics to characterize the average amplitude of a time-varying quantity. Because these and other criteria can be used to characterize pressure amplitudes, one should always specify exactly which was used in reporting sound amplitude measurements.

**Standard measures**

Up to this point, we have focused on how to quantify and report the difference in amplitude between two sound signals. What if we want to give some indication of the amplitude of a single sound relative to some standard? If everyone agrees on the standard, this would provide the same benefits as having an absolute measure of sound amplitude.

In air, the standard reference is the lowest power that a young (pre-rock music) human ear can detect. This turns out to be sounds at a frequency between 1 and 3 kHz and a pressure amplitude of $20 \times 10^{-6}$ pascals. The intensity standard in air is then about $10^{-12}$ watts/m$^2$. The scale for underwater sounds is offset from that of sound in air for two reasons. The first is that underwater researchers set a reference pressure of $1 \times 10^{-6}$ pascals, a value 20 times smaller than the reference pressure for air. Given this lower level of reference, the same pressure recorded in both media would be accorded a 26 dB higher value in water than it would in air. Second, it takes much less power to create a high pressure in water than in air because water’s characteristic acoustic impedance is 3500 times higher than air’s (see text for definitions). Thus the decibel value comparing a sound pressure in water to one in air will be inflated by an additional 35.5 dB. Overall, measurements of sound amplitudes in water will be inflated by $35.5 + 26 = 61.5$ dB relative to an equivalent measurement in air. Thus an underwater measurement of 170 dB for a sperm whale’s sounds would be equivalent in power to an airborne sound of 109 dB. This is still very loud, but much less than a 170 dB sound in air (a value commensurate with a nearby volcanic eruption). Studies of sounds propagated in solid substrates tend to use the same reference as for water, but the even higher acoustic impedance of solids requires a different adjustment than the 35.5 dB value used in water studies.

Each research community is firmly committed to its own standards, and there is little interest in adopting a common
As a result, it is very important to specify the reference level and medium when publishing standardized measurements of sound amplitude. Many workers append the expression “SPL” (for sound pressure level) as a shorthand for using the standard air reference. However, there has been a recent call for researchers to give the actual reference in all cases (e.g., “34.5 dB re 20 μPa in air”).

Finally, it is important to remember that sound waves attenuate as they propagate away from the sound source. Spreading losses alone (see text for details) will decrease the recorded sound pressure by 6 dB for each doubling of the distance between sample points. Heat losses and scattering will add additional attenuation. As a result, standardized measurements of sound amplitude are usually accompanied by both the reference level and the distance at which the sound was measured (e.g., “34.5 dB re 20 μPa at 2 m in air”).

As a general guide to the scale for airborne sounds, here are some typical values of sound amplitudes (all re 20 μPa) covering the range we are likely to encounter:

<table>
<thead>
<tr>
<th>Source Description</th>
<th>Sound Amplitude (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jet engine at 30 m</td>
<td>150</td>
</tr>
<tr>
<td>Jackhammer at 2 m</td>
<td>100</td>
</tr>
<tr>
<td>Singing birds at 2 m</td>
<td>80–90</td>
</tr>
<tr>
<td>Street traffic</td>
<td>70</td>
</tr>
<tr>
<td>Human conversation</td>
<td>65</td>
</tr>
<tr>
<td>Quiet restaurant</td>
<td>50</td>
</tr>
<tr>
<td>Whisper</td>
<td>20</td>
</tr>
<tr>
<td>Rustling leaves</td>
<td>10</td>
</tr>
<tr>
<td>Threshold of human hearing</td>
<td>0</td>
</tr>
</tbody>
</table>

**Moving between standard and relative measures**

One nice feature of using ratios for amplitude measurements is that a comparative measure can be obtained for two sounds by subtracting the standardized measure of one from the standardized measure of the other. This is equivalent to replacing the reference value in the ratio of one of the standardized measures with the second sound amplitude. For example, suppose two animal signals are recorded at the same distance from each source and one obtains a standardized level of 35 dB re 20 μPa for one and 45 dB (using the same reference and medium) for the other. The appropriate comparison between the amplitudes of the two sounds is simply 45 dB – 35 dB = 10 dB. Given the definition of a decibel, this means that one signal has a pressure 3.2 times greater than the other. One cannot, of course, go the other way: knowing that two sounds differ by 10 dB does not provide sufficient information to reconstitute the standardized values of the two original sounds.

One should also be careful about adding decibel values. If one sound has a standardized value of 50 dB and the second has a standardized value of 10 dB, their presence in the same place at the same time is not the sum of these two values. The first sound has a pressure 100 times greater than the second; having them present at the same time, even if they interfered constructively, would still produce a sound with an amplitude only slightly greater than 50 dB.

**Other scales**

Because humans (and most animals) are not equally sensitive to all frequencies, a number of psychophysical and perceptual scales have been derived for humans that correct for these sensitivity differences. The phon scale identifies the amplitude of a 1 kHz standard sound in dB that is perceived as equivalent in loudness to a test sound at another frequency. The less sensitive the human ear to that frequency, the greater the amplitude of the 1 kHz sound must be to be perceived as equivalent. If a test sound was perceived as equivalent to a 70 dB 1 kHz standard, it would be assigned a perceptual amplitude of 70 phons. The sone scale is based on the number of phons required to double the
perception of loudness. A sound estimated to be 40 phons equals one sone; each additional 10 phons doubles the number of sones, and each decrease of 10 phones halves the sone score. Perceptual scalings can and have been derived for certain animals in the laboratory. However, they invariably differ from that of humans and from each other.

**More information**

The Web is full of useful sites discussing sound amplitude measurements, both for the dB and psychophysical scales. We leave it to the reader to track down current ones: Wikipedia is always a great place to start.
Web Topic 2.4
Fourier Analysis of Animal Sounds

The challenge

Animals can produce quite complicated sounds. If we record such a sound and examine its waveform, there are a number of measurements that we can make and compare with other sounds. However, it is extremely difficult to describe the shape of that waveform in ways that allow us to compare it with sounds of other species, or even other sounds of the same species. We thus need a systematic and quantitative method for describing and comparing different animal sounds. The comparison of waveforms alone is not sufficient. What else can we recruit to the task?

The solution

Fourier’s Theorem states that we can decompose nearly any periodic (regularly repeating) signal into a set of pure sine waves. By recording the frequency, amplitude, and relative phases of these waves, we can provide a full and quantitative description of the sound which can be compared to the Fourier decomposition of any other sound. Although animal sounds are rarely entirely periodic, we can usually get around this constraint by breaking a sound into segments that are sufficiently periodic to perform a Fourier decomposition on each one and then string the results together into a spectrogram or other composite graph.

Background on analyzing animal sounds

Every tool comes with some constraints and costs. This is as true of Fourier analysis as it is of any method. Learning how to properly perform Fourier decomposition of animal sounds requires some understanding of its limits and constraints and a minimal amount of practice using the tools. Luckily, many of the measurements that we can make on the waveform of the sound have some corresponding representation in the Fourier decomposition of that sound. Making sure that the waveform and spectrogram measurements correspond is one way to check that we have done each type of measurement properly. Here, we provide materials explaining useful measurements that can be made on waveforms, how to make a spectrogram and extract measurements from it, and how to compare the results from two views of the same sound. There are two parts to the presentation:

- A Primer on Analyzing Animal Sounds
  This text web page provides an introduction to animal sound measurements
- A Primer on Analyzing Animal Sounds: Figures and Sample Sounds (PowerPoint)
  This link provides relevant figures and notes for sound analysis.

You should open up both files so you can hear and see examples in the PowerPoint that you are reading about in the text pages. You are welcome to download your own copies of both of these files and refer to them when using one of the analysis software packages listed below, or when using the free RavenViewer tools to view waveforms and spectrograms of any of the sounds in the Macaulay Library (http://macaulaylibrary.org/index.do).

A practical in sound analysis using Raven

1. If not already installed on your computer, download the Fourier Test Sounds from this site and unzip the contained folder. It contains the following sample sounds:
   - Puresine: A single pure sine wave with no modulation
   - SinusAM: A single sine wave with sinusoidal amplitude modulation
   - LowFM: A single sine wave carrier weakly frequency-modulated in a sinusoidal manner
   - HiFM: A single sine wave strongly frequency-modulated in a sinusoidal manner
- **Periodsymm**: A periodic but nonsinusoidal signal. The shape of the waveform is triangular such that it is half-wave symmetric.
- **Pulsedsine**: A single sine wave turned on and off. This can be interpreted as periodic nonsinusoidal amplitude modulation of the sine wave carrier.
- **Onepulse**: A single, very brief pulse of sound with sudden onset and offset
- **Tungarafrog**: Male advertisement call of a tungara from (*Engystomops pustulosus*) from Central America
- **Parrotduet**: Antiphonal duet by a mated pair of White-fronted Amazon parrots (*Amazona albifrons*) in Costa Rica. Try to figure out where both birds overlap.

2. This practical assumes you have access to the sound analysis program **Raven Pro** [http://www.birds.cornell.edu/brp/Raven/RavenVersions.html](http://www.birds.cornell.edu/brp/Raven/RavenVersions.html). This program requires a licensing fee although you may be able to use it in demo mode for a short time. Note that the free program **Raven Lite**, available at the same site, does not have all the features assumed to be available for this practical. The authors of **Raven Pro** will often negotiate a class license for a short period if asked.

3. The general goal of this practical is to learn the expected frequency domain patterns for each of three basic time domain waveforms, see how compound signal waveforms decompose into additive combinations of the basic three in the frequency domain, and build up confidence that whatever you see in one domain must have a corresponding presence in the other domain. You should be able to go back and forth between domains with ease after completing these exercises.

4. For each of the examples in the **Test Sounds** folder (we suggest doing them in the order above), do the following:
   - **Load the file**: Use Open Sound Files in the File menu to load the sound into the program. For now, use the default window settings when asked.
   - **Get a general look at the waveform**: Examine the waveform of the sound at both compressed and expanded time scales. Does it repeat? Is it sinusoidal or periodic non-sinusoidal?
   - **Measure periodicities in waveform**: On an expanded time scale, measure the time interval between repeats if periodic. The reciprocal of this number will predict the frequency of the signal if it is a pure sine wave or the fundamental frequency of a harmonic series if it is periodic but nonsinusoidal. If AM or FM, see steps below. If the repetition rate of the waveform varies during the course of the sound, measure the period of repeating and compute its reciprocal at each of several locations. If possible, leave these locations marked in the waveform so that you can find the same point in the spectrogram.
   - **Make a spectrogram**: You may need to fiddle with the frame size/bandwidth of the analysis. Try to get an image that is relatively smooth and shows both slow frequency modulations and amplitude modulations without breaking them into sidebands. However, do use a small enough bandwidth to see harmonics as separate bands.
   - **Compare the waveform and spectrogram**: Do the measurements you made in the waveform correspond as predicted in the spectrogram? Vice versa?
   - **Make a power spectrum slice**: Create a spectrum through one segment of the spectrogram. Compare corresponding measurements in the spectrogram and the power spectrum so that you are convinced that the latter is just a slice through the former. If you have not already done so, try varying the spectrogram/power spectrum analysis bandwidth: do the peaks get fatter or thinner as expected?
   - **AM signals**: If your signal is an AM signal, try adjusting the bandwidth for the power spectrum until it is small enough to show the sidebands and the carrier. Measure the frequencies of the sideband and the carrier. Then examine the following:
     - The difference in frequency between the carrier and each sideband should be equal to the repeat rate in amplitude modulation that you can see in the waveform. Is this true?
     - Is the amplitude of any sideband greater than that of the carrier? It should not be.
**FM signals:** As with AM signals, adjust the bandwidth of the power spectrum until you can decompose it into a carrier and sidebands:

- Look at the frequency modulation pattern in the waveform. Measure how many times the frequency reaches a peak per second. This is the modulating frequency. Now compare this number to the difference in frequency between the carrier and the sidebands in the spectrogram, and between the sidebands and each other. It should be the same.
- Measure the maximum frequency and the minimum frequency in the expanded waveform. If the frequency modulation is sinusoidal, compute the average frequency by adding the maximum and minimum frequencies and dividing by two. Is this equal to the carrier in the spectrogram? It should.
- Compute the modulation index for the FM signal by dividing the difference between the maximum and minimum frequencies by the modulation frequency measured in the waveform. If it is less than 10, then you should see a power spectrum for this signal in which the sideband amplitudes are less than that of the carrier; if it is more than 20, then the sidebands will be taller than the carrier. Is this the case? Between 10 and 20, some sidebands may be similar to the carrier in amplitude.
- Now spread the time scale out with both waveform and spectrogram visible so that you can see 5–10 complete modulations. Then increase the spectrogram and power spectrum analysis bandwidth (e.g., reduce the number of samples/frame in the top right slider control) until the signal no longer breaks down into sidebands. You should now see a spectrogram that shows the frequency rising and falling in a sinusoidal way. You are actually viewing the modulation in a time domain view on a spectrogram! Are the maximum, minimum, and modulation frequencies measured on this spectrogram the same as those you measured in the waveform view?

**Periodic nonsinusoidal signals:** Is the fundamental in the spectrogram view equal to the repeat rate in the waveform? Convince yourself that the bands above the fundamental are truly harmonics. To do this, make a power spectrum through a given point and measure the difference in frequency between successive peaks in the power spectrum. This difference should be the same for each pair of adjacent peaks, and it should equal the frequency of the fundamental (at least within the accuracy of the analysis tools). Is the waveform half-wave periodic or not? If so, are the even harmonics missing?

**Two real animal sounds:** Next, do the same kinds of time domain/frequency domain comparisons for the two animal sounds provided: a tungara frog call and a parrot duet. These sounds change in structure during their course, so you will want to compare measurements at the same instant when comparing waveform, spectrogram, and power spectrum measurements. Remember that anything you measure in one domain should be consistent with general Fourier principles in the other domain. You should be able to describe the changes that occur in the signal waveform during the sound and how and why these are reflected in the spectrogram.

**More real animal sounds:** Now use what you have learned from studying the artificial and two sample animal sounds to examine other real animal sounds. To do this:

- If you do not have Quicktime installed on your computer, go to [http://www.apple.com/quicktime/download/](http://www.apple.com/quicktime/download/) and install it. Quicktime works on both Macintosh and Windows computers.
- Then go to the Macaulay Library website and download and install their browser spectrogram tool called RavenViewer ([http://macaulaylibrary.org/help/ravenViewer/index.do](http://macaulaylibrary.org/help/ravenViewer/index.do)). Then quit and restart your browser.
- Finally, go to the Macaulay Library search page ([http://www.animalbehaviorarchive.org/loginPublic.do](http://www.animalbehaviorarchive.org/loginPublic.do)) and pick some favorite species. To start, only check the Audio box before the search. Once the results show up and you have chosen a clip to examine, select the playback icon with the two gold wings on each side. When the RaveViewer opens, familiarize yourself with the various control options. Then have fun! Most of the audio and video recordings in this vast collection can be played with a window showing the waveform,
There are a variety of software packages available for sound animal analysis. They differ in whether they are free or must be purchased, which operating systems are supported (Linux, Macintosh, and Windows), which measurements and features they support, and how they look and feel. Below, we list some software options, where to download or order it, and whether or not we have provided a primer on using the package on this site. Note that if we have not provided a primer, the programs themselves often come with detailed instructions and tutorials.

- **RavenViewer**: This is the free sound analysis player made available by the Macaulay Library at the Cornell University Lab of Ornithology. Users can look at online waveforms, spectrograms, and power spectra of any of the thousands of animal sounds available in this archive. It works equally well on Macintosh and Windows computers. You do need to have a fast browser and the most recent version of QuickTime to use it.
  - **Getting the viewer**: [http://macaulaylibrary.org/help/ravenViewer/index.do](http://macaulaylibrary.org/help/ravenViewer/index.do)
  - **Getting started**: This is the easiest program to use. See the Help menu in the program for details.

- **Raven Lite**: This is also a free sound analysis package made available by the Bioacoustics Research Program at the Cornell Lab of Ornithology. It works on Linux, Mac, and Windows platforms, and generates both waveforms and spectrograms in ways that facilitate accurate measurements. It does not currently provide power spectra.
  - **Getting the player**: [http://www.birds.cornell.edu/brp/Raven/Raven.html](http://www.birds.cornell.edu/brp/Raven/Raven.html)
  - **Getting started**: We have not provided a separate guide to Raven Lite, but many of its functions are similar to those for the larger Raven Pro program (see below). Get our [Raven Primer here](http://www.birds.cornell.edu/brp/Raven/Raven.html).

- **Syrinx**: This free program is written and designed by Dr. John Burt of the University of Washington. It only runs on Windows platforms and currently does not provide power spectra. However, it is extremely user-friendly, has good measurement tools, and can also be used for playbacks and signal extraction. It is currently widely used by researchers on bird vocalizations.
  - **Getting the player**: [http://syrinxpc.com/](http://syrinxpc.com/)
  - **Getting started**: Use our [Syrinx Primer](http://syrinxpc.com/).

- **Sound analysis**: This program was written by Dr. Ofer Tchernichovski (City College of New York) and colleagues. To create spectrograms, this program uses a slightly different set of algorithms, which makes for improved measurements. It also provides a very sophisticated set of tools that are now widely used by neurobiologists to study animal sound generation and perception. It is free, but only works on Windows platforms.
  - **Getting the player**: [http://ofer.sci.ccny.cuny.edu/sound_analysis_pro](http://ofer.sci.ccny.cuny.edu/sound_analysis_pro)
  - **Getting started**: See program documentation

- **Ishmael**: This sound analysis package was written by Dr. David Mellinger of Oregon State University, and was initially aimed at marine animal sound analysis. However, it is completely effective for any animal sound and is free. It is only Windows-based.
  - **Getting the player**: [http://www.pmel.noaa.gov/vents/acoustics/whales/ishmael/](http://www.pmel.noaa.gov/vents/acoustics/whales/ishmael/)
  - **Getting started**: See program documentation

- **Raven Pro**: Raven Pro is the successor to the popular Canary program that was widely used in the 1990s. It runs on Linux, Windows, and Macintosh platforms, and creates waveforms, spectrograms, and power spectra that can be measured with a variety of useful tools. There is a charge to acquire the software, but this varies depending on the client and context. Special educational discounts are available. Check the
URL below for details.

- **Getting the player:** [http://www.birds.cornell.edu/brp/Raven/Raven.html](http://www.birds.cornell.edu/brp/Raven/Raven.html)
- **Getting started:** Use our [Raven Primer](http://www.birds.cornell.edu/brp/Raven/Raven.html).

**PRAAT**: This program was written by Paul Boersma and David Weenink at the University of Amsterdam. It was largely designed for analyses of human speech, but many of its tools can be used on any animal sound, especially those of the many species whose signals are periodic nonsinusoidal sounds. The program is free and runs on all common computer platforms.

- **Getting the player:** [http://www.fon.hum.uva.nl/praat/](http://www.fon.hum.uva.nl/praat/)
- **Getting started:** See program documentation

**XBAT**: XBAT is a highly sophisticated program environment for animal sound analysis from the Bioacoustics Research Program at Cornell’s Lab of Ornithology. Its provision of waveforms, spectrograms, and power spectra is only the starting point for a very diverse battery of measurement and analysis tools. XBAT is free, but it must be run in the Matlab environment. Matlab in turn must be purchased, although many universities provide low cost site licensing. Users need to be somewhat familiar with how to perform operations within Matlab to use XBAT.

- **Getting the player:** [http://xbat.org/home.html](http://xbat.org/home.html)
- **Getting started:** See program documentation

**AviSoft**: This is a very sophisticated and powerful sound analysis package that is now widely used by researchers in Europe. It runs only in Windows and is fairly expensive. However, new tools and measures are continually being added, and the designers are very responsive to new needs of clients.

- **Getting the player:** [http://www.avisoft.com/](http://www.avisoft.com/)
- **Getting started:** See program documentation

**Signal/RTS**: Engineering Design has been in the business of creating sound analysis software for some time and its packages are used by many American researchers. The software is sophisticated and provides waveforms, spectrograms, and power spectra, along with a diverse number of measurement tools. It also has tools for sound synthesis and manipulation that can be useful for playback studies. The program is Windows-based, but it can be run on a Macintosh using a Windows emulation mode. It must be purchased and is the most expensive of the packages that we list here.

- **Getting the player:** [http://www.engdes.com/](http://www.engdes.com/)
- **Getting started:** See program documentation

**Other Sound Analysis Packages**: Dr. Steven Hopp (Emory and Henry College) maintains a web page that lists a wide variety of animal sound analysis software options. Check it out at: [http://zeeman.ehc.edu/envs/Hopp/sound.html](http://zeeman.ehc.edu/envs/Hopp/sound.html).
Overview

The interaction of sound at a boundary between two media can be a fairly complicated process that depends on the angle of incidence of the propagating wave, the acoustic impedances of the two media, the speeds of sound in the two media, and any textural patterns on the boundary surface. In general, some of the incident sound wave (A) will be **reflected** back (B) into the initial medium, and some will propagate across the boundary (C), where its direction of travel is likely to be **refracted** (bent) instead of traveling on in the second medium on its initial trajectory (A’). \(C_{\text{Higher}}\) and \(C_{\text{Lower}}\) refer to two different angles of possible refraction as discussed below. There are four different cases to consider: two are well-known and widely cited in most textbooks; the other two are less well-known, but could easily be encountered while studying animal sound communication.

![Diagram](image)

**Basic definitions**

**Reflection coefficients**

Consider a sound traveling in Medium 1 and encountering a boundary with Medium 2. The angle between the direction of propagation and the surface of the boundary is the grazing angle \(\phi\). The value of the reflection coefficient, \(R\), at this boundary depends on the ratio of the acoustic impedances of the two media \((Z_2/Z_1)\), the grazing angle \(\phi\), and the ratio of the velocities of sound in the two media \((c_1/c_2)\):
Reflection coefficients vary from +1 to –1. When they are +1, all of the incident energy is reflected from the surface and the reflected wave undergoes no phase shift. Except for a change in propagation direction, it is as if the boundary were not even there. When \( R = -1 \), all of the energy is reflected, but the reflected wave is phase-shifted 180° (or one half wavelength): that is, it begins a half-cycle behind that which one would have expected had there been no reflection. If the incident wave were at a maximum when it hit a boundary with \( R = -1 \), then it would begin as a minimum in the reflected wave. The smaller the absolute value of \( R \), the less energy is reflected and the more energy passes into the second medium.

**Angles of incidence**

The grazing angle of the sound, \( \phi \), can greatly affect the reflection coefficient at a boundary. It \( \phi \) varies between 0, when the sound is parallel to the boundary, and 90°, when the sound is traveling in a direction perpendicular to the boundary. Between these two extremes, there will be an important threshold value: for angles of incidence below the threshold, \( R \) will vary one way with \( \phi \), and above the threshold, it will vary in another way. When the medium with the incident sound has the lower velocity, e.g., \( c_1 < c_2 \), the threshold value is called the critical angle. It is denoted by \( \phi_c \), and is computed as \( \cos \phi_c = \frac{c_1}{c_2} \). When the incident medium has the higher velocity, e.g. \( c_1 > c_2 \), then the threshold value is called the angle of intromission, and is denoted by \( \phi_i \). It is computed as:

\[
\cos \phi_i = \sqrt{\frac{\left(\frac{Z_2}{Z_1}\right)^2 - 1}{\left(\frac{Z_2}{Z_1}\right)^2 - \left(\frac{c_2}{c_1}\right)^2}}
\]

**The four cases**

We can divide the possible relationships between \( R \) and \( \phi \) into four cases depending upon whether \( Z_1 > Z_2 \) or \( Z_1 < Z_2 \), and whether \( c_1 < c_2 \) or \( c_1 > c_2 \). Many of the situations one encounters fit either Case I or II; however, the other two cases are not uncommon and the reader should be aware of them. In the following plots, orange zones are ones with a full 180° phase shift at reflection, white zones have no phase shift, and blue zones show a continuous change in phase shift with increasing grazing angle.

**Case I**: \( Z_1 > Z_2 \) and \( c_1 > c_2 \)
If $Z_1 > Z_2$ and $c_1 > c_2$, then the value of $R$ is always negative. This would be the case if the incident sound waves were in water and the sound hit the water’s surface. The result is a 180° phase shift (orange region) regardless of incident angle. The critical angle is irrelevant in this case. As the angle of incidence increases from 0° to 90°, the value of $R$ increases from –1 toward its value for perpendicular incidence, $R_{90}$.

**Case II: $Z_1 < Z_2$ and $c_1 < c_2$**

The opposite extreme occurs when $Z_1 < Z_2$ and $c_1 < c$. An example would be sound in air hitting the surface of a body of water. The relevant threshold in this case is the critical angle $\phi_c$: if $\phi > \phi_c$, then $R$ (solid black line) is always positive (no phase shift, white region) and decreases the closer $\phi$ is to 90°. For low enough incident angles, e.g., when $\phi < \phi_c$, all energy is reflected ($|R| = 1$) but the phase shift (indicated with a dashed line) decreases from a full 180° at a 0° grazing angle down to no phase shift at a grazing angle of $\phi_c$ (blue region).

**Case III: $Z_1 < Z_2$ and $c_1 > c_2$**
Here, the incident medium has a lower impedance but a higher velocity: $Z_1 < Z_2$ and $c_1 > c_2$. Examples include sound traveling in water and striking a muddy bottom or sound traveling in air and hitting certain types of soils. The important threshold angle of incidence here is the angle of intromission, $\phi_i$. For $\phi < \phi_i$, reflected waves always experience a 180° phase shift, and the value of $R$ will increase from −1 to 0 as $\phi$ increases. When $\phi = \phi_i$, no energy is reflected: it all passes into the second medium! For $\phi > \phi_i$, there is no phase shift and the fraction of energy reflected increases with the incident angle.

**Case IV:** $Z_1 > Z_2$ and $c_1 < c_2$

This is the opposite of Case III, since $Z_1 > Z_2$ and $c_1 < c_2$. It can occur when sounds propagated in a muddy or soil substrate reach the interface with the overlying medium. Both thresholds must be invoked in this example. At incident angles less than $\phi_i$, all energy is reflected, but the phase shift varies from 180° when $\phi = 0°$ to none at $\phi_c$. Further increases in incident angle decrease $R$ without any phase shift until $\phi = \phi_i$, when no energy is reflected. Higher incident angles result in a 180° phase shift and variation in $R$ from 0 to $R_{90}$.

**Reference**

Web Topic 2.6
Sample Animal Sounds

Overview

It is often difficult to imagine what a particular animal sounds like from a text treatment. In these pages, we provide examples of most of the types of vibration production known in animals. Note that the sounds that were recorded—which you will hear—are rarely the original vibrations. Instead, the initial vibrations have been modified in various ways before being emitted, often in a frequency-dependent fashion, into the surrounding medium. Note also that some of the recordings below were recorded in air, some under water, and some by placing a special sensor against a plant stem or leaf. Most of the recordings listed below are currently archived at the Macaulay Library of the Cornell University Laboratory of Ornithology (Ithaca, New York). This is currently the world’s largest archive of animal sounds and access is free. Other sources are listed as needed.

Sources and tools

Before accessing the sounds in the table below, you will need to install a player plugin on your computer. There are two options:

- **Direct play**: Most browsers on regular computers come equipped with plugins to play movies and sounds. These are likely to be Flash or Shockwave. You will need one of these to play the Macaulay Library sounds and videos. Try it first, and if they play, then don’t worry about plugins. If they don’t play, you can get either plugin player free: Flash is available at [http://get.adobe.com/flashplayer/](http://get.adobe.com/flashplayer/), and Shockwave at [http://get.adobe.com/shockwave/](http://get.adobe.com/shockwave/).

- **RavenViewer**: This is also free and works on both Windows and Macintosh computers. It requires that you already have either a Flash or Shockwave plugin installed in your browser and a recent version of Apple QuickTime (also a program compatible with both Windows and Macintosh machines) installed on your computer. If you do not have QuickTime, install it first from the Apple Computer QuickTime website ([http://www.apple.com/quicktime/](http://www.apple.com/quicktime/)). Then go to the Macaulay Library website and install the appropriate RavenViewer plugin ([http://macaulaylibrary.org/help/ravenViewer/download.do](http://macaulaylibrary.org/help/ravenViewer/download.do)). You then need to quit your web browser and restart it by coming back to this page. Once installed, the RavenViewer will provide detailed waveforms, spectrograms, and power spectra for any sound that you play. You can customize the colors and scales and save the settings for future use. You may want to speed up playback of slow sounds or slow playback of fast ones. There is a help feature (upper right corner of the player), or you can read now about the player by clicking the following link: [http://macaulaylibrary.org/help/ravenViewer/details.do](http://macaulaylibrary.org/help/ravenViewer/details.do).

Once you have the appropriate player installed, you can begin to hear and, if you are using the RavenViewer, see the spectrograms and waveforms of any sound listed. Return to our list and select another sound. The classification of vibrational sources follows that outlined in the text:

The samples:

1. **Solid body part moved against other solid or reshaped**
   
   A. **Percussion**:
   
   - **Long-fingered fiddler crab** (*Uca speciosa*): crab taps legs against substrate
     
     - **Audio**: [http://macaulaylibrary.org/audio/116948](http://macaulaylibrary.org/audio/116948)
- **Crimson-crested woodpecker** (*Campephilus melanoleucos*): woodpecker drums bill against hollow or resonant trunk or branch of tree

- **White stork** (*Ciconia ciconia*): Mated pair clacking their bills in mutual display

- **Chimpanzee** (*Pan troglodytes*): Screams followed by drumming with hands and feet on tree buttress

- **Western diamond-back rattlesnake** (*Crotalus atrox*): Hollow buttons on end of tail whipped rapidly to create rattle
  - **Audio**: [http://macaulaylibrary.org/audio/flashPlayer.do?id=107975](http://macaulaylibrary.org/audio/flashPlayer.do?id=107975)

**B. Stridulation**:

- **Spiny lobster** (*Panulirus argus*): Stridulations produced by rubbing files and plectra on base of antennae
  - **Audio**: [http://macaulaylibrary.org/audio/flashPlayer.do?id=112458](http://macaulaylibrary.org/audio/flashPlayer.do?id=112458)

- **Northern mole cricket** (*Neocurtilla hexadactyla*): Males call from hole in the ground and stridulate using file and plectrum on forewings

- **Blue-striped grunt** (*Haemulon sciurus*): Coral reef fish grinds pharyngeal teeth to produce a stridulation
  - **Audio**: [http://macaulaylibrary.org/audio/flashPlayer.do?id=112432](http://macaulaylibrary.org/audio/flashPlayer.do?id=112432)

**C. Buckling**:

- **Treehopper** (*Vanduzeea sp.*): Tymbals produce repetitive sound by buckling solid plate; vibrations then transmitted through legs into plant stems
  - **Audio**: [http://macaulaylibrary.org/audio/141253](http://macaulaylibrary.org/audio/141253)

- **Cicada** (species not indicated): Males of this Costa Rican species use tymbals to create loud repetitive call; this usually stimulates other males to call

**D. Tremulation**:

- **Lacewings** (*Chrysoperla spp.*): Sounds produced by trembling legs, then transmitted into plant stems; different species use different patterns
  - **Audio**: [http://www.pbs.org/wgbh/evolution/library/05/2/swf_pop/l_052_01.html](http://www.pbs.org/wgbh/evolution/library/05/2/swf_pop/l_052_01.html)

**II. A body part moved against a fluid (air or water)**

**A. Pulsation**:
• **Long-spined squirrelfish (Holocentrus rufus):** Coral reef fish making swim bladder sounds
  - **Audio:** [http://macaulaylibrary.org/audio/flashPlayer.do?id=116765](http://macaulaylibrary.org/audio/flashPlayer.do?id=116765)

• **Silver croaker (Bairdiella chrysoura):** Temperate bay fish making sounds with swim bladder
  - **Audio:** [http://macaulaylibrary.org/audio/flashPlayer.do?id=116455](http://macaulaylibrary.org/audio/flashPlayer.do?id=116455)

• **Toadfish (Opsanus tau):** Territorial males use swim bladder to make loud sounds during territory defense
  - **Audio:** [http://macaulaylibrary.org/audio/flashPlayer.do?id=112914](http://macaulaylibrary.org/audio/flashPlayer.do?id=112914)

• **Midshipman fish (Porichthys notatus):** Territorial males drum with swimbladder for long periods to attract females and defend territory
  - **Audio and story:** [http://www.npr.org/player/v2/mediaPlayer.html?action=1&t=1&islist=false&id=111273282&m=111273271](http://www.npr.org/player/v2/mediaPlayer.html?action=1&t=1&islist=false&id=111273282&m=111273271)

**III. A fluid moved over a body part**

A. **Vocalization:**

• **Marine toad (Bufo marinus):** Male making repetitive call; air is recycled back into lungs from vocal
sac between calls

- **American alligator (Alligator mississippiensis):** Bellows and groans of adult alligators

- **Tree hyrax (Dendrohyrax arboreus):** Tree hyraxes (small mammals that live solitarily in top of rain forest) call at night to defend territory; callers produce long string of increasingly shrill calls

- **Franquets epauletted bat (Epomops franqueti):** Males call at night while hanging from a branch in large display territory; wings are partially opened and waved with white pouches on shoulders everted for each call. Males synchopate call when females are hovering near or before moving singing perches

- **Hammerheaded bat (Hypsignathus monstrosus):** Species forms leks of up to 150 males that hang from perches, flap wings, and call to females; larynx is enormous and males have flared lips to direct sounds

- **Bearded seal (Erignathus barbatus):** Males of this arctic species produce unworldly sounds underwater (the only sounds as bizarre are the equivalent underwater sounds of male Weddell seals in Antarctic seas)

- **Humpback whale (Megaptera novaeangliae):** Male whale producing complex songs underwater; stereo recording. (To hear and see only whale channel, go to Audio tab in RavenViewer and select Left channel.)

- **Gray wolf (Canis lupus):** Howls of wild wolves produced using larynx

- **Oilbird (Steatornis caripensis):** Vocalizations produced using bronchial syrinx; in addition to squawks, loud clicks produced to echolocate in dark caves where these birds roost and nest

- **Smooth-billed ani (Crotophaga ani):** Vocalizations produced using bronchial syrinx; these communal nesters vary duration and inflection of call depending on context

- **Yellow-naped amazon parrot (Amazona auropalliata):** Mated pair engaged in vocal duet and contact calls produced by a tracheal syrinx
• Audio: [http://www.xeno-canto.org/browse.php?query=amazona+auropalliata](http://www.xeno-canto.org/browse.php?query=amazona+auropalliata) (click on arrow at very left of entry row by recordist David Bradley to play recording)

• **Musician wren* (*Cyphorhinus aradus*): Tropical songbird with tracheobronchial syrinx that alternates between two sides for different frequencies in song

B. **Aerodynamic sounds:**

• **Madagascar hissing cockroach* (*Gomphadorhina portentosa*: Hissing produced by forcing air through spiracles (small holes)

• **King cobra* (*Ophiophagus hannah*: Cobras and other snakes hiss through mouths when threatened
Web Topic 2.7
Animations of Vocalizing Birds

Dr. Roderick Suthers and his colleagues at Indiana University have played key roles in elucidating how birds and mammals produce vibrations and modify them before emission. His group has produced several short video clips that demonstrate important components of vocalization in several songbirds. These videos show many of the basic points discussed in the text. We provide a few introductory notes below for three of these clips. You will need to have Quicktime installed on your computer to play these clips and all clips include concurrent sounds. You can stop the movie at any frame, and use the cursor or the arrows to move it back and forth a frame at a time to see the details of any stage in the process.

Production of vibrations in syrinx

The first two clips (http://www.indiana.edu/~songbird/multi/songproduction_index.html) show the production of sound vibrations in the syrinx of two common North American songbirds. The website provides some information on each clip, and we provide additional commentary below:

- **Northern Cardinal (Cardinalis cardinalis) singing**: This clip begins with a photo of a male cardinal and zooms in to show the location of the trachea, syrinx, and bronchi in the animal. It then moves in further to show a longitudinal section through the syrinx. Note the positions of the lateral and medial labia on each side where the bronchi join the trachea. Note also the position of the cartilage (yellow) just behind each lateral labium. The bird breathes in and out several times (blue) without any vocalizing. Then just before an exhalation, the cartilages on each side are rotated, forcing the lateral labium into the cavity. The cardinal is beginning to sing one of the song syllables shown in the spectrogram in the upper right. The first half of this frequency “down-swept” syllable is produced by the right side of the syrinx; the left side is closed off to any airflow. About halfway through the syllable, the right labium closes off its channel completely, and the left lateral labium opens just enough to vibrate and produce the second half of the syllable. It takes enormous coordination for the bird to produce the beginning of this second section of the syllable at just the right moment and at the exact frequency at which the right side ended its contribution. Inexperienced birds do not always make a perfect union and you can sometimes see the gap between the two parts in spectrograms of their songs. The video continues with several more of these down-swept syllables, the bird inhaling before each, and ends with several final normal breaths without vocalizing. Cardinals nearly always produce high frequencies on the right side of the syrinx, and low frequencies on the left.

- **Brown-headed cowbird (Molothrus ater) singing**: This video assumes you saw in the prior video where the syrinx is located. It thus begins with a glimpse of the whole bird and plays a short song at normal speed. The video then moves directly to the longitudinal section of the syrinx as the bird breathes in and out once and inhales a second time. During the subsequent movie, the sounds heard are at normal frequencies but the time scale has been expanded to show details. On the next inhalation, the bird closes off the right hand side of the syrinx and uses the left lateral labium to produce the first syllable in the song. This is a low relatively constant frequency. The bird inhales again and then produces four notes in rapid succession: the first and third are lower frequencies and are produced on the left side. The second and fourth are higher frequencies and produced on the right side. The bird finishes exhaling, and inhales again. It then produces 5 successive notes, again alternating so that low frequencies are produced by the left side and high frequencies by the right.

Modification of the sounds after production by syringeal vibrations

(http://www.indiana.edu/~songbird/multi/cineradiography_index.html)

This clip shows a male northern cardinal singing a song with upswept syllables. Spectrograms of the radiated song (outside the bird) show a single frequency modulated component and no higher harmonics. However, we know that the signals at the syrinx are periodic but not sinusoidal and thus should contain significant energy in the higher harmonics.
As we can see in this x-ray movie, a singing bird amplifies the fundamental and filters out the higher harmonics by inserting a resonant cavity between the vibrating syrinx and its mouth. This cavity is created by muscular expansion of a pharyngeal cavity and the upper part of the esophagus. Because each syllable’s fundamental frequency changes rapidly, the bird must keep changing the shape and volume of these cavities so that its resonance tracks the changing frequency generated by the syrinx. Clearly, this requires a lot of coordination!

More background on the Suthers’ Lab:

More information on the methods, publications, and approaches of the Suthers' lab group can be found at: http://www.indiana.edu/~songbird/index.html.
A collection of interacting forces can operate as a **linear system** or as a **nonlinear system** (Strogatz 1994; Kaplan & Glass 1995). When it acts as a linear system, its responses are proportional to the amount of change in any constituent force, and if several forces are changed at the same time, the overall response is simply the sum of the responses that we might have seen had we changed each force separately. In general, a mathematical description of the dependence of a linear system’s response to changes in forces will contain no higher order terms like squares or cubes of forces. If our collection of forces acts as a nonlinear system, any or all of these conditions can be violated. Instead of producing a response that is in proportion to a change in one of the forces (the linear expectation), we might suddenly see a response that is totally unrelated to the magnitude of that change. Alternatively, the total response when we change several forces at once might not be the simple sum expected when each force is changed by itself but some complex interaction between the forces. Finally, the mathematical descriptions of nonlinear systems will contain terms that include higher order exponents. Both linear and nonlinear systems can produce sustained oscillations (periodic orbits and limit cycles, respectively): they differ in how they respond to changes in the relative forces and their temporal alignment.

In all animal vibration sources, a number of forces are brought into play at the same time. A stridulating insect uses muscle forces to drag the comb over a sharp edge, and the tensile forces in the teeth of the comb resist being bent until the muscle forces bend the tooth enough for it to escape the edge and snap back to normal shape. Similarly opposed forces are used by terrestrial vertebrates to create vibrations in their respiratory valves. As long as these forces exert their effects out-of-phase, the system will vibrate and produce a sound. As we have seen earlier, most vibrations in animal sound sources are likely to be periodic but nonsinusoidal.

Most animal sound sources have evolved properties that cause them to act as quasi-linear systems over intermediate ranges of force magnitudes. The stable periodic oscillations that result are called **modal** sound production. However, at extreme values, the same systems reveal their underlying nonlinearities. As an analogy, consider a child’s swing. If the parent pushes the swing gently, it will oscillate back and forth at a steady rate. If they push a bit harder, the amplitude of the swing’s motion will increase proportionately and the system will be responding linearly. However, if an unwary parent pushes too hard, the response of the swing is to rotate higher than its attachment point and then suddenly drop straight down (perhaps dumping the child from his or her seat). With an even stronger push, the swing (and child) might rotate right up and around the attachment point to complete a full circle. In either case, the sudden transition from one pattern of response to another as some force is varied is called a **bifurcation**.

Bifurcations can easily occur in the dynamic behavior of animal sound sources (Strogatz 1994; Herzel et al. 1995). A common bifurcation arises when the air pressure in a terrestrial vertebrate’s respiratory tract is altered while the valve is oscillating. At normal pressures and airflows, the two sides of the valve feed back on each other’s movement until they both move at the same frequency. This is called **entrainment**. However, if the air pressure or muscle tensions on the valves are changed sufficiently from normal levels, the two sides of the valve may begin to oscillate at different frequencies that may not be harmonically related. This is known as **biphonation**. Alternatively, the changes in forces may retain the entrainment but cause both sides of the valve to begin moving in a more complex trajectory. If each cycle takes twice as long, the fundamental frequency in the resulting signal will be half of what it was before the bifurcation and one will see twice as many harmonic bands in the spectrogram. This type of effect is called the generation of **subharmonics**. Finally, even more extreme change in air pressure or muscle tension can cause the valve movements to become completely non-repetitive (aperiodic) in their movements. The resulting spectrogram will show a broad smear of energy over a wide range of frequencies. This is known as **deterministic chaos**.
Figure 1: A spectrogram of three consecutive “peow” calls of a male White-fronted Amazon Parrot (Amazona albifrons). While the first call is largely linear with clear harmonics, the last quarter of the second call shows a nonlinear bifurcation with the appearance of subharmonics. The final two thirds of the last call consists of deterministic chaos. This sequence is regularly seen in peow call sequences in this species. Nonlinear components like this appear to be a common feature of wild parrot vocalizations (© Jack Bradbury).

Animals can either adjust their sound producing forces to generate periodic (but usually nonsinusoidal) sound waves, or they can push the sound source out of the modal range and trigger one or more bifurcations. Most songbirds appear to go out of their way to keep sound production in the modal range. In contrast, parrots routinely include bifurcations in their calls (Fletcher 2000). Non-modal vocalizations have been described in a variety of mammals (Riede et al. 1997; Riede et al. 2000; Riede et al. 2001; Fitch et al. 2002; Riede et al. 2004; Riede et al. 2005). Human speech is usually modal, but humans can, and often do, push their sound producing systems out of modal patterns and into nonlinear states (Berry et al. 1994; Herzl et al. 1994; Herzl et al. 1995; Herzl and Knudsen 1995; Steinbeck and Herzl 1995; Berry et al. 1996; Fletcher 1996; Mergell and Herzl 1997; Mergell et al. 1998; Mergell et al. 1999; Mergell et al. 2000; Berry et al. 2001; Gerratt and Kreiman 2001; Hanson et al. 2001; Berry et al. 2006). Canids and primates often push their laryngeal vibrations into biphonation states creating more complex sounds with two apparently independent fundamentals (Volodina et al. 2006). Certainly more examples of these shifts will be found in other taxa as researchers are aware that they should look for them.

References Cited


Web Topic 2.9
Radiation Efficiency and Sound Radiator Size

Introduction

The size of an animal’s sound emission organs can seriously limit the size of the sound wavelengths that it can radiate efficiently: small animals cannot radiate high amplitude sounds with wavelengths much larger than they are. Here we provide a more detailed explanation for this general finding.

Frequency scaling and $ka$

The sound pressure generated by a sound source typically increases as either the frequency being produced or the size of the sound source is increased. Since the two parameters have similar effects, acousticians use a scaled version of their product to characterize the dependence of sound pressure on their values. This scaling involves converting the frequency, $f$, in Hz (cycles/sec) into a spatial measure, the wave number $k$, which is the number of cycles at the given frequency found in $2\pi$ meters of the relevant medium. This can be computed by dividing $2\pi$ by the wavelength in meters, $\lambda$, of the frequency of interest. The product we need is that between $k$ and some appropriate measure of the size of the sound source. For a spherical monopole, this would be the radius of the sphere, $a$, in meters. For a dipole such as a flat disk vibrating back and forth along a line perpendicular to its surface, it would be the radius, $a$, of the disk.

In general, $ka$ is a useful measure of the ratio between the size of a sound source and the wavelength of the sound that it is generating. For example, $ka$ for a spherical monopole is equal to $2\pi a/\lambda$, or $\pi d/\lambda$ where $d$ is the diameter of the sphere. When $ka = 1$, the wavelength is then about 3 times larger than the diameter of the sphere; when $ka = 3$, then the sphere’s diameter and the wavelength are approximately equal.

Sound pressure and $ka$

As we shall see below, there are two concurrent processes that can limit the amplitude of a sound wave as a function of $ka$. One is simple and applies to all basic sound types: as we increase either frequency or the size of a sound source, all other factors being equal, we will increase the potential pressure of the sound radiating from the source. For $ka > 1$, this potential pressure is what is realized. However, when $ka < 1$, a second process comes into play, which reduces the efficiency of radiation and thus the resulting sound pressure. We outline this secondary process for different radiating geometries in subsequent sections.

It should be noted at this point that all other factors may not be equal. For example, the amplitude of the sound pressure produced by a sound source will also depend on the magnitude of the sound source’s movement. This magnitude can be measured as the maximum change per half cycle in the radius of a pulsing spherical monopole or the maximum distance moved by a central point on a dipole per half cycle. In addition to increasing the resulting sound pressure, increasing the distance moved will also increase the costs of sound production. If the sound producer is energy-limited, they might elect to reduce the amplitude of the movement of their sound producing structure whenever they increased the frequency to keep costs constant. The predicted increase in sound pressure when an animal increases its frequency might thus be countermanded by a concurrent decrease in oscillator amplitude. All things would thus not be equal when compared to the initial condition, and thus the sound source might not produce a sound with a different amplitude as predicted when $ka$ increases. In the discussions of how $ka$ affects radiation efficiency below, we shall assume that all other parameters remain constant as $ka$ is varied.

Efficiency of monopoles with varying $ka$

Consider a completely spherical monopole that is expanding and contracting to produce sinusoidal waves at a single
frequency. As the sphere expands, it pushes against the layer of medium immediately surrounding it. This layer of medium responds to the sound source force in two ways: it will be compressed, raising the local pressure inside the layer, and at the same time, the entire compressed layer may begin to move outwards (a mass flow). This expanding first layer will collide with the next closest layer of medium and the same process will be repeated: the higher pressure of the first layer will force molecules in the second layer to move, resulting in both compression and mass flow away from the sound source. As the sphere completes the expansion part of the cycle and begins to contract, a similar process occurs but in reverse with the mass flow now moving towards the sphere’s retreating surface.

As successive layers are compressed (or rarefied), three properties of the medium resist changes in molecular velocities and thus contribute to the local acoustic impedance. All media are viscous to some degree and this exerts a frictional drag on molecule motion. This resistive acoustic impedance hinders mass flow and results in compression and the local build up of pressure. Because molecules must move faster at higher frequencies, resistive acoustic impedance is higher at higher frequencies. A second relevant property is the stiffness of the medium: as force is exerted on a given layer of molecules, the stiffness exerts a counter-force that resists both the compression and the mass flow of the medium. The longer that a given force is exerted, the more the counter-force. As a result, the hindering effects of stiffness are greater for low sound frequencies than for high ones. Finally, all molecules have inertia that resists changes in their velocities. This inertia will hinder frequent changes in molecular direction more than infrequent ones; inertial effects thus become greater as the sound frequency is increased. Close to a sound source, stiffness and inertial effects rise and fall in phase with the mass flow but out-of-phase with the resistance effects. In addition, only the frictional effects contribute to molecular concentration and rarefaction, and thus sound pressure depends only on the resistive impedance. The effects of the medium stiffness and molecular inertia are thus lumped into a single component called the reactive acoustic impedance. (Note: The terms resistive and reactive impedance initially came from analyses of electrical circuits. It turns out that the resistance to flow in a liquid has similar behavior to the resistance to electrical current in a wire. Similarly, stiffness in a fluid behaves similarly to a capacitance in an electrical circuit, and inertia produces fluid behaviors analogous to inductances. Capacitance and inductance together produce the overall reactive impedance in an electrical circuit and the overall term is used by analogy in this more mechanical situation. See Fletcher 1992 for a good discussion of the utility of electrical terms in analyzing the behavior of mechanical systems.)

The values of the resistive and reactive impedances in an acoustic system depend on the relative sizes of the sound source and the wavelength of sound that it is generating, and thus on $ka$. To see why, we note that molecules contributing to the mass flow are increasingly diluted with non-contributors at greater distances from the source. As a result, the mass flow velocity falls off with the square of the distance from the source. In practice, most of the mass flow is limited to a blanket of medium around the sound source that is about one wavelength in thickness. The volume of medium participating in the mass flow thus depends on the thickness of this blanket and the surface area of the sphere (a larger sphere with its larger surface area moves more total medium per cycle).

When the diameter of the monopole is small relative to the wavelength of the sound being generated ($ka \ll 1$), the volume of air moved per unit of surface area of the sphere is large. Most of the force exerted by the sphere’s small surface will be needed simply to overcome the inertia and stiffness effects and get the large mass of medium moving. This leaves little additional force to condense or rarely adjacent layers of medium. As a result, sound pressures generated when $ka \ll 1$ will be small.

There are two ways to increase $ka$. Let us first hold the diameter of the sphere constant and gradually increase the sound frequency being produced. This will decrease the wavelength of the sound and thus the thickness of the blanket of medium that must be moved. The volume of medium that must be moved per unit surface area on the sphere will accordingly decrease. Initially, stiffness effects will be high (due to the low frequencies) and increasing the frequency will increase inertial effects. Overall reactive impedance will thus initially rise with increasing frequency. As the frequency is increased further, stiffness effects decrease faster than inertial effects continue to increase. The result is that the reactive impedance first increases and then decreases as we increase the frequency of the sound being generated. At the same time, the decreasing thickness of the medium blanket will reduce the fraction of the force exerted by the sphere that is needed to generate the mass flow and allow for greater compression and rarefaction. This
will generate a higher sound pressure variation at the source. Eventually, at high enough frequencies \((ka > 1)\), it will take so little force to move the thin blanket of medium that all of the sphere’s force will go into compression and rarefaction and the efficiency of radiation will approach 100%. Further increases in frequency will not affect the efficiency of the radiator significantly.

What if we hold frequency constant and vary the diameter of the sphere? When the wavelength is much larger than the sphere’s diameter, any increase in that diameter will reduce the volume of medium that must be set into motion per unit area of sphere surface. This will allow a larger fraction of the applied force to be used for compression and rarefaction and thus create a larger pressure amplitude of the sound. Again, as \(ka\) is increased by increasing the sphere size, efficiency of radiation gradually arises and asymptotes towards 100% as \(ka\) becomes larger than 1.

We can look at this process in a different way by noting that when \(ka << 1\), the resistive impedance of the blanket of medium surrounding the sphere (which is the only part of the impedance that can increase sound pressures), is very small when compared to the characteristic acoustic impedance (which is entirely resistive) of the medium far from the source. Because of the difference in resistive acoustic impedances between this blanket and layers of medium further from the source, we would not expect much sound energy to be transferred into the more distant medium. As \(ka\) is increased, the resistive impedance of the blanket increases until it is essentially identical to the characteristic acoustic impedance of the medium. At this point, efficiency of radiation from the blanket to more outlying layers is nearly 100%.

It is useful to consider some quantitative limits on these efficiency effects. When the ratio between the diameter of the monopole and the sound wavelength is less than 1/3 (e.g., \(ka < 1\)), radiation efficiency will increase monotonically with increases in the ratio. When the ratio equals 1/3, the resistive and reactive impedances around the monopole are roughly equal and efficiency is 50% of what it could be. To achieve an efficiency of 90% or more, the wavelength must be the same size or smaller than the diameter of the sound source.

![Figure 1](image.png)

**Figure 1**: Change in impedances (relative to the characteristic impedance of the medium) and efficiency of spherical monopole sound source as a function of \(ka\). Red line shows resistive
impedance of medium surrounding monopole, which is essentially equal to the efficiency of the sound pressure output. Blue line shows reactive impedance which governs mass flow around sound source. Note that both axes are logarithmic scales. (Computed from equations in Fletcher 1992.)

Efficiency of dipoles and quadrupoles with varying $ka$

Dipoles (which oscillate back and forth along a single dimension) and higher order sound sources (which oscillate with trajectories using two or more dimensions) experience losses in efficiency in part due to the reasons outlined above for monopoles. However, an additional factor reduces efficiency in these sound sources even further. This is called acoustic short-circuiting.

Because of the way a dipole operates, it generates a condensation on one side of the sound source at the same time that it creates a rarefaction on the other. If the dipole is oscillating slowly enough (e.g., at a low enough frequency), a condensation generated at one end might have sufficient time to propagate to the opposite end of the dipole and interfere negatively. One way to avoid this short-circuiting is to use sufficiently high frequencies that a condensation cannot get to the rarefaction before it is complete and already radiating into the medium. Another way is to insert the dipole in a baffle so that condensations have to travel all the way to the edge of the baffle and back before they can reach a rarefaction. The further a condensation has to travel to reach a rarefaction, the weaker its amplitude; it will thus have only a minor effect on the rarefaction and short-circuiting will be minimal.

Suppose that the shortest distance through the medium between the radiation site for a condensation and that for a rarefaction is $D$. It will take $D/c$ seconds for a condensation to reach a rarefaction site (where $c$ is the speed of sound in the medium). A sinusoidal frequency $f$ takes $T = 1/f$ seconds to complete one cycle, and therefore $T/2$ seconds to complete producing a condensation on one side and a rarefaction on the other. To minimize short-circuiting, it must be the case that the time required to create and radiate a condensation or rarefaction ($T/2$) is less than the time required to travel between the sites ($D/c$). If $T/2 < D/c$, it follows that $Tc < 2D$. Since by definition the wavelength of the sound $\lambda = Tc$, short-circuiting can be minimized when $\lambda < 2D$, or rewriting, when $2D/\lambda > 1$. If the dipole is an insect wing that is vibrating up and down, $D$ is half the diameter (to traverse the distance from the center of the wing to its outside margin) plus another half diameter (to traverse the opposite side of the wing). $D$ is thus roughly equal to one diameter of the wing. To minimize short-circuiting, twice the ratio between diameter and wavelength must be greater than one. Using the measure $ka$, where $a$ is the wing radius, this is equivalent to requiring that $ka > \pi/2 = 1.57$. Other geometries of the dipole will give slightly different values, but the general result is the same.

When short-circuiting is combined with other sources of radiation inefficiency, it takes a higher value of $ka$ to produce a given efficiency of sound radiation with a dipole than it would to achieve that same efficiency with a monopole. Put another way, the reduction in sound pressure amplitude when the wavelength being generated is much larger than the sound source is much higher for dipoles than for monopoles: monopoles are more efficient sound radiators than dipoles when $ka < 1$. At the same time, the fraction of the force exerted on the medium that goes into mass flow is much higher for dipoles than for monopoles at any given $ka < 1$. When $ka > 3$, a dipole produces twice the sound pressure that a similarly sized monopole would produce at the same frequency. This is because the two sides of the dipole effectively act as individual monopole sound sources once the short-circuiting is minimized.
Putting the frequency dependences together

We now want to combine the effects of increasing the potential sound pressure as $ka$ increases with the efficiency costs when $ka < 1$. All other factors being equal, sound pressure from a monopole increases with the square of $ka$ when $ka < 1$, and as a linear function of $ka$ when $ka > 1$. For a dipole, the exponents all increase by one: for $ka < 1$, pressure increases with $ka$ to the third power, and for $ka > 2$, it increases with the square of $ka$. For quadrupoles, just add one to each exponent again.

Directionality, type of sound source, and frequency

A spherical monopole radiates equally in all directions: in the absence of any nearby boundaries, sound pressures at all points equidistant from the monopole should be equal. A dipole has a much more complex sound field. All points equidistant from the two ends of the dipole axis will exhibit little if any sound pressure because of negative interference by waves from the two ends that are out-of-phase and of roughly equal amplitude. On the other hand, sound pressures will be maximal as one moves away from the dipole along the axis of its motion. The resulting sound field can be described as having two large lobes within which are all points with sound pressures above some minimal value. Each lobe is anchored at one end of the dipole and has its long axis parallel to the axis of the dipole. At low values of $ka$, the lobes are wide and nearly circular. As $ka$ is increased, the lobes become much narrower and additional lobes at other angles relative to the dipole may appear. Quadrupoles also generate sound fields with lobes but usually begin with more than two lobes at low $ka$, and add more and narrower lobes as $ka$ is increased. Fletcher (1992) provides illustrations showing lobe patterns at various $ka$ values for different types of sound radiators.

Figure 2: Change in impedances (relative to the characteristic impedance of the medium) and efficiency of a dipole sound source as a function of $ka$. Solid red line shows resistive impedance of medium surrounding dipole incorporating effects of short-circuiting. Solid blue line shows reactive impedance, which is similar to that of monopole but somewhat frequency dependent for $ka > 1$. Note larger gap between reactive and resistive impedances for dipole when compared to the monopole in Fig. 1 for $ka < 1$, and higher final values for $ka$ just >1. (After Fletcher 1992.)
Implications for animal sound production

While the pressure wave generated by a sound source falls off with the reciprocal of distance traveled, the mass flow around the sound source decreases with the reciprocal of the square of the distance. In practical terms, detectable mass flow is limited to about one wavelength from the sound source, whereas the pressure wave will be detectable much further away. Animals trying to communicate at distances many times the size of their own bodies must therefore rely on detecting the pressure waves and not the mass flow. As we have seen, this imposes serious constraints because sound pressures will be produced very inefficiently unless the wavelength is the same size or smaller than the sound-producing organ. The smaller the animal, the smaller any sound producing organs will be, and this limits small animals communicating over significant distances to using high frequencies. If the animals need to produce sounds with \( ka < 1 \) for other reasons, a monopole design would be better than a dipole or quadrupole design.

Animals that use sound for close-range communication could use either the pressure wave or the mass flow to detect and identify the signal. If they use the latter, lower frequencies would give greater range, and a dipole or quadrupole would be a better choice of sound source than would a monopole.

Further reading


Web Topic 3.1  
Transfer Functions

Introduction

We are often interested in how a system affects a signal as it passes into and then out of the system. For example, what happens to the sounds produced by our larynx as they pass through our pharynx and nasal cavities and then emerge from our mouth and nostrils? What happens to a light signal as it passes through water? What happens to a bird’s song as it propagates through a leafy forest? In each case, the signal is passing through a black box and emerging, usually transformed, on the other side.

If the response of the black box to introduced signals is linear and time-invariant, we can compute a transfer function for it that allows us to predict the waveform and spectrogram of any signal after it has passed through the box.

The response of a black box is linear if it meets the principle of superposition: that is, suppose that inserting a simple signal $x_1$ into the box produces an output signal $y_1$, and inserting a signal $x_2$ produces an output signal $y_2$. The box is linear if inserting the sum of the two input signals, $x_1 + x_2$, produces the output signal $y_1 + y_2$. A system is nonlinear if the output contains products of inputs, (e.g., $x_1 \times y_2$) or higher powers of terms (e.g., $x_1^2$ or $y_1^3$). Many natural systems are linear at least over some range of inputs; however, at very low or high values of $x$, the same system may become nonlinear.

A black box is time-invariant if inserting $x_1$ produces $y_1$ whether we test it now or sometime later. Again, many natural systems are time-invariant for short periods, but, later on, inserting $x_1$ into the system might produce a different output $y_3$. As an example, sound propagation near the ground early in the morning will follow the same rules until the sun begins to heat the ground. As the ground warms up, it heats the air just above it, and the patterns of sound refraction change. Thus the transfer function for sound propagation will not be time-invariant when we compare early morning to mid-morning testing periods.

Transfer functions

Suppose we limit our attention to the range of inputs and time intervals for which the response to signals of a black box of interest is linear and time-invariant. The transfer function that we can compute will have two parts: (1) the frequency response (which measures changes in the amplitude of any given frequency in the signal as it passes through the box); and (2) the phase response (which measures changes in relative phase of each frequency as it passes through the box). These two components of the transfer function are usually summarized as graphs called Bode plots. In both graphs, the x-axis is frequency.

In the frequency response Bode plot, the vertical axis indicates the relative change in the amplitude of each frequency as it passes through the box. One could use a linear scale in which the vertical axis coordinate indicated the fractional change in the input magnitude of a frequency. On this scale, a value of 1 would mean no change, a value of 0.5 would be a halving of the input amplitude, and a value of 2.0 would mean that the box amplified that frequency to twice its input value. In practice, frequency response Bode plots use a logarithmic dB scale in which 0 means no change in amplitude, −6 dB means a halving of the input amplitude, and +6 dB means a doubling of the input amplitude.
Figure 1: Bode plot for frequency response (red line) of a sample black box. Dashed line indicates no change in amplitude of an input frequency at the output of the black box. Where the red line is above the dashed line, relevant frequencies are amplified during transit in the box; where the red line is below the dashed line, relevant frequencies have reduced (filtered) amplitudes.

In a phase response Bode plot, the vertical axis indicates the change in relative phase of each frequency component. The scaling is usually linear with 0 indicating no phase change, and plus and minus values indicating phase shifts measured either in degrees (360° for one full cycle) or radians (2π radians for one full cycle).

Figure 2: Bode plot for phase response (blue line) of a sample black box. Dashed line indicates no change in relative phase. Frequencies for which the blue line is above the dashed line are advanced during passage in box, and those for which the blue line is below the dashed line are retarded relative to the reference value.

The use of transfer functions in a linear system is shown in Figure 3:
Figure 3. Application of transfer functions to real input signal with waveform shown in (A). Latter is first broken down into (B), its power spectrum (amplitude versus frequency), and (C), its phase spectrum (phase relative to one component, here marked as a dot, versus frequency). The amplitude of each
frequency component in the power spectrum of the input signal is then increased or decreased according to the frequency response in (D) to produce the output signal power spectrum (F). Similarly, the phase of each component in the input power spectrum of (C) is advanced or retarded according to the phase response graph (E). The frequencies in the output power spectrum with their adjusted amplitudes (F) are then added together using their new phases (G) to produce the output waveform (H).

This procedure allows one to predict the power spectrum, phase spectrum, and waveform of any input signal as long as the black box remains linear and time-invariant and the transfer functions have been measured previously for that box.

**Measurement techniques**

How does one measure the transfer functions for a black box? In most cases involving animal communication, we are only interested in the frequency spectrum of the output signal and can ignore its phase spectrum and waveform. This simplifies our task considerably.

The basic approach is to broadcast a signal of a known power spectrum through the black box and compare the power spectrum of the output signal to that of the input signal. In principle, any test sound could be used. In practice, most natural sounds lack certain frequency bands, which makes it impossible to know what would have happened to them during propagation had they been present. Frequency responses are usually measured using signals that cover all frequencies at approximately similar amplitudes. There are four basic ways this can be done:

- **Multiple frequency testing**: Here, one introduces a single pure tone of a known frequency at a known amplitude, and measures the amplitude of this tone as it emerges from the black box. The frequency response is then constructed by repeating this process for many different frequencies and combining them to create the Bode plot.
  - **Pros and cons**: This method can be used in noisy environments because one can make each tone long enough to be detected at some point above the background. On the other hand, this method is the most tedious of the four listed here. In addition, long duration signals may create standing waves due to interference between outgoing sound and returning echoes.

- **FM signals**: In this method, one plays back a frequency-modulated signal of constant amplitude that sweeps through the bandwidth of interest.
  - **Pros and cons**: A single FM pulse can provide a good initial sense of the frequency response of a black box very quickly. However, if the box has resonant properties with a high Q, the speed at which the signal sweeps through the frequencies may be too high to excite the natural resonance modes. Some frequency response measurement instruments actively slow down the sweep when the output indicates a resonant mode is being measured. As a minimum rule, the rate of sweep for FM test signals must be adjusted so that the duration of the FM signal is as long, or longer than the reciprocal of the bandwidth covered by the sweep.

- **Impulses**: The Fourier decomposition of an instantaneous pulse signal consists of all frequencies at equal amplitude at once. Inserting a pulse into a black box and measuring the relative amplitudes of all emerging frequencies can produce the full frequency response very quickly.
  - **Pros and cons**: While the theoretical instantaneous impulse consists of all frequencies at equal amplitude, real impulses are finite in duration and this creates biases in the amplitudes of different frequencies that may need to be corrected later. The recording instrument must also have a fast enough response to record the output sound accurately, and, in noisy environments, it may be difficult to pick out a very brief impulse from background noise.

- **White noise**: White noise, like impulses, theoretically consists of all frequencies at similar amplitudes. Exposing the black box to a segment of white noise should produce an output sound with a power spectrum that is a
reasonably good replica of the frequency response of the system.

- **Pros and cons:** In practice, one can only generate relatively even amplitudes of all frequencies within a given bandwidth. Also, white noise is more difficult to generate than “pink noise,” in which the relative amplitudes of the component frequencies decrease with increasing frequency.

More elaborate devices and techniques have been developed for measuring frequency responses, and a wide variety are discussed on the Internet. See also Wikipedia “Bode Plot.”

**Potential problems**

The major problems with measuring frequency responses in nature are noise and non-linearities. If noise is to be considered part of the black box response, then including it in the measurements is appropriate. However, in many cases we want to know how a specific process such as refraction, reflection, or scattering varies with habitat. In this case, we want to measure the frequency response of the system without noise. There are several ways to do this. One way is to measure noise without test signals and subtract average values from the frequency spectrum measured with test signals. There are also sophisticated instruments and statistical methods to extract noise from spectra once recorded.

Non-linearities are to be expected in nearly all natural systems. In many cases, exposing the system to test sound amplitudes above the range in which it responds linearly will generate harmonics of the test signal. Where one is using pure tones, this is easy to detect and correct. On the other hand, if natural signals occasionally reach levels at which the system responds non-linearly, then knowing that this results in the generation of harmonics of the input signal is useful. Other types of non-linear responses may be more difficult to quantify and characterize.
Web Topic 3.2
Dispersive Sound Propagation

Introduction

Most animal sounds and human speech consist of many different frequencies summed together. Each frequency has a given amplitude and phase relative to the others, and it is the particular combination of frequency, amplitude, and phase values that results in the waveform of the signal. In unbounded air and water, complex sounds propagate as a unit—all frequency components move at the same velocity and this preserves their initial alignment right to the ear of the receiver.

However, the component frequencies in a complex sound do not always propagate at the same velocity in all contexts. In certain situations, they travel at different speeds and thus get out of alignment. This changes both the spectrographic structure and the waveform of the signal. A medium in which different frequencies propagate at different velocities is said to be dispersive. The basic principles outlined here can be applied to light wave and sound wave propagation.

Group versus phase velocity

When a complex sound is generated in a dispersive medium, the entire ensemble of component frequencies initially radiates away from the source as a unit. However, because component frequencies propagate at different individual speeds, known as phase or wave velocities, the alignment of the components also changes as the signal propagates. If the medium were not dispersive, the part of the signal hosting the peak amplitude (the “envelope peak”) would propagate at the same speed as each component frequency. In dispersive media, the changes in component alignment due to different phase velocities causes the speed at which the signal peak moves, known as the group velocity, to be different from that of any component’s wave velocity. In some media and contexts, the group velocity of propagation for the signal is slower than that of any component phase velocity; in other contexts, the group velocity exceeds that of any frequency’s phase velocity. Signal propagation in dispersive environments is thus highly dependent on the relevant context.

Examples of dispersive propagation

Consider the following contexts, in all of which sound propagation is dispersive:

- **Solid bars and rods**: Phase velocities in solid bars and rods increase with the square root of the component frequency and group velocities are greater than phase velocities.

- **Surface waves on Earth**: Both Rayleigh and Love surface waves on the Earth’s surface are dispersive. In contrast to solid bars and rods, higher frequencies have phase velocities that are slower than those for low frequencies. Group velocities also tend to decrease with increasing frequency.

- **Surface waves on water**: The water’s surface is kept flat in calm conditions by two forces: gravity and surface tension (a result of the attraction between water molecules). Low-frequency disturbances (< 14 Hz) in the water’s surface are restored to a flat condition by gravity forces; propagation is dispersive with higher frequencies (as long as they are less than 14 Hz) having slower phase velocities. High-frequency disturbances (>14 Hz) are restored by surface tension and phase velocities increase with frequency (just the opposite as for gravity effects). Group velocities are lower than phase velocities for low-frequency disturbances, and higher than phase velocities for high-frequency disturbances. Intermediate frequencies have the lowest phase velocities with the minimum occurring at the breaking point of 14 Hz. For this single frequency, group and phase velocities are equal. Dispersion can be reduced in shallower waters.

- **Bending waves in plants**: Bending waves in plants move the plant tissue back and forth along a line
perpendicular to the direction of propagation (which is along the stem or branch). While these waves are thus similar to transverse waves, they differ from them in that they cause a rhythmic bending back and forth of the entire stem. Phase velocities in plant stems are similar to those in other solid rods in that they increase with the square root of the component frequency. Group velocities are greater than component phase velocities.

- **Waveguides**: Although waveguides can produce reduced spreading loss during propagation, and thus technically foster long-range communication, they are dispersive and thus can change the structure of propagated sound signals significantly. Both phase and group velocities increase with frequency in waveguides.

### Waveform changes with dispersion

How the waveform of a complex signal changes as it propagates dispersively depends on the signal’s initial composition and the phase and group velocities in the propagating medium. Some movies of dispersive propagation of simple waveforms can be found at:

- Phase Speed and Group Speed—Dr. Dan Russell, Kettering University: [http://www.kettering.edu/~drussell/Demos/PhaseGroupSpeed/PG.html](http://www.kettering.edu/~drussell/Demos/PhaseGroupSpeed/PG.html)
- Group and phase velocity simulation: [http://publicliterature.org/tools/group_and_phase_velocity/](http://publicliterature.org/tools/group_and_phase_velocity/)

### Follow-up references


Web Topic 3.3
Animal Communication and Anthropogenic Noise

Introduction

Human activities add anthropogenic noise to all environments, including, of course, those in which animals are attempting to communicate using sound. There are four basic ways that this noise may impact animals and their communication systems:

- **Distraction**: Although the frequencies of anthropogenic noise may not overlap with an animal’s own sound communication system, loud and sudden noises may distract it from its necessary activities. Distracting sounds can interrupt foraging, reproduction, growth, territorial defense, predator vigilance, proper hygiene, sleep, provisioning, nursing, etc.

- **Masking**: In this case, the frequency distributions of anthropogenic noise and an animal’s signals are sufficiently overlapping that some signals are masked. Receivers thus cannot detect or evaluate signals and normal communication exchanges are interrupted.

- **Stress**: Loud and disturbing noises may induce hormonal and neural responses in animal receivers that are physiologically expensive.

- **Damage**: The most typical damage is injury to sensitive auditory organs. At extreme amplitudes, ambient noise can destroy other tissues as well.

The effects wrought on animals by any of these perturbations can be short-term or long-term. An animal may be able to recover its hearing acuity after a short bout of damage-level noise, but not after continued exposure over extended periods of time.

Research approaches

Current research to assess the effects of anthropogenic noise on natural populations of animals takes several tacks:

- **Audiograms and ambient noise measurements**: This approach involves measuring the frequency composition and amplitudes of ambient noise and comparing it to the range of frequencies that focal animals can hear. Unless noise occurs at tissue-damaging levels, distraction, masking, or stress are only likely when there is overlap between noise frequencies and auditory sensitivities.

- **Behavioral shifts**: Here, one compares behaviors of animals in the presence and in the absence of anthropogenic noise. If the noise is sufficiently intermittent, one can compare behaviors when it is absent to when it is present. If the noise is persistent, one must find a control study site as similar to the noisy site as possible except for the presence of the noise. Behavior shifts that might be monitored when noise is present include any cessation of normal behaviors, shifts in animal signal frequencies or amplitudes to reduce masking, altered activity time budgets in the presence of noise, relocation to less noisy sites for signaling, etc.

- **Health and demographic shifts**: Again, here, one needs to compare a site with noise to one without noise to identify changes due to noise. Long-term health and demographic shifts might include higher mortality rates due to increased disease or predation, reduced reproductive success, lower recruitment of dispersing young, greater emigration of all age classes, etc.

- **Physical damage**: Autopsies of dead animals that are known or suspected to have had exposure to high noise levels can usually reveal which organs or tissues, if any, are damaged, and estimate whether this damage could have been caused directly or indirectly by ambient noise.

Current research examples
• **Cross-taxonomic reviews**: General reviews of the impacts of anthropogenic noise on ecosystems and conservation strategies can be found in Barber et al. 2010 and Laiolo 2010.

• **Terrestrial invertebrates**: Despite the vast numbers of terrestrial insects that communicate with far-field sounds, very little effort has yet been expended on the possible impact of anthropogenic noise on insect behavior. It is worth noting that most insects use signal frequencies higher than the more common causes of loud far-field anthropogenic noise on land (e.g., vehicular traffic). In fact, insects such as cicadas, crickets, and katydids are themselves the major source of ambient noise at these frequencies. Insects, scorpions, and spiders that communicate using substrate signals propagating in plants, webs, or the ground may be more susceptible to disturbance since they use lower frequencies that are frequent components of anthropogenic substrate noise. However, this has yet to receive much research attention.

• **Aquatic invertebrates**: Because sound attenuation is so much lower in water than in air, aquatic anthropogenic noise can carry long distances. In addition, humans generate some extremely loud sounds in water. Examples include the use of air-guns for seismic mineral exploration, pile driving, long-range military and climate monitoring communication signals (ATOC, ACME), and sonar sounds. Cargo ships, Jet Skis, racing boats, and military vessels all produce loud noise in water. Again, the impact of any of this noise on aquatic invertebrates has been little studied. Lovell et al. (2005, 2006) measured audiograms in marine prawns and argued that they may well be affected by ambient shipping and related human noises. The communication sounds of lobsters are inherently low volume and are likely to be masked by ambient sounds except very close to a signaling animal (Patek et al. 2009). A large number of other crustacean species on reefs are sensitive to sound and use this sensitivity both during larval settlement and as adults to avoid predators (Simpson et al. 2011). In both cases, anthropogenic noise may aggravate and hinder the normal uses of sounds in the animal’s biology.

• **Fish**: Although they can be highly sensitive to low amplitude sounds, the hearing of most fish, which are hearing generalists, is limited to frequencies considerably less than 1 kHz. Hearing specialists such as carp, goldfish, and catfish can hear up to several kHz, with even higher sensitivities. Whereas hearing generalists can probably hear racing boat noise only at short ranges, hearing specialists can detect it at distances of several hundred meters (Amoser et al. 2004). A similar pattern shows up for air-gun noise (Mann et al. 2007). There is growing concern about the effects of human noise on fish (Slabbekoorn et al. 2010). Aquatic anthropogenic noise has been shown to perturb the normal behavior of several fish species (Popper 2003; Popper et al. 2003; Purser and Radford 2011), induce hormonal responses indicative of stress (Smith et al. 2004; Wysocki et al. 2006), and, in the case of air-guns and loud shipping noise, produce short-term (Smith et al. 2004) and permanent damage to fish ears (McCauley et al. 2003; Wysocki and Ladich 2005). Hearing specialists appear particularly vulnerable to ear damage (Scholik and Yan 2002a,b).

• **Amphibians and reptiles**: Frogs, toads, lizards, and turtles are all potentially vulnerable to both airborne and substrate propagated anthropogenic noise. Some species of frogs increase calling rates, and consequently reduce both evening chorus durations and seasonal calling activity in the presence of anthropogenic noise (Sun and Narins 2005; Kaiser and Hammers 2009; Kaiser et al. 2011). Other species increase the pitch of their calls or reduce calling altogether when exposed to nearby traffic noise (Lengagne 2008; Parris et al. 2009). Frogs and turtles may also be exposed to aquatic noise. Studies of aquatic anthropogenic noise in a New York estuary frequented by marine turtles found significant amplitudes of noise within the known auditory sensitivities of the turtles (Samuel et al. 2005). Whether this noise causes any level of effect remains unstudied.

• **Birds**: Lab studies clearly show that current levels of ambient noise can significantly mask the communication signals of birds (Lohr et al. 2003; Pohl et al. 2009). Species vary in their responses to this problem. Nightingales increase their song amplitude and decrease the number of notes per song in high ambient noise (Brumm and Todt 2002; Brumm 2004). House finches, song sparrows, blackbirds, and great tits living in noisy urban environments sing louder and faster, include fewer notes, and shift the minimum frequencies of their songs to higher values than birds outside urban areas (Slabbekoorn and Peet 2003; Fernandez-Juricic et al. 2005; Slabbekoorn and den Boer-Visser 2006; Wood and Yezerinac 2006; Bermudez-Cuamatzin et al. 2009; Mockford and Marshall 2009; Nemeth and Brumm 2009). In a study comparing relatively silent and noisy but otherwise matched sites, male ovenbirds showed significantly lower pairing success in the noisy sites (Habib et al. 2007).
European robins close to urban areas decrease singing during noisy daytime hours and increase singing at night (Fuller et al. 2007). Anthropogenic noise also seems to affect bird dispersal and settlement patterns. Birds that normally sing at higher frequencies are more likely to settle in urban areas than those with lower frequencies (Hu and Cardoso 2009). Shifts in songbird dialect distributions may also arise from changing urbanization (Luther and Baptista 2010). European birds show significantly lower nesting densities in zones along highways and the effect increases with the level of vehicular traffic (Reijnen et al. 1996; Reijnen et al. 1997). The densities of nesting passerines in otherwise similar boreal forests were 50% higher when no gas compressors and other noisy facilities were nearby (Bayne et al. 2008). Reviews discussing other possible effects of anthropogenic noise on birds can be found in Katti and Warren (2004), Patricelli and Blickley (2006), and Warren et al. (2006).

- **Terrestrial mammals:** There are currently few data characterizing the impact of anthropogenic noise on terrestrial mammals. Captive marmosets increase the amplitude and duration of their calls when exposed to continuous white noise (Brumm 2004). Ground squirrels near air turbines increase the amplitudes of their alarm calls and shift energy to higher harmonics; those near turbines are also more wary given the shorter distances that alarm calls can be detected (Rabin and Greene 2002; Rabin et al. 2003; Rabin et al. 2006). Captive bats actively avoid foraging in high levels of ambient noise (Schaub et al. 2008) and wild bats tend to leave to forage later when an adjacent music festival is in session (Shirley et al. 2001). Anthropogenic noise propagating in the substrate may interfere with elephant seismic communication (O'Connell-Rodwell et al. 2001).

- **Marine mammals:** Perhaps because of the ubiquity and amplitude of anthropogenic noise in the oceans, much is now known about the potential and realized impacts of such sound on the behavior of marine mammals. Field measurements indicate that air-guns used in geological exploration, pile driving, and intensive boat traffic can produce sounds loud enough to mask marine mammal communication and echolocation sounds at considerable distances (Goold and Fish 1998; Erbe and Farmer 2000a,b; Southall et al. 2003; Boebel et al. 2005; David 2006; Madsen et al. 2006; Jensen et al. 2009; Bailey et al. 2010; Di Iorio and Clark 2010; Brandt et al. 2011; Gedamke et al. 2011). While some anthropogenic sounds appear to have little effect on the behavior of nearby marine mammals (Croll et al. 2001; Costa et al. 2003; Lemon et al. 2006), other studies show increases in call duration (Miller et al. 2000; Foote et al. 2004) and call rate (Buckstaff 2004) or a temporary cessation in acoustic communication when noisy vessels are nearby (Lesage et al. 1999). Manatees shift normal behavior patterns in the presence of high levels of noise (Miksis-Olds and Wagner 2011), and a variety of pinnipeds are known to avoid noisy areas (Gotz and Janik 2010). Captive animals exposed to high levels of anthropogenic sound show significant nervous and immune system effects (Romano et al. 2004). Beaked whales (Ziphiidae) have been shown to make the deepest and longest dives known among cetaceans (Tyack et al. 2006); examination of beached Ziphiids after military tests of intense sonar have suggested that these anthropogenic sounds may disrupt the slow ascent necessary to prevent gas-bubble formation and thus cause tissue damage and death (Cox et al. 2006). Whether this is true is still under investigation. General reviews on the impact of anthropogenic noise on marine mammals can be found in Richardson et al. (1995), Simmonds et al. (2004), Nowacek et al. (2007), and Weilgart (2007).

**Mitigation**

The research so far on animal responses to anthropogenic noise indicates that at least some frogs, birds, and mammals can modify their sound signals or signaling schedules to minimize masking. However, very loud or sudden sounds can cause stress and damage to animal receivers, either directly (as with fish ears) or indirectly (as with beaked whales). Current efforts are underway to design quieter shipping, and large scale monitoring schemes using underwater sound (e.g., ATOC, ACME) are the subject of considerable current discussion. However, the popularity of air-guns for seismic exploration in marine environments makes abatement of this source of noise unlikely. Similarly, vehicular traffic in urban areas is equally hard to reduce given current human activities. On the other hand, the use of all-terrain vehicles, snowmobiles, and Jet Skis in national parks and refuges could be reduced considering that sufficient research shows that such activities affect the survival and reproduction of protected species. Clearly, more research is needed to fill gaps in our knowledge about specific taxa and levels of effect.
Literature cited


Linear kinetics

Linear kinetics apply when a force is applied for a reasonable period in a single direction. Rules for linear kinetics are insufficient to explain processes such as sound where the relevant forces reverse directions rapidly. However, linear processes form the basis from which acoustic kinetics are derived. Some useful definitions for linear kinetics are:

- **Basic kinetics**: Suppose a force, $F$, is applied to a static object for a given time period, $t$. If the force is sufficiently strong, it will begin to accelerate the object and will continue to do so until the fixed time period is completed. During this time, the object will move a given distance, $d$, from its starting point. The average velocity, $v$, achieved by the object is equal to the distance moved divided by the time that the force was applied: $v = \frac{d}{t}$. Velocity and distance traveled are thus proportional to each other for a fixed time interval.

- **Mechanical impedance**: If the object is resistant to being accelerated by this force, it will not have moved very far during the time $t$. This resistance to being accelerated is called the mechanical impedance, $z$, of the object. It can be measured by dividing the force, $F$, by the velocity achieved, $v$: $z = \frac{F}{v}$. The further the object moves under force $F$ in time $t$, the lower the impedance.

- **Work**: The total work, $W$, done by moving the object is the product of the force $F$ and the distance moved $d$: $W = Fd$.

Vibratory kinetics

Sounds and other vibratory processes experience a recurrent reversal of the direction of the relevant forces. The response of the system being forced may differ depending upon the frequency of these reversals. This requires some modifications of the definitions used in linear kinetics as follows:

- **Vibratory mechanical impedance**: When a sinusoidally varying force, $F$, acts on an object and induces a sinusoidally varying velocity, $u$, of that object, the mechanical impedance is defined as $z = \frac{F}{u}$. The variable $z$ is measured in N·s/m. If the phase of the induced velocity differs from that of the force, $z$ is given as a complex quantity (i.e., includes both real and imaginary components).

- **Characteristic acoustic impedance**: Characteristic acoustic impedance is a property of a medium. When an acoustic plane wave travels in an effectively unbounded medium (e.g., in a large volume of air or water), the acoustic pressure, $P$ (measured in Pascals), at each point is proportional to the average particle velocity, $u$ (measured in m/sec). The characteristic impedance of an unbounded volume of medium is then defined as the proportionality constant $Z_c$. Thus, $P = Z_c u$, or it can be rewritten as $Z_c = \frac{P}{u}$. Here, $Z_c$ is measured in units Pa·s/m. Note that the characteristic acoustic impedance does not depend on frequency. In air at 20° C, $Z_c$ is around 420 Pa·s/m, whereas in water it is about 1500 Pa·s/m.

- **Acoustic impedance $Z_a$**: In bounded contexts, such as at the opening of an animal’s ear canal, inside the canal, at an eardrum, or in the terrestrial vertebrate middle ear bones that conduct sounds into the inner ear, the ability of a localized region of medium to propagate an oscillating wave will be different from that in a large unbounded volume of the same medium. The response will differ depending on the area, $S$, of responding medium exposed to the oscillating sound pressures. During one cycle of the oscillating pressure, medium in the area $S$ will be moved a distance $d$ in the direction of the force for $t$ seconds. The total volume moved will be $S \times d$ and the average volume velocity will be $U = S \times \frac{d}{t}$. Since $\frac{d}{t} = u$, the particle velocity, we can also write this as $U = S \times u$. The acoustic impedance of this patch of medium is then $Z_a = \frac{P}{U}$. This measure of impedance is measured in units of Pa·s/m$^3$. Because there is usually a phase difference between the pressure variations and the associated volume velocity, $Z_a$ is written as a complex number.

**Specific acoustic impedance $Z$**
This measure controls for the area of medium responding in the bounded case above. It thus gives one a measure similar to the characteristic acoustic impedance for an unbounded volume of medium, but this time is based on the response of a bounded volume. The specific acoustic impedance is $Z_s = S \times Z_a$ and has the same units ($\text{Pa} \cdot \text{s/m}$) as characteristic acoustic impedance.

Levers as impedance transformers

For both linear and vibratory systems, levers (or their fluid equivalents) are used to convert one ratio of force/velocity (e.g., one impedance) into another such ratio (another impedance). To see the general process, consider a solid beam resting on a balance point (fulcrum) at some location between the two ends of the beam. When one applies a force and moves one end of this lever a certain distance, one does work. By the conservation of energy principle, the same amount of work must simultaneously be done at the other end of the beam. Suppose the fulcrum is not at the midpoint of the beam so that one end of the beam passes through a larger arc than the other end when moved. For the work at the two ends to be the same, it follows that the force applied to the large-arc end must be less than that applied at the other end. By positioning the fulcrum off-center, a lever thus becomes an impedance transformer: large displacements at low force at one end can be turned into low displacements at high force at the other end, and vice versa.

Types of mechanical levers

Type 1 mechanical lever

This is a classical see-saw device. One balances the beam on a fulcrum at some location between the two ends of the lever. If the fulcrum is placed exactly in the middle of the beam, exerting a force and causing movement on one end of the beam (effort end) is replicated exactly (but in the opposite direction) on the other end of the beam (load end). There is no mechanical advantage to this geometry. If the fulcrum is moved toward the load end of the beam, a large movement at little force on the effort end results in a small movement but high force on the load end. Thus, someone on the effort end can use a type 1 lever to lift a heavy load more easily than if they try to lift it without a lever. The claw on the rear of a hammer works this way to remove nails. Scissors or pliers consist of a double-beam type 1 lever.

Type 2 mechanical lever

Here, the fulcrum is placed at one end of the beam. The effort is applied to the other end, and the load is placed at some point in the middle. A classic example is a wheelbarrow. A nutcracker is a two-beam type 2 lever system.
Type 3 mechanical lever

As with a Type 2 lever, the fulcrum is again placed at one end of the beam, but now the load and effort points are reversed, with the load on the opposite end of the beam from the fulcrum and the effort applied at some intermediate location. Many muscles that operate animal limbs work as type 3 levers. A pair of tweezers is a two-beam example.

Folded see-saw mechanical lever

This is a variation of a type 1 lever in which the beam is folded at the fulcrum so that the two resulting beam segments (called arms) maintain a fixed angle between them. When one arm is forced to rotate around the fulcrum in a particular direction, the other arm rotates in the same direction. While the two arms rotate at the same angular velocity (degrees/sec), the tip of the longer arm sweeps through a longer path than does the tip of the shorter arm. The mechanical advantages of the folded type 1 lever are exactly the same as for an unfolded one: if one arm is twice as long as the other, it will rotate through an arc twice as long as that for the smaller arm, and the shorter arm will exert twice the force of the longer arm.
Hydraulics

Hydraulics function in a manner analogous to solid levers. The difference is that fluids are used to exert the forces. In a typical hydraulic system, two surfaces of different area are connected by a relatively incompressible fluid inside a tube or cavity which is also incompressible. When pressure (force/unit area) is applied to one of these surfaces, the total force applied is equal to the product of the pressure and the area of the surface. This same force is applied by the fluid to the second surface. If the force generates movement, both surfaces will move in the same direction, a distance $d$, and thus both will do the same work as expected for a lever. Note, however, that if the two surfaces have different areas, the smaller surface will experience a higher pressure (e.g., the same force divided by a smaller area) than will the larger surface. Thus, this type of device can act as a transformer for acoustic impedance.

A similar principle is used to create hydraulic brakes and jacks, but in these cases, the effort is applied to the smaller surface and the larger surface carries the load.
Two cylinders are linked by a tube at the bottom, filled with some incompressible fluid, and each is equipped with a piston (the two surfaces). When one pushes down on the small piston, the fluid in the cylinders is compressed and the pressure is increased. The value of this pressure depends upon the ratio of the force exerted and the area of the small piston. Because pressure must be the same throughout a static fluid, the large piston now experiences an increased pressure from the fluid. The total force it experiences is the product of the fluid pressure and its surface area. Because that area is larger than the area of the small piston, the large piston experiences a greater force and thus a mechanical advantage for lifting heavy weights. As it moves upward due to this force, it brings the total volume of the cylinder fluids back to the value they had before the small cylinder was depressed. This lowers the fluid pressure back to its starting point and the large piston stops moving. Because each increment of movement by the large piston increases cylinder volume faster than similar movement by the small piston, the distance traveled by the large piston is shorter than that moved by the small piston. In the end, the result is just like a type 1 mechanical lever: one end of the system moves a long distance at low force, and the other end moves a short distance at high force.

**Catenary levers**

When a cable is attached at both ends to some fixed points and allowed to sag, the points close to the attachment experience the greatest force, since the whole cable is pulling down on them, and the least mobility, since they are closest to the attachment point. Points in the middle experience the least force and have the greatest mobility. A similar effect occurs when a circular membrane is attached at its margins and is forced to bend inward or outward. The high force and low mobility at the margins of the cable or membrane are transformed into low force and high displacements at the center. A small force applied to the center of the cable or membrane results in major displacement at the center, but a smaller displacement at considerable force near the margins.
Auditory levers

Animal ears often face impedance mismatch problems. For example, all terrestrial vertebrate inner ears are filled with fluids that must be set in motion to stimulate the auditory sensory cells. The high acoustic impedance of the fluid-filled inner ear requires a source that is high pressure, low velocity, and low displacement to set the fluids into motion. The available stimuli are sound waves in air that are low pressure, high velocity, and high displacement. Without some form of acoustic impedance matching, most of the incident sound energy would be reflected away from the animal's tympana. Terrestrial vertebrates use type 1 and/or type 2 mechanical levers, hydraulic levers, and catenary levers to achieve effective impedance matches. Some examples follow.

Frogs and toads

A typical frog ear is diagrammed anatomically on the left and mechanically on the right:

Frog and toad middle ears contain three articulated cartilaginous or bony elements. The **extrastapes** (or extracolumella) acts approximately like a folded type 1 lever. One arm of the lever connects with the inside of the tympanum. An ascending process extends from one side and attaches to the skull. This provides a fulcrum for this element. The other arm of the lever attaches to the outside end of the **stapes** (or columella). The stapes is a long thin element that attaches on its inner end to the middle of the **footplate** element. The latter is hinged to the inner ear capsule. Together, the stapes and hinged footplate form a type 3 mechanical lever. The footplate presses in on the oval window and transfers its motions to the fluid-filled cavity on the other side of the membrane. The two lever systems are given different colors in the diagram on the right. White circles indicate articulations, the vertical dashed line shows the tympanum, and green triangles indicate fulcra. Given the orientation of the two sequential levers, the footplate and oval window membrane move outward when the tympanum moves inward. The tympanum and oval window thus move 180° out-of-phase in frogs and toads (Mason and Narins 2002; Werner 2003). The mechanical advantage of this ossicular system in male bullfrogs is about 5.7:1. Given the relative sizes of the tympanum and oval window in male bullfrogs, the
hydraulic advantage is 50:1. An additional mechanical advantage may accrue from some flexibility and springiness in the extrastapes element (Mason and Narins 2002), and there is some evidence of a catenary advantage due to the bending of the tympanum in frogs (Moffat and Capranica 1978).

**Birds and reptiles**

Reptiles and birds also use a three-ossicle lever system, but the ossicles are arranged to form a single type 2 lever (Saunders et al. 2000). The extracolumella is a rod-shaped element anchored to the skull at the inner edge of the tympanum. The other end extends over and past the center of the inside surface of the tympanum. The second element, the columella, is a long thin bone that articulates with an intermediate point along the length of the extracolumella. On its other end, it attaches to the footplate which, again, sits over the oval window. It is held in place by soft tissues and by a ligament. The diagram below shows a simplified lizard middle ear:

![Diagram of lizard middle ear](image)

In birds, the extracolumella spreads three arms instead of one over the inner surface of the tympanum; two of these can be hinged to form fulcra on the periphery of the tympanum (Saunders et al. 2000). In addition, the columella in birds connects to the extracolumella at an acute angle (less than 30° relative to the plane of the tympanum). This change in angle significantly changes the directions of motion of the lever elements. In addition, the ligament holding the footplate over the oval window may act as a hinge on one side (as with frogs and toads), adding a second lever system (Gaudin 1968).

The mechanical advantage of the middle ear ossicles in living reptiles and birds is typically about 2:1 to 4:1. The hydraulic advantage given the difference in tympanum and oval window size ranges is 13:1 in collared lizards, and from 11:1 to 40:1 in birds (Saunders et al. 2000).

**Mammals**

In mammals, three ossicles are again used to link the tympanum to the oval window. However, these have evolved independently from those acquired by amphibians, reptiles, and birds, and use a different leverage action:
The first ossicle (the **malleus**) has a long arm that is attached to the inside of the tympanum. The malleus is then linked to a second ossicle (the **incus**) at an angle so that together they form a folded type 1 lever. An inward movement of the tympanum causes the folded lever to rotate counter-clockwise, and the internal arm of the incus forces the footplate (the **stapes**) to press in on the oval window. In contrast to frogs, the tympanum and oval window are in-phase in mammal middle ears. Over a wide range of terrestrial mammals, the ratio of effective tympanum to oval window size is roughly constant at an average of 19:1. The mechanical advantage created by the middle ear bones is also relatively constant with an average around 2.4:1 (Rosowski 1994; Hemilä et al. 1995).

**Impedance and transformer ratios**

The goal of an acoustic impedance transformer is to convert a propagating sound wave from one set of pressure and particle velocities to another set. Airborne sounds propagate at low pressures and high particle velocities and displacements; inner ear fluids propagate sounds at high pressures and low particle velocities and displacements. The transformer must thus be able to vibrate in concert with airborne waves on its input side and with water waves on its output side. The hydraulic lever system is a pressure converter: low pressures hitting the tympanum are concentrated onto the smaller oval window resulting in higher pressures. The ratio of output to input pressures from this mechanism is equal to the ratio of the effective surface area of the tympanum ($A_1$) to that of the oval window ($A_2$). The ossicle chains can usually be reduced to an equivalent type 1 lever where the length of the beam between the input end and the fulcrum is $L_1$, and that of the output side is $L_2$. Since the ossicular levers both increase the output pressure, and decrease the output velocities, the ratio $L_1/L_2$ affects the acoustic impedance twice and is thus squared in computations. The resulting ratio between the output and input impedances of such an acoustic transformer is (Dallos 1973):

$$\frac{Z_2}{Z_1} = \frac{A_1}{A_2} \left(\frac{L_1}{L_2}\right)^2$$

Thus if the average hydraulic ratio for mammal ears is 19:1 and the average lever ratio is 2.4:1, then, excluding any additional effects such as catenary leverage, the output impedance of the middle ear system at the oval window will be 109 times that experienced by the tympanum. The combination of terms on the right hand side of the equation is called the **impedance transform ratio** and varies from less than 100 in some birds and mammals to 500 or more in some lizards and frogs (Mason et al. 2003). These higher ratios do not necessarily imply improved hearing, and in some cases may be due to multiple functions of the ear structures (e.g., male call radiation from the tympanum in bullfrogs).

**Literature cited**


Web Topic 3.5
Auditory Amplification

Basic principle

Although individual auditory mechanoreceptors can be extremely sensitive, there are distinct advantages to cohesive stimulation of adjacent cells with similar characteristic frequencies. When multiple cells are stimulated, multiple nerves will be activated and jointly send impulses to the animal’s brain. Small differences in characteristic frequencies of receptor cells can also be averaged out by pooling the responses of multiple cells. Both arthropods and vertebrates have hit on a similar way to accomplish this coordination: motile sensory cells.

Near-field sound propagation (including transmitted sounds inside an inner ear) involves a tidal oscillation back and forth of a medium. The dendrites (in arthropods) or stereocilia (in vertebrates) of the sensory cells are then stimulated either directly by these fluid movements or indirectly through motions induced in an overlying membrane, otolith, or other structure. Once stimulated, some of the sensory cells respond by changing their shape (in mammals) or physically moving their dendrites or stereocilia (in other vertebrates and arthropods) in concert with the sound oscillation. This active movement by the sensory cells increases the amplitude of motion in the overlying medium or structures, and this, in turn, generates even greater stimulation of the sensory cells. The resulting feedback loop amplifies very small sound levels and provides very high sensitivities. It allows adjustable tuning of the resonant frequencies of the entire ensemble by varying how and when motions are induced. In both arthropods and vertebrates, motion in the feedback loop is most often found when the animal is exposed to very low level sounds; at high levels, active sensory cell movement is minimal and the system responds largely according to its physically determined natural modes.

One consequence of this active feedback loop is that tiny amounts of random noise may be perceived as a very low amplitude signal. This will cause the motile sensory cells to move in an attempt to amplify a very faint signal even though no sound is really present. The result is spontaneous motion of the sensory cells. In vertebrates, these movements generate artifactual sounds in the inner ear that can be propagated back to the ear drum and detected as otoacoustic emissions. They can also stimulate the sensory cells and produce the sensation known as tinnitus, or “ringing of the ears.”

Another consequence of the feedback loop is that the auditory system is no longer linear. In fact, the acute sensitivity of many auditory systems to very low-level sounds results from changes in the resonant properties of the system due to active feedback. Anesthetized or recently dead animals do not show these nonlinearities: their ears act as simple mechanical systems with natural modes set by their physical properties.

Auditory amplification in fruit flies

The common fruit fly (Drosophila melanogaster) detects near-field sounds using a pair of antennae on their heads (Göpfert and Robert 2002). Each antenna hosts a plume-like terminal segment called the arista that is anchored immovably to an elliptical segment called the funiculus. The funiculus articulates with a third segment, the pedicel, which is firmly attached to the fly’s head. The funiculus has a small hook on its proximal end that fits into an invagination in the pedicel wall. The tip of the hook then connects via a flexible hinge to the pedicel. The pedicel is a hollow cavity filled with several hundred sensory scolopidia forming the Johnston’s organ. The scolopidia are divided into two groups; the dendrites of each group attach via thin threads to one side of the funiculus hook.
Because the arista is located at a point off of the natural axis of motion of the funiculus, movements of the medium caused by near-field sound propagation generate a torque on the arista and an alternating rotation of the funiculus-arista in concert with the near-field oscillations. These movements alternately stretch one group of scolopidia while compressing the other. Stimulated scolopidia then send nerve impulses back to the fly’s brain that indicate the presence and frequency of the incident sound.

Based on the physical mechanics alone, the fly antenna has a resonant frequency of about 800 Hz. However, living flies can change the resonant properties of the system by active movement of stimulated scolopidia in a feedback loop (Göpfert & Robert 2003): lower ambient sound intensities result in lower resonant frequencies. The graph below illustrates this.
In addition to changes in the resonant frequency due to scolopidial movements, the amplitude of antennal oscillations actually increases at low stimulus levels:

As with vertebrate otoacoustic emissions, fruit fly antennal scolopidia move spontaneously in silence, and these spontaneous movements can be exaggerated by giving the fly drugs that break the feedback loop by blocking output from the sensory cells. These spontaneous movements disappear except for those induced by thermal molecular motion if the fly is dead:
In Drosophila, the feedback loop is mechanical; there are no efferent nerves from the brain that regulate the strength of the feedback (Kamikouchi et al. 2010). Similar studies have shown that other insects, such as mosquitoes, employ sensory cell motion to create auditory amplification (Göpfert et al. 1999; Göpfert and Robert 2000, 2001; Robert and Göpfert 2002; Robert 2005).

**Auditory amplification in vertebrates**

Auditory amplification has now been demonstrated in each of the major terrestrial vertebrate groups. Fish may also exhibit this behavior, but it remains to be examined. Among the terrestrial vertebrates, mammals have two kinds of cochlear hair cells: a single row of inner hair cells that runs the entire length of the organ, and 3–5 additional parallel rows of outer hair cells, as shown below.
Most afferent innervation (from sense organ to central nervous system) in the mammalian ear involves synapses with the inner hair cells; the outer hair cells mostly receive efferent input (from central nervous system to sense organs). At any point along the length of the cochlea, outer hair cells are tuned to the same characteristic frequencies as adjacent inner hair cells. When both are stimulated, the outer hair cells change shape rhythmically to amplify stimulation of the nearby inner hair cells (Robles and Ruggero 2001). Unlike Drosophila, where the feedback is solely mechanical, the efferent nerves in mammals regulate the amount of movement and thus control the feedback level (Fettiplace 2006; Ashmore et al. 2010). This feedback loop is mostly used to improve reception of higher frequencies and low stimulus levels. As with other feedback systems, otoacoustic emissions are well known in mammals and are often used to study the functioning and health of mammalian hearing.

In amphibians, reptiles, and birds, auditory amplification is achieved by active waving of the hair cell stereocilia (Manley 2000; Manley et al. 2001; Fettiplace 2006; Strimbu et al. 2010). Stereocilia are linked to each other, making coordinated movement feasible. Most taxa have at least two types of hair cells, with one type being the main source of motile responses.

Lizards, like other taxa with active auditory amplification, produce spontaneous otoacoustic emissions that can be monitored at the eardrum in silent conditions. Lizards that have a single tectorial membrane over their hair cells produce spontaneous sounds with a few dominant frequencies. However, some lizards, such as geckoes, divide up their tectorial membrane into a pinnate shape rather like a fern leaf (Manley 2000, 2002). Whereas the otoacoustic emissions from lizards with a single continuous tectorial membrane show frequency spectra with 1–2 main peaks (see Figure 6A below), lizards with subdivided tectorial membranes show many different small peaks (see Figure 6B below),

**Figure 5:** Diagrammatic cross-section through mammalian cochlea (inner ear). Sensory cells are suspended between three parallel cavities filled with fluid and sandwiched between a tectorial membrane and basilar membrane. Sounds propagated in fluid channels move the tectorial membrane relative to the basilar membrane and thus bend stereocilia on hair cells. While both inner and outer hair cells are stimulated by sounds, 95% of sound input to the central nervous system comes from the inner hair cells. Outer hair cells receive input from the central nervous system, and largely function by changing shape in oscillating fashion to amplify stimulation of adjacent inner hair cells at low sound intensities.
apparently reflecting different frequencies of spontaneous movement in different zones of the inner ear.

![Figure 6](image)

**Figure 6**: Frequency spectra of spontaneous otoacoustic emissions of (A) a dwarf tegu lizard (*Callopistes maculates*), and (B) a gecko (*Gekko gecko*). (After Manley 2002.)

### Literature Cited


Web Topic 3.6
Animations of Vertebrate Ears

Introduction

The terrestrial vertebrate ear is a complicated device that converts sounds from a low acoustic impedance to a high acoustic impedance (function of the middle ear), and then breaks the converted complex waveforms down into their component frequencies (function of the inner ear). Static images of the ear do not do justice to either process. Luckily, a number of websites provide excellent animations of the middle ear or the inner ear or both in operation. Below, we list some suggested sites. All focus on the mammalian (especially human) ear, but the processes are largely the same for reptiles, birds, and mammals. Frog ears are a bit different in geometry, but the principle is also the same.

Suggested sites

- **University of Pennsylvania Health System**: This site provides an excellent animation showing both the middle ear and the cochlea in humans. A voice track explains what is being shown.
  

- **Human Anatomy (McKinley and O’Laughlin)**: This website is a complement to a human anatomy text. It shows a nice color animation with associated voice track of the human ear at each stage. There is even a quiz to take afterward.
  

- **Rockefeller University Lab of Sensory Neuroscience**: This site provides a graphic simulation of how different sounds cause the traveling wave in the cochlea to produce maxima at different locations.
  
  http://www.rockefeller.edu/labheads/hudspeth/graphicalSimulations.php

- **Wada Laboratory (Japan)**: This site provides some interesting animations of the different components typical in vertebrate ears.
  
  http://www.wadalab.mech.tohoku.ac.jp/contents-e.html - middle-ear

- **Other sites**: At present, we could find no suitable animations of Drosophila ears, but there are numerous other videos of vertebrate hearing mechanisms available on Vimeo and YouTube.
Web Topic 3.7
Measuring Auditory Resolution

Introduction

The text lists multiple reasons why different animals may differ in the limits and resolutions of their hearing organs. Airborne, waterborne, and substrate-propagated signals confer different constraints on suitable receiver mechanisms, and body size imposes limits at all stages of the communication process. There is also the problem that animals usually want to extract more than one type of information from receipt of a signal: improving ear resolution for one type of information invariably reduces resolution for another type of information. The impact of ambient noise on sound communication depends critically on the range of frequencies that a receiver can hear and on the spectral distribution of energy in the noise. Whether one is interested in the consequences of anatomical differences, the physics of sound signal exchanges, the physiology of hearing, or the behavioral ecology of sound communication, knowledge about the limits and resolutions of auditory organs in particular species can be very important. How can one measure auditory performance, and which measures are most useful in comparing taxa (such as animals versus humans)?

Level of measurement

Measuring the limits and resolution of an animal’s acoustical abilities is a challenging task. A first step is deciding at which stage in the sound perception process one should make the measurement. Options include measurements at:

- **Sensory cell stimulation**: In many animals, one can use neurobiological methods to record the slow depolarization of auditory sensory cells during and after stimulation. In vertebrates, the summed slow potential changes of many hair cells creates a microphonic potential, which can be used as an index to test for bandwidth limits and sensitivity of the ear. While receptor potentials can assess the tuning responses of individual cells accurately, it can take many such probings to establish the bandwidth limits of an organ. In addition, the resolution of individual sensory cells is only one part of the final resolution seen at the level of behavior.

- **Otoacoustic emissions**: In insects such as fruit flies and mosquitoes and all terrestrial vertebrates, active amplification of sounds results in the generation of otoacoustic sound emissions that can be monitored to characterize auditory sensitivities. These emissions are currently used as diagnostics for testing human hearing (see the article, Otoacoustic Emissions: [http://emedicine.medscape.com/article/835943-overview](http://emedicine.medscape.com/article/835943-overview)), and have been used in various animal studies similarly to assess limits and resolutions.

- **Sensory nerve activity**: Recordings of single fibers or entire nerve bundles can be used to monitor the responsiveness of an ear to different frequencies and amplitudes. These use standard neurobiological techniques and have the advantage that one is looking at the coded information being made available to the brain.

- **Brain activity**: Recordings at the level of the brain have been widely used in both invertebrate and vertebrate acoustic studies. Where the sequential pathways are well understood, researchers can track successive analyses by the animal.

- **Behavior**: The behavior of the receiver is the ultimate test of whether a given sound has been perceived or not. Unfortunately, a lack of response could be due either to an inability to perform a discrimination or a lack of motivation to respond. Signal detection methods allow one to separate these two factors (see Web Topic 8.10). The best context is conditioned learning (psychoacoustics) in which captive subjects are rewarded for correct discriminations and are thus highly motivated to detect and classify signals if they can.

Not all levels are feasible for all taxa. For example, researchers have had very poor luck applying conditioned learning techniques to lizards, and there are ethical reasons to exclude certain salient species such as higher primates from destructive neurobiological techniques. The result is that the mix of levels that has been studied varies for different
taxa. The fact that measures obtained at different levels may produce different results even in the same species needs to be kept in mind when making cross-taxon comparisons. The strongest results are those that are consistent across levels in the same species.

**Basic measures**

Below, we define the logic and goal of some commonly applied auditory measures. The list is not meant to be exhaustive, as researchers are continually devising new features to measure or new ways to measure well-studied features.

**Audiograms**

The audiogram (also called audibility curve) is a graph with stimulus frequency as the horizontal axis (usually on a log scale), and the minimal stimulus amplitude (usually in dB re some reference) required to evoke a response at a given frequency on the vertical axis. Audiograms (red line below) are usually U-shaped, meaning that very low and very high frequencies require a high amplitude stimulus to evoke any response:

![Audiogram graph](image)

The frequency that requires the lowest amplitude stimulus (the low point of the U-shaped curve), is often called the **best frequency**. The threshold amplitude required at the best frequency is a commonly invoked measure of the overall **sensitivity** of the auditory system. The lower and upper frequencies in the audiogram at which thresholds are some specified number of dB above that at the best frequency define the **frequency range** of the system, and the difference in frequency between them is the system’s **bandwidth**. Note that the bandwidth is equivalent to measuring the $Q$ (quality) factor for a resonant system: the narrower the bandwidth, the higher the associated $Q$ (see Figure 2.34 in Chapter 2). $Q$ values may be reported as a measure of the tuning of an auditory system.

**Critical ratios**

A critical ratio is a measure of how much greater the amplitude of a single frequency tone must be to elicit a response in a given level of ambient noise. In most birds and mammals, critical ratios increase with the frequency being considered (usually 2–3 dB more tone power is required for each doubling of frequency). Thus, higher frequency signals must be received at higher amplitudes to be detected against noise. In both birds and mammals, there are exceptions to this rule. Parrots, for example, show a decrease in critical ratio with increasing frequency at least up to the higher frequencies used in their long range contact calls (Dooling et al. 2000).

**Critical bands**

Tonotopic inner ears are often considered as a bank of filters: any filter in the sequence produces a response only when a component frequency in a sound stimulus falls within the **critical bandwidth** (usually abbreviated to **critical band**) of that filter. Critical bands can be measured by varying the bandwidth of the noise masking a pure tone, by examining critical ratios at different frequencies, or by varying the width of a “notch” in the noise used to mask a single pure tone. Critical bands are a measure of the frequency resolution of the ear, since an ear with many narrow filters will
differentiate between more frequencies than will one made up of only a few wide filters. In general, critical bandwidths increase with the center frequency of the relevant filter. This is what one would expect given Weber’s Law (see Chapter 8 and Web Topic 8.6). In humans, each successive filter representing one critical bandwidth occupies about 1 mm of basilar membrane along the cochlea. Since frequencies are distributed logarithmically along the basilar membrane in tonotopic vertebrate ears, a constant 1 mm generates larger bandwidths as one moves to the high frequency end of the cochlea. Again, parrots are an exception: they have their most narrow critical bandwidths at the intermediate frequencies used in their contact calls (Dooling et al. 2000).

**Frequency discrimination**

The goal here is to measure the minimal difference that can be discriminated between the frequencies of two pure tones. It can be measured by decreasing the difference in frequency between successive stimuli until no differential response is obtained. Another method habituates the subject with a constant frequency signal and looks for the minimal difference in frequency that causes a renewed (dishabituated) response. As a rule, the minimal differences in frequencies of pure tones that can be differentiated are 10–20 times smaller than the critical bands obtained by masking a single pure tone with noise. Frequency discrimination, like critical bands, tends to follow Weber’s Law with threshold differences ranging from 0.5–4% of the compared frequencies for birds and mammals.

**Comodulation masking release**

If the band of noise masking a pure frequency tone is amplitude-modulated with the same pattern as another band of noise centered on one or more other regions of the audible spectrum, the critical ratio required to detect the noise can decrease by 10 dB or more. This is because the ear and brain can identify general noise by its shared modulation pattern and correct for its presence when masking a pure tone. This phenomenon has been demonstrated in both mammals and birds.

**Intensity discrimination**

This simple test is similar to frequency discrimination: two pure tones of similar but slightly different amplitude are made increasingly similar until differential responses to them are lost. This test can be performed at all levels to identify the degree to which higher level processing augments or decreases intensity resolution.

**Temporal integration**

The ability to detect a sound depends on the ear receiving a minimal amount of energy. For sounds less than about a quarter of a second in duration, the same energy could be supplied at low amplitude over a long period or high amplitude over a short period. This principle sets the stage for plotting the minimum amplitude that is required to obtain a response to a sound of a given duration. The rate at which threshold amplitude decreases with increasing signal duration is a measure of the integration constant of the system. Birds and mammals show similar constants for this measure.

**Gap detection**

This measures the ability of a receiver to detect a break between two consecutive sounds. Where the consecutive sounds are both noise, birds and mammals show similar abilities; where they are pure tones, humans and birds perform similarly if the two sounds have similar frequencies, but birds do better than humans when the sounds have different frequencies.

**Duration discrimination**

The goal here is to identify the threshold difference—usually scaled as a relative percentage difference between the means of the stimuli—in the durations of two stimuli.

Excellent reviews of these and additional measures can be found in Dooling et al. (2000) and Dooling (2004) for birds.
and Long (1994) for mammals.

Pattern measures

Animal signals usually involve patterned structure in the distribution of energy across the available frequencies and time segments in the signal. There is thus considerable research interest in assessing how well receivers of different species can classify patterns into a priori categories and discriminate between slightly different patterns in two or more signals. Because so much is known about human pattern processing of sounds, human capabilities are often used as a standard reference and animal abilities are then examined in comparison. Some pattern measures that are of recent research interest are detailed below.

Acoustic scene analysis

Humans routinely parse an acoustically complicated environment into individual acoustic objects that can be tracked individually by varying attentive focus (Deutsch 1999). This is known as “acoustic scene analysis.” A number of studies have now shown that birds perform similar acoustic classifications (Gentner and Hulse 2000), and can use these to track individual objects in noisy contexts (Hulse et al. 1997; Wisniewski and Hulse 1997; MacDougall-Shackleton et al. 1998; Hulse 2002; Appeltants et al. 2005).

Missing harmonics

As we note in the text, most animal sounds are likely to consist of harmonics; it is only with great effort and anatomical specialization that animals can produce single pure frequencies. Human speech is a classical example: each person’s speech consists of a harmonic series with different vowels having different relative amplitudes of the same component harmonics. When humans are presented with two tones that could be harmonically related, (e.g., with frequencies having a ratio of 3:2), they typically perceive a third tone equal to the presumed (but actually absent) fundamental. This perception of the “missing fundamental” can be used to create the illusion that a low frequency is present when in fact the sound only contains higher harmonics. Examples include the perception of bass notes by organs and by small home speaker systems. Laboratory experiments have shown that starlings also infer “missing fundamentals” when presented with suitable pairs of pure tone frequencies (Cynx and Shapiro 1986). This raises the interesting but unstudied question of whether this, or similar auditory illusions, might be exploited by animals to extend their perceived frequency ranges despite physical constraints on sound production.

Consonance and dissonance

When two sound frequencies both stimulate the same critical band in a human cochlea, the perception is of a “rough” and “unresolved” sound (Plomp and Levelt 1965). Such sounds are said to be dissonant. Maximal dissonance occurs when one frequency is just 25% of the critical band higher in frequency than the other. If the two frequencies are sufficiently different that they stimulate different critical bands, the mixture of two pure tones will sound smoother and is said to be consonant.

The sounds of animals (and of most musical instruments) are not pure tones: they are usually complex sounds containing many harmonically related frequencies. When two complex sounds with fundamental frequencies that excite different critical bands are played together, they are no longer necessarily perceived as consonant. In fact, there is considerable variation in the perceived consonance depending upon the ratio of the two fundamental frequencies (Rossing 1990; Deutsch 1999). The most consonant complex sounds have fundamentals that have a frequency ratio of 2:1 (called octaves). The next most consonant combinations are sounds that form a perfect fifth (fundamental frequency ratio of 3:2) or a perfect fourth (ratio of 4:3). As observed by Pythagoras before 500 BC, fundamental frequency ratios that require higher integer values, e.g., major sixths (5:3), major thirds (5:4), minor sixths (8:5), etc., are perceived as increasingly dissonant despite the absence of overlap of their fundamentals in the same critical band. Two non-exclusive explanations have been proposed. One is that while the fundamentals of complex sounds may not fall in the same critical band, the logarithmic scaling of frequency along the cochlea can easily result in higher harmonics of the two sounds falling into the same critical band (Terhardt 1974a,b). The number of harmonics that
overlap this way will generally increase as the fundamental frequency ratios require higher integers. The second explanation is that the waveform of complex sounds will be more regularly periodic for pairs of sounds with low integers relating their fundamental frequencies (Tramo et al. 2001; Tramo et al. 2005). This periodicity is known to be conveyed to the brain by the auditory nerves in humans and cats, and is highly correlated with perceptions of consonance. It becomes less noticeable with a lower frequency of repeating for mixtures of complex sounds having higher ratios like sixths and thirds.

Interestingly, the most visible periodicity in mammalian auditory nerve activity is the “missing fundamental” implied by the fundamentals of the two complex sounds. If one complex sound has a fundamental of 440 Hz and the second has a fundamental of 660 Hz (making them, with a ratio of 3:2, a perfect fifth), the auditory nerve will exhibit a periodicity in its pooled nerve impulses equal to 220 Hz even though this frequency is not present in either complex sound. This “inferred” component is the fundamental of a harmonic series in which the 440 Hz component of one sound is the second harmonic, the 660 Hz component is the third harmonic, and all other components in either complex sound are also higher harmonics. It has been suggested that consonance for complex sounds whose fundamentals excite different critical bands depends largely on whether a missing fundamental can be identified that is consistent with all components in the complex sounds.

Humans attend to consonances both when several complex sounds are heard at the same time and when separate sounds are presented serially (e.g., melodies). Since birds and other mammals also have tonotopic inner ears divided into critical bands, they may well attend in similar ways to whether complex sounds, heard either simultaneously or serially, are perceived and discriminated as dissonant or consonant. One intriguing study by Hulse et al. (1995) suggests that this is probably the case for starlings. Birds such as motmots, penguins, and oilbirds produce independent sounds on each side of their syrinx that are very similar, but not identical in frequency. To a human ear, these sounds are very harsh and dissonant. Do these birds produce such sounds because their dissonance is jarring and thus demands a receiver’s attention (Owren et al. 2010)? Do animals favor dissonant sounds for aggressive signals (Morton 1975)? Similarly, humans have remarked for millennia on the musical nature of many passerine songs. Are there selective pressures for male songbirds to use more consonant sounds when attracting females? Hopefully, future studies will examine these possibilities.

**Relative versus absolute pitch**

Most humans can listen to a tune and then recognize it as the same melody even after it has been transposed into another key. Transposition involves either raising or lowering the frequencies of all notes in the tune while retaining the ratios between them. In practice, a melody in the key of C that consisted of notes C-E-G can be transposed into the key of D as the sequence D-F#-A. As a rule, most humans cannot identify, nor are much interested, in which frequencies are actually used to play back a melody. A minority of humans have absolute (or perfect) pitch which allows them to identify a given note in a melody on an absolute scale. However, these individuals, like other humans, still recognize a tune as the same even if it is transposed to another key. The easiest transposition is to double or triple each frequency in the melody (e.g. move it to a higher octave). Even human infants are capable of recognizing an octave transposition as the same melody. The emphasis on the frequency ratios of successive notes when learning or recognizing a melody, rather than the absolute frequencies of each note, is called relative pitch.

Studies on a variety of birds suggest that many birds may not discriminate between sound signals using relative pitch and melodic pattern, but instead memorize the frequencies of successive notes using absolute pitch. Starlings, pigeons, and zebra finches can be taught to discriminate rising or falling patterns of successive notes within a familiar frequency range, but cannot then recognize the same relative sequence transposed to a frequency range outside of that in which the training occurred (Hulse and Cynx 1985, 1986; Page et al. 1989; Cynx 1993, 1995). Octave shifts are particularly devastating to generalization in these birds. Field sparrows did not recognize their own species songs even when transposed by small amounts (Nelson 1988). On the other hand, pet shama thrushes and bullfinches have been reported to transpose human melodies easily into higher keys (Tretzel 1997; Guttinger et al. 2002). Wild chickadees engaged in song contests routinely transpose their “feebee” songs up and down while holding the ratio of frequencies for successive notes constant (Ratcliffe and Weisman 1985; Christie et al. 2004). Veeries and white-throated sparrows
also vary the initial frequencies of their songs while holding frequency ratios between successive notes constant; despite the variations, these birds have no difficulty recognizing conspecific vocalizations (Weary et al. 1991; Hurley et al. 1992). Finally, rhesus macaques can recognize melodic (consonant) sequences when transposed an octave, but cannot do so with random notes having no serial harmonic relationships (Wright et al. 2000). The take-home message from research to date is that species vary in their usage of relative versus absolute pitch in recognizing and classifying different sound signals. While we humans take for granted our natural ability to recognize a transposed melody, it may be naïve to assume that other species, especially songbirds, share that skill.

Evolutionary roots of music (“biomusic”)

There has been considerable interest in recent years as to whether the roots of human music can be found in behaviors and sound signals of other animals (West and King 1990; Krause 1992; Gray et al. 2001; Huron 2001; West et al. 2004; Baptista and Keister 2005; Fitch 2005; McDermott and Hauser 2005a,b; Fitch 2006). As we can see above, some of the most interesting pattern measures appear to show at least some similar processes in animals and humans. There are multiple levels at which comparisons can be made. Below, we list some of the levels of comparison and a few comments about whether or not that level currently shows suggestive links between taxa.

- **Ancestry:** There are few behaviors in our primate relatives that appear to be phylogenetic precursors to human music. One possible exception is percussion (see below). Some of the authors listed above have argued that music in humans arose de novo without any clear antecedents in other primates.

- **Development:** Much has been made of the fact that most songbirds, like infant humans, must learn their vocal signals by imitating the vocal signals of adults and only rarely by innovation. The claim that bird song sheds light on the evolution of human speech has fueled a highly successful research establishment working on the possible parallels. A similar claim has been made for music. In fact, song-learning in passerine birds differs from that in humans in that it is often limited to males, may occur only during a limited period early in life, and occurs in limited contexts (usually territorial defense and mate attraction), whereas humans engage in musical expression in a much wider variety of situations. Parrots may be an interesting contrast in that, like humans, both sexes must learn most of their vocal repertoire, learning is open-ended throughout life, and learned vocal signals are used in a much wider variety of contexts than just mate attraction or territorial defense. Whether learning should be invoked as a necessary condition for calling a signal music is problematic. Once one looks at a variety of animal taxa, one is faced with deciding how much of the acquisition must be learned versus innate or innovative, how open-ended the learning period must be, what fraction of the population must have the capability, etc. If the relevant learning must be exactly like that in humans, few other taxa meet the condition. If a more relaxed requirement is invoked, then it is not clear if the learning component helps much in explaining the evolution of human speech or music. While some learning is involved in passerine bird acquisition of songs and human acquisition of musical patterns, what does this tell us?

- **Physiology:** As we have seen, sound producing organs can draw on a limited set of mechanisms, and all taxa are limited by their body size in the range of frequencies that they can produce efficiently. Humans share these constraints with other taxa. Similarly, auditory organs have access to a limited number of mechanisms, and are also constrained by body size and the sound-propagating medium. It is thus not surprising that there would be convergences in the ways frogs, lizards, crocodiles, birds, and mammals make and process sounds. How finely these parallels can be drawn will require further comparative research. The discussion above of the role of critical bands in producing consonance versus dissonance suggests that birds and humans may process complex sounds in similar ways. On the other hand, the emphasis on relative pitch in humans and absolute pitch in many songbirds when learning or discriminating between note sequences suggests that the role of melody patterns, rather than memorized sequences of notes, may be a major difference between these taxa.

- **Signal structure:** Traditional human music usually consists of multiple notes produced serially in somewhat stereotyped patterns. This is a trait shared with the acoustic signals of many other animals, including birds, sacc-winged bats, and humpback whales. The selective forces that have favored structured patterns in sound signals include improved receiver discrimination between signals and ambient noise, receiver discrimination between
signals of different species and conspecific signals having different functions, provision of different kinds of information in different signals, and better tests of performance for potential mates when perfecting a difficult pattern. Whether any of these benefits accrue to human music remains unclear. Traditional human music is often melody- and harmony-based. While some bird songs appear to follow fixed frequency ratio rules like human melodies, many songs do not. On the other hand, some modern music is as dissonant as the calls of motmots and penquins, and we currently have no clear hypotheses about why any of these species do or do not limit consecutive notes to consonant alternatives. Percussion is a mechanism for sound signal generation that is common to humans and many animals. The drumming of membracid insects on their plants and woodpeckers against trees are clearly patterned signals not unlike rhythmic percussion by humans. Percussion may be one area in which parallels between humans and animals in the structure of sound signals are marked and worthy of more quantitative comparisons.

**Performance mode:** While individual musicians are common in many human cultures, group performances of music are also widespread. Similarly, many birds and mammals vocalize individually without any inter-individual cooperation, whereas others regularly produce sounds in groups. In most lek and male chorusing species, each displaying individual is competing with conspecifics for attention by potential mates, and any apparent synchrony or anti-synchrony is likely an emergent consequence of individual display rules (Greenfield 2005). However, there are species in which coordinated chorusing by entire groups does occur. This includes lions, coyotes, chimpanzees, hyaenas, and wolves among mammals, and greater anis, kookaburras, barbets, Australian magpies, and quail among birds. The level of coordination with individual group choruses is often low, however. Perhaps the most coordinated group performances are the duets of tropical birds. These include the joint displays of male *Chiroxiphia* manakins and vocal duets by a wide variety of mated pairs in various bird taxa (boubou shrikes, wrens, parrots, etc.). These duets are typically highly coordinated temporally, must be learned and perfected over time, and often consist of specific roles assigned to each partner. They thus show striking parallels to the duets of human musicians and vocalists.

**Ecology:** The environment in which signals propagate becomes very important for long-range signaling, but is often less critical for short-range signals. Many of the sounds suggested as examples of animal music (whale and passerine songs) are highly adapted for long-range propagation in the relevant medium. Most human music is performed at close range. These animal signals and human music are thus more likely to show parallels at other levels (e.g., physiology or function) than through common selective forces for propagation. While there is clearly competition among sympatric species of animals for an “acoustic niche” (Krause 1992), there is little evidence that sympatric species collaborate with each other to produce a given “symphony” of joint sounds. Mutual avoidance of using the same frequency ranges at the same times of day often spreads out the calls of species in a given habitat across the possible times and frequency bands. One possible exception may be mixed species flocks of birds or monkeys that forage as a unit and respond to each other’s alarm calls. Some, such as drongos, may even mimic another species’ calls when a predator is spotted. However, most of the vocalizations used by these species in this context would hardly be considered “musical.” While there is clear evidence of allomimesis (copying of another species’ signals) by a variety of bird species, there is little evidence to date that species other than humans indulge in alloesthetics (the sensory or psychological enjoyment of listening to other species’ sounds independent of any specific signal function). However, one never knows until one looks whether salient species such as parrots, chimps, elephants, or dolphins might not have evolved this capability.

**Function:** While behavioral ecologists have become quite adept at identifying the specific functions of animal signals (see the list of options in Chapter 1 and subsequent chapters in text), the functions of music in humans remain unclear. Some authors have suggested that music arose to promote sexual advertisement much as the displays of lekking birds and mammals function to advertise male condition and quality. Others suggest that human music evolved to promote social cohesion with competing groups. Fitch (2006, 2010) suggests that some forms of music may have been the antecedents of language. Until some consensus is reached about the current and—even more challenging—original functions of music in humans, comparative contrasts with animals will be difficult. On the other hand, the extensive amount of information we have and continue to accumulate on the functions of sound signals in animals will provide a relatively exhaustive list of possibilities to be considered when
discussing human music.

This short list is designed only to outline possible points of overlap between the patterns of human music and animal sounds. More details on specific levels may be found in the citations that began this section. Clearly, data suggest significant overlap for some levels, whereas other levels have been little studied. This field is in its early stages and many surprises may appear with time.

**Literature Cited**


Cynx, J. 1993. Auditory frequency generalization and a failure to find octave generalization in a songbird, the European starling (*Sturnus vulgaris*). *Journal of Comparative Psychology* 107: 140–146.


Web Topic 4.1
Light Wave Meets Boundary

Introduction

The wave theory of light gives us precise equations for computing the reflection and refraction of light at the boundary between two materials with different indices of refraction. This unit provides a brief overview of these equations, and links to a number of web resources containing animations and tutorials for these principles.

Huygens wave principle of reflection and refraction

It is far easier to understand the processes of reflection and refraction by viewing a movie of a light wave as it encounters a boundary. At the website listed below, you can find an animation of Huygens wave principle of reflection and refraction. A few sample points along the boundary of the two media are shown in pink and the wave is shown being reflected and refracted off these points. In reality, there would be an infinite number of points along the way. Christian Huygen (1629–1695), a Dutch scientist, was the first major proponent of the wave theory of light.

http://www.sciencejoywagon.com/physicszone/otherpub/wfendt/huygens.htm

Quantitative expressions for the effects of different refractive indices on either side of a boundary

The difference in the refractive indices of the two media at a boundary affects reflection and refraction in several ways.
Most of these effects can be described quantitatively. In the equations below, subscript 1 refers to the first medium (from which the wave is coming) and subscript 2 describes the second medium (after the wave crosses the boundary). The refractive indices are indicated by $n$, and angles relative to the normal are indicated by $\theta$.

**Angle of refraction:** The angle of refraction depends on the incident angle and the indices of refraction for the two media, according to Snell's law, where the subscript 1 refers to the first (incident) medium and the subscript 2 refers to the second medium:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2$$

The application at the following website enables the user to vary the angle of incidence and the refractive indices of the two media to compute the angle of refraction: [http://www.fisica.ufpb.br/prolicen/fendt/physengl/refraction.htm](http://www.fisica.ufpb.br/prolicen/fendt/physengl/refraction.htm).

**Amount of reflected light:** More light is reflected at a boundary the greater the contrast in refractive indices. The proportion of light reflected, $R$, is given by the Fresnel equation:

$$R = \left( \frac{n_1 - n_2}{n_1 + n_2} \right)^2$$

**Snell's window:** When light from a relatively dense medium (high index of refraction) hits the boundary of a medium with a lower index at a high angle of incidence, it is completely reflected. This point, called total internal reflection, occurs when the angle of refraction is $90^\circ$ or larger and Snell's law is undefined. One familiar consequence of this phenomenon, called Snell's window, is the restriction of the visual field to low angles of incidence when under water looking up at the surface toward objects in the air. As illustrated below, an animal in water that looks up at the surface sees the entire hemisphere above the water condensed into a solid angle of $97^\circ$ due to refraction. At angles of incidence greater than $48.5^\circ$ (Brewster's angle for water), light is reflected back into the water at the surface and this region appears mirror-like.
Index of refraction: The index of refraction can also be expressed as:

\[ n = \frac{c}{v} = 1 + \frac{N \text{ (constant)}}{(\omega_0^2 - \omega^2)} \]

Here, \( N \) is the density of atoms in the medium and the constant includes the weight and charge of electrons. The density of atoms decreases the speed of propagation, \( v \), (and thus increases \( n \)), because the more molecules or atoms there are available for polarization, the larger the counter-field that can be induced in a material by an external electric force. The frequency difference \( (\omega_0^2 - \omega^2) \) reflects the degree to which the driving frequency is similar or not similar to the resonance frequency. When the two are equal, the denominator in the above expression is zero and thus \( n \) is infinitely large. This is equivalent to saying that the speed of propagation in the medium at this frequency is zero: everything is absorbed and nothing is transmitted. At driving frequencies close to \( \omega_0 \), transmission will be poor but not zero and there will be complicated phase effects. As the difference in frequencies increases, \( n \) drops off quickly. In fact, because the frequency terms are squared, it will not take a very large difference between \( \omega \) and \( \omega_0 \) before \( n \approx 1 \).

Glass and water are transparent substances with indices of refraction greater than 1 (1.5 and 1.33, respectively). Both materials have resonant frequencies close to the visible light range. Glass has a resonant frequency of about 15.0\( \times 10^{14} \) Hz in the UV range. Shorter wavelength radiation in the visible range (e.g., violet and blue) is closer to this resonant frequency and is therefore refracted at a greater angle when entering and leaving boundaries with air than is longer wavelength radiation (e.g., yellow and red). A beam of white light directed at a glass prism at an angle is refracted at two boundaries and is split into a spectrum of colors, a process called dispersion. The illustration below shows this process and provides the wavelengths and frequencies of different visible light hues. Violet light is clearly closer in frequency to the resonant frequency, and is bent more at both boundaries.
Brewster's angle: When light from a less dense medium strikes the surface of a denser dielectric medium at an angle, the reflected ray is partially polarized in the plane parallel to the surface. At a critical angle, called Brewster's angle, it is completely polarized. This phenomenon occurs when the incident and refracted angles add up to 90°. Brewster's angle can be computed by:

\[ \tan \theta_B = \frac{n_2}{n_1} \]

In the illustration below for light incident in air hitting a glass boundary, Brewster's angle is about 56°. The incident ray contains electric vectors oriented in all possible directions but only two are shown, \( E_1 \), oriented in the plane of the page, and \( E_2 \), oriented perpendicular to the page. At the critical angle, only vector \( E_2 \) will be present in the reflected wave and \( E_1 \) will predominate in the refracted wave.
Polarized sunglasses that transmit only vertically polarized light are designed to cut the glare of the horizontally plane-polarized light that reflects off of flat substrates without reducing other light substantially. Take your glasses off and rotate them 90°. Now you will see primarily the reflected glare!

Scattered light is also plane polarized when viewed from a 90° angle from the incident beam for similar reasons. The natural world contains complex patterns of polarized light generated by reflection and scattering. Many animals can perceive the plane of polarized light and use the patterns for a variety of functions, including navigation, orientation toward water, prey detection, and even signaling. These issues are taken up in Web Topic 5.2.

Other website links

**Light and color**: An excellent resource with good explanations and many useful applets:  

**Reflection and refraction**: More animations of Huygens waves:  

**Radiation and the human body**: A good explanation of how different frequencies of radiation interact with biological organisms:  
[http://hyperphysics.phy-astr.gsu.edu/hbase/mod4.html](http://hyperphysics.phy-astr.gsu.edu/hbase/mod4.html)

**Other useful physics websites**:

- The Physics Zone website has useful tutorials on general physics, motion, forces, work and energy, momentum, electricity and magnetism, light and sound waves, and nuclear physics at the high school and introductory college levels. Find it here:  
The Hyperphysics website contains more advanced tutorials in physics and astronomy at the college level. Find it here: http://hyperphysics.phy-astr.gsu.edu/hbase/hph.html.
Radiance versus irradiance

It is important to distinguish two types of light measurements: radiance and irradiance. Both are measures of radiant flux (energy per unit time). The two differ in the acceptance angle of the sensor (Figure 1).

Irradiance is the total amount of light incident on a surface, and includes scattered (diffuse) light as well as direct light. It is measured by an instrument that collects the light from a 180° solid angle. Radiance, on the other hand, is the flux of energy emitted from a specific radiant area such as the sun or an animal's body or signal patch. Only the light that travels directly from the source area to the receiver is measured using a tube-like or telescoping instrument that cuts out scattered light. The solid angle over which the instrument is measuring must be specified. The units for these different measures are shown in the table below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Measurement</th>
<th>Units(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiant flux</td>
<td>Flux (flow per unit time)</td>
<td>Photons sec(^{-1})</td>
</tr>
<tr>
<td>Irradiance</td>
<td>Flux density at a surface</td>
<td>Photons sec(^{-1}) m(^{-2})</td>
</tr>
<tr>
<td>Radiant intensity</td>
<td>Flux per unit solid angle</td>
<td>Photons sec(^{-1}) sr(^{-1})</td>
</tr>
<tr>
<td>Radiance</td>
<td>Flux per unit solid angle per unit area</td>
<td>Photons sec(^{-1}) sr(^{-1}) m(^{-2})</td>
</tr>
</tbody>
</table>

\(^a\) The measure “sr” means steradian, the unit of solid angle. There are 4\(\pi\) steradians in the complete solid angle of a sphere.

The general term for light measuring instruments is radiometer. These photoelectric devices record radiant energy or power (in joules sec\(^{-1}\), or watts), whereas the relevant unit for animal vision is quantum flux, the number of photons per
unit time. Since a photon’s energy is related to its wavelength, an energy unit of red light represents more photons than the same unit of blue light. It is relatively easy to convert an energy flux measure into a photon flux measure with a wavelength-specific equation. The photon flux for a single wavelength in micro moles of photons m\(^{-2}\) s\(^{-1}\) is given by \(Q(\lambda) = 0.00835191E(\lambda)\), where \(E\) is energy flux in watts m\(^{-2}\) (Endler 1990). This conversion may not be necessary, since most light measurements are given as percentages or proportions relative to a standard; the standard and any other components of a visual signaling system must also be measured in the same units.

**Wavelength-specific measurements**

Both radiance and irradiance measures can be wavelength-specific by making separate serial measurements on a narrow range of wavelengths. The instrument that does this job is called a **spectrometer**. While there are several types of spectrometers available, most animal communication researchers now use the **diode-array** type because it is fast, lightweight, possesses no moving parts, and has low power requirements, so is easy to run on batteries for field use. A popular model is illustrated in **Figure 2**. The light signal from the sensor is brought into the instrument with a fiber optic cable, conditioned, and spread into a rainbow of colors with a diffraction grating. The dispersed light falls on an array of photodiodes, each of which responds only to the narrow range of wavelengths impinging on it. The diodes are connected to a charge-coupled device (CCD) that produces a voltage. Voltages from each diode are converted to digital counts and sent to a dedicated computer with spectrum analysis software. The acquired spectrum is immediately displayed on the screen and saved in standard computer files.

**Figure 2. Diode-array spectroradiometer (Ocean Optics USB4000).** Light enters the spectrometer (1) from a fiber optic cable connected to a probe. The amount of entering light is controlled with an adjustable slit (2) and a filter (3) restricts the wavelength range. The filtered light then encounters a collimating mirror (4) that focuses the beam on a diffraction grating (5). Like a prism, this grating scatters the different wavelengths onto a focusing mirror (6) that directs the light via a collector lens (7) onto the detector (8). The detector converts the optical signal to digital format. (From Ocean Optics 2008, [http://www.oceanoptics.com](http://www.oceanoptics.com).)
Photodiodes exploit the photovoltaic effect at the interface between a semi-conductor and a metal. Electron transitions generate either a voltage in the cell or a decrease in resistance in an electrical circuit with an external voltage source. Diodes measure accumulated radiant energy (joules) and can be made very small. A silicon photodiode permits linear responses over 10 orders of radiant power between 100 and 1100 nm. A typical diode-array instrument contains an array of up to 1000 photodiodes. The wavelength interval that can be measured is fixed by the width of the diodes and therefore cannot be varied. However, a single scan takes a fraction of a second, and the diodes can then be discharged and prepared for a second scan. Several scans may be made and averaged in a matter of seconds. For further information see Andersson and Prager (2006).

### Sampling setup

To make accurate and reliable measurements of a color spectrum from a live animal color patch or sample, several important decisions and steps must be made.

First, an appropriate **light source** must be obtained. The light source must provide strong, stable, and spectrally smooth light over the desired range of wavelengths. A tungsten-halogen source works well for a wavelength range of 380–700 nm. If the color patch turns out to have a strong reflectance in the UV range, then a specialized UV source should be used. Currently available UV light sources require quite a bit of power and cannot be run from batteries.

Second, the **geometry of angles** for the light source and the reflection probe relative to the sample surface must be determined. This geometry is of crucial importance for the measurement of structural colors, which depends on the angle of illumination and the angle of viewing. There are several possible configurations, illustrated in Figure 3. The most reliable and recommended arrangement for pigment-based colors is coincident normal (CN), where both the light source and probe are directly above the sample at normal angle. Coincident oblique (CO) moves the both light and probe together to an angle of 45° relative to the sample surface. In oblique normal (ON) the light source derives from a 45° angle but the probe is directly above at the normal angle. For some structurally-based color mechanisms, one may desire to measure the specular reflectance at different angles, so the light source and probe are set at equivalent angles to either side of the normal, e.g., oblique opposite (OO). Finally, a diffuse light source can be established with either multiple light sources, as in diffuse normal (DN), or with an integrating sphere (IS).
The third decision involves the **white standard**. All reflectance measurements are made relative to a standard and reported as proportional or percentage values. The white standard is a reference surface that is measured with the same illumination set-up as the sample. The two spectra from the white reference and the sample are then compared. The reference surface should have a high reflectance (close to 100%) over all wavelengths of interest. To make highly accurate measurements, it is also important to remove the effects of electronic noise in the system from both the sample spectrum and the reference spectrum by subtracting what is called “dark currents.” The reflectance of the sample is then computed as $R = (AR - D) / ARr - D$, where $AR$ is the illuminated sample, $ARr$ is the illuminated white reference, and $D$ is dark current noise (Andersson and Prager 2006).

**Quantifying brightness, hue, and saturation**
Once one has obtained a reflectance spectrum from an animal patch, the next step is to extract useful measures of brightness, hue, and saturation. Researchers use a variety of methods for extracting these values, so they are not always comparable.

Raw reflectance spectra are often jittery curves, and repeated measurements from the same or different individuals may show some variation. Individual curves may be smoothed using a moving average or cubic spline, and several curves may be averaged. Subsequent measurements may be based on these smoothed or averaged curves.

The commonly used methods and formulae for computing brightness, hue, and saturation are illustrated in Figure 4. For more details, the reader is referred to Montgomerie (2006). Brightness is typically the area under the curve over a specified range of wavelengths, but mean brightness and maximum brightness values may also be used. Hue is the location of the wavelength with the greatest reflectance value. Saturation is the most difficult property to measure consistently. Four popular measures include the slope, the ratio of the sum of the low-reflectance region to the sum of the whole region, the ratio of the maximum reflectance to the lowest reflectance, and the difference between the maximum reflectance and lowest reflectance.

\[
\begin{align*}
S_1 &= \frac{X}{X+Y} \\
S_2 &= \frac{R_{\text{max}}}{R_{\text{min}}} \\
S_4 &= b_{\text{maxneg}} \\
S_6 &= R_{\text{max}} - R_{\text{min}}
\end{align*}
\]
Figure 4: Formulae for extracting color measures from spectrographs. Saturation measures: $S_1$ (reflectance ratio) is the sum of reflectance values integrated over the lowest reflectance region ($X$) divided by sum of reflectance integrated over the entire wavelength range ($X + Y$) = $B_1$; $S_2$ (spectral saturation) is the ratio of highest to lowest reflectance values; $S_4$ (spectral purity) is the greatest negative slope on the curve; and $S_6$ is the difference between maximum and minimum reflectance values. Brightness measures: $B_1$ (total reflectance) summed over the appropriate range of wavelengths; $B_2$ is the average of these values or $B_1$ divided by $n_w$, the number of wavelength intervals included in the sum; and $B_3$ is the maximum reflectance value. The best measure of hue for a unimodal curve is the wavelength with the maximum reflectance. For cutoff pigments such as carotenoids, the hue is measured as the wavelength at which reflectance is halfway between its minimum and maximum values within the range of 450–700 nm, often called $\lambda[50]$. Other measures
It is often informative to plot the absorbance (or absorbance) curve for a color patch, especially one derived from a pigment mechanism. Absorbance is again a relative measure, and can be generated by subtracting the reflectance curve from the white standard curve.

There are some caveats on the meaning of these measurements for structural versus pigmentedary colors (Andersson 1999; Andersson and Prager 2006). Carotenoid and other pigments are subtractive colorants, in that they absorb certain wavelengths within the visible range. Structural colors are produced by additive and spectrally selective modifications of surface reflectance using nanoscale structures. Differences in the quality of these two types of color signals will have different consequences for some of the color property measurements. For example, as the concentration of a color-producing pigment increases, more light will be absorbed, and overall brightness will actually decrease. On the other hand, saturation should increase, and the hue may change as well. But, for very highly saturated pigments, further increases in concentration will not change the saturation value. For structural colors, increasing the thickness of multi-layer stacks should result in an increase in brightness, and increasing the regularity of the stack should increase the saturation of the color. Changes in the size and spacing of scatters will change the hue. Therefore, the types of color characteristics we measure as potential indicators of sender quality are expected to differ for the two color mechanisms.

Using photography

A very simple alternative to the expensive and computationally intensive spectroradiometer technique is to use digital photography for quantifying color parameters. The only standardizing strategy is to take comparable digital photographs of animals from a specified distance and under a controlled source of illumination (or flash). Photographs can then be analyzed using graphic software applications such as Adobe Photoshop™, Canvas™, SigmaScan Pro™, Corel Paint™, or ImageJ™ to give mean values of hue, saturation, and brightness for user-specified color patches. This method will work only for animal color signals that fall within the human vision color range, where RGB color space is appropriate. A good example of the effective use of this technique can be found in Kilner's (1997) study of mouth coloration in nestling birds.

Literature Cited


Web Topic 4.3
Dimensionality of Structural Color Mechanisms

Introduction

Coherent scattering of a narrow range of wavelengths requires spatial periodicity in refractive index in one or more dimensions. Figure 1 illustrates idealized structures with periodicity in one, two, and three dimensions.

![Figure 1](image)

Figure 1. One, two, and three-dimensional array structures composed of high and low refractive index materials. Arrows show the vector directions with periodic variability.

Examples of natural one-dimensional nanostructures include single-layer systems, with a thick keratin layer over a melanin layer, and multi-layer quarter-wave stacks. An example of a two-dimensional structure is the collagen fiber arrays in the colored skin of many birds and mammals. Three-dimensional structures include diffraction lattices and other crystalline arrays. The effective dimensionality of some complex natural structures is often difficult to determine. Given the basic principles of interference physics, the size and spacing of the scattering structures must be consistent with the observed color of coherently scattered light. Electron microscopy is an essential tool for describing the dimensionality of nanostructured surfaces. Traditional transmission electron microscopy of very thin tissue slices only provides a two-dimensional view, although slices can be taken at different angles. Other techniques such as scanning electron microscopy and higher-voltage electron microscopy of thicker slices with tomographic reconstruction give a three-dimensional view of the structure. In this Web Topic unit we show how these observational techniques, along with mathematical modeling, have been used to quantify the nanostructure properties of colored surfaces and define their dimensionality.

Reflection modeling

One-dimensional structures typically produce iridescent colors, defined as colors that change as a function of viewing angle or source lighting angle. Their properties can therefore be described with TEM images and basic optical ray reflection and refraction equations. The angular dependence of the reflectance spectrum is described with an instrument called a goniometer, as shown in Figure 2.
**Figure 2: A goniometer.** The object (feather, wing, carapace, etc.) is secured to a flat platform. A protractor centered over the object shows the normal angle (0°) at the top, and gradations up to 90° on each side. A point light source (left) is aimed at the object from a specified angle, and a directional light receptor (right) is also aimed at the object on the other side. The light source and detector are often placed at matching angles, and measurements of brightness and hue are made at a series of angles. Alternatively, the light source can be stationary at zero degrees while the platform is rotated.

If the object is a flat multilayer stack, the bright hue is observed only over a very narrow range of angles, and outside of these angles the object should appear black. In other words, there is a single highly saturated brightness peak. The hue of the colored region should match that predicted by the Bragg equation—\( \lambda/4 = n_l d_l = n_h d_h \)—where \( n \) is the refractive index of the layer, \( d \) is its thickness, and the subscripts \( l \) and \( h \) refer to the low- and high-density alternating layers. The linear dimensions would be measured with a TEM image along the \( z \)-axis; once the wavelength is known, refractive indices from this equation can be estimated and compared to other estimates for these materials derived from similar kinds of measurements. The brilliantly colored throat and chest feathers of hummingbirds typically show this type of structure, and are only visible when the bird is directly facing the receiver. The striking flash effect of this signal can be seen in this video clip at [http://macaulaylibrary.org/video/flashPlayer.do?id=40898](http://macaulaylibrary.org/video/flashPlayer.do?id=40898).

If the object has a simpler two-layer nanostructure, consisting of one thick and relatively transparent layer of a material such as keratin over another layer of melanin granules, then thin-film optical interference principles apply. Such objects are often basically dark-colored, with one to six brightness peaks of different hues which may shift as the angle of incidence or viewing is changed (Figure 3).
The peaks for a given curve are harmonically related and specified by this formula:

$$\lambda_m = \frac{2d \sqrt{n^2 - \sin^2 \theta}}{m + 1/2}$$

It is not immediately obvious which interfaces are contributing to the iridescent color. Figure 4 shows the possible alternative reflection models for such two-layer systems. Model A involves only the two sides of the keratin layer, Models B and C involve the melanin layer with either the top or bottom keratin interface, and Model D combines reflection from all three interfaces.
To figure out which model best applies, predicted reflectance spectra are computed for each model based on the thicknesses of each layer, their refractive indices, attenuation within each layer, and occurrence of half-wave phase shifts at Interfaces 1 and 2. The reflected components in each model are combined additively. For the avian species subjected to this type of analysis, Model A, involving only the keratin layer, with reflection from the top and bottom interface, appears to provide the best fit (Brink and van der Berg 2004; Doucet et al. 2006; Yin et al. 2006). The melanin layer is too dense for a significant amount of light to pass through it. Figure 5 shows these results for the satin bowerbird, whose plumage is black with ultraviolet iridescence.
Figure 5: Observed reflectance curve and the four predicted model curves for the iridescent male satin bowerbird feather. The solid line is the observed curve; the dashed lines are modeled curves. Model A fits the observed curve best, suggesting that only the keratin layer is responsible for the iridescence while the melanin layer serves to define the bottom keratin interface. (From Doucet et al. 2006. Reproduced with permission from the Journal of Experimental Biology.)

Fourier analysis

The periodicity of any alternating or oscillating pattern can be described with Fourier analysis. In Web Topic 2.4, we show how simple Fourier analysis of sound waves can be used to break down a complex wave into the sum of component sine (or cosine) waves, each with a specific frequency, amplitude, and relative phase. The amplitudes of each of the component sine waves in the Fourier transform give the relative contributions (weighting) of those frequencies to the periodicity of the original data. A plot of the square of the amplitudes for each of the Fourier components versus their frequency is called the Fourier power spectrum. The power spectrum helps us identify the loudest or dominant frequencies in a sound. Fourier analysis can also be applied to a graphical image of patterned light and dark areas. If there is some regularity to the pattern, coherent scattering of a limited range of wavelengths should occur, and the Fourier power spectrum should exhibit a peak corresponding to the wavelength of the observed dominant hue.

Although optical Fourier principles were developed in the 19th century, Benedek (1971) derived new physical models of coherent light scattering based on electromagnetic principles to examine transparency of the human cornea. This approach has since been adopted by biologists to study color production in animal tissues (Prum et al. 1998, 1999, 2004, 2006; Prum and Torres 2003a, b, 2004; Shawkey et al. 2009). Because any graphical representation of a spatial pattern has at least two dimensions, the Fourier analysis must be expanded to encompass two dimensions. The first step for making this measurement is to obtain a high-contrast gray-scale TEM image of a nanostructured tissue region such as the collagen arrays of bird and mammal skin shown in Figure 6A. The digital image is essentially a matrix of dark and light pixels. Imagine a transect line drawn anywhere across this image. Along this line, one encounters a series of dark pixels followed by a series of light pixels in a repeating pattern. The dark and light cycles are converted into the peaks and troughs of a wave relative to the mean gray level in the whole image. The cyclical pattern of this waveform is measured by distance in units of cycles per nm (spatial frequency), rather than by time in cycles per sec as in a sound wave.

Next, one measures a series of spatial waves radiating out in all directions from each point in the image. Vectors from one such point are shown in Figure 6A. The Fourier component amplitudes of the vectors oriented in the same direction are averaged over all points, and this calculation is repeated for all vector angles. These results are then plotted in the two-dimensional Fourier power spectrum shown in Figure 6B. The center of this graph corresponds to variation at a spatial frequency of zero. Each point reports the magnitude of the periodicity in the original data of a specified spatial frequency in a given direction from all points in the original image. Darker pixels correspond to higher amplitude Fourier components. In this particular example, it is easy to see the regularity in the size and spacing of the dark fibers in A, which results in a circular ring around the origin in B. The two-dimensional plot can now be summarized by taking the radial average of concentric radial bins, or annuli, over all vectors. The outcome of this calculation is graphed in Figure 6C. The high peak at about 0.0065 nm⁻¹ corresponds to the dense ring in the two-dimensional plot in Figure 6B.

The final step is to convert this combined power spectrum into a predicted reflectance spectrum. According to optical Fourier principles, coherently scattered wavelengths from an ordered array should be equal to twice the wavelength of the predominant components of the Fourier transform. Following this model, the spatial frequency averages for each wavelength are inverted and multiplied by twice the average refractive index of the medium and expressed in terms of wavelength. The result is a theoretical prediction of the relative magnitude of coherently scattered light based on the spatial variation in refractive index of the tissue, as shown in Figure 6D. The power spectrum predicts a peak of coherent reflectance at 410 nm, which is consistent with the dark blue color of the tissue (Prum and Torres 2003a, 2004).
Figure 6: Two-dimensional Fourier analysis of a collagen fiber array in the dark blue rump skin of the male mandrill (*Mandrillus sphinx*). (A) TEM cross-section of the skin showing darker collagen fibers surrounded by lighter mucopolysaccharide (scale bar = 250 nm). White lines show a few of the vectors for Fourier transform analysis emanating from one point. (B) The two-dimensional Fourier power spectrum of the sample in (A) showing the ring-like appearance expected of a periodic two-dimensional structure; note the nearly symmetrical pattern around the origin in the center of the graph. (C) The radial average of power spectra originating from the origin in (B); the gray zone indicates the spatial frequency range visible to mammals. The low peak at a spatial frequency of 0.014, corresponding to the faint outer ring in (B), is in the UV wavelength range. (D) Predicted reflectance spectrum (black) based on the two-dimensional Fourier spectrum in (C) showing a strong peak in the blue wavelength region; the measured reflectance spectrum is in gray. (From Prum and Torres 2004. Reproduced with permission from the Journal of Experimental Biology.)

The performance of this two-dimensional methodology on a one-dimensional laminar array structure is shown in Figure
Figure 7: A two-dimensional Fourier analysis on a one-dimensional multilayer array structure. The tissue comes from the barbule of a green back-feather of the Splendid Sunbird (*Nectarinia coccinigrastra*). (A) A TEM through the barbule showing multiple layers of melanin and keratin. (B) A two-dimensional Fourier power spectrum; note the lack of circular ring. (C) The power spectrum predicted by Fourier analysis. (D) Measured reflectance of the green feather. (From Prum and Torres 2003a.)

Two-dimensional Fourier analysis applied to a TEM of the spongy matrix of blue and green barb tissues of bird feathers generates a Fourier power spectra such as the one shown in Figure 8. While there is a hint of a ring structure, it is more like a disk. Even disordered materials can produce circular plots.
Figure 8: A two-dimensional Fourier analysis of the blue feather barbs of the rose-faced lovebird, *Agapornis roseicollis*. (A) A TEM of the quasi-ordered, spongy medulary keratin and air structure in the interior of the barb. Scale bar = 200 mm. (B) A two-dimensional Fourier power spectrum of the same feather barb, showing a slightly oval disk. The magnitude of the power spectrum is given by the color scale bar on the right. (From Prum et al. 1999.)

This spongy tissue actually has a quasi-ordered three-dimensional structure, which cannot be accurately described with conventional two-dimensional visualization and analysis techniques. Newer tomographic techniques have facilitated a more accurate three-dimensional approach. Using intermediate-voltage electron microscopy, a high-power electron beam penetrates thicker slices of tissue and takes images at a series of angles, which then allows the third dimension to be reconstructed. Three-dimensional Fourier analysis can then be used to quantify the spatial frequencies on three dimensions and convert them to a predicted Fourier power spectrum that more precisely matches the observed reflectance spectrum of the tissue (Figure 9). Movies for visualizing the three-dimensional structure can be found at this URL: [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2706473/- supplementary-material-sec](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2706473/- supplementary-material-sec).
Figure 9. A three-dimensional analysis of the blue feather barbs of the eastern bluebird *Sialia sialis*. (A–C) Two-dimensional planes of the tomographic reconstruction of the spongy medulary tissue. The dark areas are keratin and the light areas are air. (D–F) Averaged projections of the three-dimensional Fourier power spectra from the tomographic reconstructions along different the $x$-, $y$-, and $z$-axes, respectively. The ring shape indicates periodic order over short spatial scales. Relative magnitude is indicated by the color map on the right of each graph. (G) Frequency spectrum predicted by the Fourier power spectrum data. The bar color shows corresponding hue. (H) Measured reflectance spectrum from spongy tissue (blue). Black dots with Gaussian fit are predicted spectrum points from (G). (From Shawkey et al. 2009.)
A second interesting application of the Fourier method was undertaken to examine the wavelength components of the transparent and opaque skin of the eye, i.e., cornea and sclera, respectively. For the cornea to be transparent, the collagen fibers must be very small and ordered to coherently reflect very small wavelengths well outside of the visual range while transmitting light in the visible range. The Fourier spectrum plots shown in Figure 10 illustrate the difference in collagen fiber periodicity very clearly. See Figure 5.14A in the main text for cross section TEMs of these two tissues.

**Figure 10: Fourier components for transparent and opaque eye tissues.** A comparison of the Fourier components for two regions of the eye: the transparent cornea and the opaque sclera. The cornea reflects only radiation of very small wavelengths, around 75 nm, while transmitting all wavelengths in the visible range. The sclera, in contrast, reflects all wavelengths in the visible range. (After Vaezy and Clark 1994 and Johnsen 2000.)

**Photonics and bandgap modeling**

Photonics is the study of light generation, transmission, modulation, amplification, and detection from the particle perspective. The field began around 1960 with the invention of the laser. It arose from solid-state physics and the study of crystal structure using X-ray diffraction patterns. Analogous to the way in which atomic crystals control the movement and spread of electrons through the structure, photonics views the way in which dielectric materials and nanostructured lattices control the propagation of photons of light. The photonics approach uses Maxwell's electromagnetic wave equations to model light wave transmission, in the same way that electronics uses Schroedinger's equations of quantum mechanics to analyze electric currents (Joannopoulos et al. 1995).

Any material with periodic variation in high and low dielectric regions over one, two, or three dimensions can be
analyzed from this perspective, but the term photonic crystal was first used in two seminal articles by Yablonovitch (1987) and John (1987) to describe two- and three-dimensional structures. Photonic crystals can guide and trap light traveling through the matrix. Depending on the dielectric materials and their arrangement, certain wavelengths or energy bands can pass through the material, called modes, whereas other wavelengths or energy bands are forbidden to pass, called bandgaps. Differential transmission and reflection of wavelengths in the visible range is what produces color. In fact, it is the forbidden wavelengths that are scattered, and this scattering is coherent. Because these higher-dimension structures can selectively pass light waves in different (disjunct) energy regions and directions, some unusual color effects can be produced. Biologists have recently taken up this way of thinking about highly ordered arrays of different dimensions to predict which wavelengths would be scattered and which transmitted. Natural photonic crystals and structures have now been identified for many structurally colored animal integuments. Figure 11 illustrates a bandgap analysis of the iridescent feathers of the peacock’s tail, a relatively simple two-dimensional square structure of melanin rods and airspaces (see also Figure 4.24D in the main text).
Figure 11: Bandgap model of iridescent peacock feathers. (A) Reflectance spectra of the blue, green, yellow, and brown regions of the eye spots of the green peacock *Pavo muticus* (see Figure 4.24D in main text for photo of the feather and SEM of the barbule nanostructure with a two-dimensional square array of melanin rods and air spaces. (B) Bandgap analysis. The \( y \)-axis is frequency in units \( c/a \), where \( c \) is the speed of light and \( a \) is the lattice constant (140, 150, and 165
nm for blue, green, and yellow regions, respectively, based on measurements of rod and airspace periodicity. The $x$-axis indicates different wave vectors (the planes of different directions photons can travel through the structure). The $\Gamma$–$X$ plane represents waves traveling in a direction normal to the barbule cortex surface, as shown in the bottom right. The lines in the graph show the calculated band structure for electric polarization (solid lines) parallel to the rod direction and magnetic polarization (dashed lines); electric and magnetic polarizations differed little in this two-dimensional structure. The orange blocks show the single partial bandgap region for this structure (for E polarization). A complete bandgap would have extended across all plane directions. This bandgap region corresponds to a reflectance peak in the middle of the visible light range. (C) Reflectance of blue, green, yellow, and brown feather regions predicted by the bandgap model. The peak positions match reasonably well, although the curve shapes are somewhat different. The bandgap position shifts to a higher frequency with increasing angle of incidence, indicating that the color is iridescent. (After Zi et al. 2003.)

More complex bandgap models and advanced visualization techniques are required to analyze three-dimensional structures. One notable study involving bandgap modeling coupled with a novel three-dimensional visualization technique characterized the brilliant green color of a weevil. The three-dimensional structure was assessed by consecutively milling away thin layers (about 30 nm thick) of the structure followed by coating the surface and taking SEM images. The color is generated by a diamond-structured photonic crystal with three bandgap regions, which yield three overlapping spectral peaks in the green frequency range (Galusha et al. 2008). Other studies have used bandgap modeling to explain structural colors produced by several butterfly species (Argyros et al. 2002; Michielsen and Stavenga 2008; Poladian et al. 2009; Saranathan et al. 2010). The Saranathan paper employed small angle X-ray scattering to obtain the three-dimensional structure, a promising technique that yields more precise descriptions of the structure compared to the electron tomographic reconstruction method described earlier.

The details of bandgap modeling are beyond the scope of this online unit. For a good overview of the photonics approach, see the review by Joannopoulos et al. (1997). Also useful are McPhedran et al. (2003), Vukusic and Sambles (2003), Parker (2004), Welch and Vigneron (2007), and Parker (2009). Links to several good web resources, tutorials, and lectures on the physics underlying photonic crystals are listed below.

http://ab-initio.mit.edu/photons/tutorial/

http://www.wisegeek.com/what-is-a-photonic-crystal.htm

http://wwwrsphysse.anu.edu.au/nonlinear/meeting/about.shtml

Literature Cited


Princeton University Press.


Web Topic 4.4
Movement Displays

Introduction

In Chapter 4, we reviewed the diversity of mechanisms that self-advertising animals have adopted to produce colors that are conspicuous against their normal backgrounds. However, most animals do not stop at color production, but add specialized motions that make their color patterns even more conspicuous. Luminescent species are usually nocturnal or live in deep ocean: many turn their lights on and off with species-specific timings to communicate with conspecifics in the dark. Species with neurally controlled chromatophores can vary their color patterns rapidly during a display. Species with iridescent colors typically use movement to shift the viewing angle relative to a receiver and thus change the apparent color. Other species cover colored regions when at rest but expose them for short periods during a display. In this module, we provide examples of each of these different strategies for combining color with motion.

Examples of movement and color/light displays

Cephalopods

Soft-bodied cephalopods produce and modulate their body color patterns using chromatophores (see Figure 4.33, Chapter 4). Although this ability was likely first favored as a camouflage adaptation (see first clip below), species in suitably lighted environments have co-opted the system for social and mating signals.

- **Common Octopus** (*Octopus vulgaris*): Individual changing camouflage colors rapidly to match changing backgrounds. (http://www.arkive.org/common-octopus/octopus-vulgaris/video-10a.html)
- **Giant cuttlefish** (*Sepia apama*): Pair changing colors during social interactions. (http://www.arkive.org/giant-cuttlefish/sepi-a-apama/video-09b.html)
- **Flamboyant cuttlefish** (*Metasepia pfefferi*): Individual displaying and feeding; note waves of black and white bands moving on back. (http://www.arkive.org/flamboyant-cuttlefish/metasepia-pfefferi/video-01a.html)
- **Common reef squid** (*Sepioteuthis sepiodea*): Rapid color changes during social interactions; good closeup of chromatophores in skin. (http://www.arkive.org/caribbean-reef-squid/sepioteuthis-sepioidea/video-12.html)

Fiddler Crabs

Males advertise themselves and their burrows to nearby females by waving their enlarged, and usually contrastingly colored, claw in a stereotyped display. Since some species overlap, each species has its own waving pattern:

- **American red fiddler crab** (*Uca heteropleura*): male on mud flat at low tide performing claw wave display. (http://www.youtube.com/watch?v=Gwet0JLuqWY)
- **Chinese fiddler crab** (species not indicated): Good views of males displaying and fighting at low tide. (http://www.youtube.com/watch?v=PylpZbx-m6Y&feature=related)
- **Multiple fiddler crabs**: a collection of videos of different fiddler crab species organized according to activities; check out the waving list. (http://www.fiddlercrab.info/uca_video.html)

Insects

- **Fireflies** (species not indicated): Short clip of firefly males advertising in flight during early evening in Hudson
Valley, New York.  
(http://www.youtube.com/watch?v=9QJpcBGPSEs)

- **Helicopter damselfly** (*Megaloprepus caerulatus*): This species has special markings on the wing tips that generate a fluttering image to attract mates (Schultz and Fincke 2009).  
(http://www.arkive.org/helicopter-damselfly/megaloprepus-caerulatus/-video-00.html)

- **Butterflies**: Butterflies host spectacular color patterns that are made more conspicuous by their fluttering flight. Some species have ultraviolet and iridescent colors that are only visible at certain viewing angles. Species also differ depending on whether both the top and bottom sides of the wings are brightly colored, and which side is exposed when resting. This clip shows slow motion videos of various British butterflies in flight or at rest and provides examples of each of the combinations noted.  
(http://www.bbc.co.uk/nature/collections/p00bf3g7#p00cnt8g)

- **Morpho butterfly** (*Morpho spp*): As documented in Chapter 4 (see Figures 4.18F and 4.29B), Morpho butterflies have iridescent blue coloration on the upper surface of the wings. This color shifts to white/grey at glancing viewing angles and appears to flash in and out when the insect is in flight. Two views are provided below: the first shows a live Morpho opening and closing its wings with a consequent appearance and disappearance of the blue color; the second shows a simulation of how the color changes with viewing angle:
  - Live Morpho *peleides*.  
    (http://www.youtube.com/watch?v=InKl9Nc1hpQ&feature=related)
  - Simulated Morpho.  
    (http://www.youtube.com/watch?v=aGsSrgg16l0&feature=related)

**Lizards**

Most lizards have fixed colors in places that are most clearly visible to conspecifics during territorial and mating interactions. Chameleons are unusual in that they can change color using chromatophores. Some species like anoles hide colored patches when not displaying. Some examples:

- **Panther chameleon** (*Furcifer pardalis*): aggressive display by one to another:  
  (http://www.arkive.org/panther-chameleon/furcifer-pardalis/video-12b.html)

- **Minor’s chameleon** (*Furcifer minor*): male female interaction; slow paced but coloring interesting. See more about this species in Figure 4.32, Chapter 4).  
  (http://www.arkive.org/minors-chameleon/furcifer-minor/video-00.html)

- **Broadley’s flat lizards** (*Platysaurus broadleyi*): two males with fixed colors adopt various body postures taht reveal the different color patches during dominance interactions.  
  (http://www.arkive.org/broadleys-flat-lizard/platysaurus-broadleyi/video-09a.html)

- **Anole pushup displays** (*Anolis spp*): Anoles expand their normally hidden colored dewlap and incorporate it in species-specific head bobbing displays:
  - *A. carolinensis*: one of several patterns in this species.  
    (http://www.biol.vt.edu/faculty/jenssen/videos/A-Display.avi)
  - *A. distichus* (http://lacertilia.com/?p=1112)
  - *A.cristatellus, A. pulchellus, A. stratulus, and A. gundlachi*.  
    (http://www.youtube.com/watch?v=_nmYyKA91hc)

- **Frilled lizard** (*Chlamydosaurus kingii*): This Australian species has a conspicuously colored but normally collapsed collar that it can inflate to startle possible predators. This clip shows the collar both deflated and extended. You may want to silence the sound track.  
  (http://www.youtube.com/watch?v=XAo09yYOpcU)

**Birds**
There are a large number of bird species that combine spectacular colors and motion in their displays. However, many perform their displays in flight (e.g. hummingbirds and manakins) making it hard to video-record displaying individuals at sufficiently close range to see the colors. Also, many species include sounds during their displays (e.g., icterids) making it difficult to know how much the colors and motion are important to receivers. Below, we focus on male courtship displays in one group of birds that routinely display while perched at a fixed site, making close-up videography feasible, and are usually silent once females settle nearby. These examples thus show the various ways that color and motion can be combined into a purely visual signal.

**Birds of Paradise (Paradisaeidae)**

Most species in this group of New Guinea and northern Australian birds are promiscuous and with no paternal care. The resulting sexual selection has favored unusually colorful and structurally complex male plumage. Males show off their coloration with stylized courtship movements. Most species call loudly to attract females from afar, but then switch to silent displays when females are close by. Each male prepares a display site by clearing all leaves from a viewing perch for females near to their cleared display saplings or ground court. In the examples below, note how the coloration patterns and movements work together to make the display highly conspicuous:

- **Parotia**: Species in this genus all display on cleared ground courts in deep forest and usually provide a cleared overhead branch for watching females. Males have a number of wire-like plumes ending in small knobs on their heads. Their display involves a silent ballerina-like dance with head wagging that is exaggerated by the head wires, at the end of which the male pauses briefly, and then rapidly flashes his iridescent throat patch by moving it from side to side.
  - *P. sefilata*.  
  - *P. carolae*.  
  - *P. wahnesi*.  
  - *P. lawesii*.  

- **Seleucidis melanoleucus**: The 12-wired bird of paradise lives in lowland coastal swamps. Males usually display in a spiral around a cleared vertical branch.  
  ([http://macaulaylibrary.org/video/62682](http://macaulaylibrary.org/video/62682))

- **Cicinnurus**: The two species of this genus listed below both display on vertical saplings or low horizontal branches above a cleared ground court. They first call loudly to attract females, but are relatively silent during displays when females are close by:
  - *C. magnificus*: The magnificent bird of paradise lives in low altitude primary forest and secondary forests. Males cyclically expand and contract their greenish-black chest shield when females are nearby but not close; once a female lands on his sapling, the male expands a white neck cape and vibrates his body so that the tail wires amplify the movements.  
  - *C. respublica*: The male Wilson’s bird of paradise sports striking colors on its back and head, as well as an expandable and green iridescent breast shield.
    - Male calling  
    - Court clearing, male calling, male displays.  

- **Lophorina superba**: The superb bird of paradise lives halfway up the sides of steep mountains in primary forest. When not displaying, males are largely black with little visible color. When displaying, they erect an iridescent
turquoise chest shield and surround it with a large black oval generated by their wings and neck cape. Two light-colored iridescent spots are also present in the center of the combined plumage. The male then dances in front of the female. This is such an amazing display that we provide two different clips showing it:

- Male display 1
- Male display 2
  (http://macaulaylibrary.org/video/flashPlayer.do?id=58003)

- **Ptiloris magnificus.** Male magnificent riflebirds have a hidden iridescent turquoise throat patch that can be expanded into a moderately sized chest shield. As in *Lophorina*, they then surround this with a black oval, here formed by the wings while they dance before the female. During the dance, the male points his beak upwards and tosses his head from side to side to change the iridescent patch viewing angle and make it flash. Females solicit dancing in males by approaching with their heads also pointed up. Two clips of this display:
  - View from front and back
    (http://macaulaylibrary.org/video/flashPlayer.do?id=55444)
  - View from side
    (http://macaulaylibrary.org/video/flashPlayer.do?id=55199)

- **Astrapia:** Most members of this genus have green or blue-green iridescent throat and crown patches. This close-up of a male *A. mayeri* shows how such iridescent plumage can change color depending upon the viewing angle.
  (http://macaulaylibrary.org/video/flashPlayer.do?id=65313)

- **Paradisaea:** Males of this genus tend to display high in treetops and they may form leks with multiple males displaying very closely together. Most have highly elaborate and plumose tails they show off by bending forwards on the branch while extending the flapping wings.
  - *P. apoda:* This clip of the greater bird of paradise shows multiple males on a lek displaying to a visiting female. The display involves a fluffing of the brightly colored tail plumage and softly beating wings that jiggles the tail plumes. Here, the female starts to favor one male who shifts to a second type of display eventually leading to female solicitation and copulation.
    (http://macaulaylibrary.org/video/flashPlayer.do?id=65338)
  - *P. minor:* This clip shows a male lesser bird of paradise calling and then performing several bowing displays that again show off its elaborate tail plumage.
    (http://macaulaylibrary.org/video/flashPlayer.do?id=55265)
  - *P. rubra:* The red bird of paradise lacks the highly plumose tail of the above species, but instead has two very long wiry feathers extending from the tail. The forward bow has largely been replaced by continuous wing waving and side-to-side movements that flip the tail wires back and forth. This sequence also ends with a copulation.
    (http://macaulaylibrary.org/video/flashPlayer.do?id=64532)
  - *P. rudolphi:* Male blue birds of paradise have taken the bow display to its limit: they hang upside down when displaying. As they rotate forward on a perch, they spread out their plumose tails, expand a dark chest shield, and while hanging by their feet, vibrate their whole bodies.
    (http://www.arkive.org/blue-bird-of-paradise/paradisaea-rudolphi/video-00.html)

**Further examples**

The prior examples provide a sense of the diverse ways that animals have combined motion and color into signals. As noted earlier, we have focused here on strictly visual displays, and skipped over examples in which visual components are combined with auditory, olfactory, or short-range modality components. Some examples of multimodal signals with both color and motion components can be found in Web Topics 10.3, 12.2, and 13.2.
Web Topic 5.1
Constructing Color Models

It has only recently become feasible to compute the detectability of a target object from an animal receiver’s perspective using the ARTS model, due to the large amount of data and information required to estimate each of the four spectra. In some cases, the computation can be simplified a little by making a few assumptions or focusing the question. For example, if one is interested in comparing hue contrast for different color patches in the same environment or for the same color patch against different backgrounds, then the environmental transmission step can be omitted. If one is interested in comparing the minimum detection distance for different signal patches, then some of the complex hue components can be reduced. In this unit we will take a closer look at the methods and equations used to estimate each of the four ARTS steps, and then describe two studies that successfully used this approach to answer meaningful questions about object detectability.

Ambient irradiance

The first step is to determine the available light spectrum. This must be an irradiance measurement using a spectroradiometer and a cosine-correcting hemispherical sensor or integrating sphere, as described in Web Topic 4.2. The key point to add here is that this measurement must be made in the precise habitat in nature in which the target object or signal is made. Many such spectra are presented in the main text in Figures 5.3–5.6. There are also several published standard spectra available for daylight (Vorobyev et al. 1998; Chiao et al. 2000). For reflected light signals in both terrestrial and aquatic environments, available light is always assumed to be downwelling light. We will refer to this spectrum as $A(\lambda)$ and it is measured over the entire range of wavelengths relevant to the receiving animal in question (usually 320–700 nm to include most organisms).

Target and background radiance

The second step is to measure the reflectance spectra for both the target and the background against which the target is viewed. This must be a radiance measurement taken through a tubular, angle-restricted light probe connected to a spectroradiometer. As described in Web Topic 4.2, the measurement must be taken relative to a white standard and corrected for dark noise. We will refer to these as $R_t(\lambda)$ and $R_b(\lambda)$, for target and background spectra, respectively. Some target and background spectra are shown in main text Figures 5.3, 5.9, and 5.28.

Receiver sensitivity

To adequately evaluate the receiver’s sensitivity, we need to know how many types of photoreceptor cells the receiver possesses, which ones are integrated for use in color vision and which for brightness contrast, and the absorbance spectrum for each one. It is of course important to know the absorbance curve for the photopigment associated with each receptor type. In addition, we need to know the effects of any filtering materials, such as oil droplets or corneal pigments that may be associated with each photocell type. There are several ways to obtain this information, either by directly measuring physiological responses by each cell type to presentations of a series of monochromatic light stimuli directed into the eye, or by measuring the components and multiplying their effects. For example, the spectral sensitivity for the $i^{th}$ cell type, $S_i(\lambda)$ can be computed as:

$$S_i(\lambda) = \int_{320}^{700} P_i(\lambda)D_i(\lambda)M(\lambda)d\lambda$$

Here, $P_i(\lambda)$ is the photopigment absorbance, $D_i(\lambda)$ is the transmission spectrum of colored oil droplets, and $M(\lambda)$ is the transmission spectrum of the ocular media (which would be the same for all cell types in a single-chambered eye but
could be cell-type specific in a compound eye). This measurement or calculation is made for each photocell type, resulting in a series of spectra. Examples are given in the main text in Figures 5.26, 5.27B, and 5.28A.

**Photon catch**

Computation of the photon catch for each photocell type is a bit more complex because we need to account for the fact that the eye is viewing a complex world and may undergo some adaptation to the amount and color of the ambient light and to the background hue. This physiological phenomenon is called *color constancy*, and requires us to correct the photocell sensitivity estimate above for this color adaptation. We must also normalize the spectral sensitivity of each photocell type at each wavelength. This is done by computing $k_i(\lambda)$, which incorporates the background reflection $R_b(\lambda)$ and ambient light $A(\lambda)$ spectra as follows:

$$k_i(\lambda) = \frac{1}{\int_{320}^{700} R_b(\lambda)S_i(\lambda)A(\lambda) d\lambda}$$

Now we can compute photon catch for each photocell:

$$Q_i(\lambda) = k_i \int_{320}^{700} R_i(\lambda)S_i(\lambda)A(\lambda) d\lambda$$

The next step is to normalize the maximum excitation of each photoreceptor cell type to unity using the following formula:

$$E_i = \frac{Q_i}{(Q_i + 1)}$$

Using the relative excitation of each photocell type, we compute the three-dimensional coordinates of each color (target and background) in the color space. For a trichromat, this color space is a two-dimensional triangle, with photocell types designated as S (short), M (medium), and L (long). The $(x, y)$ location of a color within the triangle is computed as follows:

$$x = \sin 60^\circ(E_L - E_S)$$

$$y = E_M - 0.5(E_S + E_L)$$

For a tetrachromat, we add the VS (very short) photocell type and a third dimension:
Now we can compute hue contrast, $D$ (also called chromatic contrast, $C_c$), by taking the difference between the target and background values for each coordinate, i.e., $\Delta x = x_b - x_t$:

$$
D = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2}
$$

The $z$ term would, of course, be omitted for the trichromat. The variable $D$ is the Euclidian distance between the target and background points within the color space. The longer the distance, the greater the hue contrast perceived by the particular receiver. Achromatic contrast can be computed in the same way if the photoreceptor types that the receiver uses for brightness contrast are known. The calculations would be made using only these photocell types. For example, bees use their green receptors for achromatic contrast, while the trichromatic primates use combined output from the medium- and long-wavelength cones (Spaethe et al. 2001; Mullen and Kingdom 2002). Birds possess specialized double cones for achromatic vision using their medium- and long-wavelength cones with the oil droplets removed (Hart et al. 2000; Osorio and Vorobyev 2005).

Vorobyev and Osorio (1998) suggest that a better estimate of color discrimination abilities is obtained if receptor noise is also incorporated into the calculation of $D$. There are three sources of receptor noise. **Dark noise** is caused by a constant low level of spontaneous firing by the photoreceptor cells, even in absolute darkness, which is sometimes called “false photons.” Dark noise is only likely to pose a problem for object detection under extremely low levels of ambient light. **Photon noise** is caused by statistical fluctuations in the capture of photons by the photoreceptors. It increases approximately as the square root of ambient light intensity, and is most likely to be a constraining factor only under intermediate levels of light illumination. **Neural noise** is caused by the effects of neural transduction in the analysis of visual input. At very high light levels, a small difference in contrast may not be detectable. Given the context in which $D$ is being estimated, the most constraining of these noise factors will set the threshold limit on the receiver’s ability to discriminate two hues. In practice, the combined effects of error can be measured physiologically and specified with an error term, which can then be incorporated into the computation of hue and contrast discrimination. For more details, refer to Vorobyev and Osorio 1998, Kelber et al. 2003, Goldsmith and Butler 2003, Montgomery 2006, and Lind and Kelber 2009.

An excellent example of the use of this method for calculating hue contrast (Figure 1) can be found in Théry et al. (2005), who asked whether female crab spiders waiting on flowers to ambush pollinator prey (bees) were cryptic to
both their prey and their predators (birds). To humans, the yellow spiders match the yellow centers of marguerite daisies very well, and white spiders match the white peripheral petals, but bees and birds have different color sensitivities so the spiders may not be as well camouflaged to these receivers. The spiders can change their colors to match at least some flower backgrounds (Chittka 2001; Insausti, 2008). Measurements of chromatic and achromatic contrast were made for yellow females against both the yellow center and white peripheral petals of marguerite daisies. The yellow color matched the yellow flower center very well, i.e., $D$ was not greater than the hue discrimination threshold for either bee or bird vision. But, the yellow spider did contrast significantly with the white petals, and achromatic contrast was also significant for both visual systems.


![Figure 1: Contrast between crab spiders (Thomisus onustus) and flower backgrounds.](image)

Spiders can change their body color depending on the background on which they are residing (see Figure 14.1D in the main text). Bar graphs show chromatic and achromatic contrast between a yellow female spider and the yellow center and white periphery of marguerite daisy flowers computed from photoreceptor sensitivities of a Hymenopteran eye and bird eye. The dashed lines show discrimination thresholds for Hymenopteran and bird visual systems. Spiders match the hue of the yellow center for both types of receivers, but do not match the petals, and are darker than both center and periphery from the perspective of achromatic brightness contrast. (After Théry and Casas 2002; Théry et al. 2005.)

**Transmission**

Studies that specifically model the transmission of visual signals or target objects are usually interested in evaluating the distance at which an object is just barely detectable. This distance is called the sighting distance or the maximum...
**detection distance.** One might ask questions that compare the maximum signal transmission range for alternative color signals, or that compare crypticity values of different coloration strategies. The visibility of an object generally depends more on its contrast than on its size, although one might need to set the size boundary above the threshold at which the visual system’s resolution can just distinguish an object, or develop different models for large and small objects.

Transmission models begin with the inherent contrast of the object against its background, $C_0(\lambda)$ measured at a very short distance. Contrast can be measured as Euclidian distance (provided earlier in this unit), or more simply as the difference between the reflectance of the target minus the background reflectance, divided by either background reflectance or background plus target reflectance:

$$C_0(\lambda) = \frac{R_t(\lambda) - R_b(\lambda)}{R_b(\lambda)}$$

We then examine how the perceived contrast falls off with increasing distance $d$ using the following type of equation:

$$|C(\lambda)| = |C_0(\lambda)| \times e^{-\alpha(\lambda)d}$$

Here, $\alpha(\lambda)$ is the beam attenuation coefficient of the medium. The reciprocal of this value is attenuation length, $L_a$, the maximum distance at which a large and contrasting object can just be detected. It is both medium- and wavelength-specific. Typical values for attenuation lengths are shown in Table 1. The beam attenuation coefficient is empirically measured in a specified ambient light context and environment. Radiance measurements of a contrasting object would be taken using a narrow probe at varying distances away from the object. The value $\alpha$ is then the slope of the regression line for the log-log plot of radiance versus distance. We can use $\alpha$ to calculate the falloff in radiance over any distance $d$.

**Table 1. Beam attenuation lengths in various environments.**

<table>
<thead>
<tr>
<th>Environment</th>
<th>Wavelength (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td>Clear air</td>
<td>3.8 km</td>
</tr>
<tr>
<td>Moderate fog</td>
<td>50 m</td>
</tr>
<tr>
<td>Pure water</td>
<td>?</td>
</tr>
<tr>
<td>Oceans</td>
<td>?</td>
</tr>
</tbody>
</table>


If we let $C_{\text{min}}(\lambda)$ be the minimum contrast for object detection for a given visual system, then we can substitute $C_{\text{min}}(\lambda)$ for $C(\lambda)$ in the equation above and solve the equation for $d$:

$$d(\lambda) = \frac{\ln\left(\frac{C_0(\lambda)}{C_{\text{min}}(\lambda)}\right)}{\alpha(\lambda)}$$

This distance $d(\lambda)$ is the maximum distance at which the object is detectable, and is referred to hereafter as the sighting distance, or $d_{\text{sighting}}(\lambda)$ (Johnsen 2002). This distance depends on the product of two factors: (1) the
relationship of the inherent contrast of the object to the minimum contrast threshold of the viewer, called the contrast factor, given by

\[ \ln \left( \frac{C_0(\lambda)}{C_{\text{min}}(\lambda)} \right) \]

and (2) the attenuation length or the penetration factor, given by

\[ \frac{1}{\alpha(\lambda)} = L_a \]

An interesting application of this sighting distance model can be found in Johnsen and Sosik (2003), where the effectiveness of two potentially cryptic color strategies of pelagic fish, diffuse colored reflectance and specular mirrored reflectance, were examined under different viewing situations. Johnsen used the general model described above, but the penetration factor required two components to deal with the effects of viewing a target while looking into the sun versus viewing a target with the sun behind the viewer. Their sighting distance model was therefore:

\[ d(\lambda) = \frac{\ln \left( \frac{C_0(\lambda)}{C_{\text{min}}(\lambda)} \right)}{\alpha(\lambda) - K_L(\lambda)} \]

Here, \( \alpha(\lambda) \) is the beam attenuation coefficient of the object in water and \( K_L(\lambda) \) is the attenuation coefficient of the background radiance (\( K_L \) is greatest for downward viewing and zero for horizontal viewing). They used the cod (\( Gadus morhua \)) as the model viewer, which has a measured \( C_{\text{min}} \) of 0.02, which is well within the range of other marine fish, and known cone sensitivity curves. The viewing conditions they considered are illustrated in Figure 2A and include looking down on the fish from above, looking up from below, looking horizontally into the sun so the fish is backlit, and looking horizontally away from the sun so the fish is sidelit. Figure 2B shows a schematic illustration of the way the two types of reflectance—diffuse and specular—were modeled. Downwelling and upwelling radiances were measured directly at different depths (equivalent curves in Figure 5.6 in the main text showed some examples of these measurements). The target object luminance (\( L_t \)) was computed for each case by multiplying ambient irradiance by the reflectance strategy.
(A)

θ

from above

$L_t$ $L_b$

into sun

$L_t$ $L_b$

from below

away from sun

$L_t$ $L_b$
(B)

Diffuse

\[ L_t \] \arrow{\rightarrow} \text{I}_s \arrow{\rightarrow} \]

\[ L_b \] \arrow{\rightarrow} \]

Specular

\[ L_t \] \arrow{\rightarrow} \]

\[ L'_b \] \arrow{\rightarrow} \]

\[ L_b \] \arrow{\rightarrow} \]
In contrast to benthic (deep water) marine fish, for which light is generally dimmer and less directional, pelagic (surface water) fish must contend with changing conditions that render them cryptic under some circumstances and highly visible under others. Crypsis was generally better with mirrored reflection than diffuse reflection, especially when looking horizontally away from the sun at the fish, because the background radiance is high and the mirrored sides are bright and highly reflective. Colored surfaces were more cryptic when looking horizontally into the sun and looking down from above. Both strategies were very conspicuous when viewed from below because they cast a strong silhouette against the bright downwelling irradiance. Both strategies were also quite visible at very shallow depths. An important insight from these studies is that a prey fish cannot control the viewing angle and lighting conditions in which it is viewed. Predators, therefore, can overcome many of these crypsis strategies of the prey. For example, predators can move in circular patterns to view prey from a series of angles and sun angles. They can also drive prey into shallower water where they are more visible. It appears that the best strategy for maximizing crypsis in pelagic organisms is transparency (Johnsen 2003).

Further reading


**Literature cited**


A beam of pure polarized light consists of a population of electromagnetic waves with parallel electric vectors. The natural world contains complex patterns of polarized light that humans cannot perceive but which are very useful for animals that can distinguish different electric vectors of light. In Web Topic 4.1, we described some of the basic sources of natural polarized light from scattering and reflection. In this section we shall first clarify how polarized light is quantified and measured and then describe the patterns of polarized light in nature. Next, we show how photoreceptors can be made sensitive to the planes of polarized light and outline some of the uses animals make of polarized light.

Quantifying light polarization

Natural polarized light is rarely pure (i.e., comprised of 100% parallel e-vectors). Instead, it is partially polarized to varying degrees because it contains some fraction of e-vectors oriented in other directions. Four components are needed to completely describe polarized light are:

- $I$ = the overall intensity of the light beam
- $p$ = the degree of linear polarization, or the fraction of the overall light intensity that is linearly polarized parallel to a reference plane, which can range from 0–100%
- $\alpha$ = the angle of strongest linear polarization, expressed in angles from 0–180º
- $\epsilon$ = the fraction of circularly polarized light

Circularly polarized light is relatively rare in nature and often ignored, but later in this unit we describe some examples of animals that do in fact reflect circularly polarized light. The other three parameters are essential components for describing a beam of polarized light. The photoelectric device for measuring and computing these components is called a polarimeter. It consists of a radiometer that measures light intensity (in watts or photon flux per unit time, see Web Topic 4.2) with a rotating linear polarizer in front of the light-receiving lens. Three simultaneous or rapidly sequential measurements of light intensity must be made at three independent angles of rotation, typically 0, 60, and 120º. Using these values, one can compute the three essential components as follows:

$$I = \frac{2(I_0 + I_{60} + I_{120})}{3}$$
$$p = \frac{(Q^2 + U^2)^{1/2}}{I}$$
$$\alpha = 0.5 \arctan \left( \frac{U}{Q} \right)$$

Here, $Q = 2(I_0 + I_{60} + I_{120})/3$ and $U = -2(I_{120} - I_{60})^{3^{-1/2}}$. These three components—intensity, degree of polarization, and angle of polarization—are analogous to the color components of brightness, saturation, and hue, respectively. Also analogous to color measurements, polarized light can be measured for a very small point in space using a narrow light acceptor, called point-source polarimetry, or globally over a very large area, called imaging polarimetry. In presenting the results of imaging polarimetry, it is best to provide separate images of the three components (Figure 1). Overall intensity is typically displayed with a normal color image of a scene, degree of polarization is displayed with a grayscale image of the same scene where white may represent either 0% or 100% polarization depending on the situation, and angles of polarization are depicted with different hues (Horváth and Varjú 2004).
Environmental patterns of polarized light

There are two main sources of polarized light in nature—reflection off of smooth surfaces and scattering in the sky by atmospheric particles. In both cases, the degree of polarization is very high at certain viewing angles, but the fraction of light that is polarized decreases gradually as the viewer or the light source moves away from this angle, and may become essentially nonpolarized at other angles. Figure 2 illustrates this concept for scattering in the sky.
Figure 2: Polarization of scattered light. Light from the sun, depicted by yellow arrows, is scattered by small particles (black point in center) in the atmosphere (note: The sun is not visible, but is located beyond the upper left corner of this illustration). This scattered light is vertically polarized when viewed from an angle of 90° relative to the sun. The plane of polarization of the scattered light is perpendicular to the plane defined by the incident and the scattered ray. Light that is scattered at an angle of 0° is nonpolarized and at intermediate angles it is partially polarized. (After Wehner 2001.)

Scattering creates a band of strongly polarized light across the sky that changes position as the sun moves (Figure 3). At sunrise and sunset the polarized band arcs across the middle of the sky with respect to the observer and can reach a maximum of 90% polarization on a clear day, but is usually less than this value. At noon the band is close to the horizon. Wavelengths in the middle of the visual range (blue and green) exhibit the strongest polarization intensity. The polarization intensity of red wavelengths is reduced by dust in the atmosphere, and UV intensity is reduced by multi-path scattering of the short wavelengths. Hazy and overcast skies increase this secondary scattering and may completely obliterate the polarization cues. However, conditions of patchy clouds with at least some open areas of sky will still retain their polarization pattern with respect to the sun and enable animals that rely on sky polarization to continue their activities (Wehner 1976, 2001; Cronin and Marshall 2011).
A similar pattern persists when the sky is viewed from underwater, but the degree of polarization is lower (maximum $p$ of about 70%) because of multi-path scattering of light by the water. Figure 4 illustrates the pattern of polarized light entering through Snell’s window (see Web Topic 4.1) for a viewer within about 15 m of the surface. The position of the band is tilted at lower sun angles as a result of refraction of light at the air–water boundary. Outside of Snell’s window and at increasing depths, there is still a considerable amount of polarized light from extensive scattering in the water, but the sky pattern is attenuated and the e-vector orientation is primarily horizontal. At depths between 100–200 m below the surface, light is still linearly polarized at about 30% at noon, and long (red) and short (blue) wavelengths are more strongly polarized than middle wavelengths (Cronin and Shashar 2001; Sabbah et al. 2006; Cronin and Marshall 2011.)
Light that is reflected off of the smooth surface of a dielectric material is completely horizontally polarized at Brewster’s angle of incidence and becomes increasingly less polarized at incident angles above and below Brewster’s angle. In nature, this source of polarized light arises from the reflection of light off of a smooth water surface such as a pond or lake (Figure 5A), and from specular reflection off of silvery scales (Figure 5B, C).

Figure 5: Polarization of reflected light. (A) Light incident on a smooth dielectric surface such as water is horizontally polarized (Wehner 2001). (B) Cross-section of a fish’s body. Downwelling light reflecting off of silvery fish scales containing guanine crystals is polarized at certain viewing angles where Brewster’s angle conditions are met. (C) Underwater photo of a fish with normal lens (top) and with an imaging polarizer (bottom), where white regions indicate stronger degree of polarization. Dorsal side of fish reflects 70–95% horizontally polarized light (upper arrow), whereas side flanks reflect 30–50% polarized light (lower arrow) (Denton and Rowe 1994; Shashar et al. 2000.)

Receptor sensitivity to polarized light

Sensitivity to polarized light has evolved many times independently and is found in a wide range of animals, including many orders of insects, spiders, crustaceans, cephalopods, fish, amphibians, reptiles, and birds (see Roberts et al. 2011 for a phylogenetic tree of taxa showing polarization sensitivity). Humans can perceive the linear polarization of light but this ability is believed to be a byproduct of the ocular media or foveal region and has no biological function. Species with well-developed polarization vision possess specialized sets of receptors sensitive to different e-vectors of polarized light. The neural output from several differently-angled cells or units are combined in additive and subtractive ways in higher-level neurons that comprise a polarization-opponent system, analogous to the color-opponent system. Thus different e-vectors of polarized light may be analyzed and perceived as if they were different colors. Behavioral experiments and neurophysiological studies support the view that the degree and orientation of polarized light is perceived as a graded stimulus (Horváth and Varjú 2004).

The visual pigment rhodopsin is inherently sensitive to the plane of polarized light. The crucial double bond of the retinal chromophore must be aligned in the same direction as the plane of polarization (e-vector) of the light for the molecule to absorb a photon. Dichroism is the general term for molecules or photoreceptor cells with selective sensitivity to the angle of polarization. Rhabdometric photoreceptors are inherently more sensitive to polarized light than ciliary receptors, but both types of photoreceptor cells possess mechanisms to make them more dichroic (Roberts et
In a normal rod or cone cell, the long axes of the retinal molecules are maintained in a fixed horizontal position parallel to the plane of the disc membrane and perpendicular to the direction of incoming (axial) light shining on the photoreceptor (Figure 6A). The opsin molecules are relatively free to move along the disc membrane and can rotate around at random compass angles when viewed from above. No matter what the orientation of the e-vector of polarized light is, a similar number of retinal chromophores will be oriented parallel to it, and the response of the cell will be the same. However, if one were to shine light transversely from the side of the rod or cone, the response of the cell would become dependent on the polarization angle of the light—horizontally polarized light would be absorbed by retinal molecules oriented perpendicular to the ray’s direction, but vertically polarized light would not be absorbed by any retinal molecules. The cell is now selectively sensitive to just one plane of polarized light, i.e., it is dichroic.

Rhodopsin molecules are anchored on the tubular membranes of microvilli in rhabdometric photoreceptors. Even if the rhodopsin molecules were randomly oriented on this rolled surface, as on a ciliary cell disc (Figure 6B), the cell would respond more strongly to axial light polarized in a direction parallel to the microtubules. Rhodopsin molecules on the sides of the tubules are more likely to be properly oriented to absorb this e-vector of light, compared to rhodopsin molecules on the tops and bottoms of the tubules, which are equally likely to respond to all e-vector orientations. The theoretical dichroic ratio of the cell’s response to parallel versus perpendicular polarized light is around 2. In fact, many rhabdomeric cells have dichroic ratios much higher than this value, up to 15. The orientation of rhodopsin molecules in specialized polarization-sensitive receptors is often strongly aligned parallel to the tubules, as shown in Figure 6C. This alignment is caused by connections between rhodopsin molecules and the cytoskeleton within the tubules, and by linkages between rhodopsin molecules on adjacent tubules (Roberts et al. 2011).
Figure 6: Sensitivity to polarized light in rhabdomeric and ciliary photoreceptors. (A) Ciliary receptors do not respond differently to light of different polarization orientations when shone from above (axial rays), but they will selectively absorb horizontally polarized light shone from the side (transverse rays). (B) Rhabdomeric receptors with rhodopsin chromophore molecules oriented randomly on the plasma surface, as shown here, would respond more strongly to axial rays polarized parallel to the microtubules than to rays polarized perpendicular to the microtubules. (C) Rhabdomeric cells specialized for polarized light detection contain retinal chromophores oriented parallel to the tubules, resulting in strong absorption of parallel polarized light.

The fine structure of polarization-sensitive rhabdomeric receptors has been studied most extensively in terrestrial insects (field crickets, honey bees, desert ants, and house flies) that detect the oscillation plane of polarized skylight with a group of specialized ommatidia situated at the dorsal rim area (DRA) of the compound eye. The dorsal rim ommatidia have properties that make them especially suitable for polarization vision. All of the ommatidia in the DRA, as well as the individual rhabdomere cells within each ommatidium, are sensitive to the same wavelength of light so that true color analysis is not compromised. Various adaptations are employed to make the visual field (light
acceptance angle) of each ommatidium (or group of ommatidia) very broad for viewing a large portion of the sky. Each ommatidium contains two sets of strongly polarization-sensitive photoreceptors with orthogonally-arranged orientations. Neural outputs from the two sets are compared antagonistically so that each cell reports the intensity of incident polarized light of a given angle. Within the DRA, the main microvillar directions of adjacent ommatidia are rotated from front to back in a fan-shaped pattern. It is believed that output from at least three ommatidia types sensitive to different angles are combined and compared in polarization-opponent neurons to provide the insect’s brain with information on the orientation of celestial polarization, much like the measurement procedure of a polarimeter (Labhart and Meyer 1999; Horváth and Varjú 2004). Figure 7 shows cross sections through the specialized ommatidia compared to the regular ommatidia for these major groups of polarization-sensitive insects.

![Figure 7: Optical, spectral, and structural characteristics of the specialized ommatidia in the dorsal rim area of six insect species.](image)

Top row: these optical specializations generally increase the visual field of the ommatidium. Second row: polarization-sensitive ommatidia are often associated with a single hue type. Third row: the number of rhabdomeres in each ommatidium that contribute to polarization sensitivity. Fourth row: the ratio of these that are oriented in the two orthogonal positions. Fifth row: cross-sections through specialized dorsal rim ommatidia, colors approximate the hue sensitivity of the receptor (violet represents UV). Sixth row: regular dorsal ommatidia, which are used for color and spatial analysis; they lack the strict orthogonal orientation, and have other adaptations such as long and twisted rhabdons, misaligned microvilli, and/or randomly oriented retinal chromophores so they are not differentially sensitive to the e-vector of light. (Labhart and Meyer 1999.)

Sensitivity to polarized light has been documented in vertebrates, primarily teleost fish, but also in some amphibian and reptile species (Horváth and Varjú 2004). How can ciliary photoreceptor cells become differentially sensitive to polarized light? Four mechanisms have been proposed. The most evident mechanism is to tilt the outer segment of the cone cell receptor on its side, so that the flat surfaces of the discs are parallel to the direction of incoming light cells (Fineran and Nicol 1978; Novales Flamarique and Hawryshyn 1998; Novales Flamarique 2011). This adaptation has
been well-documented in anchovy fish (Figure 8). The same strategy has been employed in the extraocular polarized light detectors of amphibians and reptiles; the intracranial pineal body and frontal organ of amphibians contains cone-like receptors with longitudinal disc orientation, and the parietal eye of reptiles contains a ring of cone-like receptors lying on their sides (Hamasaki and Eder 1977; Adler 1976). Coho salmon (*Oncorhynchus kisutch*) have achieved a degree of polarization sensitivity by partially tilting the discs within the cones (Roberts et al. 2004).

**Figure 8: Anchovy fish cones specialized for polarization sensitivity.** (A) Side-on view of bay anchovy (*Anchoa mitchilli*) cones. Cones are arranged in rows of alternating long cones (LC, pink) and short cones (SC, blue) with bilobed outer segments. The disc stacks in the outer segments of these cones are tilted longitudinally so they are parallel to the incoming light direction (upward from the bottom). The discs in the two cone types are oriented orthogonal to each other, as shown in the top-down view in (B). This view also shows their flat-sided shape, which facilitates orthogonal packing. The cross-hatched areas are zones in which the discs of the two cone types overlap. Several mechanisms ensure that the long cones respond selectively to perpendicular e-vectors ($E_{\text{perp}}$, perpendicular to the plane of the page and indicated by the dot-filled circle; the orthogonal e-vector $E_{\text{par}}$ is parallel to the plane of the page and indicated by double-headed arrows). Unpolarized incident light undergoes dichroic absorption when it passes through the bilobed outer segments of the short cells, and is further perpendicularly polarized as it reflects off the multilayered guanine platelets (yellow) and on to the lone cone outer segments. Likewise, light passing through the dichroic-absorbing short cones is further polarized as it reflects off the multilayered tapetum and back
A second mechanism for making vertebrate cone cells dichroic is to guide light transversally onto adjacent cone cells by scattering or reflection. This idea arose from the observation that the polarization sensitivity in some species is found in double cones. In fish such as trout, the cones for color vision are arranged in highly regular mosaics with a square arrangement of red and green double cones, UV cones between them, and a blue cone in the middle, as shown in Figure 9A, B. The partitioning membrane between the red and green cones has a bulge that directs light transversally to the red cone and on to the UV cone, which is the primary polarized light sensor (Novales Flamarique, Hawryshyn and Hárosi 1998). Birds also possess double cones, which are arranged in orderly mosaic patterns of four or six double cones surrounding one or two single cones. The principle cone contains clear oil droplets and there is no screening pigment between the two cones, leading to the possibility that sideways scattering from the droplets could direct polarized light transversely to the secondary cone (Young and Martin 1984; Waldvogel 1990). However, this proposal is by no means proven (Muheim 2011).

Figure 9: Reflection and scattering in double cones. (A) Side view and (B) top view of the cone mosaics of the rainbow trout (*Oncorhynchus mykiss*), showing UV-, blue-, green- and red-sensitive cones. Axial incident light is reflected from the tilted partitioning membrane surface in two directions (single-headed arrows). The UV receptor receives transverse rays. The quadrilateral arrangement of four double cone units in a square mosaic leads to orthogonal e-vector reception; double-headed arrows in (B) indicate the dominant plane of polarization of the reflected light ($E_h$ = horizontal e-vector, $E_v$ = vertical e-vector). (A, B after Novales Flamarique, Hawryshyn and Hárosi 1998.)
A third possible mechanism that could facilitate selective e-vector absorption in vertebrate, as well as invertebrate, photoreceptors is to provide each receptor with a polarizing filter. Invertebrates with a cuticular carapace can produce surface structures that selectively reflect or transmit certain e-vectors of light (dichroism or birefringence, respectively). The lenses of ommatidia may contain polarizing filters that either enhance the selectivity or the polarization sensors, or act like polaroid glasses to filter out glare in insects that hunt on the water’s surface (Horváth and Varjú 2004). Birds, reptiles, and amphibians can detect the Earth’s magnetic field by optical means with specialized photoreceptors (Philips et al. 2001; Wiltschko and Wiltschko 2006). The design of these receptors could make them simultaneously sensitive to the plane of polarized light if they cause the alignment of rhodopsin molecules within discs, but the true mechanism for polarized light detection remains unknown (Muheim 2011). Birds appear to primarily use their sensitivity to polarized light to recalibrate their magnetic detectors before they fly at night, taking advantage of the strong overhead arch of polarized light at sunset (Moore and Phillips 1988; Phillips and Moore 1992; Muheim et al. 2006).

Finally, some recent research has suggested that cone cells might be able to align adjacent rhodopsin molecules in a parallel fashion using underlying protein-protein interactions, similar to the protein-cytoskeleton interactions demonstrated in rhabdomeric photoreceptors (Nair et al. 2002; Roberts and Needham 2007; Elliott et al. 2008). These interactions allow rhodopsin to oligomerize and form rafts of parallel chromophores, which would make the photoreceptor cell potentially sensitive to axially oriented polarized light (Roberts et al. 2011).

### Functions of polarized light vision

The functions of polarized light vision include celestial cues for compass orientation, detection of aquatic habitat, reduction of flare from the water surface, increased contrast and prey detection underwater, and social communication. We take these up in turn.

Bees, ants, crickets, burrowing beetles, spiders, and probably many other arthropods that routinely return to the points of departure of their foraging journeys (e.g., nests, burrows, webs) use the patterns of polarized light in the sky for navigation. The natural sky patterns provide the animal with a compass if it knows the time of day, and with the time of day if it knows the direction. In fact, receivers can infer any particular compass direction from any particular sector of the sky at any time of day (Rossel 1993; Wehner 2001). Experimental evidence has recently been obtained demonstrating that honeybees truly use polarized light information to navigate to a food source (Kraft et al. 2011). The point to emphasize is that the polarization sensors do not provide the animal with individual e-vector directions, but with the compass direction of head orientation derived from global processing of e-vector gradients in the sky (Heinze and Homberg 2007). Behavioral and neurophysiological studies on the desert ant *Cataglyphis* suggest how this compass might work. Output from the polarization-sensitive photoreceptors in the dorsal rim area of the eye converge onto sets of at least three large-field polarization-sensitive interneurons, called POL neurons, located within a restricted area of the second visual ganglion. Each point of the compass is characterized by a particular response ratio of three POL neurons. There is some kind of neural network translating the broadband compass responses of the POL neurons into narrowly tuned responses of particular ‘compass neurons.’ A particular compass neuron should be activated whenever the animal is heading in a particular compass direction. **Figure 10** illustrates the proposed compass model based on neurophysiological data. For further details on the neurophysiological basis of polarization analysis, see the review by Homberg et al. 2011.
Figure 10: Proposed mechanism for the e-vector compass in insects. (A) The e-vector pattern in the sky with sun (yellow point) at 60° elevation; orientation and size of the blue bars represent the angle and degree of polarization. (B) Array of polarization detectors in the dorsal rim area (DRA), showing left (L) and right (R) visual fields. The position of the pink bars shows the fan-like orientation of tuned e-vector sensors; only a few of the 55–75 polarization (POL) detectors per eye are shown here. Each detector consists of a pair of orthogonally arranged photoreceptors (see Figure 7). The dashed line in the center indicates the animal’s longitudinal body axis. (C) Response ratios of three large-field POL neurons, represented here by false colors. If the animal rotates relative to the skylight pattern (see (B), black arrow), different false colors show up (see (C), white arrow). (D) Hypothetical compass neurons arranged in a circular array. Each compass neuron encodes a particular response ratio based on input from the broadly tuned POL neurons. The filled pink circle indicates the compass neuron that is maximally excited when the animal faces the solar azimuth. (From Wehner 2001. Reproduced with permission from the Journal of Experimental Biology.)
Insects such as water beetles, bugs, dragonflies, and butterflies that seek water for breeding can make use of the polarized light reflected off the water’s surface during dispersal flights to detect aquatic habitat. Because light reflected off of water is horizontally polarized, these insects are especially attracted to horizontally polarized light sources on the ground and may possess photoreceptors that are selectively sensitive to this plane of polarization. Backswimmers Notonecta glauca are a prime example, as well as dragonflies, mayflies, and waterstriders. Similarly, butterflies may take advantage of reflection off of smooth, shiny leaves to detect optimal oviposition sites. Insects that hunt on the water’s surface, such as Dolichopodid flies and waterstriders, have vertical polarizers to reduce glare (Horváth and Varjú 2004).

Another major function of polarization sensitivity is improved underwater vision and foraging. The marine environment imposes a greater challenge for visual predators and prey than the terrestrial environment because of the strong scattering and absorption of light by the water. Maximum beam attenuation length in water is approximately 15 meters, compared to 6 kilometers in air. Scattering degrades the contrast between objects and the background by interposing veiling light between the observer and the object, much like a dense fog (Lythgoe 1979). As we saw above, the ocean is also characterized by partially polarized light. The plane of polarization is mostly horizontal, but the pattern of polarized light varies greatly as a function of time of day and viewing angle. A vertically polarized visual analyzer reduces the amount of scattered light perceived and greatly increases visibility and contrast (Figure 11). This effect is analogous to the benefit we obtain by wearing Polaroid glasses to reduce street glare (horizontally polarized reflections).

Underwater visual animals can gain much more than haze reduction if they possess sensitivity to the full range of e-vectors with a polarization-opponent system. Biological tissues reflect and scatter light with different polarization characteristics. Just as with color, animals that are able to perceive polarization differences have additional contrast cues they can use to detect objects (Wehner 2001; Cronin et al. 2003; Sabbah and Shashar 2006). Two particular cryptic strategies by marine prey organisms—transparency and mirrored reflection—can be broken by predators with polarization sensitivity. Transparent animals are never completely transparent, although they can be very difficult to see from a distance. However, transparent tissues may modify the light transmitted through the body. Some tissues are birefringent, in which the refractive index varies with the plane of polarized light and light waves are split into unequally reflected or transmitted waves. Other materials, called quarter-wave retarders, are able to depolarize incident waves that are initially polarized. Thus some objects in the ocean medium may be less polarized along some vector than the background, while other objects may be more polarized. In addition, transparent animals may reflect and scatter some of the unpolarized downwelling light so that the prey animal stands out against the horizontally polarized background light (Figure 12) (Johnsen et al. 2011). The squid Loglio pealei is a polarization-sensitive visual
A predator that can detect zooplankton prey at 70% greater distance under partially polarized lighting than under nonpolarized lighting (Shashar et al. 1998). Moreover, under normal lighting, the squid were far more likely to attack transparent glass beads that were made polarization-active by heat stressing compared to transparent beads that were not polarization-active. Similar studies have also discovered polarization sensitivity in predatory crayfish (Tuthill and Johnsen 2006).

**Figure 12: Polarization breaks transparency.** Two views of a transparent prawn (*Lucifer* spp.). The photo on the left shows the prawn under normal lighting with no polarizing filters on the lens. The photo on the right shows the same animal viewed through cross-polarizers. Two orthogonal filters are positioned in front of the lens so that the background illumination is minimized. The animal is bright because it is scattering unpolarized light at other angles. The transparent tissues (cuticle and muscles) are also birefringent and modify the horizontally polarized transmitted light to various oblique angles. (Photo courtesy of N. Shashar.)

The logic behind the crypticity strategy of mirrored reflection is to match the background illumination. Many pelagic fish use silvery specular reflectance from guanine crystals in their scales as camouflage to reduce detection. The schooling fish in Figure 11 are attempting to use mirrored reflection to match their brightness to the background, but because they scatter more unpolarized light relative to the background, they become more visible when viewed through a vertical polarizing filter. In addition, silvery fish produce strong horizontally polarized specular reflectance at certain viewing angles (see Figure 2B,C), and a polarization-sensitive predator can perceive this reflection along the dorsal side of the prey (Shashar et al. 2000). Schooling fish may also be able to detect each other’s complex polarization reflections and use this information along with other senses to maintain school integrity (Rowe and Denton 1997).

The final function of polarization-sensitive vision is **social signaling to conspecifics**. A dramatic example occurs in the cuttlefish (*Sepia officinalis*) which uses controlled reflection of polarized light to produce species-specific signals on the arms, eyes, and forehead (Figure 13) (Shashar et al. 1996; Boal et al. 2004; Mäthger et al. 2009a). In contrast to the nonpolarized achromatic visual signals which are used in aggressive and courtship contexts, the polarized signals appear to be general indicators of sex and species identity. Females in particular respond differentially to images of conspecifics with and without the polarized signals. The patterns are turned on during normal alert activities, but disappear when the animal lies camouflaged in the sand and when it is engaged in aggressive interactions, attacking prey, copulation, and egg-laying. The signals arise from reflecting iridophores in a chromatophore organ. Iridophores contain flat guanine platelets that produce partially polarized reflections. Groups of iridophores are oriented at different angles so the patterns are visible under a range of horizontal viewing angles (Chiou et al. 2007). As in other cephalopods, these iridophores are dynamic cells, capable of undergoing ultrastructural changes on neural command (Cooper et al. 1990; Shashar et al. 2001; Cronin et al. 2003; Mäthger et al. 2009b). Such changes shift the iridophores between organized and disorganized alignments, which enables the animals to change their polarization reflectances on time scales of a second or less. Cuttlefish eyes, like those of many other cephalopods, have a horizontal band of specialized orthogonal polarization-sensitive photoreceptors (Talbot and Marshall 2011). As mentioned previously, cuttlefish also use their polarization sensitivity while foraging to break the prey camouflage strategies of transparency.
and specular reflectance.

**Figure 13: Frontal display of a cuttlefish, *Sepia officinalis.*** The left panel shows an alert animal under normal white light illumination with no camera filter. The right panel shows the same photo with a horizontal polarizing light filter, which reveals a striking polarized light pattern on the forehead, eyes, and arms. (From Cronin et al. 2003. Reproduced with permission from the Journal of Experimental Biology.)

As we saw in the main text, mantis shrimp possess eyes with phenomenal color discrimination abilities. One possible reason for their system of many narrow but overlapping photoreceptor absorbance curves is their need for color constancy given the large changes in light quality at different water depths (Cheroske et al. 2009). They seem to have one more backup system for visual communication under conditions of variable light quality: the use of polarization signals. Polarization is much more predictable and stable with increasing depth than spectral quality. A large fraction of their compound eyes—the dorsal and ventral hemispheres—are devoted to spatial resolution and achromatic brightness contrast, and these ommatidia are polarization-sensitive. The color receptors are located in rows 1–4 of the midband, while midband rows 5–6 all have the same photopigment and are specialized for polarization detection. Their location adjacent to the wavelength detectors, paired arrangement with receptors having perpendicular e-vectors, and neural wiring similar to the color detectors suggests that they are specialized for a polarization-opponent system. Many stomatopod species have body parts that are obviously specialized for the reflection of strongly polarized light and are used in behavioral contexts that seem clearly linked to intraspecific communication (**Figure 14**). Species-specific patterns based on differential reflection of partially linearly polarized light could be unusually direct and easy to interpret since no other objects in the scene are likely to have a similar appearance. They would also be private and invisible to animals lacking polarization sensitivity. The polarization of specific body parts must be produced structurally in the carapace, as they do not change over time and are even present in the molt casts (Marshall et al. 1999; Kleinlogel et al. 2003; Cronin and Marshall 2004).
Polarized-light signaling has been found in terrestrial forest environments, where the intensity and spectral composition of light varies greatly (see Figure 5.4 in the main text). Although natural polarization is limited because of strong filtering by the canopy, contrasting polarized light reflectance can provide a more consistent pattern for species recognition than colored signals. The butterfly *Heliconius cydno* reflects iridescent colors from the wings. The reflected light is both chromatically saturated and 90% polarized (Sweeney et al. 2003). Males of this species appear to recognize females based on this polarization. When the reflected light from females is artificially depolarized, males approach them much less frequently. A survey of the presence of polarized light reflectance from 144 species of nymphalid butterflies found 75 species with polarized patterns. These species were significantly more likely to inhabit forest habitats than open habitats (Douglas et al. 2007).

Finally, a few animal taxa have been found to reflect circularly polarized light, in which the electric vector rotates either clockwise (right-handed) or counterclockwise (left-handed) while the wave travels. Although very rare in the natural environment (starlight becomes partially circularly polarized, as well as underwater backscattered light outside of Snell’s window), it is not difficult to produce with manmade materials, and circularly polarized reflection (CPR) is widespread among scarab beetles. In a survey of 16,650 species from 1320 genera, 89% showed some level of CPR ranging from very low degrees of polarization up to a maximum of 97% (Pye 2010). Most cases had left-handed rotation. These beetles are generally green in color, but they can change in appearance from brilliant, metallic green to black when viewed with left- and right-handed polarizing filters (Hegedüs et al. 2006; Goldstein 2006; Pye 2010). The fine structure of the exoskeleton consists of tightly packed hexagon cells with a cone structure (Jewell et al. 2007; Sharma et al. 2009). Under the light microscope, the cones appear yellow in the center with a green surround. Concentric nested arcs encircle the cones to form a helical structure (Figure 15). Behavioral tests demonstrate that the
beetles respond selectively to objects of different degrees and rotations of CPR (Brady and Cummings 2010). Visual reception mechanisms for CPR have not been studied in the beetles yet, but in another group that also shows CPR, the mantis shrimps (stomatopods), a potential visual mechanism has been described (Chiou et al. 2008). In a linearly polarized light wave, the x and y vibrational e-vector components are in phase; when these vectors are out of phase, an elliptical wave results, and when the vectors are 90° out of phase, a circular wave results. If a circularly polarized light wave travels through a birefringent material with a thickness and refractive index that slows the wave in one e-vector orientation by 1/4 of a wavelength (called a quarter-wave retarder), the two vectors are brought back into phase, and the wave becomes linearly polarized. A few species of stomatopods (Odontodactylus) with the usual midband linear-polarized light detectors have placed such quarter-wave retarders in a layer over these photoreceptor cells to filter the incoming light. These species have been shown to distinguish behaviorally between left-CPR and right-CPR objects. Moreover, three species were found to have sex-specific circular reflectance patterns on body parts used for behavioral displays in males (Chiou et al. 2008). Circularly polarized light signals thus create a private communication channel in both scarab beetles and stomatopods.
Figure 15: Reflection of circularly polarized light from the scarab beetle *Crysina gloriosa*. (A, B) Photo of a beetle under unpolarized light illumination using a left-hand and right-hand circular polarizing filter. (C) Light microscope photo of the same beetle’s exoskeleton showing the packed, conical cells, approximately 10 mm in diameter, with yellow centers and green surrounds. (D) An x-y section of a confocal microscope image showing concentric rings that form the helical reflective surface (Sharma et al. 2009).

Further reading


The entire July 2001 issue of the *Journal of Experimental Biology* contains excellent review articles from the “Second Workshop on Ultraviolet and Polarization Vision.”
The March 2011 issue of *Philosophical Transactions of the Royal Society of London – B* contains a series of articles on “New directions in the detection of polarized light” stimulated by a small international meeting in 2008 on Heron Island, Australia.

**Useful websites**

http://polarization.com/


http://micro.magnet.fsu.edu/primer/java/scienceopticsu/polarizedlight/filters/

http://hyperphysics.phy-astr.gsu.edu/hbase/phyopt/polarcon.html


**Literature cited**


Web Topic 5.3
Evolution of Primate Color Vision

Introduction

A set of at least two photoreceptor types with pigments having different but overlapping absorption curves is essential for color vision, and larger sets with three or four photoreceptor types are even better. We now have a good understanding of how small changes in the amino acid sequence of the opsin protein component of different rhodopsin visual pigments can shift the wavelength of peak absorption to produce these photoreceptor types. Scientists have also identified the DNA regions that house the genetic code for the different opsins. Comparative studies of the opsin genes in a range of species provide us with invaluable information about the evolution of photopigments and color vision across the ancient and more recent animal taxa. In this Web Topic, we will briefly review the evolution of photopigments in the early vertebrates, outline the subsequent evolutionary history of mammals, and then examine the interesting examples of primate trichromacy and the source of our own color vision ability.

Early vertebrates

Vertebrates arose shortly after the Cambrian explosion, approximately 505 million years ago (mya). The first ancient ancestral vertebrates were the Agnathans, or jawless fish, represented today by a few extant species of lamprey and hagfish. Instead of jaws, they possess a suction mouth with rasping teeth and feed by attaching themselves to larger prey. They do swim around, live in shallow, clear water, and they have excellent eyes. Recent sequencing of their opsin genes indicates that they have four cone pigment types and good color vision (Collin et al. 2003). It is believed that these pigment genes all evolved from a single ancestral opsin gene which was duplicated and then subjected to different mutations and selection that favored a broad range of overlapping pigments. Rod opsins and vision adapted to low light conditions did not evolve until the jawed fish (Gnathostomata) arose around 485 mya. This rod gene, along with the four cone genes, persisted in subsequent vertebrate groups, including teleost fish, reptiles, amphibians, and birds. Table 1 below lists these basic vertebrate opsin classes.

Table 1: Basic vertebrate photoreceptor pigment classes and the range of peak wavelengths for each class.

<table>
<thead>
<tr>
<th>Pigment name</th>
<th>Photoreceptor</th>
<th>Wavelength range</th>
<th>Color range</th>
</tr>
</thead>
<tbody>
<tr>
<td>LWS</td>
<td>Cone</td>
<td>495–570</td>
<td>Green–red</td>
</tr>
<tr>
<td>SWS1</td>
<td>Cone</td>
<td>355–450</td>
<td>UV–violet</td>
</tr>
<tr>
<td>SWS2</td>
<td>Cone</td>
<td>415–480</td>
<td>Violet–blue</td>
</tr>
<tr>
<td>RH1</td>
<td>Rod</td>
<td>460–530</td>
<td>Green</td>
</tr>
<tr>
<td>RH2</td>
<td>Cone</td>
<td>470–530</td>
<td>Green</td>
</tr>
</tbody>
</table>

Source: After Bowmaker 2008.

Species can modify these opsin genes via several mechanisms. First, mutations within an opsin gene can lead to a spectral shift in the sensitivity of the pigment. A single nucleotide substitution may lead to the replacement of an amino acid that alters the interaction between the chromophore and opsin, leading to a spectral shift. There are a relatively small number of amino acid substitution sites that lead to meaningful spectral shifts, and the effects of different substitutions are usually additive (Yokoyama and Radlwimmer 2001; Yokoyama 2002). Selection will favor substitutions that fine-tune a species’ color vision system to its environment and specific visual tasks. Second, a species may modify the ancestral vertebrate pattern of four spectrally distinct cone classes by the loss of one or more of the cone classes, followed by shifts in the wavelength peak. Third, new pigments in the same class can be generated by gene duplication. Mutations in the duplicated genes can then lead to the divergence in their absorption peaks, creating two or more spectrally distinct pigments within a single opsin class (Bowmaker and Hunt 2006). This is the classic
“duplicate and diverge” strategy. Finally, the sequential order of these last two steps can be reversed in a “diverge and duplicate” strategy, where several alternative alleles with different absorption peaks arise at a gene locus, followed by a duplication event that fixes one allele at a new locus (Surridge et al. 2003).

Mammals first arose in the early Triassic period about 250 mya from a Therapsid reptile ancestor. Most early mammals were small and nocturnal, being dominated ecologically by the larger dinosaurs. This situation persisted until the end of the Cenozoic era 65 mya, when the dinosaurs went extinct. Although birds and modern reptiles clearly retained tetrachromatic color vision from their fish and reptilian ancestors, the early nocturnal mammals lost two of their cone pigments, specifically RH2 and SWS2. They retained the two spectrally extreme classes, but shifted their absorbance peaks inward toward the middle of the visible wavelength range. For several mammalian species whose genome has been sufficiently well-sequenced, the retained short-wave SWS1 gene is located on an autosomal chromosome while the long-wave LWS gene is located on the X chromosome (Ahnelt and Kolb 2000). This chromosomal pattern could be widespread and ancient in mammals, and, as we shall see below, has led to some bizarre sexual differences in color vision abilities in a few species. A rod-rich retina with dichromatic vision was probably the norm for the ancient nocturnal species, although marine mammals and a few terrestrial mammals lost another cone pigment (SWS1) and became monochromats (Ahnelt and Kolb 2000). Interestingly, the echidna and platypus (monotremes) share the LWS and SWS2 genes with reptiles, birds, and fishes, suggesting that the mammalian loss of SWS2 and RH2 occurred in the common ancestor of marsupial and placental mammals (Wakefield et al. 2008). When mammals finally radiated in the Pleistocene, becoming larger, diurnal, and carnivorous, most did not improve their color vision capabilities by evolving new long-wave pigments. Trichromatic color vision has evolved only a few times in select groups, but these independent evolutionary events enable us to examine the potential factors favoring better color vision (Bowmaker 1998; Yokoyama 2000; Bowmaker and Hunt 2006).

Marsupials

Recent studies of the spectral characteristics of photoreceptors in four Australian marsupial species—the fat-tailed dunnart (*Sminthopsis crassicaudata*), the bandicoot (*Isodon obesulus*), the honey possum (*Tarsipes rostratus*), and the quokka (*Setonix brachyurus*)—suggest that trichromacy may be present in these species (Arrese et al. 2002; Arrese et al. 2006). The first two species belong to the polyprotodont marsupial taxonomic division; the second two belong to the diprotodont division. Microspectrophotometry has indicated that these species possess three classes of cone photoreceptors maximally sensitive in the UV, green, and red ranges. Molecular analysis has shown that the UV pigment belongs to the ancestral SWS1 gene class and that the red pigment belongs to the LWS class; the middle wavelength pigment may be derived from the rod RH1 pigment (Cowing et al. 2008). The intriguing possibility is, therefore, that in some marsupials—in marked contrast to placental mammals—the RH2 opsin gene has been retained and is expressed (Arrese et al. 2002). Middle wavelength cone types seem to be present in some species but absent in others (e.g., wallabies *Macropus eugenii*, which are clearly dichromats in behavioral tests; Deeb et al. 2003).

A color vision model was developed for the honey possum to test three alternative hypotheses for the selective advantage of its trichromatic receptor tuning. This mouse-like mammal is a crepuscular nectarivore that feeds primarily on yellow and red flowers. The absorption peak for the longer wavelength cone type is 557 nm, more red-shifted than the LWS cones of other marsupials. The results of the model suggested that the visual task selecting for improved color vision was not detecting red and yellow against a green leaf background, for which an even further red-shifted peak would be optimal. Instead, the tuning may be designed for discriminating the stages of maturity (green to yellow) of the animal’s major nectar food source, the flowers of *Banksia attenuata* (Sumner et al. 2005).

Primates

The Old World Catarrhine primates, including macaques, baboons, guenons, great apes, and humans, possess three cone photopigment types (*Figure 1*) and true trichromatic vision, often called routine trichromatic vision to distinguish it from the New World primate system described below. The short wavelength pigment belongs to the SWS1 class and the two longer wavelength pigments both belong to the LWS class. The most likely scenario for the evolution
of trichromacy in this taxon is a duplication of the ancestral LWS gene followed by divergence of the spectral peaks (Surridge et al. 2003; Jacobs 2007, 2008). This duplication event must have occurred at the base of the catarrhine lineage around 30–40 million years ago. The green photopigment gene appears to be the duplicated one, inserted immediately downstream of the red gene and its locus control region (LCR) on the X chromosome. The red gene also shows greater homology with the ancestral mammalian LWS gene (Dulai et al. 1999).

Figure 1: Photopigment absorbance curves for Catarrhine primates, including humans. The dashed line shows the rod opsin absorbance curve (R), and peak wavelengths for the blue (S, or short), green (M, or medium), and red (L, or long) cone types are indicated at the top of each curve.

Color vision is much more variable in New World platyrrhine primates, a monophyletic group that includes owl monkeys, capuchins, marmosets, squirrel and spider monkeys, howlers, and others. Most of these species exhibit allelic trichromacy, which is based on only two opsin gene loci, an autosomal SWS1 gene as in Old World primates, and a polymorphic LWS gene located on the X chromosome. The LWS locus has several allelic forms that encode pigments with different wavelength peaks between about 535 and 565 nm. The occurrence of this polymorphic X-linked photopigment gene leads to individual and sex-based differences in color vision capabilities. Males possess only one allelic form of the LWS gene located on their single X chromosome. Together with the autosomal SWS gene, males have two cone pigment types and are always dichromats. Females have two X chromosomes. Heterozygous females that inherit different LWS alleles from each parent possess two long-wave cone pigments in addition to the short-wave pigment and are trichromats; homozygotic females at this locus are dichromats like the males. Even more variation is introduced by the presence of three LWS alleles in most species, so there are three different types of heterozygous female trichromats plus three different dichromat phenotypes (Mollon 1989; Boissinot et al. 1998; Talebi et al. 2006; Jacobs 2007, 2008). Some members of the lemur family (strepsirrhine primates) also exhibit allelic trichromacy (Tan and Li 1999).

There are two interesting exceptions to this New World primate color vision pattern. The howler monkeys, *Alouatta*, have evolved routine trichromacy with virtually the same photopigment absorption curves as the Catarrhine primates. Even the critical amino acid substitutions are the same. The howler monkey opsin genes are otherwise very similar to other platyrrhine opsin alleles, indicating that routine trichromacy in howlers is a recent and independent evolutionary event. It is believed that they diverged from a platyrrhine ancestor with the allelic trichromacy system by duplicating one of the
X chromosome alleles and then fixing it at an adjacent new locus—the diverge and duplicate strategy (Surridge et al. 2003). As in the catarrhines, male and female howlers possess the same visual capabilities (Kainz et al. 1998; Dulai et al. 1999). The second exception is found in the nocturnal owl monkeys (Aotidae). Their rod-rich retinas are specialized for nocturnal vision and their SWS cone pigment gene has mutated to the point of losing functionality. These New World monkeys are monochromats with a single X-linked LWS gene (Jacobs et al. 1996; Silveira et al. 2001; Jacobs 2007, 2008).

**Evolution of two color-opponent systems**

The evolution of a third cone pigment is only one of the critical steps that must occur to achieve true trichromatic color vision from a dichromat ancestor. In addition, a second color-opponent system must be established. Two additional steps or conditions must be met: (1) the new visual pigment must be expressed in a distinct class of photoreceptors, and (2) patterns of neural wiring must develop that can extract chromatic information by comparing the degree of excitation of the new and preexisting classes of photoreceptors. How did this happen in trichromatic primates, especially in the context of the X-linked location of two of the opsin genes?

All trichromatic primates appear to possess two color-opponent systems. The ancestral dichromatic mammals had a single color-opponent system based on two cone pigments, a short-wave pigment (S) absorbing maximally in the blue region and a long-wave pigment (L) absorbing in the yellow-green region. These opsin genes were independent, located on separate chromosomes with their own locus control regions (LCR). Prior selection for a functional photoreceptor mosaic would have favored genetic control over opsin gene expression in specific cone types arranged in an adaptive configuration in the retina (Ahnelt and Kolb 2000). Antagonistic (inhibitory) interactions between the neural outputs of these two cone types led to the ancestral color-opponent system called the blue–yellow chromatic channel. With the duplication and divergence of the long-wave pigment gene into middle- (M) and long-wave pigments, the second color-opponent system had to involve antagonistic interactions between cones separately containing these two pigments. Even though the absorption peaks for these pigments are not very far apart (green and yellow-green), this opponent system permits greater sensitivity in the green to red region of the light spectrum, and is called the red–green chromatic channel. **Figure 2** shows the classic evidence for two color-opponent systems in the macaque, a catarrhine primate. Similar neural evidence for two color-opponent systems has also been described for individual platyrrhine monkeys with allelic trichromacy (i.e., heterozygotic females) (Yeh et al. 1995).
How did the M and L cone types develop the ability to express just one opsin pigment gene per cell? The mechanism appears to be different for each of the three independent origins of primate trichromacy. In the case of allelic trichromacy, the process of random X-chromosome inactivation was already established to prevent overdosing of X-linked enzymes in females. In each cell of a female mammal, one of the X chromosomes is de-activated on a random basis. A given photoreceptor cell therefore expresses only one of the two opsin alleles in a heterozygous female. In the case of the catarhine primates, the gene duplication event involved only the duplication and head-to-tail insertion of the opsin coding region of the green (M) gene, and it shares the locus control region with the red (L) gene. The two alleles apparently compete for access to the joint control region, so that only one allele is expressed in a given cell on a random basis (Smallwood et al. 2002). Finally, in the howler monkey case, the duplication and insertion event involved the entire opsin gene plus control region, so the M and L genes each possess their own LCR. Selection must have then operated to regulate the expression of one or the other gene in each cone cell (Smallwood et al. 2002; Bowmaker and Hunt 2006).

How did the neural connectivity become established for a second color-opponent system? In particular, how can the wiring mechanism work in a species in which some individuals only have two pigment genes while others have three? The answer lies in the pre-adapted primate fovea. Nonprimate mammals sum the outputs of several cone receptors onto ganglion cells before transmission to the brain. This pooling not only entails loss of spatial resolution, but it also means that it would not be possible to derive an additional chromatic signal from a third (new) cone type without parallel modification of neural connections. However, early primates evolved a specialized fovea for fine-scale resolution of spatial details. The fovea has an additional class of retinal ganglion cells, the midget ganglion cells that receive primary input from a single cone cell. It is these midget ganglion cells that encode the red–green signal in trichromatic species. Thus prior selection for finer achromatic resolution may have established the neural pathways that pre-adapted the primates for separate use of M and L cone opsins and trichromatic vision (Ahnelt and Kolb 2000; Surridge et al. 2003).
the background were computed for the visual phenotypes. With three pigment gene alleles, there are six different visual phenotypes, and the researchers were particularly interested in evaluating the color discrimination ability of the anomalous trichromats, with their M and L pigment peaks very close together compared to the normal trichromatic phenotype with more widely separated M and L peaks. They also examined the effects of bright versus dim ambient light on the discrimination abilities of the different visual phenotypes. Figure 3 shows some of their results for squirrel monkeys (results were very similar for the tamarins). Normal trichromats, with the broadest spread of pigment peaks, were better at distinguishing target fruit under all conditions, but especially so under dim light. Anomalous trichromats were surprisingly good under bright light but less so under dim light, and were clearly better than dichromats. Dichromats performed better if they had longer-wavelength pigments, but under no conditions did they outperform the trichromats.
Figure 3: Color model of food detection by squirrel monkey visual phenotypes. (A) Measured spectral curves for the background (B) and target fruit (T). Photoreceptor quantum catches by receivers are estimated from color models combining responses for each receptor type, plus noise, for the target and background. (B) The target and background stimuli are located in a Cartesian space whose axes are given by the responses of S, M, and L cones. For T and B spectra, the estimated cone excitations locate the centers of ellipsoids whose dimensions are given by the standard deviation of noise in each cone mechanism. (C) The ellipsoids are projected onto a two-dimensional chromatic surface to estimate discriminability, the difference, $d$, between targets and various backgrounds (B$_1$–B$_3$). Only differences in hue and saturation (but not brightness) are incorporated in the model. (D) Pigment wavelength peaks for the six phenotypes. (E) The percentages of fruit items that are distinguishable (greater than threshold) against the background for the main phenotypes under bright illumination and dim illumination. (After Osorio et al. 2004).
Several studies have compared the foods taken by species with different color vision systems. Lucas et al. (2003) collected the fruit and leaves eaten by five catarrhine species and three platyrrhine species, measured their color and nutritive value, and modeled the hue contrast for these items relative to the background. Routine trichromats ingested leaves that were red-shifted compared to background foliage more frequently than allelic trichromats. They did not find any differences in fruit color between the two groups, and argued that the consumption of young leaves at the optimal nutritional state seemed to be the primary visual task selecting for routine trichromacy. Leonhardt et al. (2009) compared fruit colors and foraging performance for four lemur species and found a moderate shift toward redder fruit in ruffed lemurs which have allelic trichromancy, but also found that dichromatic collared lemurs were very efficient at retrieving red and green food items under camouflage conditions. Stoner et al. (2005) compared the color of fruits taken by sympatric howler and spider monkeys, and found that spider monkeys actually took more red fruit, while howlers took more green fruit (Figure 4). They concluded that dichromats were not at any particular disadvantage when searching for colored fruit, and that howler monkeys may have evolved routine trichromacy to detect young leaves better.

**Figure 4: Primate fruit color selection.** Chromaticities of dietary fruit items taken by spider monkeys *Ateles geoffroyi* (*n* = 25 fruit species) and howler monkeys *Allouata palliata* (*n* = 12 fruit species). The *y*-axis represents computed chromaticity on the yellow–blue channel (higher values are more blue) and the *x*-axis represents the red–green channel (higher values are more red), modeled from the color-sensitive perspective of each species separately. The sampled fruits comprise 70% of the total fruit diet for both monkey species. The fruit species taken by both monkey species are included and indicated by purple circles. Circle size represents the percentage of each fruit species in the diet of each primate (the bigger the circle, the greater the percentage it represents). The dashed rectangle represents the chromaticities of a mature-leaf background. (After Stoner et al. 2005.)

Studies on captive primates (and humans) given various types of discrimination tasks find that while trichromats generally perform better with colored objects against more or less contrasting backgrounds, trichromats can actually be confused when presented with cryptic stimuli or with colored objects on a colorful background (Morgan et al. 1992; Caine and Mundy 2000; Caine et al. 2003; Smith et al. 2003a; Saito et al. 2005a, b; Rowe and Jacobs 2007; Prado et al. 2008; Caine et al. 2010). Contrary to Orsorio et al. (2004), dichromats were found to perform better under low light conditions. These results suggest that dichromats rely more on achromatic contrasts.

With the ability to genotype individual monkeys in the field using blood or feces, it is now possible to determine whether
there are any fitness or foraging strategy differences based on visual phenotype in allelic trichromatic species. Field studies can address the question of how the polymorphism of dichromats and trichromats remains stable in these species. Several hypotheses have been proposed. One hypothesis is **heterozygote advantage**, whereby heterozygous trichromatic females maintain the polymorphism via some fitness advantage. In this case, some type of balancing selection would operate to keep the number of L/M cone opsin alleles stable at three (Riba-Hernandez et al. 2004). Assuming Hardy-Weinberg equilibrium, the percentage of heterozygous females is 50, 67, and 75 with two, three, and four alleles, respectively. But the number of useful alleles may be limited by the available “color space” in the 535 to 565 nm range. Three alleles may represent a compromise that maximizes the distance between spectral tuning peaks (creating a useful color-opponent system) while maximizing the number of alleles to yield the greatest possible frequency of heterozygotes (Cropp et al. 2002). Real populations show a reasonable fit to these expectations, with most species having three alleles but a few having two and one unusual species (titi monkeys, *Callicebus moloch*) having 5 (Surridge and Mundy 2002; Surridge et al. 2005a; Jacobs and Deegan 2005; Hiramatsu et al. 2005; Jacobs 2007; Hiwatashi et al. 2010). In tri-allelic species, the three alleles are often not equal in frequency in the population, implying some selection on different types of heterozygotes or possibly inbreeding. For example, in tamarins (*Saguinus labiatus*), the middle-wavelength allele is rare, and field studies suggest that females with the two more extreme-wavelength alleles have greater longevity and prefer mates that would give their offspring the widest divergence in opsin phenotypes (Surridge et al. 2005b).

A second hypothesis is **frequency dependence**, whereby dichromats have an advantage over trichromats under some conditions. Human dichromats can detect patterns based on lightness that are indistinguishable to trichromats (Morgan et al. 1992), and, as mentioned above, experimental studies show that primate dichromats are able to select the correct visual stimulus under camouflaged conditions better than trichromats. Recent field studies designed to assess whether dichromats and trichromats differ in foraging strategies have yielded mixed results. A series of studies on white-faced capuchins (*Cebus capucinus*) have found a number of differences: trichromats spend more time foraging visually, select more red fig and high-quality fruits, and preferentially feed on colonial insects that require extraction (ants, wasps); dichromats rely more on taste and smell, spend more overall time foraging (possibly because they are not able to select the highest quality fruit), and are more successful in finding exposed but cryptic insects (Melin et al. 2007, 2009, 2010; Vogel et al. 2007). A study of black-handed spider monkeys (*Ateles geoffroyi*) found no differences in foraging efficiency and suggested that brightness contrast was the most important fruit identification cue, not hue (Hiramatsu et al. 2008). Finally, it has been suggested that the social nature of group foraging in these primates places no disadvantage on dichromats if their trichromatic group mates can assist them in finding food. However, one field study of tamarins found no tendency for trichromatic females to lead the progression of foraging groups (Smith et al. 2003b). The two hypotheses for opsin gene polymorphism maintenance in platyrrhine monkeys—heterozygote advantage and frequency dependence—are not mutually exclusive, so both processes may be operating simultaneously.

The unequal spacing of the spectral tuning peaks in all trichromatic mammals, compared to the more even distribution in other trichromatic and tetrachromatic taxa, may be an adaptive solution for making certain types of hue discriminations. The primate spectral tuning appears to be optimized for detecting red and orange fruit and leaves against a background of mature leaves, but is not well optimized for discriminating degrees of fruit ripeness, for which a greater spread of pigment peaks would be better (Sumner and Mollon 2000b, c; Osorio and Vorobyev 2005). This differs somewhat from the case of the marsupial honey possum mentioned earlier, where the pigment spacing it best tuned for discriminating degrees of flower maturity rather than contrast of flowers against the vegetation (Sumner et al. 2000). One possible cost of this uneven distribution, however, may be less effective color constancy, the ability to adapt to strongly colored ambient light. Models have shown that this ability is better when there are a greater number of narrowly tuned curves that are evenly spaced (Osorio et al. 1997). Certainly any human can attest to the frustrating difficulty of matching colors under different qualities of ambient light!

**Further reading**


Literature cited


Caine, N. G., D. Osorio, and N. I. Mundy. 2010. A foraging advantage for dichromatic marmosets (Callithrix geoffroyi) at


Mollon, J. D. 1989. Tho she kneeld in that place where they grew ... the uses and origins of primate color vision. *Journal of Experimental Biology* 146: 21–38.


Web Topic 6.1
How Are Pheromones Identified?

Introduction

Identification and analysis of chemical signals has lagged behind the analysis of visual and auditory signals because our own sense of smell is rather poor. We generally cannot smell most animal pheromones, so we are not aware of most chemical senders. In addition, chemical signals are complex blends in many species, especially in insects and mammals. To compound matters, the amount of chemical released is often small, which makes it difficult to collect a sufficiently large sample for chemical analysis. Nevertheless, new techniques have improved our ability to collect the secretions and analyze the chemical components. Although it is not possible to conduct “playback” experiments in digital form from a computer as we do with acoustic and video signals, one can release controlled amounts of specific chemical components to assess differential responses. In this Web Topics unit we shall briefly outline the typical steps taken to identify chemical signals, drawing in large part from Chapter 2 of Wyatt (2014).

Step 1: Finding a bioassay

The first step in the process of identifying a chemical signal is observing the animals and discovering behaviors or interactions that appear to be influenced by chemicals. As we mentioned in the main text, ritualized marking behaviors and release postures are clear indicators of the sending of chemical signals, and these behaviors also provide clues to the likely location of secretory glands. However, many chemical signals are released without such visual correlates. It is more important to identify receiver responses to putative chemical signals because these will provide the bioassay needed to test the salience of chemical components in all of the subsequent steps below. Olfactory investigation behaviors such as increased sniffing (mammals) or tongue flicking (lizards and snakes), and very obvious responses such as flehmen in male mammals, aggregation in bark beetles, and wing fluttering in male moths, make excellent bioassays. Some chemicals signals produce specific emotional responses, such as the agitation, mobbing, and retreat responses from alarm pheromones. Chemical signals with a priming function have more subtle effects on receivers, such as changes in hormonal profiles and reproductive condition. The most useful bioassay for further analysis of a putative pheromone or chemical signal is a reliable behavior or other response associated with receipt of the natural chemical product.

Simple bioassays that are easily detected in a laboratory situation will be most useful. For instance, the wing fanning behavior of male moths is a far simpler test of responses to female pheromone products than flight tests in a wind tunnel. However, there are tradeoffs here, because males will readily wing fan in response to partial blend components, whereas upwind anemotactic flight depends on recognition of the appropriate mixture (Charles Linn, personal communication). The ability to discriminate among odors can also be tested with various types of laboratory tasks, such as reward-based operant conditioning discrimination tasks, habituation tasks, and free-field natural presentation of stimuli. Care must be taken to ensure that all other test conditions are controlled for, such as time of day, animal age, reproductive state, and sex. Nevertheless, the results may differ depending on the methodology used (Schellinck et al. 1994, 1995; Brown et al. 2000; Schellinck et al. 2001). Finally, because many chemical signals are blends that may contain redundant information, mere removal of a specific component may not diminish the response, even though it is an important component. Thus care must also be taken to present different combinations of the blend components to look for synergistic effects, using a bioassay that is highly sensitive to the quality of the mixture.

If the olfactory receptors of a species are sufficiently well understood, one can also use the animal’s own sense organs as direct bioassay detectors. This procedure has been well developed in the study of moth pheromones, where the investigator implants electrodes in the base of the antennae to record the electrical responses (electroantennogram, EAG) to various airborne odors (van der Pers and Denotter 1978; Groot et al. 2005). Simple portable units with moth antennae as the sensor have been developed to monitor presence of pheromone in the wild (van der Pers and Minks 1993, 1998). This method will not detect pheromone components for which there are only a few receptor cells, so a more accurate method is single-sensillum recording (Baker et al. 2004; Linn et al. 2007; Inoshita et al. 2011). Neurophysiological studies are also widely used to understand pheromone detection in vertebrates (Su et al. 2009; Tirindelli et al. 2009; Touhara and Vosshall 2009).
Step 2: Collecting the chemicals

Pheromone chemicals are often collected by squeezing or swabbing potential glands or by excising the whole gland. Unfortunately, such crude extracts of gland products risk gathering many non-pheromone components, including carrier compounds and chemical precursors of the pheromone. For volatile pheromones, a better method is to trap and concentrate the chemicals actually released by the animal (or plant) into what is called the headspace area above it. With this method, clean air (or water) is passed over the animal and the chemicals in the exhaust are then trapped by special chemical “sponges” or “wicks”. These are porous polymer resins that adsorb organic molecules (e.g., Porapak-Q™ and Tenax™). They release the chemicals when heated or washed with organic solvents. Exhaust air can also be cooled to -195° C with liquid nitrogen. Washing a gland or polymer resin with solvent results in a relatively diluted solution with low resolution for subsequent analyses. An alternative method is Solid Phase MicroExtraction (SPME). This technique avoids the use of solvents by employing an inert fiber coated with an absorbing polymer, which can then be directly exposed to the input air or water directed into the gas chromatography analyzer (see below) (Pawliszyn 1997, 1999; Ouyang et al. 2011).

Step 3: Separating and identifying the chemicals

Once the chemicals have been collected, the goal is to find the active components. To make some initial inroads into figuring out the pheromone compound, a traditional chemical technique called fractionation may be undertaken. This process separates the components based on their physical characteristics, such as solubility in different solvents. The sample is typically mixed in water and a non-water-soluble organic solvent and shaken well. Once the polar water phase and non-polar solvent phase separate out, one can then use the bioassay test to determine which fraction contains the active component. This process narrows down the general chemical class of the active component into lipophilic alkanes, oils, and fats, versus hydrophilic acids, alcohols, aldehydes, and esters, and also purifies and concentrates the pheromone.

Next, the sample is submitted to analysis by gas-liquid chromatography (GC). The development of this technique has revolutionized the study of chemical signals, and it is especially useful when there is a mixture of components and small amounts of sample. It can only be used for chemicals that vaporize upon heating without decomposing. The gas chromatograph (Figure 1) separates the chemicals on the basis on their physical characteristics. The instrument inputs a steady flow of inert gas, usually nitrogen or helium, into a long thin tube called the column (although it is coiled for space efficiency). The column is then heated, and a small sample of gas or solution containing the pheromone is then injected into the tube with a syringe. The inside of the tube is coated with a thin layer of liquid or polymer that interacts with the chemicals as they pass through. The compounds in the sample are separated based on their vapor pressures (boiling points) and their relative solubility in the inert gas. Smaller, more volatile components travel faster through the tube and exit first, followed by larger and less volatile components. The retention time in the tube is the key output variable that distinguished the components. A known reference chemical is included in the sample so that the unknown components can be assessed relative to the reference peak.
Some type of detector is needed to record the spikes as each chemical component exits the tube. A common detector is the **flame ionization detector** (FID). The FID ionizes the chemicals as they emerge from the column using a small hydrogen and air flame. It burns the chemicals, producing positively charged ions and negatively charged electrons. Two electrodes are used to measure the potential difference between them. The measured current is proportional to the number of reduced carbon atoms in the flame, so it is sensitive to the mass of the chemical compound. Another detector is the **mass spectrometer** (MS). This instrument also ionizes the sample, but with an electron beam, and the ions are separated according to their mass-to-charge ratio based on the electromagnetic fields they generate. This method is especially useful for analyzing larger molecules such as peptides.

The GC analysis can identify the chemical components of a mixture, but it does not reveal which ones are the active ones, and certainly does not allow the investigator to collect aliquots of the components for further testing with a bioassay. However, an ingenious method has been devised to obtain this information at least from insect subjects. It involves coupling the GC analysis with electroantennogram detection (EAD) or single-sensillum recording, as also shown in Figure 1. Basically the gas sample is split once it has passed through the column, with parallel streams entering the FID or MS detector and the neurophysiological set-up with the insect’s antenna. The timing of the spikes in the neural responses can then be linked to specific peaks in the detector output. Some examples of studies on a variety of insects using this technique can be found in the following articles: Toth et al. 1991; Leal et al. 1997; Gunawardena et al. 1998; Kalinova et al. 2006; Leal et al. 2006; Burger et al. 2008; Kim et al. 2009; Vitta et al. 2009; Zhang et al. 2011.

When it is not feasible to conduct parallel neurophysiological and gas chromatograph recordings, especially for vertebrates, there are two alternatives. First, one can pursue the fractionation process to separate and purify chemical components using the chemist’s tool kit. This strategy requires obtaining a sufficiently large sample of the pheromone-containing substance from the gland or resin washing. At each step, the separated aliquots are tested for bioactivity using the bioassay. Second, if gas chromatography has resulted in identification of the chemical structures of the components, these compounds can then be synthesized. Again, each synthesized component must be tested.
with the bioassay to make sure it is the correct chemical. This process can be tedious if there are several active components acting synergistically along with a variety of inert gland components.

A good example of the use of these techniques is the discovery of the sex attractant of female red-sided garter snakes (Thamnophis sirtalis parietalis). Early field studies showed that males were attracted to the trails left by moving females and they required an intact vomeronasal organ to accomplish this task. This indicated that a non-volatile component of the female’s skin was the likely source of a chemical signal. Once males approach a gravid female, they begin to court them by rubbing their chin along the female’s dorsal side; this chin-rubbing behavior provided the reliable bioassay. Organic solvent washes of the female’s skin revealed a large amount of lipid material, but there are no obvious skin glands. Serum and liver extracts from receptive females applied to unreceptive females made them attractive to males, so it was initially hypothesized that the pheromone source was the egg-yolk protein vitellogenin, which is produced by females under the control of estrogen (Garstka and Crews 1981). However, field trials and other evidence did not support this hypothesis. Gas chromatograph analysis of the skin lipids (coupled with MS and nuclear magnetic resonance spectroscopy) then revealed the presence of a series of 12 non-volatile saturated and monounsaturated long-chain (29- to 39-carbon) methyl ketones. These compounds were then successfully synthesized, and the mixture of synthetic compounds produced the full courtship bioassay response when applied to towels (Mason et al. 1989; Mason et al. 1990). This was the first pheromone identified and synthesized in reptiles (Mason and Parker 2010). Several amphibian and many rodent pheromones have now been identified (Beynon and Hurst 2004; Houck et al. 2007; Houck 2009; Woodley 2010). Many of these pheromones are proteins, requiring additional techniques to identify them such as polymerase chain reaction (PCR) amplification of DNA and amino acid sequencing.

**Further reading**


**Web resources**


Oregon State University pheromone labs: [http://plethodon.science.oregonstate.edu/index.html](http://plethodon.science.oregonstate.edu/index.html)

and [http://masonlab.science.oregonstate.edu/index](http://masonlab.science.oregonstate.edu/index)

**Literature cited**


Web Topic 6.2
Guide to Organic Compounds and Biosynthetic Pathways

Basic compounds and nomenclature

Organic molecules are based on a chain of carbon atoms with 1–3 attached hydrogen atoms. The carbon backbone usually forms a zigzag shape because of the tetrahedral arrangement of the four possible bonds around each carbon atom. The hydrogen atoms attached to the backbone lie in two planes, one above the plane depicted below by the solid wedges, and one below the plane depicted below by the dashed wedges.

![Carbon Backbone](image)

The structure is often simplified to show just the carbon backbone.

When other atoms such as oxygen (O), nitrogen (N), sulfur (S), or other functional groups are present, their letter symbols are inserted in the chain. The naming of organic compounds indicates the length of the carbon chain and what and where the important functional groups are located.

Functional Groups

The table below shows the formulae and names of the main functional groups.

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Prefix</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol</td>
<td>–OH</td>
<td>Hydroxy-</td>
<td>-ol</td>
</tr>
<tr>
<td>Aldehyde</td>
<td>–CH=O</td>
<td>Formyl-</td>
<td>-al</td>
</tr>
<tr>
<td>Ketone</td>
<td>&gt;C=O</td>
<td>Oxo-</td>
<td>-one</td>
</tr>
<tr>
<td>Carboxylic acid</td>
<td>–COOH</td>
<td>Carboxy-</td>
<td>-oic acid</td>
</tr>
<tr>
<td>Ester</td>
<td>–COOR</td>
<td>R-oxycarbonyl-</td>
<td>-R-oate</td>
</tr>
<tr>
<td>Amine</td>
<td>–NH₂</td>
<td>Amino-</td>
<td>-amine</td>
</tr>
<tr>
<td>Thiol</td>
<td>–SH</td>
<td>sulfanyl-</td>
<td>-thiol</td>
</tr>
</tbody>
</table>


Hydrocarbons

Hydrocarbons contain only carbon and hydrogen. They vary in length and the presence, number, and position of
double or triple carbon bonds. The basic classes are shown in the table below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alkane</td>
<td>$C_nH_{2n+2}$</td>
<td>-ane</td>
</tr>
<tr>
<td>Alkene</td>
<td>$C_nH_{2n}$</td>
<td>-ene</td>
</tr>
<tr>
<td>Alkyne</td>
<td>$C_nH_{2n-2}$</td>
<td>-yne</td>
</tr>
</tbody>
</table>

The longest continuous chain of carbons ($n$) is the parent structure and specifies the prefix: $1 = \text{meth}-$, $2 = \text{eth}-$, $3 = \text{prop}-$, $4 = \text{but}-$, $5 = \text{pent}-$, $6 = \text{hex}-$, $7 = \text{hept}-$, $8 = \text{oct}-$, $9 = \text{non}-$, $10 = \text{dec}-$, $12 = \text{dodec}-$, $15 = \text{pentadec}-$, and so on. The position of a functional group or double/triple bond is specified by counting the carbons from whichever end gives the lowest possible position number. For example, the compound shown below, which is the pheromone of the pink bollworm (*Pectinophora gossypiella*), is (Z,E)-7,11-hexadecandien-1-ol acetate. The “dien” specified that there are two double bonds, and the numbers specify that they occur at carbons 7 and 11 counting from the functional group (acetic acid) on the right (from Wyatt 2014).

![Compound Diagram](image)

**Isomers**

Molecules with the same formula can be assembled in different ways. The arrangement of atoms affects the shape of the molecule and therefore affects its binding (ligand) properties with receptor and binding proteins. There are several types of isomers. Positional isomers have functional groups or double bonds in different positions, as in the examples shown below (from Wyatt 2014).

![Isomers Diagram](image)

The carbonyl unit occurs at a different position in these two ketones.
Stereoisomers have the same connectivity but differ in the arrangement of atoms in space. Geometrical isomers have a spatial twist around a double bond and are indicated by E and Z (formerly, trans and cis, respectively). E stands for entgegen (opposite) and Z stands for zusammen (together), and refers to the location of the “high-priority” substituents on the same or different sides of the double bond.

Optical isomers, also called enantiomers, are mirror images of each other. Solutions of pure enantiomers rotate the plane of polarized light in opposite direction. Variants are indicated by L and D (or left and right, or – and +).
Classes of larger organic compounds

Amino acids, peptides and proteins

Amino acids contain both amine and carboxyl functional groups. Different amino acids have a variety of other functional groups and side chains attached, including cyclic hydrocarbon units. There are 20 standard amino acids used by cells in protein biosynthesis that are specified by the DNA genetic code. These twenty amino acids can be biosynthesised from simpler molecules, but organisms differ in how many they are able to produce and essential amino acids must be obtained in their diet. A few are shown below, and more can be found at [http://en.wikipedia.org/wiki/List_of_standard_amino_acids](http://en.wikipedia.org/wiki/List_of_standard_amino_acids).

There are two naming conventions for amino acids—single-letter abbreviations and three-letter abbreviations, as follows:

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>1-letter name</th>
<th>3-letter name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>A</td>
<td>Ala</td>
</tr>
<tr>
<td>Arginine</td>
<td>R</td>
<td>Arg</td>
</tr>
<tr>
<td>Asparagine</td>
<td>N</td>
<td>Asn</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>D</td>
<td>Asp</td>
</tr>
<tr>
<td>Cysteine</td>
<td>C</td>
<td>Cys</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>E</td>
<td>Glu</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Q</td>
<td>Gln</td>
</tr>
<tr>
<td>Glycine</td>
<td>G</td>
<td>Gly</td>
</tr>
<tr>
<td>Histadine</td>
<td>H</td>
<td>His</td>
</tr>
<tr>
<td>Amino Acid</td>
<td>3-letter Code</td>
<td>1-letter Code</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>I</td>
<td>Ile</td>
</tr>
<tr>
<td>Leucine</td>
<td>L</td>
<td>Leu</td>
</tr>
<tr>
<td>Lysine</td>
<td>K</td>
<td>Lys</td>
</tr>
<tr>
<td>Methionine</td>
<td>M</td>
<td>Met</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>F</td>
<td>Phe</td>
</tr>
<tr>
<td>Proline</td>
<td>P</td>
<td>Pro</td>
</tr>
<tr>
<td>Pyrrolysine</td>
<td>O</td>
<td>Pyl</td>
</tr>
<tr>
<td>Selenocysteine</td>
<td>U</td>
<td>Sec</td>
</tr>
<tr>
<td>Serine</td>
<td>S</td>
<td>Ser</td>
</tr>
<tr>
<td>Threonine</td>
<td>T</td>
<td>Thr</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>W</td>
<td>Trp</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y</td>
<td>Tyr</td>
</tr>
<tr>
<td>Valine</td>
<td>V</td>
<td>Val</td>
</tr>
</tbody>
</table>

**Peptides** are formed by linking various amino acids together. The link between one amino acid residue and the next is an amide bond, also called a peptide bond. Polypeptides are a single linear chain of amino acids. Figure 1 shows the structure of oxytocin, a neuropeptide with nine amino acids.

![Figure 1: Oxytocin](image)

A peptide of nine amino acids linked in the sequence cysteine-tyrosine-isoleucine-glutamine-asparagine-cysteine-proline-leucine-glycine. The cysteine residues form a
Proteins are basically very large peptides. A general convention is that proteins are molecules with more than 50 amino acids. Although the amino acid sequence is still read out by the DNA genetic code as a linear chain, it is folded and twisted as a consequence of salt bridges, hydrogen bonds, and disulfide bonds. Proteins have complex three-dimensional structures that affect their function. Many proteins are enzymes that catalyze biochemical reactions. Others have structural or mechanical functions, such as the proteins in the cytoskeleton that form a scaffolding system to maintain cell shape. Proteins are also important in cell signaling, immune responses, cell adhesion, and the cell cycle. Visual and olfactory sensory cells possess G protein coupled receptor proteins responsible for the critical first step in sensory transduction. Finally, binding proteins play an essential role in chemical communication, both at the production stage (transporting pheromone components from internal organs to external release sites) and at the reception stage (transporting incoming pheromone molecules through the mucus to the receptor).

Figure 2: Amino acid sequence and proposed structural model for a rat olfactory G protein coupled receptor protein anchored in a ciliary cell membrane. The 7 trans-membrane helices have been stretched out. Circles with letters indicate the different amino acids. The gray circles highlight the domains that are highly variable among different receptor cell types that bind different odorants. (From McClintock 2003.)

Lipids: fatty acids and glycerides

Lipids are carboxylic acids or esters with a long hydrocarbon tail. This tail makes them generally non-polar and thus insoluble in water (hydrophobic). Fatty acids have a single carboxylic acid at the end, and tail chain lengths between 12–24 carbons with the general formula CH₃(CH₂)ₙCOOH. Very small fatty acids are volatile and can serve as pheromones. Glycerides are esters formed from glycerol and fatty acids. They have three hydroxyl functional groups, which can be esterified with one, two, or three fatty acids to form monoglycerides, diglycerides, and triglycerides. The hydrocarbon tails may be saturated or unsaturated. Lipids are used for energy storage and some are important hormones; they form the lipid bilayer of cell membranes. The tails of...
fatty acids may be saturated or unsaturated. A few examples are illustrated below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Butyric acid</td>
<td><img src="image1.png" alt="Butyric acid" /></td>
</tr>
<tr>
<td>Caprylic acid</td>
<td><img src="image2.png" alt="Caprylic acid" /></td>
</tr>
<tr>
<td>Oleic acid</td>
<td><img src="image3.png" alt="Oleic acid" /></td>
</tr>
<tr>
<td>α-Linolenic acid</td>
<td><img src="image4.png" alt="α-Linolenic acid" /></td>
</tr>
<tr>
<td>Triglyceride</td>
<td><img src="image5.png" alt="Triglyceride" /></td>
</tr>
</tbody>
</table>


**Carbohydrates**

Carbohydrates are the most common type of biomolecule. They are based on a ring unit containing 4 or 5 carbons and one oxygen. Carbohydrate molecules vary enormously in size, from one to several thousand of these units. **Monosaccharides** are the smallest. They contain carbon, hydrogen, and oxygen in a ratio of 1:2:1, with the general formula $C_nH_{2n}O_n$ (where $n = \text{at least 3}$). **Disaccharides** contain two units. These small carbohydrate molecules are commonly called sugars. They have a sweet taste and provide readily accessible fuel for cellular metabolism.

**Oligosaccharides** contain 3–10 ring units. Smaller molecules consisting of short chains of fructose units are found in fruit and can only be partially digested by humans, hence their use as sugar substitutes. Oligosaccharides are often combined with proteins to form glycoproteins. **Mucins** are the **glycoproteins** secreted in the mucus of the respiratory and digestive tracts. The sugars attached to mucins give them considerable water-holding capacity and also make them resistant to proteolysis by digestive enzymes. Some hormones are glycoproteins, such as follicle stimulating hormone, luteinizing hormone, thyroid stimulating hormone, and human chorionic gonadotropin.

**Polysaccharides** are increasingly larger molecules often referred to as complex carbohydrates. They contain less carbon relative to oxygen and hydrogen, depending on how they are connected. They have no taste. Cellulose is a polymer built from glucose units. Plants use it as the main structural component of their cell walls. Animals can neither manufacture it nor digest it (unless they obtain the aid of microbes). Glycogen is an animal polysaccharide used for energy storage.
### Carbohydrate class

<table>
<thead>
<tr>
<th>Carbohydrate class (# ring units)</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monosaccharide (1)</td>
<td>Glucose</td>
</tr>
<tr>
<td>Disaccharide (2)</td>
<td>Sucrose</td>
</tr>
<tr>
<td>Oligosaccharide (3–10)</td>
<td>Chitosan oligosaccharide from crab carapace</td>
</tr>
<tr>
<td>Polysaccharide (&gt;10)</td>
<td>Cellulose polymer based on glucose units</td>
</tr>
</tbody>
</table>

Terpenoids

Terpenes are a large and varied class of hydrocarbons based on the 5-carbon isoprene unit. They were once believed to be produced primarily by plants, hence the name *terpene* (from the word *turpentine*, a product derived from plant resins that contain these compounds). However, as the biosynthetic pathways became better understood, it was discovered that animals can synthesize some plantlike terpenes. Moreover, isoprenoid building blocks give rise to a wide range of metabolically important compounds. Isoprene units can be connected head-to-tail, tail-to-tail, or assembled into rings, and they can be oxygenated with different functional groups. As chains of isoprene units are built up, the resulting terpenes are classified sequentially by size as hemiterpenes, monoterpenes, sesquiterpenes, diterpenes, sesterterpenes, triterpenes, etc. The all-inclusive class of terpenoids thus includes the longer terpene hydrocarbons along with isoprene-containing compounds modified by oxidation, methylation, and cyclization. The table below shows examples of terpenoids from each of the size classes.

Steroids

A steroid is a terpenoid lipid containing a carbon skeleton with four fused rings. Different steroids vary in the functional groups attached to these rings. Hundreds of distinct steroids have been identified in plants, animals, and fungi. All steroids are derived either from the sterol lanosterol (in animals and fungi) or the sterol cycloartenol (in plants). Both sterols are derived from the cyclization of the triterpene squalene. The figure below shows squalene laid out in way that reveals its similarity to the steroid skeleton.
The vertebrate sex hormones are all derived from cholesterol.

Pathways for pheromone biosynthesis

A few animal species ingest and modify plant compounds to produce their pheromones. In the main text we described the use of secondary plant compounds such as pyrrolizidine alkaloids to synthesize the male sex attractant in Utetheisa moths. Other moths in the Geometridae, Arctiidae, and Noctuidae families use plant-specific linoleic or linolenic acids to produce their pheromones, which characteristically have desaturated carbon chains with two or three double bonds, respectively. However, the majority of animals produce their pheromones de novo using biosynthetic pathways of normal metabolism. These products are then altered to yield species-specific pheromone components. There are three main sources of pheromones: isoprenoid hydrocarbons, fatty acids, and amino acids. We briefly review the biosynthetic pathways for the first two sources, but for further details the reader should consult Tillman (1999) and Jurenka (2004).

Isoprene pathways

Isoprene itself does not undergo the building process. Instead, activated forms, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), are the key building blocks (Figure 3). IPP is formed from acetyl-CoA via the MEVALONATE pathway, which involves production of the intermediate compound mevalonic acid.
using HMG-CoA reductase enzymes. An alternative, totally unrelated biosynthesis pathway of IPP, called MEP/DOXP pathway, has been discovered recently in some some eubacteria, green algae, and the plastids of plants. This route is initiated from pyruvate and glyceraldehydes (C$_5$ sugars), with 2-methyl-D-erythritol-4-phosphate (MEP) and deoxyxylulose phosphate as intermediates. In both pathways, IPP is isomerized to DMAPP by the enzyme isopentenyl pyrophosphate isomerase.

Synthesis of all higher terpenoids proceeds via formation of geranyl pyrophosphate (GPP), farnesyl pyrophosphate (FPP), and geranylgeranyl pyrophosphate (GGPP). Organisms differ in the possession of these enzymes and which pathways they employ. Although plants routinely produce hemi- and monoterpenes, the latter comprising the fragrant essential oils, some insects have evolved enzymes to synthesize similar compounds. Two key pheromones of bark beetles, ipsdienol and frontalin, are derived from geranyl diphosphate. The addition of an IPP unit to GPP leads to farnesene and other C$_{15}$ sesquiterpenoids in both plants and animals. Insects derive their main developmental hormone, juvenile hormone, from this compound (Dipterans are an exception, producing a steroid as in vertebrates). Most organisms, but not insects, combine two FPP molecules by their tails to produce a C$_{30}$ triterpenoid, often squalene. Squalene then undergoes cyclization to yield phytosterol in plants and cholesterol (and its many important derivatives) in animals. Only plants add an IPP unit to produce a diterpenoid, and then combine two of these molecules to yield C$_{40}$ tetraterpenoids such as carotene.
Figure 3: Isoprene pathway. Key building block components are shown in red, enzymes in blue, and major terpenoid classes comprising increasing numbers of isoprene units in green. Plants and animals can both produce the key double isoprene compound geranyl-PP, but do so via different pathways: animals through the mevalonate pathway on the left, and plants through the MEP/DOXP pathway on the right. Major signaling compounds produced via this pathway include juvenile hormone, the bark beetle pheromones ipsienol and frontalin, farnesene pheromones in rodents and many insects, steroids in animals and plants, and carotenoids in plants. (After Dewick 2001; Seybold and Tittiger 2003.)
Fatty acid pathways

The pheromones of many Lepidopterans are based on fatty acids and make use of the enzymes that produce normal fatty acids found in all organisms. The basic fatty acids—palmitic acid (16-carbon) and stearic acid (18-carbon)—are constructed from the base substrate acetyl coenzyme A (CoA). Malonyl-CoA then provides 2-carbon units to elongate the chain in steps. Fatty acids always have an even number of carbons. These two basic saturated fatty acids then undergo a series of modifications to generate the species-specific pheromone (Morse and Meighen 1986). The steps may occur in any order. The saturated fatty acid acquires a double bond at some point along the chain under the control of a desaturase enzyme. The chain is often reduced in length in 2-carbon steps under the control of Acyl-CoA oxidase. The compound then typically is converted to an acyl alcohol via a reductase step. The alcohol may then be converted to an aldehyde via an oxidase enzyme or to an acetate ester via an acetyl transferase enzyme. A large variety of mono-unsaturated compounds with different double bond positions, chain lengths, functional groups, and stereoisomers can be generated from this basic process (Figure 4).

Figure 4: Fatty acid pathway. (A) Chain elongation with acetyl-CoA and concatenated additions of two carbons from Malonyl CoA yields palmitic and stearic acid, both saturated fatty acids. (B) Palmitic and stearic acid undergo chain shortening, desaturation, and reduction to an alcohol, and the functional group may be modified to aldehyde or acetate ester. (C) Production of the components of the cabbage looper moth (Trichoplusia ni) pheromone blend. Z8–12:CoA is illustrated. (After Jurenka 2004.)

Literature cited


Web Topic 6.3
Marking Behaviors and Displays

Introduction

Overt scent marking behaviors are much more conspicuous in terrestrial than in aquatic animals. Aquatic animals do release pheromones, but they are more likely to exude these over much of their external body surface instead of from a dedicated organ (e.g., many non-arthropods), or pump them into excurrent respiration channels (e.g., crustaceans). As noted in the text, sea hares (Aplysia spp.) scent mark their fertilized egg masses while laying them, and barnacle larvae leave pheromonal “footprints” while seeking a site to settle. Many fish release pheromones in their urine, which being liquid like the surrounding water, is not visually conspicuous. As a result, most videographers have paid little attention to the pheromone dispersing actions of aquatic animals and we were unable to find many examples for this module.

In contrast, terrestrial species often have specialized organs and behaviors that they employ to disperse odorants into air. Many male moths have eversible brushes and hair-pencils that they can wave to increase pheromonal dispersal (Birch and Poppy 1990). Most wasps, ants, and bees have specialized abdominal, foot, or head glands that they use to mark substrates. Some lizards rub femoral glands inside each thigh on substrates to defend territories and attract females. Urine and feces are often used by mammals as vehicles for pheromonal compounds; they also have pheromonal glands on the head (antelopes), wrists (lemurs), chests (koalas), and near the anus or genitals (most carnivores). Nearly all of these species have specific behaviors that they use to deposit urine, feces, and/or gland secretions on specific sites as territorial markers or during courtship to females. These behaviors are quite conspicuous, and it was thus much easier to find good video examples of pheromone dispersal for terrestrial taxa.

(Note: Windows users may need Quicktime installed to see some examples; Mac users may need to install Flip4Mac [free version] to view some Windows-based formats).

Examples of pheromonal dispersal behaviors and displays

Below, we provide short descriptions of marking behaviors in a variety of taxa and links to corresponding videos. See Chapters 6, 12-13 for more details on the contexts in which these behaviors are performed.

Fish

- **Combtooth blenny** (*Ecscenius bicolor*): Nesting males of the blenny *Scartella cristata* have a large gland near their anal fin (Neat et al. 2003). It is not clear how the pheromone is applied, but the gland is enlarged only in nesting and hole-defending males; sneaker males have only a rudimentary gland. This video shows a nesting male of a related species repeatedly entering and exiting its nesting hole in a way that could apply pheromones to the substrate, in addition to fanning the eggs.
  (http://www.youtube.com/watch?v=MtQgb64Dyuc)

Insects

- **Indian luna moth** (*Actias selene*): In this species, the female releases pheromones from the tip of her abdomen that attract males. Whereas this female lacks the scent-pencils or brushes of males in other species, the source gland is under pressure and the video shows the forced extrusion of pheromone.
  (http://www.arkive.org/indian-moon-moth/actias-selene/video-09a.html)

- **Leaf-cutter ants** (*Atta cephalotes*): Colonies of this species cut fresh leaves and return them to their nests as nutrient for their cultivated fungus gardens. Ants follow pheromone traits to and from current leaf sources. Based on numbers and quality of leaves being brought back, the trails are reinforced with pheromones by the tiny workers seen here not carrying leaves.
Army ants (*Eciton burchellii*): Raiding army ants stream out in all directions from their current bivouac, and then kill, dismember, and return parts of encountered prey (largely other arthropods for neotropical species) to the colony. Streams follow pheromone marked trails with outgoing ants in lanes on each side of the stream, and returning ants in the center lane (see Figure 14.30, Chapter 14). Incoming ants do most of the trail marking. Having three lanes instead of two minimizes the tendency for outgoing ants to keep veering towards the incoming lane (where trail pheromone is strongest), and thus pushing the incoming lane further and further to one side.

Euglossine bees scrape fragrant oils from flowers and other plant material and store it in their enlarged hind tibia. Males establish small territories and display to females in a hovering dance while dispersing the scent from their legs. This website from the laboratory of Thomas Eltz shows how males pull the scent out of their tibial pockets and disperse it by fanning with their wings.

Reptiles

- **Caiman lizard** (*Dracaena guianensis*): male in captivity marking his territory by rubbing his femoral and genital areas on the substrate.

Mammals

**Marsupials**

- **Opossum** (*Didelphis virginiana*): adult marking vegetation with chin and throat.

- **Sugar glider** (*Petaurus breviceps*): urination and head gland marking by multiple individuals.

- **Koala** (*Phascolarctos cinereus*): individual marking tree with chest gland and good view of gland.

- **Eastern grey kangaroo** (*Macropus giganteus*): male scent marking on grass and dirt.

**Chiroptera (bats)**

- **White-lined bat** (*Saccopteryx bilineata*): Each afternoon, adult males refill the sac in each wing with urine and secretions from several glands. Later, males hover in front of females in the male’s territory, snap open the sacs to release odorant that is blown into the female’s face by the draft of the wings, and exchange vocalizations with the female. Males will also open the gland on one wing when roosting and shake it at either other males or females.

**Rodentia**

- **Capybara** (*Hydrochoerus hydrochaeris*): between the male fighting and copulation sequences, a male marks an overhanging branch with glands on nose.

- **Grizzled giant squirrel** (*Ratufa macroura*): Adult marking branch with anogenital drag.
**Primates**

- **Ring-tailed Lemur** (*Lemur catta*): Good examples of lemur rubbing its wrist glands on its tail and then waving it to disperse odorants.  

- **Golden tamarin** (*Leontopithecus rosalia*): Adult scent marking branch with ventral body.  

**Carnivora**

- **Giant otter** (*Pteronura brasiliensis*): Adult female mixing urine into mud to mark territory.  

- **Wolverine** (*Gulo gulo*): Marking with anogenital region on substrate.  

- **Fossa** (*Cryptoprocta ferox*): Scent marking a log.  

- **Civet** (*Civettictis civetta*): First part of long clip shows adult African civet marking various nonliving and living objects with anogenital glands.  

- **Bush dog** (*Speothos venaticus*): Group successively urine marking a tree.  

- **Cheetah** (*Acinonyx jubatus*): Urination marking of tree.  

- **Serval** (*Leptailurus serval*): Individual urination marking its territory.  

- **Snow Leopard** (*Panthera uncia*): Brief shot of animal urination marking.  

- **Brown Hyaena** (*Hyaena brunnea*): Scent marking bush with anal glands.  

- **Giant panda** (*Ailuropoda melanoleuca*): Anogenital marking of tree.  

- **Brown bear** (*Ursus arctos*): One of two courting adults urination marks.  

**Perissodactyla**

- **Black rhinoceros** (*Diceros bicornis*): Adult female defecation marking.  

- **Sumatran rhinoceros** (*Dicerorhinus sumatrensis*): Urination spraying of trees.  

**Artiodactyla**

- **Sulawesi babirusa pig** (*Babyrousa celebensis*): Male sniffs female genital area and female then urinates in front of male.  

- **Pygmy hippopotamus** (*Choeropsis liberiensis*): Adult dung marking territory. Note how it rotates stubby tail rapidly to spray feces in all directions.  
• **Vicuña** (*Vicugna vicugna*): individual marking at dung heaps. (http://www.arkive.org/vicuna/vicugna-vicugna/video-12a.html)

• **Okapi** (*Okapia johnstoni*): adult scent marking bushes. (http://www.arkive.org/okapi/okapia-johnstoni/video-12.html)

• **Thomson’s Gazelle** (*Eudorcas thomsonii*): male engaged in urination, defecation, and marking of plants with eye gland. (http://www.arkive.org/thomsons-gazelle/eudorcas-thomsonii/video-12a.html)

• **Pronghorn antelope** (*Antilocapra americana*): individual marking a bush with head glands. (http://www.youtube.com/watch?v=kAoBIWT0AkA)

• **Moose** (*Alces alces*): Bull moose performing flehmen assessment of female moose’s urine. (http://www.youtube.com/watch?v=4OXCV2ungHt)

**Hyracoidea**


**Proboscidea**

• **African elephant** (*Loxodonta africana*): adult male in full musth; notice wet inside of rear legs from dribbling urine and dark streaks from temporal glands on head. (http://www.youtube.com/watch?v=I03YOS28R4)

**Literature cited**


Web Topic 6.4  
Chemical Transmission Models

This unit provides the quantitative details of the chemical transmission models presented graphically in the text. It includes the classic work by Wilson and Bossert from the 1960s on diffusion principles and design rules for chemical signals, which we now know is limited in its applications to very small organisms. For most animals communicating over even moderate spatial scales, current flows overwhelm the action of diffusion. The unit also briefly describes the history of attempts to model the spread of chemicals under current flow conditions.

Reynolds number

The Reynolds number, named after Osborne Reynolds (1842–1912) who first proposed it, is a dimensionless number that estimates the tendency of a flowing medium to develop a turbulent versus laminar flow pattern, based on the medium’s viscosity and density. More specifically, it is the ratio of inertial forces to viscous forces in the flowing medium, so Reynolds number quantifies the relative importance of these two forces for given flow conditions. Quantitatively:

\[
Re = \frac{\text{inertial}}{\text{viscous}} = \frac{U \rho}{\mu / L} = \frac{UL}{\nu}
\]

where \( U \) = mean flow velocity  
\( L \) = characteristic length  
\( \mu \) = absolute dynamic fluid viscosity  
\( \rho \) = fluid density  
\( \nu \) = kinematic viscosity = \( \mu / \rho \)

Inertial forces can be envisioned as the tendency of an object to continue moving when given a push, or its momentum. In fluid mechanics terminology, this is expressed as \( U \rho \) (mean flow velocity times the medium’s density). Note that the term fluid can refer to either air or water. Viscosity is the stickiness of the medium, or how difficult it is for one molecule to slide past its neighbors, and is expressed as \( \mu / L \) (the force required to push an amount of medium over a specified distance in a specified period of time). The ratio of viscosity to density, \( \mu / \rho \), is an important variable in fluid dynamics called kinematic viscosity, \( \nu \). The Reynolds number is essentially the inverse of \( \nu \) times the mean or “bulk” flow velocity of the medium over some specified distance or area. Laminar flow occurs at low Reynolds numbers (\( Re < 2000 \)), in which viscous forces dominate, and is characterized by a smooth, even flow at all layers. Turbulent flow occurs at high Reynolds numbers (\( Re > 3000 \)), in which inertial forces dominate, producing eddies, vortices, and other random flow variations. Water is 70 times as viscous as air, but about 830 times as dense, so its kinematic viscosity is lower by a factor of 8–15. Thus water currents will develop turbulence at lower velocities and shorter distances compared to wind flow in air (Denny 1993). Figure 1 illustrates the boundary conditions between laminar and turbulent flow in air and water and the tradeoff between distance and speed at which this switch occurs.
**Diffusion**

Molecules move down their concentration gradients in a process called diffusion. The rate of diffusion depends upon: (1) the steepness of the concentration gradient, and (2) the ease with which a particular type of molecule moves in a particular medium. The slope of the concentration gradient can be described as the change in concentration of a molecule over a given distance. The **diffusion constant** is a measure of the ease of movement of a particular molecule type in a particular medium; it depends on the size of the molecule, how the molecule interacts with the medium, and how the medium molecules interact with each other. **Fick’s first law** provides a quantitative description of the rate of movement of molecules diffusing through a small window per unit time. Armed with Fick’s Law, we can predict the concentration of diffusing molecules at any time and at any distance away from the source of the molecules for a variety of emission strategies and ecological conditions. Quantitative expressions for these processes are described below. In the following discussion of chemical signal transmission strategies, we shall refer frequently to the following quantitative variables:

- \( C \) = concentration of odorant molecules per unit of volume (molecules/cm\(^3\))
- \( K \) = minimum concentration of odorant molecules required for receiver detection
- \( Q \) = total number of molecules released
- \( D \) = medium-specific diffusion constant for a given type of molecule (cm\(^2\)/sec)
- \( r \) = distance from the chemical release source for circular transmission (cm)
- \( x \) = distance from the chemical source for longitudinal transmission (cm)
- \( t \) = time from the onset of emission (sec)

**Law of Diffusion**

Fick’s first law describes the rate of movement of molecules down a concentration gradient. This rate depends on the
diffusion constant $D$ and the slope of the concentration gradient. If the concentration of the molecules at any point $x$ is $C(x)$, then the slope of the gradient is the derivative of $C(x)$ with respect to $x$, or

$$\frac{dC(x)}{dx}$$

The number of molecules diffusing through a small window per unit time, $J$, is therefore given by:

$$J = -D \frac{dC(x)}{dx}$$

where $J$ is in units of molecules/cm$^2$ × sec, $D$ is in units of cm$^2$/sec, $C$ is in units of molecules/cm$^3$, and $x$ is in cm. The sign of the concentration is negative since molecules are moving from a higher to a lower concentration region as $x$ increases. Adding the minus sign to the expression therefore makes $J$ a positive number of molecules moving down the gradient.

**Single Puff in Still Air**

If a single instantaneous puff of odorant is released quickly from a point well away from any boundaries, the odorant will diffuse outward in all directions. Using Fick's Law, it can be shown that the concentration $C(r,t)$ at any distance $r$ and time $t$ is:

$$C(r,t) = \frac{Q}{d} \frac{d}{(4\pi D t)^{d/2}} \exp\left[-\frac{r^2}{4Dt}\right]$$

where $Q$ is the number of molecules released and $D$ is the diffusion constant. The variable $d$ is a dimensionality constant: $d = 1$ for diffusion occurring in one dimension as in a pipe, $d = 2$ for diffusion occurring in two dimensions away from a point, and $d = 3$ for diffusion in three dimensions from a point source (the usual case for animal signals). If the animal is on the ground, the same number of molecules must diffuse into half as much space, so $C$ is then twice that predicted above for each $r$ and $t$ (Sutton 1953; Wilson and Bossert 1963; Dusenbery 1992). The active space is the region in which the concentration $C$ is equal to or larger than the detection threshold concentration $K$. The size (radius) of the active space, $r_A$, can be computed at any time $t$ as follows:

$$r(t) = \sqrt{4Dt \log\left(\frac{2Q}{K(4\pi D t)^{3/2}}\right)} \quad \text{for } 0 \leq t \leq \frac{1}{4\pi D} \left(\frac{2Q}{K}\right)^{2/3}$$

If we plot the profile of odorant concentration with distance from the source in a series of time snapshots, as shown in Figure 2, we would see a spherical (or hemispherical) cloud that spreads outward in time, and then shrinks back down as the odorant diffuses away.
Figure 2: Spread of a single olfactory puff. Each graph shows a snapshot of odorant concentration \(C\) versus distance from the source \(r\) at different times since emission \(t\). At the instant the puff is produced \((t_0)\), all of the odorant is highly concentrated at the point of release \((r_0)\). At each subsequent time interval, the odorant molecules diffuse out from the source and the local concentration (molecules per unit volume) is reduced. The active space, \(r_A\), is the enclosed region in which the concentration of the odorant is above the threshold detection concentration, \(K\). The active space first increases, then decreases to zero.

The maximum radius of the active space when the puff is released near the ground can be calculated as:

\[
r_{\text{max}} = \sqrt{\left(\frac{2Q}{K}\right)^{\frac{2}{3}} \times \frac{3}{2\pi\varepsilon}} = 0.527 \left(\frac{Q}{K}\right)^{\frac{1}{3}}
\]

The time required for the single puff to expand to \(r_{\text{max}}\) is:

\[
t_{r_{\text{max}}} = \frac{1}{4\pi D \varepsilon} \left(\frac{2Q}{K}\right)^{\frac{2}{3}} = 0.046 \left(\frac{Q}{K}\right)^{\frac{2}{3}}
\]

Finally, the time to fadeout of the signal is:

\[
t_{\text{fadeout}} = \frac{1}{4\pi D \varepsilon} \left(\frac{2Q}{K}\right)^{\frac{2}{3}} = 0.126 \left(\frac{Q}{K}\right)^{\frac{2}{3}}
\]
The relative change in $r_{\text{max}}$ and the time points are illustrated in Figure 3.

**Figure 3**: Change in the radius of the active space of a single puff over time. The active space increases and reaches maximum size at the time $t_{\text{rmax}}$ (dashed line), and then decreases to zero. The time to reach $r_{\text{max}}$ is always 0.37 times the time to fadeout, or $t_{\text{fadeout}} = 2.72 \ t_{\text{rmax}}$. (After Bossert and Wilson 1963.)

**Continuous Emission in Still Air**

If the sender continuously emits $Q$ molecules/sec from a surface position, the concentration is given by:

$$C(r,t) = \frac{Q}{2\pi D} \text{erfc}\left(\frac{r}{\sqrt{4Dt}}\right)$$

where $\text{erfc}(\chi)$, the error function complement, is the area under the normal curve out to infinity. Here, $r_{\text{max}}$ increases and levels off at a value of approximately (Figure 4):

$$r_{\text{max}} = \frac{Q}{2\pi KD}$$

where $D$ is measured in units of distance moved per second. The time to reach 95% of $r_{\text{max}}$ is:
Continuous Release From a Moving Source

The case of an animal releasing a trail of odorant while moving can be modeled as a linear series of single puffs (Figure 5). If \( Q \) is the number of molecules of pheromone released per second and \( u \) is the animal’s velocity, the total length of the active space, \( L \), will be

\[
L = \frac{0.160Q}{DK}
\]

The location of the maximum radius of the active space occurs 0.37 along the active space axis from the point near the animal:

\[
L_{r_{\text{max}}} = 0.37L
\]

The maximum diameter of the active space at this location is

\[
r_{\text{max}} = \sqrt{\frac{2Q}{eK\pi u}}
\]

where \( e \) is the base of the natural logs. The time it takes before the trail at any given location has dropped below \( K \) is

\[
t_{\text{fadeout}} = \frac{L}{u}
\]
Figure 5: Active space of an ant trail. A trail is essentially a sequential series of small single puffs along a linear transect. $L$ is the length of the active space, $r_{\text{max}}$ is the width of the active space, and $L_{r_{\text{max}}}$ is the distance to the maximum radius point from the ant. (After Bossert and Wilson 1963.)

**Advection in a current flow**

In the discussion below we will use the Cartesian coordinate system to reference the spatial dimensions of the flow. The source of chemical release occurs at the origin: downstream distance is specified by $x$, horizontal spreading perpendicular to the flow direction by $z$, and vertical distance by $y$.

**Laminar Flow**

When there is laminar current flow, the spread of a continuously emitted chemical can be modeled with the same logic as the trail above with a moving source, but in this case the source is stationary and the center of the active space moves with the flow. The concentration of the substance at any point in space is given by

$$C(r, \theta) = \frac{Q}{4\pi Dr} \exp \left[ \frac{-(1 - \cos \theta rU)}{2D} \right]$$

where $U$ is current velocity, $r$ is the straight-line distance from the source to the position of interest, $\theta$ is the angle between this line and the downstream direction, and $D$ is the diffusion constant. In this model, the maximum distance of odorant transmission is not actually increased in the downstream direction compared to a flow velocity of zero, but transmission is faster and the active space is narrower (Figure 6). This type of model would only be relevant for small animals living in the viscous boundary layer next to a surface and communicating over short distances—several centimeters in air and even less in water. Most animals must cope with turbulent flow conditions.
Figure 6: The active space for a constantly emitted source in a laminar flow field. Concentration intensity decreases as 1/r with distance from the source. (After Dusenbery 1992.)

**Turbulent Flow**

In a turbulent flow, there is a range of scales of the fluid motion. A single packet of fluid moving with a bulk velocity is called an eddy. The size of the largest eddies is set by the overall geometry of the flow, and the size of the smallest eddies is set by the Reynolds number. As the Reynolds number increases, smaller and smaller eddies occur. The Reynolds number is therefore an indicator of the range of eddy size scales in the flow. Under large Re conditions, inertial forces predominate over viscous forces, and the smallest scales of fluid motion are undamped—there is not enough viscosity to dissipate their motions. The kinetic energy must “cascade” from these large scales to progressively smaller scales until a level is reached for which the scale is small enough for viscosity to become important (i.e., viscous and inertial forces become approximately equal). It is at these small scales where the dissipation of energy by viscous action finally takes place. Therefore, although the energy dissipation is produced by a viscous mechanism, the rate at which it occurs is dictated only by large-scale characteristics of the flow, while viscosity only determines the size of the smallest eddies at which the energy will be dissipated (from [http://en.wikipedia.org/wiki/Reynolds_number](http://en.wikipedia.org/wiki/Reynolds_number)).
Figure 7: A dark-field photograph of smoke rising from a cigarette. This photo illustrates several important points. First, it shows the initial laminar flow over a short distance and the turbulent distribution of particles over larger distances. Second, it shows the wide size range of vortices and eddies. One can see large-scale meandering of the primary plume, as well as fine-scale eddies and filaments.

Sutton Model

To cope with the variable and random pattern of odorant density within turbulent plumes, modelers attempted to estimate a smoothed plume shape by averaging the patchiness over time. The first such time-average model was developed by Roberts 1923 and Sutton 1953. They used the same principle of laminar current flow described above but redefined the diffusion constant as the rate of molecular advection by the current flow. The model specifies the
average concentration of a continuously released odorant at points downwind from the source as:

\[ C(x, y, z) = \frac{2Q}{\pi D_y D_z U x^{(2-n)}} \exp\left[-x^{(n-2)}\left(\frac{y^2}{D_y^2} + \frac{z^2}{D_z^2}\right)\right] \]

where \( D_y \) and \( D_z \) are diffusivity constants in the \( y \) and \( z \) planes measured empirically at a wind speed \( U \), \( Q \) is release rate in molecules/sec, and \( n \) is a parameter ranging from 0 to 1 determined by wind speed, atmospheric conditions, and terrain slope (typically \( n = 0.25 \)). The 2 in the numerator again reflects the fact that the odorant is released from the ground. An application of this model to animal signalers (moths) is shown in Figure 8, where a threshold concentration \( K \) was incorporated to determine the active space, and the maximum downwind detection distance was estimated as:

\[ x_{\text{max}} = \left(\frac{2Q}{k\pi D_y D_z U}\right)^{\frac{1}{2-n}} \]

Figure 8: Active space for an olfactory signal released into wind of different velocities. Active space dimensions estimated from the Sutton are given. Transmission distance is better for low wind speeds compared to higher wind speeds where molecules are whisked away from the surface faster.
Gaussian Model

The Sutton model did not seem to capture the way odor plumes spread out in a wedge as distance from the source increases. The Gaussian time-average model was developed to remedy this short-coming by evaluating horizontal and vertical spreading as a function of distance and wind speed. The Gaussian model is largely the same as the Sutton model, except the diffusivity constants $D_y$ and $D_z$ are replaced by $\sigma_y$ and $\sigma_z$, standard deviations of the cloud dimensions in the horizontal and vertical directions, respectively:

$$C(x,y,z) = \frac{2Q}{2\pi\sigma_y\sigma_z U} \exp\left[-\left(\frac{y^2}{2\sigma_y^2} + \frac{z^2}{2\sigma_z^2}\right)\right]$$

and the standard deviation terms are functions of downwind distance:

$$\sigma_y = \frac{1}{2} D_y x^{(2-n)/2} \quad \text{and} \quad \sigma_z = 2 \frac{1}{2} D_z x^{(2-n)/2}$$

This model has also been modified to take into account the release of odorant from some specified height off the ground, and to include the absorption of molecules on the ground surface. The Gaussian model does yield a more broadly spreading plume (Figure 9), but is it still often not as wide as real measurements of active spaces in the field (Fares et al. 1980; Elkinton et al. 1984; Murlis et al. 1992).

Figure 9. A Gaussian plume. It has a wider spread at increasing distances from the source than a plume estimated via the Sutton model.

Meandering Plume Model

Another step toward reality involves incorporating the meandering path of the main plume. Such models use the same basic logic as the Gaussian model, but allow the center of the clouds to meander with distance from the source (Figure 10).
Figure 10. Increasing complexity of plume models. (A) The basic Gaussian plume model, in which the spread of the cloud increases with distance from the source, estimated with a Gaussian standard deviation. (B) A meandering plume model, where the clouds both increase in size and follow a meandering path as they move away from the source. (C) The filamentous structure of a real plume, showing the three-dimensional regions in which the plume may meander. (After Murlis et al. 1992.)

Computational Models

The most recent attempts to generate realistic models of turbulent plumes involve statistical models and simulations of the cascading eddy behavior of turbulent flows. Figure 11 shows an example of such a model. The model uses a direct numerical simulation (DNS) technique to compute momentum with continuity constraints, taking into account the conservation of energy and the incompressibility of the medium. The model is able to produce the plume meandering pattern observed in an unstably stratified open channel flow (Liu and Leung 2006a, 2006b).
There are several other approaches to modeling turbulent flows, including: the **Reynolds-averaged Navier-Stokes (RANS)** approach, which provides a time-averaged solution like the Gaussian model; **large eddy simulation (LES)** approach, which removes the smallest scales of the flow to focus on the major eddy forms; the **detached eddy simulation (DES)**, which focus on the small-scale eddies; **coherent vortex simulation**, which decomposes the flow into the coherent vortex motion component and the incoherent background flow using wavelet filtering; and the **Reynolds stress model (RSM)**, which attempts to solve transport equations for the Reynolds stresses. See recent books on this topic listed in Further Reading below.
Useful websites on turbulence

http://turb.seas.ucla.edu/~jkim/sciam/turbulence.html

http://www.env.leeds.ac.uk/~ralph/dispersion/index.html

Further reading


Literature cited


Web Topic 6.5
Vertebrate Dual Chemosensory System

Introduction

Most amphibians, reptiles, and mammals have a dual olfactory system: the main olfactory system (MOS) with input through the main olfactory epithelium (MOE) and neural connections to the main olfactory bulb (MOB), and the vomeronasal system (VNS) with input through the vomeronasal organ (VNO) and neural connections to a separate accessory olfactory bulb (AOB). Vertebrates have other chemical sensors as well, including taste bud receptors on the tongue, the septal organ of Masera, the Grueneberg ganglion, and the trigeminal nerve system (Table 1, Figure 1). It was once thought that the VNS’s function was restricted to reception of pheromones and the MOS’s function was mainly detection of food, predators, and other ambient odors, but this division is not at all so distinct. Some pheromones are detected only through the MOS, some species follow prey using VNS input, and there are many examples in which input from both systems is integrated. The VNS and MOS differ in three main ways: (1) the receptor proteins and their genes, (2) the specificity of these receptors to different stimuli, and (3) the connection patterns between sensory cells, glomeruli, olfactory bulbs, and brain processing centers. We will first describe these differences in more detail, and then review recent findings on the relative roles of the two systems. This overview draws from several recent reviews which should be consulted for more details (Dulac and Torello 2003; Wyatt 2014; Brennan and Zufall 2006; Müller-Schwarze 2006; Ma 2007; Touhara and Vosshall 2009; Su et al. 2009; Munger et al. 2009; Tirindelli et al. 2009; Mucignat-Caretta 2010; and Ferro and Liberles 2010).

Table 1. Chemoreception systems in vertebrates.

<table>
<thead>
<tr>
<th>Name</th>
<th>Location</th>
<th>Presumed function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main olfactory system (MOS)</td>
<td>Nasal cavity</td>
<td>Detection and discrimination of a large range of ambient odorants from food, predators, and conspecifics.</td>
</tr>
<tr>
<td>Vomeronasal system (VNS)</td>
<td>Between nasal cavity and mouth</td>
<td>Detection of water-borne (also urine and mucus) pheromones and proteinaceous prey/predator chemical cues.</td>
</tr>
<tr>
<td>Taste buds</td>
<td>Tongue and mouth</td>
<td>Discrimination of sweet, sour, bitter, salty, and savory (umami) chemicals in potential food items.</td>
</tr>
<tr>
<td>Trigeminal nerve (TN)</td>
<td>Innervates the face and eyes</td>
<td>Responsive to touch, pain, temperature, and irritating and noxious chemical stimuli</td>
</tr>
<tr>
<td>Nervus terminalis (NT)</td>
<td>From hypothalamus to nasal epithelium</td>
<td>Modulation of olfactory epithelium in response to hunger state</td>
</tr>
<tr>
<td>Septal organ of Masera (SOM)</td>
<td>Base of nasal cavity near entrance to nasopalatine ducts</td>
<td>Modulation or synchronization of responses of olfactory nerves with respect to respiratory airflow</td>
</tr>
<tr>
<td>Grueneberg ganglion (GG)</td>
<td>Dorsal to nares</td>
<td>Detection of alarm pheromone, functional in neonate pups.</td>
</tr>
<tr>
<td>Guanylyl cyclase-D cells (GC-D)</td>
<td>Dispersed ciliary cells in olfactory epithelium (mice)</td>
<td>Responsive to uroguanylin, guanylin, cues in urine, CO₂ connections to necklace ganglia.</td>
</tr>
</tbody>
</table>

Sources: Breer et al. (2006); Su et al. 2009; Munger et al. (2009); Fleischer et al. (2009).

Receptor proteins and their genes

The majority of receptor cells in the MOE have G-coupled receptor proteins and are therefore called G-protein-coupled receptors (GPCRs). The genes that encode these receptors belong to large multi-gene families. Most genes specific to the olfactory system are referred to as ORs (which stands for Olfactory Receptor gene).
Vertebrate species often have up to 1,000 different OR gene variants, originating from gene duplication events, and sensitive to different medium-borne chemicals as a consequence of different amino acid variation in the ligand-binding regions (Buck and Axel 1991). A second family of receptor genes present in the MOE, called trace-amine-associated receptors (TAARs), are sensitive to biogenic amines (Liberles and Buck 2006; Hashiguchi and Nishida 2007). Both classes of genes are found throughout the vertebrates, including fish. A third class of receptors, sparsely distributed throughout the MOE of most vertebrates but not in primates, comprises the guanylyl-cyclase-D (GC-D) system; these cells project axons to the large necklace glomeruli in the MOB.

![Figure 1: The mouse dual olfactory system](image)

There are two main classes of receptor genes in the VNO, referred to as V1R and V2R. Both are also multi-gene families from the GPCR superfamily. The sensory cells expressing these two types of genes are segregated into two layers of the vomeronasal organ, and they send their axon projections to two distinct regions within the AOB (Figure 1). V1R-expressing sensory cells are located in the apical half of the VNO (yellow area), and V2R-expressing cells occur in the basal half (red area) (Johnston 2000; Halpern and Martinez-Marcos 2003). A final type of receptor cell found in a restricted area of the basal VNO expresses a formyl peptide receptor (FPR), which mediates cell responses to disease as part of the general immune system and in the VNO may confer sensitivity to disease-related molecules (Liberles et al. 2009; Riviere et al. 2009). This cell type is absent in primates (Yang and Shi 2010).

Figure 2 illustrates some of the key differences among the receptor proteins. OR, TAAR, V1R, and V2R proteins are all 7-transmembrane G protein-coupled receptors embedded in the membranes of the cilia (OR) or the microvilli (VRs) of the bipolar sensory cells. The G protein for OR and TAAR receptor cells is distinguished as $G_{\alphaolf}$. These cells employ cyclic nucleotides (cAMP) as a second messenger to open cyclic nucleotide gated ion channels (CNG) and facilitate the signaling cascade for nerve impulse generation. In contrast, VRs (and a few ORs) employ a unique gated cation channel to facilitate or modulate transduction, called a transient receptor potential (TRP) channel. The V1R protein superficially appears very similar to the OR protein, although they do not share any sequence motifs. Its G protein is called $G_{\alpha i2}$. The V2R receptor molecule is quite distinct, with a very large hydrophobic N-terminal extracellular domain (Figure 2) and a different G protein, $G_{\alpha o}$. Unlike the single-expressed-gene-per-cell rule in all other vertebrate chemical sensors, each basal VNO sensory cell usually expresses two different V2R genes. In rodents, some V2R receptor cells co-occur with a non-classical major histocompatibility complex molecule, either M10 or M1, and an associated molecule of $\beta 2$-microglobulin ($\beta 2m$), shown as a purple strand in the figure (Ishi et al. 2008; Leinders-Zufall et al. 2009). Classic MHC molecules are also expressed in association with $\beta 2m$ at the surface of most cells in the body, where they signal the presence of foreign invaders. The unique structures of V1R and V2R proteins and their associated molecules are believed to determine receptor specificity to certain types of ligands, as discussed below.
Figure 2: Five types of vertebrate olfactory receptor proteins. Illustrations show the basic shape of each type of receptor protein and some of the coupled transduction components: the ciliary or microvillar cell membrane in which the receptor is imbedded is shown in gray; the cell exterior is above, the cell interior below. All but GC-D receptors are 7-transmembrane G protein-coupled receptors, shown schematically in their bundled shape. All have a nitrogen (N) terminal on the cell exterior and a carboxylic acid (C) terminal extending into the cell interior. An FPR is not shown here, but is very similar to the V1R receptor in basic shape. (After Dulac and Torello 2003; Spehr and Munger 2009.)

The identification and sequencing of these olfactory receptor genes presents the opportunity to analyze the evolutionary relationships among chemical receptor systems. TAAR receptors belong to the same multigene family as the serotonin and dopamine receptors in the brain. The two VR gene families are completely unrelated to each other, and neither one shares any sequence homology to the OR family of genes except the regions specifying the helical structure (Ryba and Tirindelli 1997). The V1R genes share some sequence motifs with the T2R bitter taste receptors and with opsin genes (Adler et al. 2000). The V2R genes are closely related to glutamate and GABA_B neurotransmitter receptors in the brain, as well as to the T1R sweet and umami taste receptors on the tongue, which all possess the large N-terminal (Alioto and Ngai 2006; Bjarndottir et al. 2005). Thus the four main vertebrate chemoreceptor gene classes arose and diverged independently, on different time scales (Grus and Zhang 2006, 2008, 2009).

All of these gene families are present in the olfactory epithelium of fish (Hashiguchi and Nishida 2006; Grus and Zhang 2006; Liberles and Buck 2006; Hino et al. 2009; Saraiva and Korshing 2011). Even though fish do not possess a vomeronasal organ, different cell types with their associated receptor gene types form layers within the olfactory
The ciliated receptor cells expressing OR-type receptor genes have cell bodies located in the basal or middle layers of the epithelium, whereas microvillous and crypt cells containing V1R or V2R receptor genes are located in middle or surface layers of the epithelium (Hansen et al. 2004; Pfister and Rodriguez 2005; Hamdani and Døving 2006). The move onto land by the amphibians is associated with a morphological subdivision between the microvillous VR receptor cells and ciliary OR cells into separate but interconnected chambers of the nasal organ. Access of odorants occurs through grooves on the sides of the external nares, even when the main air-breathing chamber is closed while underwater (Døving et al. 1993; Døving and Trotier 1998; Petti et al. 1999). Complete separation into two chambers arose with the fully terrestrial reptiles and mammals, and in most cases access of odorants shifted to the roof of the mouth (rodents re-evolved odorant access through the base of the nasal chamber). Birds, as well as marine mammals, Old World monkeys, and some bats and reptiles, have completely lost the vomeronasal organ and its associated genetic components (Figure 3).

During these shifts between terrestrial and aquatic habitats, different sets of olfactory receptor genes either expanded or decreased. With the complete genomic sequencing of about a dozen vertebrate species, recent analyses of the
receptor genes have revealed a fascinating history of olfactory system evolution. What is especially interesting is that the “lost” genes are still mostly present in the genome as pseudogenes, whose functionality is gone as a result of missing or disrupted sequences (pseudogenization). Figure 3 also shows the number of intact versus non-functional genes for OR, TAAR, V1R, and V2R receptor proteins in different vertebrate species. The rodents, opossum, platypus, and frog have substantial gene repertoires for all four receptor types. Many terrestrial mammals (dog and cow) have lost the V2R genes, humans and birds have essentially no VR genes and a reduced OR repertoire, and fish possess mainly OR and V2R genes (Niimura and Nei 2006; Shi and Zhang 2007; Liman 2006).

These shifts in the relative importance of different receptor types make sense in light of the different types chemicals each one is designed to detect. Finding the active ligands for specific receptor types is painstaking work, but numerous matches have now been made. OR receptors bind to a wide variety of small volatile molecules, mainly food and environmental chemicals, but also to a few species-specific pheromones. TAAR receptors respond selectively to biogenic amines, such as those found in urine that vary as a function of sex, status, and stress. V1R receptors respond narrowly to small volatile water-soluble pheromones in urine such as sulfated steroids, and in rodents to specific known male urinary pheromones. V2R receptors bind to nonvolatile peptide and protein pheromone molecules in urine and tears, and must be received by direct contact (Zhao et al. 1998; Boschat et al. 2002; Leinders-Zufall et al. 2004; Mombaerts 2004; Kimoto et al. 2005; Su et al. 2009; Tirindelli et al. 2009; Munger et al. 2009; Mucignat-Caretta 2010).

Figure 4 shows the results of an analysis of the relative gene-class repertoire sizes for some of the species shown in Figure 3. The ratio of the number of intact V1R genes to V2R genes for a given species has been plotted as a function of aquatic, transitional, and terrestrial environment (red bars). There is a preponderance of V2R genes in the two aquatic groups (fish and frogs) and a reduction or complete loss of V2R genes in favor of V1R genes in the terrestrial vertebrates (Shi and Zhang 2007).

![Figure 4: Vertebrate evolutionary patterns during aquatic to terrestrial transition.](image)

A similar analysis has been made for OR genes. Olfactory genes cluster into many related subfamilies, but a study of
OR genes in the African clawed frog (Zenopus laevis) provided a critical insight (Freitag et al. 1995). Its genes clustered into two distinct groups, one closely related to fish OR genes (Class I) and the other more closely related to mammalian OR genes (Class II). This species lives in both aquatic and terrestrial environments as an adult, and its nasal chamber is divided into two pockets, one for breathing in air and the other for water (the VNO is a third pocket close to the water-breathing pocket). A toggle-like valve opens one or the other to ambient media. The air-pocket epithelium contains only cells with Class II (mammal-like) receptor genes, while the water-pocket epithelium contains only cells with Class I (fish-like) receptor genes. Figure 4 also presents the ratio of OR Class II / Class I genes for the different vertebrate species (black bars). Not surprisingly, Class I genes predominate in the fish and Class II genes in the terrestrial species, but the frog, which has both, in this case shows the terrestrial pattern (Shi and Zhang 2007). One difference in the structure of Class I and II receptors is a longer extracellular loop between several of the transmembrane helices in the Class I proteins. This structure may enable these receptors to bind with water-soluble ligands, while Class II proteins bind with hydrophobic volatile ligands (Freitag et al. 1995). A possible scenario for the VR and OR patterns is that fish use OR Class I receptors for detection of small water-soluble odorants such as food-derived amines and conspecific alarm pheromones, and V2R receptors for detection of large water-soluble molecules such as conjugated steroids, peptides, and proteins. Amphibians continued to use V2R receptors for large water-soluble molecules, but evolved new OR (Class II) receptors to detect volatile airborne odors. Terrestrial vertebrates subsequently expanded their V1R repertoire to detect small water-soluble odorants in urine.

Connection patterns

The wiring patterns of vomeronasal sensory neurons also differ significantly from that seen in the olfactory sensory neurons, further supporting important functional differences between the two systems. Differences are apparent at the first level of connection to glomeruli, the spherical neuropils in the corresponding olfactory bulbs (Figure 5). Olfactory neurons in the MOE that express a given OR receptor all project to the same one or two glomeruli in the MOB. Several thousand sensory cells converge on a given MOB glomerulus. The spatial pattern of glomeruli in the olfactory bulb is highly conserved between individuals, and adjacent glomeruli often respond to similar kinds of chemical stimuli. Each glomerulus synapses with a unique mitral cell, which projects its dendrite to the brain. Lateral inhibitory connections between mitral cells by interneurons may sharpen the tuning of stimulated glomeruli, but there is little integration of signals from different receptor types at the level of the MOB. In contrast, sensory neurons in the VNO project to multiple (10–30) glomeruli in the AOB, and each glomerulus receives sensory cell input from a few hundred neurons that express several different receptor types. However, the two zones within the VNO remain separate and send axons only to the corresponding zone within the accessory bulb, and for apical zone cells at least, neurons expressing closely related subfamilies of V1R genes converge their axons on spatially clustered glomeruli. VNO glomeruli in general are smaller and more variable in size, and their positions are variable within and among individuals, compared to MOB glomeruli. At the next level, AOB mitral cells send dendrites to multiple glomeruli as well. Clearly far more integration between sensory cell types is occurring at the primary input stages of the vomeronasal system compared to the main olfactory system (Dulac and Torello 2003; Mombaerts 2004; Dulac and Wagner 2006).
Brain circuits also differ for the main olfactory and vomeronasal systems (Figure 6). Input from the MOE and MOB projects mainly into the higher cortical areas of the brain and the lateral amygdala. Input from the VNO and AOB projects primarily to the medial and anterior amygdala, which along with the hypothalamus comprises the limbic system. The limbic system controls reproductive, aggressive, and parental care behaviors. The hypothalamus is a critical control center, integrating internal and environmental cues, ensuring organismal homeostasis, and orchestrating long-lasting endocrine changes as well as short-term behavioral effects elicited by chemical signals (Yoon et al. 2005; Martinez-Marcos 2009; Mucignat-Caretta 2010).
Figure 6: Brain pathways for the vomeronasal and main olfactory systems. Major brain regions shown include amygdala (orange), cortex (green), and hypothalamus (darker violet). The vomeronasal system (VNO, AOB, and red pathways) projects to several areas of the amygdala, as well as to bed nuclei of the accessory olfactory tract and stria terminalis. The main olfactory system (MOE, MOB, and blue pathways) projects primarily to several cortical areas, including the piriform cortex, olfactory tubercle, entorhinal cortex, anterior olfactory nucleus, and tenia tecta (not shown); these nuclei then project to higher levels in the brain, as well as back to the main olfactory bulb (not shown). The olfactory system also connects to several nuclei in the amygdala. Areas in light purple indicate nuclei that receive input from both the VNO and MOE. Importantly, both pathways connect to the hypothalamus (medial preoptic area and ventromedial hypothalamus). (After Dulac and Wagner 2006; Tirindelli et al. 2009.)

The main olfactory system and VNO system are clearly designed for different tasks. Olfactory receptors show broader selectivity over a large range of concentrations and there are a large number of different sensory receptor
types in the MOE, so they can collectively perceive a huge range of odorants. The mitral cells are uniglomerular and receive inhibitory connections on lateral dendrites, so they can compare and analyze the pattern of inputs from many glomeruli and process the information in higher cognitive brain centers. The main olfactory system thus seems designed to detect and analyze a wide range of environmental chemical compounds. Nevertheless, some single-chemical pheromones shown to trigger specific behavioral responses operate through the MOE system, so the main olfactory system, with its uniglomerular input pattern, can operate in a labeled line fashion (see discussion of labeled line versus across-fiber pattern strategies for neural coding in main text Chapter 6). Subsets of the MOE may possess specialized circuits for detecting critical odorants and pheromones, such as the TAAR receptors that detect stress products in urine, the GC-D receptors that detect several peptides and CO2, and the Gruenberg ganglion neurons that may respond to an alarm pheromone (Stowers and Logan 2010). The vomeronasal system possesses fewer, more narrowly tuned, and very sensitive sensory cells. The mitral cells are multiglomerular and interconnected in ways that lead to integrative processing at early stages. The VNO system projects to lower levels in the brain that regulate genetically programmed behaviors and autonomic responses. This system seems to be designed in part to analyze and compare blends of structurally-related compounds involved in individual, sex, status, and species recognition, and aspects of reproductive behavior (Dulac and Wagner 2006; Trindelli et al. 2009).

Relative roles of vomeronasal and main olfactory systems

As previously mentioned, the classical view of the VNS as a pheromone detection system and the MOS as a general odorant detection system is now regarded as far too simplistic. In the prior sections, we outlined a few new views on the functions of these two systems, one a phylogenetic view arguing that both systems had components designed to operate in terrestrial versus aquatic environments, and a neural structure view arguing that the MOS is designed to detect and analyze volatile environmental odors and single-component pheromones while the VNS is designed to compare chemical blends. In this section we examine the evidence for different behavioral effects mediated by the two systems. Over the past few decades, numerous studies have evaluated the consequences of “removing” one or the other system. A system can be blocked by plugging or stitching closed the input ducts, by severing the nerves, or more recently, by genetically “knocking out” a specific second messenger transduction component, such as the TRP or a CNG gated ion channels. The emerging picture is that some effects are strongly dependent on the VNS, others on the MOS, and quite a few either require both systems or can be compensated for by the other one. See Table 2 in Tirindelli et al. (2009) for a summary of VNS- and MOS-removal studies.

Main role by VMS

Modulation of estrous in rats and mice is strongly affected by input of urine signals from males and other females through the vomeronasal organ (Ma et al. 1999; Keverne 1999; Musignat-Caretta 2010). These priming effects include acceleration of puberty in young females, pregnancy block produced by the odor of a strange (non-sire) male, and suppression of estrous by the odors of group-living females. Table 2 below summarizes the probable pheromones involved in mediating these effects in mice. Most of the volatile pheromones are MUP ligands. In prairie voles and *Mondelphis* opossums, the VNS is required to induce estrous. Some aspects of maternal behavior in rats, mice, and voles are also impaired by VNO removal, including maternal aggression, recognition of offspring, retrieval of pups, licking, and lactation.

Normal reproductive behaviors in a large number of male vertebrates, such as sexual arousal and copulatory behavior, are dependent on VNO input, although VNO-knockout males will still mate and prior mating experience before VNO removal can sometimes compensate. Male ungulates, carnivores, rodents, and snakes detect estrous females via chemical signals in their urine or skin secretions. In rodents, especially mice, VNO removal causes reduced aggression towards other males and copulation attempts towards both males and females. The reason for these effects seems to be that males (as well as females) require VNO input to distinguish individuals, assess status, and identify maleness. Many of the pheromones responsible for these effects are the same as those that affect female estrous.

Table 2. The structure and function of mouse urinary pheromones

<table>
<thead>
<tr>
<th>Name</th>
<th>Chemical structure</th>
<th>Origin</th>
<th>Possible signal function in female mice</th>
<th>Possible signal function in male mice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compound</td>
<td>Type of Urine</td>
<td>Effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>------------------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,5-Demethylpyrazine</td>
<td>Female urine</td>
<td>Suppression of estrous cycle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Sec-butyl-4,5-dihydrothiazole</td>
<td>Male bladder urine</td>
<td>Estrus synchrony, puberty acceleration, male aggression, female attraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,3-Dehydroexo-brevicomin</td>
<td>Male bladder urine</td>
<td>Estrus synchrony, puberty acceleration, male aggression, female attraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>α- and β-Farnesenes</td>
<td>Male preputial gland</td>
<td>Puberty acceleration, male territorial status, female attraction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Heptanone</td>
<td>Female or male urine</td>
<td>Estrus extension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-Hydroxy-6-methyl-3-heptanone</td>
<td>Male bladder urine</td>
<td>Puberty acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n-Pentyl acetate</td>
<td>Female or male urine</td>
<td>Suppression of estrous cycle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isobutylamine</td>
<td>Male urine</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Snakes use VNO input to mediate conspecific interactions such as copulation and aggressive behavior, as well as for trailing of conspecifics, aggregation, and shelter selection. In addition, they also employ this modality for prey discrimination and location. Strike behavior is reduced in rattlesnakes without a functioning VNO, and they fail to ingest envenomated prey. They also fail to recognize and take defensive action against predatory king snakes. Lizards and salamanders probably also use the VNO for prey detection. These non-conspecific chemical cues are often proteinaceous compounds inadvertently deposited as footprints by the prey (Halpern and Kubie 1984).

**Main Role by MOS**

The primary function of the main olfactory system certainly is detection of general odorants, but there are several very clear examples of pheromones detected only through this channel and not the VNO. One example is the rabbit mammary pheromone that guides pups to the nipple. The young of the European rabbit (*Oryctolagus cuniculus*) are fed only once per day for a short period of time, and when the female appears the pups must respond quickly and find a nipple. A pheromone emitted from the female’s nipple area and detected only by the MOS guides and stimulates the pups to search for the nipple (Schaal et al. 2003; Luo 2004; Moncomble et al. 2005). The second example is the androsterone compounds release by the male boar in the presence of a female, which attracts and induces her to stand while the male mounts (Dorries et al. 1997). Similarly, in sheep the ewes are attracted to the smell of ram wool, and detection of this odor via the MOS elicits a surge of luteinizing hormone. In mice, a highly volatile component of male urine, (methylthio)methanethiol (MTMT), is attractive to females and appears to be detected by the main olfactory system, in contrast to most of the other components of male urine (Lin et al. 2005). A final example is the suppression of estrous and reproduction in subordinate marmosets (*Callithrix jacchus*), which is clearly not mediated by the VNS. Olfactory, visual, and behavioral cues and signals from the dominant female appear to mediate this effect (Barrett et al. 1993).

**Interaction between VNS and MOS**

Both the vomeronasal and main olfactory systems may operate together to elicit some behavioral responses and primer effects. Several examples have been described in the golden hamster (*Mesocricetus auratus*) (Johnston 1998). Hamsters live solitarily, probably with the usual small mammal pattern of exclusive female territories and larger overlapping male home ranges. Females mark their territories with fluid vaginal secretions that are detected from a distance by the male via the main olfactory system. The volatile pheromone appears to be methyl disulfide. Other secretions from the flank gland are critical for individual recognition in both sexes and are also detected via the main olfactory system. Once a male has approached a female, a large protein component of the vaginal mark appears to be critical for a surge of androgen that induces further male sexual behavior, including copulation. Removal of the vomeronasal organ eliminated the androgen surge in males, but sexually experienced males were usually still able to copulate. Dual lesions eliminated male mating behavior, while lesions of either system alone had little influence on mating. Similarly, both male and female will call to each other after an initial encounter. The calling
is stimulated by odor input to both the VNO and MOB, and is reduced if either system is removed.

In addition to the primary role of VNO input for normal reproductive behavior in mice, blocking or knocking out the main olfactory system greatly reduces reproductive performance in both sexes. MOE-knockout males fail to investigate females’ anogenital region, mount, copulate, and behave aggressively towards other males (Mandiyan et al. 2005). Neurophysiological responses in the MOE to two known pheromones, 2-heptanone and farnesene, are absent in knockout mice but present in wild type mice (Wang et al. 2006). Similarly, MOE-knockout females fail to retrieve pups, construct normal nests, or show maternal aggression, and lack EOG responsiveness to known urinary pheromones (Wang et al. 2011). This evidence indicates that both MOE and VNO input are required for normal reproductive behavior.

The mouse olfactory recognition system is another example of integration of both volatile and non-volatile chemical components (Hurst et al. 2001; Hurst and Beynon 2004). Male mice place numerous urine marks around their territory. Mouse urine contains a significant amount of protein, most of it lipocalin-binding proteins from the multi-gene MUP family. As mentioned above, MUPs bind most, if not all, of the pheromones listed in Table 2, as well as a variety of other compounds in urine. The ligands are gradually released over many hours and become airborne volatiles, while the proteins remain stable for weeks. The proteins provide fixed genomic information about species, sex, and individual identity and can only be detected by contact through the vomeronasal system. The volatiles provide variable metabolic information concerning owner’s recent social, reproductive, and health status and his food resources and are detected by the main olfactory system. Figure 7 shows a model by which males might integrate information from the volatile and nonvolatile sources to learn about their male neighbors and competitors. The volatile components are more readily detected at a distance and would not require close investigation if the information in the volatile mix was familiar. If unfamiliar, the receiver could approach and contact the mark to obtain the associated stable information. In this way an owner could update his knowledge about the health and status of known or new intruders. Females also respond to these male marks, but the estrous-modulating effects (Table 2) only operate when they contact fresh urine marks in which the ligands are still bound to the MUPs. These ligand-MUP complexes are detected by the vomeronasal system: the protein component is apparently detected by the V2R/basal layer and anterior AOB part, and the ligand component by the V1R/apical and posterior AOB part. Variation in the suite of MUP proteins themselves provide information on individual identity and male heterozygosity. There are also MHC proteins in the urine that may add further useful information for both male and female receivers, especially with regard to genetic relatedness and strain differences. Either the proteins themselves, or their peptide products, or MHC-based metabolites bound to MUPs, may affect mating decisions in females. But the MHC associated odors are neither required nor sufficient for producing scent ownership recognition.
Figure 7: Associative matching between the volatile and involatile components of a urine mark.

(A) When a mouse encounters the volatile signature of an unfamiliar mouse (orange), it moves up the concentration gradient to the source to make contact with the scent mark. An association is built between the volatile (more variable) and involatile (genomically hard-wired) information. (B) If the same pattern of volatiles is encountered, the receiver knows the identity of the scent owner and no further investigation is needed. (C) If the scent donor’s volatile signature includes new components (green) because of a change in diet, social status or infection status) then the receiver must update its association for this donor. The receiver can thus retain an image of other individuals in the population even in the context of a shifting metabolic profile. (After Hurst and Beynon 2004.)

There is ample evidence that neural output from the VNS and MOS combine in certain regions of the brain to integrate information from the two sources (Martinez-Marcos 2009; Tirindelli et al. 2009). In the hamster, some outputs of the MOS and VNS both converge onto single neurons in the amygdala in hamsters (Licht and Meredith 1987), in the hypothalamus in rats (Han and Swanson 2010), and in the telencephalon in salamanders (Roth and Laberge 2011). New techniques for mapping active circuits in the brain also demonstrate how sexually experienced animals can use inputs from the MOS in place of the VNO, and where these link to outputs from the vomeronasal system (Martinez-Marcos 2009). Inexperienced hamster males show low levels of activation in the medial preoptic
area (MPOA) after they have investigated vaginal fluid, whereas sexually experienced males show a high level of activation after equivalent exposure to vaginal fluid. Removal of the VNO in experienced males does not reduce this activation. Experience seems to sensitize the MPOA to chemosensory input and to re-route input from the main olfactory system so that it can substitute for vomeronasal input driving the MPOA (Meredith 1998).

What about humans? Although a vomeronasal organ is present in human embryos, in the adult it becomes a blind-ended diverticulum in the septal mucosa, and contains no functional sensory nerve cells that extend to the brain. Furthermore, despite the existence of 5 intact V1R genes in humans, several of these may not be functional. One of them is expressed in the main olfactory epithelium, but actual cells expressing this gene that make connections between the epithelium and the brain have not yet been found (Rodriguez et al. 2000; Mundy and Cook 2003; Young et al. 2005; Witt and Hummel 2006). Thus it appears that humans do not possess any semblance of a vomeronasal sensory system. Nevertheless, humans do show sex-specific behavioral and physiological responses to various odors received through the MOE that likely qualify as pheromones (Stern and McClintock 1998; Jacob et al. 2001). Androstadienone is a male-produced product that activates preoptic and ventromedial nuclei of the hypothalamus in women and affects their endocrine levels, physiological arousal, mood, and sexual orientation (Savic et al. 2001; Saxton et al. 2008; Maraziti et al. 2011). The OR receptor for this chemical has recently been discovered (Keller et al. 2007). Estrogen-like substances activate the paraventricular and dorsomedial nuclei of the hypothalamus in men (Savic et al. 2001). Apocrine sweat glands of the human axilla are the likely source of these steroid pheromones (Beier et al. 2005). Such odors affect mate choice and preferences, as described in main text Chapter 16. As in other mammals, the olfactory system of humans is well-designed to detect some volatile pheromones.

Literature cited


Cilia and Sensory Receptors

Introduction

Cilia and flagella are widely used by eukaryotic organisms for locomotion, stimulus reception, or both. Although separate names were originally assigned to these organelles depending upon their size and the number per cell, the internal structure and functions of cilia and flagella have turned out to be identical. We shall thus refer to all such organelles as cilia. One of the most intriguing questions is why cilia have been recruited so often as sensory receptors. Below, we provide a general review of how ciliary structure appears to be correlated with function, and outline some possible reasons for their use as receptor devices.

Cell skeletons and cilia

Animal cells require an internal scaffolding or cytoskeleton to hold their shape. A mesh of actin protein filaments usually underlies most cell membranes. This cell surface support is complemented by a network of intermediately sized proteins that maintains the three-dimensional structure of the cell’s cytoplasm. Finally, the centrosome of the cell, consisting of two perpendicularly oriented centrioles, produces a third meshwork of microtubules throughout the cell that is used for additional support and as “rails” for the transport of internal cell components. The centrosome network also mediates the partitioning of cellular components during cell division. These microtubules are largely composed of tubulin proteins. Each centriole is a barrel-shaped organelle whose walls consist of parallel microtubules arrayed into nine clusters with three tubules per cluster.

Animal sensory cells respond to stimuli by varying the permeability of specific ion channels in their membranes. This changes the ionic composition inside the cell’s cytoplasm, produces a change in electrical fields across the cell membrane, or both. Either effect is maximized when the area of the responding cell membrane is large relative to the volume of cytoplasm that it encloses. There are several ways sensory cells can achieve this high surface area/volume ratio. One is to elaborate the membrane surface exposed to stimuli into a large number of small fingers called microvilli. The membrane surrounding each microvillus is continuous with the overall cell membrane, and actin filaments that extend from the mesh under the cell membrane into the cytoplasm of the microvilli provide the necessary support (Cooper and Hausman 2007). An alternative is to place one or more cilia on the exposed cell surface. Like microvilli, the membranes enclosing each cilium are continuous with the adjacent cell membrane. The cytoplasm inside the cilium is usually somewhat isolated from that in the rest of the cell by a terminal plate at its base (Singla and Reiter 2006). Cilia differ from microvilli in that they are typically larger in both diameter and length, and their support is provided by parallel microtubules generated by adjacent centrioles. Whereas the centrioles consist of nine triplets of parallel microtubules, the cilia attached to them usually have nine or more pairs of microtubules forming an internal cylinder of support. The ensemble of parallel pairs of tubules in a cilium is called its axoneme.

Structural types of cilia

Cilia can usually be assigned to one of two classes depending upon their axoneme structure:

- **9+2 Cilia**: The axoneme of most motile cilia and flagella contains nine pairs of parallel tubules arranged around the periphery of the ciliary interior. These tubules are continuous with those in one of the centrioles (called the basal body) just beneath the base of the cilium. A tenth pair of tubules runs along the center of the cilium and is usually surrounded by a thin sheath. This configuration is thus called a 9+2 (or $9 \times 2 + 1$ pair) design. Each of the peripheral pairs of tubules extends inner and outer arms made of the protein dynein towards the next pair of tubules. There is usually a spoke connecting each peripheral pair of tubules to the sheath surrounding the central tubules. A cross section of a 9+2 cilium near its base thus looks like this:
Closer to the tip of the cilium, the pairs of peripheral tubules may merge into a single tubule and the central pair may disappear. Although discovered first, these are often called secondary cilia.

- **9+0 Cilia**: Members of this class of cilia lack the central tubules, spokes, and dynein arms described above. It is thus described as a 9+0 configuration. Some species have a 9+0 configuration at the base of the cilium, but this gradually turns into 8 pairs of tubules in the periphery and one central tubule (e.g., 8+1) (Zakon 1986; Whitfield 2004). At the very tip, 9+0 cilia often contain only a few remaining tubules and their relative disposition is highly variable. 9+0 cilia often have conspicuous links between the two underlying centrioles with the more internal one generating a large root into the cytoplasm (Yack 2004). Cilia with the 9+0 structure are often called primary cilia because they appear so widely in vertebrates and many other animal taxa. Nearly all vertebrate cells except ova, including nerve cells, host a single primary cilium with a 9+0 configuration and usually no dynein arms or spokes at some point in development (Whitfield 2004; Praetorius and Spring 2005; Singla and Reiter 2006; Christensen et al. 2007).

**Ciliary function**

As noted earlier, cilia can have either or both of two functions: (a) propelling the organism and/or the adjacent medium relative to each other by beating rhythmically, and (b) acting as sensory receptors. It was originally believed that 9+2 cilia were always motile and locomotory organelles, whereas 9+0 cilia were always immotile and sensory in function (Satir 1977). Subsequent studies have shown that a variety of combinations of structure and function exist in nature (Ibañez-Tallon et al. 2003). We give examples below of some of these combinations:

- **Locomotory (motile) cilia**: These cilia (and flagella) provide propulsion for small organisms, or create currents of adjacent medium for larger and/or sessile ones. Cilia are very widely distributed on the external surfaces of aquatic invertebrates. However, even terrestrial vertebrates may use internal cilia to keep airways clear of dust and particles, and to move gametes around in reproductive organs. Most motile cilia have a classic 9+2 structure, and their physiology is well understood. Their major task is to beat by bending first in one direction and then the other in a repetitive manner. Studies have shown that the dynein arms on the outer tubules of their axonemes are the biochemical motors that generate the bending (Karp 2007). They do this by grabbing the nearest adjacent tubules and “burn” ATP fuel to power a ratcheting movement along the length of the other tubule. Since they are rooted to a separate pair of tubules, their movement causes adjacent pairs of tubules to slide past each other. The central pair of tubules are asymmetric and appear to coordinate the temporal patterning of a beating stroke (Porter and Sale 2000). Their action is communicated through the spokes to the inner dynein arms of each pair of peripheral tubules that then define the amplitude and waveform of the stroke. The outer dynein arms respond by doubling the frequency of beating and adding power to each stroke.
Although the central tubules, spokes, and dynein arms were all thought to be essential for rhythmic beating, the embryos of many vertebrates have special 9+0 nodal cilia on their ventral surface that are essential for normal development. These cilia lack central tubules and spokes, but they do have special dynein proteins that allow them to beat in a rotational manner. This beating causes currents that are necessary to establish the left-right asymmetry of the developing embryo (Ibañez-Tallon et al. 2003; Praetorius and Spring 2005).

- **Sensory cilia**: Multiple examples of ciliary receptors exist for every sensory modality used in animal communication: vision, audition, olfaction, touch, hydrodynamic reception, and electroreception. The relevant axoneme structures vary with both modality and taxon:
  - **Mechanoreception, audition, and hydrodynamic detection**: The external hairs and trichobothria, scolopale ears, and substrate sensitive mechanoreceptors of arthropods typically contain a 9+0 ciliary segment (Keil 1997; Yack 2004). Those in some insects may be motile despite the lack of a central pair of tubules (Göpfert and Robert 2003). The detectors inside mammalian kidneys that monitor fluid flow also rely on 9+0 cilia for stimulation. In contrast, the kinocilia of vertebrate lateral lines, vestibular organs, and ears are all 9+2 ciliary structures (Popper and Fay 1999). No known vertebrate touch receptors rely on ciliary components for stimulation.
  - **Electroreception**: Only the ampullary electroreceptors of primitive fish have a ciliary component; the ampullary-like and tuberous receptors in teleosts have microvilli instead. Where examined, the cilia of primitive fish eye electroreceptors show a 9+0 configuration at the base that changes into an 8+1 design for most of its length (Teeter et al. 1980; Zakon 1986).
  - **Photoreception**: Photoreceptors in jellyfish (Cnidaria) include a ciliary structure with a 9+2 axoneme (Eakin 1982). More advanced animals may have either microvillar (rhabdomeric) or ciliary photoreceptors, and some species have one type in their eyes and the other type located in the brain or some other tissue for monitoring circadian cycles (Arendt 2001; Arendt and Wittbrodt 2001; Arendt et al. 2004). Where ciliary photoreceptors are present, most have a 9+0 structure (Eakin 1979, 1982), but there are exceptions such as the 9+2 receptors in the larval eyes of snails (Blumer 1994). Both rhabdomeric and ciliary photoreceptors begin development with a 9+2 cilium: rhabdomeric photoreceptors entirely lose the cilium as they mature, whereas the ciliary photoreceptors tend to retain at least the outer pairs of axoneme tubules (Yamada 1982; Arendt and Wittbrodt 2001).
  - **Olfaction**: Whereas some crustaceans have 9+2 cilia in their chemoreceptors, most insect chemoreceptors use 9+0 cilia (Grünert and Ache 1988). Vertebrate chemoreceptors favor 9+2 cilia, and many are known to be motile as well as sensory (Lidow and Menco 1984).

- **Primary cilia and development**: Primary cilia have turned out to have critical signaling functions during vertebrate development, and possibly in other metazoans (Goetz and Anderson 2010; Louvi and Grove 2011; Vincensini et al. 2011). As noted earlier, most vertebrate cells host a primary cilium, at least early in development, and these appear to be the main “sensory” organelles by which embryonic cells respond to the hedgehog signaling pathway. Development creates different concentrations of hedgehog proteins in different parts of the embryo, and these influence what type of tissue and organ each cell will become as well as differentiating the main body axes. Mutants with defective ciliary functions exhibit major deformities and disfunctions as a result. The intrinsic sensory properties of cilia (see below) make them ideal targets for this type of developmental regulation.

### Cilia as preadaptations for sensory receptors

Several factors, either singly or in concert, appear to have pre-adapted cilia as sensory receptors:

- **Phylogenetic history**: As discussed in Chapter 7, cilia of single-celled eukaryotes are often responsive to touch and other stimuli. This requires the presence of suitable ion channels in their membranes that can be coupled to appropriate stimuli (Hegemann 1997; Machemer et al. 1998; Govorunova et al. 2004). In most cases, stimulation
triggers the admission of calcium ions and either chemical cascades and/or electrical field changes across the cell membrane. These mechanisms of single-celled eukaryotes were retained in early multicellular organisms, and thus provided a significant preadaptation for subsequent specialization of somatic sensory cells (Praetorius and Spring 2005).

- **Internal transport system**: All cilia, whether motile or immotile, have a system for transporting small particles and intracellular components inside the ciliary cytoplasm (Scholey 2003; Praetorius and Spring 2005; Inglis et al. 2006; Singla and Reiter 2006). This transport system uses the axoneme as a scaffolding: kinesin motors move components from the base to the tip of the cilium, and dynein motors move components in the opposite direction. Since most organisms resorb their cilia or flagella before cell division, ciliary reconstruction is a frequent event (Quarmby and Parker 2005). All materials for building and repairing cilia must come from the main body of the cell (usually the Golgi apparatus), and are passed through the selective pores of the terminal plate at the base of the cilium. They are then attached to a protein transport particle and moved along the axoneme to their site of usage. After discharging their cargo, the transport particles are carried back to the base of the cilium and readied for another cycle (Rosenbaum and Witman 2002). In addition to the building of a cilium, the transport system provides fuel for motile cilia and flagella, and transports signals stimulated by ciliary membrane receipt of developmental regulators (such as the hedgehog proteins), down to the host cell body where they modulate cell activities.

The ciliary transport system also plays a critical role in sensory receptors. Most sensory organs respond to outside stimuli continuously. Once stimulated, a sensory cell must restore itself to its prior sensitive state as soon as possible. This will invariably require rapid and massive transport processes: ions that entered the cell upon stimulation must be moved back out; photoreceptor pigments denatured by absorbing light must be restored at some energetic cost; chemical cascades begun when olfactory or light stimuli hit a cell must be retriggered for the next stimulus. The internal transport system originally evolved to provide fuel for beating cilia was an excellent preadaptation for restoring sensory receptors back to pre-stimulus conditions quickly. This is surely one reason why cilia have so often been recruited into sensory organs (Christensen et al. 2007).

- **Ubiquity**: Motile cilia occur on the external body surfaces of nearly all aquatic invertebrates (Brusca and Brusca 2003). Terrestrial arthropods usually do not have external cilia, but they use them widely inside their bodies. The sensory functions of primary cilia after development is complete are only beginning to be appreciated. At a minimum, primary cilia may act as mechanoreceptors to detect local flows of medium or movements of adjacent cells and thus coordinate activities of cells in a given region. There is also evidence that they may act as chemosensory aerials by absorbing extracellular chemical signals released by other cells and conveying them, using their internal transport system, to the cytoplasm of their own cell. These broad sensitivities have clearly been exploited by vertebrates to regulate differentiation during development. Whatever the function(s) that have generated it, the ubiquity of primary cilium in vertebrates and motile cilia in other taxa clearly enhances the chance that some will be recruited over evolutionary time into new locations and types of receptors.

### Ciliary versus non-ciliary receptor systems

Given the abundant reasons why cilia might be recruited into sensory organs, why are they not the only such source? The fact is that they are not. While all vertebrates so far examined have primary cilium sensitivity to hedgehog proteins, this is not the only way that cells can respond to hedgehog proteins, and the latter are not the only way that cell differentiation during development is regulated (Goetz and Anderson 2010; Vincensini et al. 2011). *Drosophila* development proceeds with a type of hedgehog proteins, but cilia do not play an important role in signaling. We noted above that photoreceptors exist in both ciliary and rhabdomic configurations, and that each has its own set of photoreceptor (opsin) proteins and associated genes (Arendt 2001; Arendt and Wittbrodt 2001; Arendt et al. 2004; Fernald 2006). We noted in Chapter 6 that chemoreceptive organs may have receptor cells that are ciliary (olfaction), microvillar (taste), or both (vomeronasal organs). And as discussed in Chapter 7, mechanoreceptors can rely on either of two widely distributed but distinct stimulation mechanisms, each having its own depolarizing ion (calcium or sodium), ion channel proteins (TRP or degenerin/ENaC), and associated genes. For each modality, the two alternative mechanisms seem to be equally ancient in the animal lineage. Why should most sensory modalities have evolved two
alternative ways of doing the same thing? While there may be differences in sensitivities of the two alternatives in any
given modality, the same receptor cells never seem to employ both mechanisms: if the dual alternatives are present in
the same organism, they are invariably assigned to different kinds of cells in different parts of the body. There is clearly
more to the story of when and why cilia are recruited as sensory receptors that remains to be discovered.

References cited


Arendt, D. and J. Wittbrodt. 2001. Reconstructing the eyes of Urbilateria. Philosophical Transactions of the Royal

Blumer, M. 1994. The ultrastructure of the eyes of the veliger larvae of Aporrhais sp and Bittium reticulatum (Mollusca,


Christensen, S. T., L. B. Pedersen, L. Schneider, and P. Satir. 2007. Sensory cilia and integration of signal transduction

Associates.


Reviews Genetics 11: 331–344.


Ibañez-Tallon, I., N. Heintz, and H. Omran. 2003. To beat or not to beat: roles of cilia in development and disease.
Human Molecular Genetics 12: R27–R35.


Web Topic 7.2
Hydrodynamic stimuli

Introduction

Hydrodynamic stimuli are created when an object moves through a fluid or a fluid moves past an object. The geometries of the stimuli and their persistence after being created vary with the size of the object, the relative velocities of the fluid and object, and the viscosity of the fluid. Although this sounds like a simple relationship, increasing one or more of these parameters do not simply scale up the stimuli, but instead can produce discrete changes in their geometries and life times. Fluid hydrodynamics turns out to be a fairly complicated area of physics. It is not an esoteric topic, however, as it is crucial to our understanding of weather patterns, the design of airplanes, helicopters, and sailboats, the mechanisms by which aquatic invertebrates and fish swim and birds, bats, and insects fly, the shapes of pine cones and flowers, the formation of spiral galaxies, and the behavior of black holes in space.

Reynold’s numbers

Although the interaction between fluids and objects can produce quite varied hydrodynamic stimuli, molecular fluids such as air or water behave similarly once the appropriate scale corrections are made. This scaling is achieved by computing a dimensionless Reynold’s number for the situation. The magnitude of the Reynold’s number allows us to provide at least a qualitative prediction about what kinds of hydrodynamic stimuli will be created by the interaction of an object and a fluid when they are moving at different velocities.

As discussed in Web Topic 6.4, the Reynold’s number is a ratio of the inertial forces in a fluid (which tend to maintain movement of the fluid) to viscous forces (which tend to slow down fluid movements). It is computed as the product of the effective size $L$ of an object (in meters) and the relative velocity $v$ of the object and the fluid (m/sec) divided by the kinetic viscosity of the fluid (m$^2$/sec). The latter is a measure of the resistance of a fluid to spreading at the ambient temperature. It is computed by dividing the absolute (dynamic) viscosity of the fluid $\mu$, (the forward momentum of a given area of moving fluid measured in Newton-seconds/m$^2$) by the density of the fluid $\rho$ (kg/m$^3$). Algebraically, the Reynold’s number, $Re$ is thus equal to

$$Re = \frac{\rho v L}{\mu}$$

Water versus air

The density of water at room temperature is about 800–900 times that of air. The absolute (dynamic) viscosity of water is about 51 times higher than that of air. As a result, the same relative velocity of fluid and object and the same object size result in a Reynolds number in water that is about 15–17 times higher than that in air. Put another way, we expect air and water to show similar hydrodynamic behaviors when either the object or the flow rate (or more accurately, their product) is 15 times smaller in water. In practice, typical flow rates in air (e.g. wind) are 10–15 times faster than in water (e.g. stream currents). Animals of similar size in air and water thus tend to experience similar patterns of hydrodynamic stimuli (Vogel 1994).

Some examples of Reynold’s numbers for moving objects in air (blue) and water (red):
Characterizing hydrodynamic stimuli

Hydrodynamic stimuli are traditionally generated by inserting an object of interest into a steady and linear flow of fluid. Wind tunnels are often used for studies with air, and water sluices for aquatic studies. In more recent studies, live animals are allowed to move through a static fluid (air or water), and the hydrodynamic effects of these movements studied. In either case, the challenge is to make visible the resulting eddies, vortices, and other movements of the fluid. Ideally, one would track the path of an individual fluid particle over time, and then repeat this effort with additional particles that began at different initial locations in the fluid. When no object is present in a steady fluid flow, each such particle should trace out a parallel straight line as it flows through the tunnel or sluice. These trajectories are called pathlines. Placing an object in a fluid flow will distort what otherwise would be straight and parallel pathlines. It will also create a mosaic of locations in front, around, and downstream from the object where fluid pressures and fluid velocities (both magnitude and direction) differ from each other and from the average for the fluid as a whole. Movement of an animal in a static fluid has similar effects. How can these be characterized?

There are a number of ways to visualize hydrodynamic perturbations in fluids (http://en.wikipedia.org/wiki/Flow_visualization). Several currently popular methods include:

- **Thread tracking**: Position fine threads in a grid and use their shape and/or direction pointed by their free ends to map the local patterns of fluid motions.
- **Particle tracking** (http://en.wikipedia.org/wiki/Particle_image_velocimetry): Add smoke, dyes, fine oil vesicles, tiny beads, or other particles to the flow and use visual recording methods to a) map areas of higher or lower concentration of the markers, or b) track the trajectories of individual particles over time (Yen & Strickler 1996; Videler et al. 1999; Hanke et al. 2000; Videler et al. 2002; Hanke & Bleckmann 2004; Videler 2005; Catton et al. 2007; Schulte-Pelkum et al. 2007).
- **Schlieren optics** (http://www.fas.harvard.edu/~scidemos/LightOptics/SchleirenOptics/SchleirenOptics.html): Heterogeneity of density and velocity in the fluid produces heterogeneity in its refractive properties. Light is passed through the fluid and local differences in refractive index produce an image with correspondingly lighter
or darker regions (Hwang & Strickler 2001).

- **Laser Doppler Anemometry** ([link](http://www.aoe.vt.edu/~devenpor/aoe3054/manual/expt4/index.html)): Light from a single laser is split into two beams emanating from different points but focused by a lens on a common point. When a reflective particle being moved in a fluid flow passes through this focal point, the two beams are reflected slightly out-of-phase depending upon the velocity and direction of movement of the particle. The reflected beams are recombined and the velocity of the particle is computed based on the level of beam interference (Bleckmann et al. 1991).

## Vortices

When water in a flowing stream encounters an obstacle, the result is often a swirling eddy. Eddies are one example of a vortex. In a vortex, pathlines become circular or spiral instead of linear or gently curvilinear. There are several kinds of vortices:

- **A rotational vortex** is one in which all contents within the vortex rotate around the central axis at the same angular velocity. Imagine a pan of water placed on a round turntable. A point a short distance from the center of the resulting vortex will move over a shorter curvilinear distance in a given time interval than will a more distant point. While angular velocity (complete circuits/sec) is the same everywhere,

  ![Angular velocity](image1)

  more peripheral points in the vortex travel at a higher curvilinear (tangential) velocity than do more central ones:

  ![Curvilinear velocity](image2)

Suppose a leaf falls onto the right side of a rotational vortex with its twig end closest to the vortex center and its tip facing the periphery of the vortex. Because the tip end of the leaf will move faster in a curvilinear trajectory than the stem end, by the time the leaf gets to the left side of the vortex, the stem end will now point to the right (still towards the center of the vortex), whereas the tip end will point left (towards the periphery). Thus an object borne by a rotational vortex will rotate about its own axis as it moves around one full circuit of the vortex.
An **irrotational vortex** also rotates around a central axis, but points closer to the center have a higher angular velocity than do more distant points:

We can create such a vortex by putting a spinning beater or cylinder into a pan of water. Water closest to the spinning object will have the highest angular velocity, while more distant regions of water will rotate around the pan more slowly. And here, the tangential velocity is also higher near the center:

In fact, the product of tangential velocity and radius will be a constant. Suppose a leaf drops into the right side of this vortex with its stem aimed towards the center and its tip towards the periphery. As it rotates around the pan, its stem will move more quickly along its circular path than will the leaf tip move along its path. The result is that the leaf will continue to have its stem aimed to the left as it moves around the vortex; it does not rotate around its own axis as in a rotational vortex.

Hence the name “irrotational” vortex.

- A third example is called a **vortex ring**. Here, the vortex takes the shape of a doughnut (torus):
Fluid circulates around the cross-section of the torus such that all fluid entering the center of the torus from a given side is going in the same direction. The result is a “jet” of fluid moving through the torus center parallel to its axis. A cross-section of the torus would thus show the direction of circulation to be opposite on opposing segments of the torus. This has to be the case if all fluid passing through the center is going in the same direction. A smoke ring is a well-known example of a torus. Vortex rings are usually moving as a whole along a line parallel to their main axis while passing fluid circulates through the center and around the periphery to enter again on the opposite side.

- **Complex and multiple vortices:** Most irrotational vortices actually contain a rotational vortex at their core and thus are a combination of the two types. Although it is possible to generate a single vortex (e.g. when water or air flows over the ground and encounters a hole or trench in the surface inside of which a vortex is often created), most vortices are created in pairs (rotational and irrotational vortices) or even larger numbers of the same kind of vortex (ring vortices). Pairs of vortices created at the same time usually rotate in opposite directions. Persistent vortices circulating in opposite directions tend to repel each other; those circulating in the same direction will attract each other and may fuse into one larger vortex. Note that vortices do not need to be planar. In fact, many are three dimensional. Examples of three dimensional vortices include water circulating around and through an open drain, tornadoes, dust devils, some kinds of waterspouts, and hurricanes and cyclones. Large animals moving in fluids tend to produce strings of successive ring vortices as a wake behind them. These tend to be aligned such that the jet passing through one is continuous with the jet passing through the one created before it and the one created just after it.

Examples of naturally occurring vortices can be found online at [http://www.engineering.uiowa.edu/fluidslab/gallery/vortex.html](http://www.engineering.uiowa.edu/fluidslab/gallery/vortex.html).

**Boundary layers**

When a volume of fluid and a solid object are moving relative to each other, those molecules of fluid closest to the object surface cannot move differentially from the object. They are forced to move with the object and thus act as if “stuck” to the surface. Fluid molecules far enough away from the object will move at the same relative velocity as other nearby molecules of the fluid volume. In between, there is a gradient of relative fluid velocity which goes from zero at the object surface to that shared by most of the molecules in the fluid volume. This intermediate region showing a gradient of velocity around the object is called the boundary layer.
As fluid slows down and collects in a thin boundary layer on the upstream side of an object, molecules that are not too close to the surface flow along pathlines that track the surface shape of the object. At some point along each side of the object’s surface, this fluid stops following the object shape and simply heads off downstream. These are known as the separation points. For very low Reynold’s numbers, the separation points are located well on the rear (downstream) side of the object. As Reynold’s numbers are increased, the separation points move forwards toward the object’s upstream side. This allows an increasing amount of fluid to pool on the downstream side of the object where it can even backflow towards the object, move along lines parallel to its surface, and finally join the downstream flow at the separation points. This circular movement thus generates eddies or vortices downstream from the object. At high enough Reynold’s numbers, the separation points detach from the object and move downstream. This drastically changes the composition and properties of the downstream boundary layer.

Patterns of hydrodynamic stimuli

Consider a static object in a continuous flow of fluid. As noted above, low Reynold’s numbers are obtained when the kinematic viscosity is much greater than the product of relative velocity and object size. Put another way, the resistance of the fluid to spreading in this case exceeds the inertial forces imposed on the fluid by its encounter with the object. When the object is large and/or the relative velocities are high, then inertial forces easily exceed the viscous resistance of the fluid. This is the case for large Reynold’s numbers. Intermediate values result in a more even match between viscous and inertial factors. Depending on the relative influences, there are also two intermediate cases that are easily distinguished. Each of these four situations generates a qualitatively different type of hydrodynamic stimulus (Cf. Feynman 1964 and Vogel 1994):

- **Very Low Reynold’s Numbers (<10):** These conditions produce unidirectional flow of the fluid despite the presence of the object. Either the object is so small or the relative velocity of object and fluid so minimal that pathlines that would otherwise intersect the object are bent so that the fluid simply sweeps past the object without causing eddies or other effects:

  ![Diagram](image)

  The viscous forces in this situation quickly attenuate any perturbations in the fluid as heat. No vortices are formed. In the case of a fish in water or insect in air with such a low Reynold’s number, the passage of the animal leaves no detectable wake to the side or downstream from it.

- **Low Intermediate Reynold’s Numbers (10–50):** With larger objects and/or higher relative velocities between fluid and object, fluid begins to pile up on the upstream side of the object faster than it can flow around the object to relieve the pressure. As it works its way around the object, it creates eddies (vortices) on the downstream side. The typical result is a pair of vortices circulating in opposite directions and remaining “attached” (e.g. fixed in location relative to the object).
High Intermediate Reynold’s Numbers (50–200,000): With still larger objects or higher relative velocities, the generated vortices become very large and are “shed” into the downstream fluid. Instead of two attached vortices at any time, a single vortex forms downstream from the separation point on one side of the object until it is shed, and then a second vortex forms with the opposite circulation pattern behind the opposite separation point. The two sides of the object thus alternate in creating shed vortices. The result is two parallel trails of downstream vortices called a von Kármán vortex street:


The regular frequency with which an object sheds successive vortices can be computed from another dimensionless parameter called the Strouhal number. This depends on the size of the object and the relative velocity of object and fluid. Shed vortices are usually ring vortices and successive vortices in the same track typically share a common jet (Videler et al. 2002). Note that while relative flow velocity sufficiently downstream from the object is relatively constant everywhere at low Reynold’s numbers, shed vortices produce very heterogeneous velocity profiles downstream of the object over a wide area and for as long as the vortices continue circulating. Since inertia greatly exceeds viscous resistance, vortices can persist for many minutes after a swimming or flying animal has moved on. These heterogeneities in velocity and fluid pressure can feed back on the object and make it vibrate in synchrony with the frequency of vortex generation. If this frequency is close to a resonant frequency of the object, significant vibrational amplitudes are possible. The interactions of both attached and shed vortices and the objects producing them are exploited by many animals and humans to facilitate flight, swimming, feeding, and other functions (Webb 1978; Vogel 1994; Videler 2005). As Reynolds numbers increase, the orderly arrays of shed vortices become increasingly disordered by turbulence. Initially, this turbulence is confined to the regions inside each vortex, but at even higher Reynolds numbers, even the vortices are disrupted by turbulence in the fluid.

High Reynold’s Numbers (>200,000): At a sufficiently high Reynolds’s number, the value of which depends largely on the object shape, the separation points for flow around the object separate from it and move downstream. This drastically reduces the drag experienced by the object and narrows the downstream boundary layer to a much smaller “tail” behind the object. Within that boundary layer, structured vortices are usually absent. Instead, the content is unstructured turbulence and chaos.
Animal wakes

Animals that move through water or air are more likely to leave behind a persistent wake if they are large and/or they move quickly. A small copepod that is slowly foraging may leave no wake at all; however, if it is startled and leaps away, it is likely to leave behind some shed vortices that could be detected by an appropriate predator (Yen & Strickler 1996; Videler et al. 2002). The higher the relevant Reynold's number, the more likely a wake will be left behind and the longer it will persist (Vogel 1994). As noted above, the spacing and persistence of shed vortices could provide information about the size and velocity of the animal creating them. The situation is complicated by the fact that most animals achieve locomotion by repetitively moving some body part: birds and insects flap their wings, copepods oscillate paired appendages, and fish undulate. The frequency composition of persistent wakes can then also be used to extract additional information about the wake's creator (Bleckmann et al. 1991).

Further Reading:

There are a number of texts that cover hydrodynamics and fluid mechanics. In order of increasing mathematical difficulty, we recommend: Vogel (1994) (an outstandingly readable treatment of fluid dynamics, vortices, and hydrodynamics from a biologist's point of view. Strongly recommended!); Videler (2005) (This book focuses on animal flight, but provides a useful general introduction to Reynold's numbers and vortices); Feynman (1964) (chapters 40–41 outline the basic logic of fluid mechanics. The book uses vector math, but Feynman is so good at explaining things that a reader need not be fully fluent with curls and gradients); and Lighthill (1986) (despite the title, this is a more advanced text using the usual vector math associated with fluid mechanics. It is widely cited in the prior books.).


Literature Cited


Web Topic 7.3
A primer on electrical signals

Basic electrostatics

- **Charge**: An object (atom, molecule, piece of material containing many molecules) with unequal total numbers of electrons and protons is said to be **charged**: each excess electron adds a charge of -1 and each missing electron (excess proton) adds a charge of +1. The net charge on the object is the sum of the charges contributed by each unpaired electron or proton. It is usually measured not in electrons or protons but in **coulombs**. One coulomb is equal to $6.25 \times 10^{18}$ unpaired electrons (or protons).

- **Coulomb’s Law**: Two nonmoving objects in a vacuum with charges $Q_1$ and $Q_2$ respectively will be attracted to each other (if $Q_1$ and $Q_2$ have opposite signs) or repelled (if $Q_1$ and $Q_2$ have the same sign) with a force $F$ (in **Newtons**) equal to:

$$F = \frac{1}{4\pi \varepsilon_0} \frac{Q_1 Q_2}{r^2}$$

where $r$ is the distance between the two objects in meters, and $\varepsilon_0$ is known as the **permittivity constant** ($= 8.85 \times 10^{-12}$ coulombs$^2$/newtons$\cdot$meters$^2$). Note that the amplitude of this force decreases with the square of the distance. It thus can become quite weak at large distances from the object.

- **Electric Field**: A small test charge moved into any location near enough to a charged object will experience a net electrostatic force. A map of the direction and amplitude of that force at all locations around a charged object defines the **electric field** around the object. For a single charged object (**monopole**), the electric field lines of force radiate away from (or towards) the object equally in all directions. If multiple charged objects are present, the amplitude and direction of the net force at any point in the electric field is the vector sum of the component forces present at that point. For two charges of opposite polarity (sign) sufficiently close together (**dipole**), the lines of force in the surrounding electric filed are curved.

- **Multipolar Fields**: The electrical fields generated by biological sources are rarely dipolar and practically never monopolar. Instead, complex arrays of charges will generate many axes around which the charges are distributed. The resulting electrical field is the sum of the effects of the multiple axes. The most important axis is usually the dipole component, the next most important axis the quadrupole component, a third axis the octupole component, etc. The relative contributions of each axis to the overall electrical field depends on the distance between the sampling point and the object. Dipole contributions fall off with distance as $1/r^3$, quadrupoles as $1/r^4$, octupoles as $1/r^5$, etc. As a result of the faster fading of higher order axis contributions, only the dipole component will be detectable at large distances; however, at close distances, all components can contribute significantly to the electric field.

- **Electric Potential**: The electrostatic **potential** at any location surrounding a charged object is the amount of work that is required to bring a small unit of positive charge from infinity (where the electric field force surrounding the object is zero) to the location. If the object has an overall positive charge, one must do work against the field’s repellent force to bring the unit charge to the location. The electrostatic potential in this case is said to be positive. If the object has a negative charge, it will exert an attractive force on the positive test charge and thus the work done to bring the test charge closer will be negative. In this case, the electrostatic potential at the final resting location of the test charge is said to be negative. Electric potential is measured in **volts** and is thus often called the “voltage” at a location. The electric field is the spatial gradient in voltage at any location. The potential difference between any two points in the electric field is simply the difference between the voltages at the two points and is often called the “voltage drop” between the two points. One can connect all points around an object that have the same voltage as isopotential lines. These are equivalent to the lines connecting all locations at the same altitude in a topographic map. Examples of the electric field lines (dashed) and isopotential lines (solid) for a monopole and a...
- **Dipole Voltage:** The voltage $V$ at a distance $r$ from a dipole and an angle $\theta$ (relative to the line joining the two point charges in the dipole) is

$$V(r, \theta) = \frac{1}{4\pi\varepsilon_0} \frac{Q\delta \cos(\theta)}{r^2}$$

where $Q$ is the magnitude of the charge on each part of the dipole and $\delta$ depends on the distance between them. Thus along the line perpendicular to that joining the two charges and midway between them, the angle $\theta$ is 90° and the $\cos(90°) = 0$; thus the voltage along that line is zero. Points outside the dipole but along the line joining the two charges will show the maximal voltage values.

## Electric fields in different media

- **General Formulation:** The formula given above for the electrical force at any location surrounding two charges $Q_1$ and $Q_2$ in a vacuum can be generalized for two charges in any medium as follows:

$$F = \frac{1}{4\pi k\varepsilon_0} \frac{Q_1 Q_2}{r^2}$$

where all terms are as before and $k$ is the **dielectric constant** of the medium. A vacuum is a perfect **insulator** in that the electric force created by the two charges cannot induce any repositioning or movement of other electrons, atoms, or molecules. The dielectric constant for a vacuum is 1.

- **Conductors:** At the other extreme, are **conducting** media in which electrons, atoms, and/or molecules are present and free to move under the influence of the electric field. Moveable electrons, for example, in the presence of an electric field will move towards the positive pole leaving their formerly paired positive charges to accumulate near the negative charge. This redistribution of elements of the medium so that opposite charges accumulate around the initial charges continues until it cancels out the electric field inside the conductor. The dielectric constant for a
conductor is thus set at infinity, and plugging this value into the above equation, one can see that inside the conductor, the force at any location is zero. By the same token, it will take no work to move a test charge around inside the conductor and thus the voltage inside a conductor is the same everywhere.

- **Dielectrics**: These are media in which movements of electrons, atoms, and molecules are constrained. However, it is still possible for electrons to move within a medium atom or molecule, or it is possible for a molecule to rotate so that its most positive side faces the negative charge and its negative side faces the positive charge. The parallel alignments of medium molecules or electrons inside an atom or molecule create thousands of tiny dipoles with lines of force opposite to those surrounding the original charges. The result is a reduction in the amplitude of the electrical field surrounding the charges: greater polarization of the medium results in greater diminution of the electrical field. Higher values of the dielectric constant reflect greater susceptibility to polarization and thus a greater reduction in the electrical field inside the medium. The dielectric constant for air is 1.00054, glass 4.7, and freshwater at room temperature about 80. Note that the dielectric constant also affects the measurable voltage at any point inside the medium. For example, the electrical potential surrounding a dipole in a non-conducting but dielectric medium is:

\[ V(r, \theta) = \frac{1}{4\pi\kappa\varepsilon_0} \frac{Q\delta\cos(\theta)}{r^2} \]

**Electric currents**

- **Ohm's Law**: Suppose we place an electric dipole in a medium which is a worse conductor than a metal, but a better conductor than most dielectrics. Water is such an example. Water invariably has dissolved materials within it, and many of these, such as salts, break up in water into their component charged ions. The presence of an electric field in water will cause positive and negative ions to move in opposite directions. The ionic trajectories follow the electric field lines. This movement of ions in water (or of electrons in a metal) is called an electric current. The magnitude of an electric current between two points (measured in coulombs/second or amperes) is proportional to the voltage difference between the points. The constant of proportionality between an applied voltage and a resulting current is called the conductance of the medium through which the current is flowing. More often, we use the reciprocal of conductance which is called the resistance. If \( V \) is the voltage difference between two points and \( R \) is the resistance (in ohms), then the current \( I \) (in amperes) depends on these variables according to Ohm's Law:

\[ I = \frac{V}{R} \]

The convention in physics is that current flows from a region of positive voltage to one of more negative voltage. Note that this is opposite to the actual flow of electrons in a metal (from a negative to positive potential location).

- **Resistivity**: Resistance in a particular context will be higher the greater the distance that the current must flow, the smaller the cross sectional area through which the current passes, and the worse the material as a conductor. The latter term is characterized by the material's intrinsic resistivity. Because of the resistance of the water in which we have placed our dipole, there will be a steady current of ions towards that part of the dipole of opposite charge to each ion. If there were no resistance, the initial current would quickly cancel the charge at each end of the dipole due to accumulations of oppositely charged ions. If the resistance is high enough, it may take some time before the dipole is fully neutralized. Alternatively, something may occur near the dipole to restore its charge. In either case, if the electric field is maintained or restored for a sufficiently long period, we can measure the electric potential at various points around the dipole and the amount of current at each location. For a stable source of current in a conducting medium, the potential at location \((r,\theta)\) from the dipole is
\[ V(r, \theta) = \frac{\rho_0 I \delta \cos(\theta)}{4\pi r^2} \]

where the medium resistivity, \( \rho \) and the current \( I \) have replaced the permittivity, \( \kappa \), and the charge, \( Q \), used for non-conducting media.

- **Varying Electric Fields and Impedance**: Water and many other materials are both conductors and dielectrics: some current will flow through them, but the resistance is high enough that electric fields are sustained and their ability to be polarized and act as a dielectric permits some build-up of counter-fields within the medium. For static electric fields, this may not be significant. If however, the electric field is changing in magnitude or direction, then the dielectric properties of the medium can become important. In a steady electric field, an electron in a conductor may move the entire length of the conductor. This is called a direct current (DC). Now suppose we apply a sinusoidally varying electric field to the conductor. Electrons will first move one direction and then back the other. This is an alternating (AC) current. The higher the frequency of the alternating field, the less distance any one electron can travel before it has to turn around and go the other way. In a non-conducting dielectric, electrons or polar molecules can move a bit, but they can never move far enough to sustain a steady DC current. However, if an alternating field is applied across such a material, the distance electrons have to travel per cycle may be within the polarizing limitations of the material: the higher the dielectric constant for the material, the slower the frequency of alternation which the material can track and thus carry current. The effective resistances of dielectrics may thus drop if the applied electric field is a varying one. To keep this notion of resistance distinct from classical DC resistivity, the term applied to such dielectrics is capacitative reactance. Capacitative reactance decreases with the dielectric constant of the material and with the frequency of the electric field oscillation. Like resistance, it is measured in ohms. Remember that even if the waveform of the electric field variations is not sinusoidal, it can be considered as the sum of a number of different sinusoids (see Web Topic 2.4). Applying such a non-sinusoidal signal to a dielectric, we will find that the dielectric will act like a high-pass filter since it can more easily track the higher frequency components than the lower frequency ones. The overall impedance of a medium like water to a varying electrical field will thus depend on both the resistivity of the water and on the capacitative reactance of the water at the various frequencies making up the waveform of the changing field.

- **Electrical Field Distortion**: We have assumed so far that media are unbounded and homogeneous. The resulting electric fields can be called free fields (by analogy with sound). However, most media have boundaries and contain objects whose dielectric and/or resistive properties differ from those of the medium. The usual situation is thus not a free field. Boundaries and objects in the medium will distort and change the field shape from free field conditions. For example, suppose we place a monopole in a medium such as water (left figure below) and then place an objects with a resistivity less than water near to the charge (right):

![Diagram of electric field distortion](image)

Objects which have lower resistivities than the medium bend the electric field lines in the region between themselves and the charge closer together and towards the object. This region of enhanced electric field magnitude corresponds to a region of very closely spaced isopotential lines and thus a steep gradient in voltage. Objects which have higher resistivities than the medium (below right) bend the electric field lines away from themselves, lowering the field magnitude in the region between themselves and the charge, and show a flattening
of the potential gradient in this region:

When many objects of differing conductivities are present, the shape of the field can become highly complex. Boundaries are also significant. If we place our charge near to the air-water interface, or near a non-conducting bottom, the electric field magnitude and potential near to the charge will be twice as great as that for a charge suspended in an unbounded volume of water. This is because current can radiate in all directions in the unbounded case, but can only radiate away from the water’s surface in the bounded example.

See Chapter 7 for examples of how differential resistance and capacitance in nearby objects can be used by some electric fish to discriminate between prey and inedible items, and to navigate through familiar locations.
Web Topic 7.4
Bioelectric field resources

Introduction

A variety of fish and a few primitive mammals have receptors that can respond to the electrical fields generated by other animals and electrochemical habitats. A subset of the fish species have also evolved special organs that can create significant electrical fields on command and invoke these electrical organ discharges for electrolocation and communication.

Passive electroreception

- **Paddlefish swimming**: This YouTube video gives a good view of paddlefish swimming: http://www.youtube.com/watch?v=fysqA0tr4qo
- **Sharks, rays, skates, sturgeon coelocanths, and echidna**: Good still images and/or movies of these passive electroreceptive animals can be found at: http://www.arkive.org/. The video of the thornback skate includes a brief look at its underside where the electric system pore openings can be seen (http://www.arkive.org/thornback-skate/raja-clavata(video-00.html)

Active electrogeneration and electrolocation

- **Phil Stoddard Lab** *(Florida International University)*: This group has measured the electric fields (as voltages) around various electric knifefishes (*Gymnotiforms*) and reconstructed the temporal variation in these fields as Quicktime movies: http://www.fiu.edu/~efish/visitors/electric_field_animations.htm
- **Mark Nelson Lab** *(University of Illinois, Urbana-Champaign)*: This site contains a series of very helpful webpages including movies and animations of electric fish foraging. Suggested links:
  - Background on electric fish: http://nelson.beckman.uiuc.edu/electric_fish.html
  - Background on electrolocation: http://nelson.beckman.uiuc.edu/electrolocation.html
  - Movies of foraging electric fish including simulations of stimulus patterns: http://nelson.beckman.uiuc.edu/movies.html
- **Malcolm MacIver Lab** *(Northwestern University)*: This group uses simulation and robotic models to study the stabilizing movements and electrical field measurement by electric fish. Additional programs may need to be downloaded to view some of these models and simulations: http://www.neuromech.northwestern.edu/uropatagium/ - RoboVid
- **James Bower Lab** *(California Institute of Technology)*: This group, including Chris Assad and Brian Rasnow, created a number of movie simulations of the electric fields of discharging fish. Pages include:
  - Electric fish Quicktime movies: http://alumnus.caltech.edu/~rasnow/index.html
  - Electric fish field simulations: http://alumnus.caltech.edu/~rasnow/sim.html
  - Responses to stimulation: http://alumnus.caltech.edu/~rasnow/behav.html

Electrocommunication
• **Carl Hopkins Lab (Cornell University):** Dr. Hopkins and his colleagues have posted a Flash movie of a spectrogram of electrical signaling with annotations. This example is typical of such interactions in electric fish. Be sure to listen to this file when you play it:
  Knifefish (*Sternopygus macrurus*): male courting female:
  http://www.nbb.cornell.edu/neurobio/hopkins/sternopygus/sternopygus_singing.htm

• **Erik Harvey-Girard:** This site (in French) has a very nice review of electrocommunication in the knifefish, *Apteronotus*: http://www.apteronote.com/. Use the directory on the left side of the Introduction page to examine various topics.

**Other topics**

• **An Expedition to Africa in honor of Mary Kingsley’s Prior Contributions to Electric Fish Biology:**
  http://www.nbb.cornell.edu/neurobio/hopkins/mkingsley.html
Web Topic 7.5
Adaptations for passive electroreception

Introduction

The early acquisition of passive electroreceptors in primitive fish was surely a key adaptation that facilitated its subsequent radiation and eventual dominance in aquatic habitats. In both marine and freshwater habitats, a number of strategies are employed to enhance passive electroreception.

Variations in the spatial distribution of receptors

Spreading many ampullary organs (or teleost equivalents) over a large area of body surface allows an animal to sample the amplitude of the electric field at many locations. Because the walls of ampullary canals are highly resistive, little current passes into or out of the canal except along its major axis. Thus electric field lines parallel to a canal will produce the largest stimulation of the associated sensory cells. Comparisons of stimulus levels for canals with nearby pores but different axis angles thus allow the animal to estimate not only the strength of the electrical field at a location but also its direction there. Pooling of inputs from many organs then permits the animal's brain to generate a fairly accurate map of the electrical field surrounding the sampled body surface (Montgomery and Bodznick 1999; Brown 2002; Keller 2004; Bell and Maler 2005; Bodznick and Montgomery 2005). This map can be extremely useful in determining the location of the electric field source and whether it is moving relative to the sampling animal.

Whereas lampreys, lungfish, and several extinct taxa of primitive fish spread their electroreceptive organs over much of their body surface (Bodznick and Northcutt 1981; Northcutt 1986; Ronan 1986; Northcutt 1997; Watt et al. 1999), the majority of passively electroreceptive animals concentrate them in relevant regions of their heads (Northcutt 1986; Zakon 1988; Jørgensen 2005). Within the head region, the distribution of the organs and their associated pores varies with the species' habitat, diet, and light levels when foraging. Because ampullary organs develop from the same embryonic tissues as the lateral line, their distribution is also affected by the disposition of the animal's hydrodynamic canals and superficial neuromasts (Northcutt 1986).

The 400–2500 ampullary receptors of sharks are concentrated entirely on their heads (Bodznick and Boord 1986). Species that forage in open ocean tend to have a more even distribution of receptors over the head’s dorsal and ventral surfaces while those that forage on benthic prey (such as skates) concentrate the receptors on the ventral side particularly around the mouth (Tricas 2001; Collin and Whitehead 2004). Sharks that feed on benthic prey as juveniles but in deeper waters as adults undergo a shift towards more even dispersion of ampullary receptors as they mature (Collin and Whitehead 2004). A more widespread distribution of receptors on the head would also facilitate the use of the earth’s electric fields for migratory species, but whether sharks can actually use electroreception for long range navigation remains unclear (Kalmijn 1974, 1988; Klimley 1993; Paulin 1995; Sundstrom et al. 2001; Collin and Whitehead 2004; Tricas and Sisneros 2004; Wilkens and Hofmann 2005).
Figure 1. Distribution of ampullary organ pores on the heads of sharks as function of foraging habitats. Ventral view (left half), dorsal view (right half), and ampullary canal openings (dots) are shown in each example. A) Hammerheaded sharks (*Sphyma spp.*) often feed just over the bottom of bays and estuaries, and young sharks live in murky waters with low visibility. Approximately 60% of their canal openings are on the ventral side of their bodies where prey are more likely to be encountered. B) Sandbar sharks (*Carcharhinus plumbeus*), like most shark species, have equal numbers of ampullary canal openings on the ventral and dorsal sides of the head. They encounter prey both above and below the body and can use vision to orient their attacks. C) Bull sharks (*Carcharinus leucas*) are unusual in their ability to tolerate brackish and freshwater habitats with
limited visibility. Like the hammerheaded sharks, bull sharks concentrate nearly 60% of their ampullary canal openings on their ventral side ahead of and to the side of their mouth. (After Collin and Whitehead 2004.)

The head of a skate or a ray merges smoothly into the flattened wings on each side of the body. Adult skates and rays can have from 400–1400 ampullary organs depending on the species (Bodznick and Boord 1986). These are usually clustered around the head but radiate their canals in all directions including several long canals that open on the rear edges of the wings. As with sharks, species that feed on benthic prey have higher concentrations of receptors and canal pores on their ventral side and around the mouth, whereas larger species that pursue fish as prey have a more even distribution on the dorsal and ventral sides of their bodies (Bodznick and Boord 1986; Raschi 1986; Tricas 2001). Large pelagic species, such as the manta rays (*Myliobatidae*), have many fewer electroreceptors than shallow water forms and these are limited to small patches on their ventral side (Bodznick and Boord 1986).

Freshwater paddlefishes (*Polyodontidae*) feed on tiny (< 5mm) planktonic crustaceans such as *Daphnia*. Each fish sports a long flat rostrum on its head:

![Figure 2. Photo of paddlefish (*Polyodon spathula*) foraging with mouth open.](image)
Paddle extends forward from head. (Photo by David Alexander.)

This rostrum is covered with up to 75,000 ampullary organs (Wilkens et al. 1997; Wilkens et al. 2001) that allow the fish to detect the passive electric fields of their tiny prey at distances of up to 10 cm and respond by turning appropriately and engulfing the crustaceans in their widespread mouths (Russell et al. 1999; Russell et al. 2001; Wojtenek et al. 2001; Wilkens 2004).

![Figure 3. Distribution of ampullary organs on dorsal side of paddlefish rostrum (paddle).](image)
Rostrum attaches to remainder of head on right of figure. (After Wilkens 2004.)

Sturgeons (*Acipenseridae*), close relatives of the paddlefish, concentrate their ampullary organs and canals on the ventral side of their heads (Teeter et al. 1980; Gibbs and Northcutt 2004). While both freshwater and marine catfish (*Siluriformes*) distribute ampullary-like receptors over their entire bodies (including the caudal fin), they also have the
greatest concentrations on their heads and especially around their eyes (Peters and Meek 1973; Finger 1986; Northcutt et al. 2000). Interestingly, catfish barbels appear to function only as chemoreceptive and tactile organs and completely lack electroreceptors. The other teleost fish with ampullary-like receptors, the knifefish, elephantfish, featherbacks, and frankfish, all concentrate these organs on the head with a gradient of decreasing density as one moves towards the tail (Zakon 1986; Jørgensen 2005).

A foraging platypus uses a similar array of passive electroreceptors on its bill to detect nearby shrimp, leeches, and other freshwater invertebrates (Scheich et al. 1986; Manger and Pettigrew 1995; Pettigrew et al. 1998; Pettigrew 1999). Both marine and freshwater catfish (Siluriformes) combine inputs from ampullary, hydrodynamic, and olfactory organs to detect, track, and capture their prey (Roth 1972; Kalmijn 1974; Finger 1986; Peters et al. 1999; Lorteije et al. 2000; Pohlmann et al. 2004), and North American catfish (Ictaluridae) may use their ampullary-like organs to navigate and identify specific locations using local electrochemical fields (Peters and Bretschneider 1972; Kalmijn 1974; Peters and Vanwijl. F 1974; Finger 1986; Kalmijn 1988). Some freshwater catfish may also use passive electroreception to localize and interact with conspecifics in a manner similar to that of marine stingrays (Peters et al. 2002).

In all of these species, the emphasis on passive electroreceptive organs in only one part of the body is surely a result of the limited range over which these stimuli can be detected. Given that most animals are moving in an anterior direction and that passive electroreception is most widely used to detect prey, concentrating the receptors in the anterior end of the animals would seem the logical strategy.

**Variations in the frequency tuning of receptors**

Another strategy for improving passive electroreception is to tune the maximal responsiveness of the passive electroreceptors to the frequency range of those electric fields that are of greatest interest to the animal. In many cases, the optimal frequencies are those generated by ventilation and locomotion of conspecifics or the animal’s prey. In addition to increasing sensitivity to desired stimuli, narrow frequency tuning also excludes irrelevant electrical fields that otherwise would be detected and constitute noise. Although some animals specialize in detecting relatively invariant electrical fields, even these species are most sensitive to stimuli whose amplitudes are at least slightly time-variant. For example, frequencies as low as 0.03–1 Hz will elicit responses in stingrays and catfish, whereas these same animals will ignore a steady field of similar amplitude if neither they nor the source are moving (Finger 1986; Tricas and New 1998; Peters et al. 1999; Tricas and Sisneros 2004). This preference for slightly varying fields is probably due to the fact that there is nearly always some relative motion between the sensing animal and the source of the electrical field. This will transform what the animal perceives from a steady field into a slowly varying one (Hofmann and Wilkens 2005). Because sharks routinely sway their heads left and right as a result of their swimming style, they experience a steady electric field as one that varies in amplitude at a rate of about 1 Hz (Kalmijn 1988). The ampullary receptors of sharks thus require stimuli that vary temporally by at least 1–2 Hz. Rays and skates also favor low frequency stimuli in the range of 0.5–2 Hz (Tricas and Sisneros 2004). Other taxa tend to focus on higher frequencies for passive electroreception: optimal frequency ranges are 2–7 Hz for paddlefish, 6–12 Hz for catfish, 30–50 Hz for knifefish and elephantfish, and 50–100 Hz for the platypus (Zakon 1988; Wilkens 2004; Bodznick and Montgomery 2005).

**Acknowledgements**

We thank David Alexander (http://www.flickr.com/photos/aworldoffish/) for use of the paddlefish photo.

**Literature cited**


Northcutt, R.G., P.H. Holmes and J.S. Albert. 2000. Distribution and innervation of lateral line organs in the channel
catfish. *Journal of Comparative Neurology* 421: 570–592.


Web Topic 8.1
Computing Red Lines

The payoff matrix

Consider a thirsty spider monkey that has to choose between two possible actions it will perform next: descend from its tree to drink, or stay in the canopy for now. The consequences of each action depend on which of two alternative conditions is currently the case: either there is a predator, such as a jaguar or ocelot, lurking in the bushes at the foot of the tree, or there is not. The monkey therefore must consider four possible outcomes (or payoffs) of her decision. We can organize these alternative outcomes into a payoff matrix as follows:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predator nearby</td>
<td>Stay in tree</td>
</tr>
<tr>
<td></td>
<td>Descend and drink</td>
</tr>
<tr>
<td>No predator nearby</td>
<td></td>
</tr>
</tbody>
</table>

Suppose we can quantify the four possible consequences using some common currency. Ignoring for the moment what currency might work for all consequence types, let us assume that the larger the payoff in this currency, the better off the monkey will be. Now suppose that we make the relevant measurements in the field and obtain the following table for this monkey:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predator nearby</td>
<td>Stay in tree</td>
</tr>
<tr>
<td></td>
<td>Descend and drink</td>
</tr>
<tr>
<td>No predator nearby</td>
<td></td>
</tr>
</tbody>
</table>

Clearly, the best consequence for the monkey is to descend and drink when no predator is present. However, if it stays in the canopy, even if no predator is present, it still does fairly well. If there is a predator present and it descends, it has a very high chance of being eaten and thus its payoff for this situation is the lowest. If it stays in the tree and there is a predator present, it avoids being eaten for now, but it also is vulnerable to being tracked by the predator when it moves through the canopy with a risk that it will be eaten at some later point. In this example, the right choice for the monkey when a predator is present is to stay in the tree, and the right choice when no predator is present is to descend and drink. The problem facing the monkey is that it does not know which condition is currently the case: is there a predator present or not?

Conditional payoffs

While the monkey is not sure whether or not predators are lurking nearby, she can usually come up with some general estimate of the probability that a predator is nearby based on prior experiences. Let us suppose that in the past, predators have turned out to be present on 20% of such occasions, or writing the probability as a fraction, 0.20, and thus the probability that no predator is nearby is $1.00 - 0.20 = 0.80$. If the monkey stays in the canopy and a predator is present, her payoff is 15; if she stays in the canopy and no predator is present, her payoff is 25. The average payoff (called “expected value” in economics) of staying in the canopy is the sum of each of these payoffs discounted (multiplied) by the probability of the relevant situation. In this case, her estimate of the average payoff of staying in the canopy is:
Such a calculation is called a **conditional payoff** because, in fact, no monkey will get a payoff of 23 if it stays in the canopy: instead, the payoff will be conditional upon whether a predator is present (when it will get 15) or not (when it will get 25). However, the average payoff of 23 is the best overall guess for what staying in the canopy will give the monkey since it cannot know for sure before it decides which condition will turn out to be true.

We can now calculate the average payoff of descending to drink when there is a 20% chance that a predator is nearby:

\[
PO_{\text{descend}} = (0.20)(3) + (0.80)(35) = 28.6
\]

We can see that when we compare the conditional payoffs for staying in the canopy versus descending to drink when the probability of a nearby predator is 20%, the monkey will do better (on average) if it descends and drinks (28.6 versus 23.0). This makes sense in that if there is a low chance of a predator lurking in the bushes, it is probably better to take a chance and descend.

But suppose the monkey thinks that there is a 70% chance that a predator is lurking nearby. What is the best average strategy then? We again compute the conditional payoff of staying in the canopy to get:

\[
PO_{\text{stay}} = (0.70)(15) + (0.30)(25) = 18.0,
\]

and the corresponding conditional payoff of descending to drink to get:

\[
PO_{\text{descend}} = (0.70)(3) + (0.30)(35) = 12.6.
\]

We see that staying in the canopy is now the better action. Again, this makes sense: if there is a good chance of a predator nearby, it is better stay in the canopy and live to drink another day.

These calculations suggest that there is some intermediate probability of a predator being present at which the optimal action switches between staying in the canopy versus descending to drink. That switch-point value should be somewhere between 20% (where descending was optimal) and 70% (where staying in the canopy was optimal). At the switch-point probability, it should not matter whether the monkey stays in the canopy or descends: it can expect to get the same conditional payoff. For probabilities less than the switch-point probability, the monkey should descend and drink, and at probabilities above the switch-point, it should stay in the canopy. The switch-point is the **red line** that we are seeking. How can we compute it?

### Computing the red line

Let us denote the monkey’s current probability estimate that a predator is nearby by \( P \), and the switch-point probability at which staying in the canopy and descending to drink give the same payoff by \( P_s \). Substituting \( P_s \) into our conditional probability computations as before, and setting the two conditional payoffs equal to each other, we get the following equation:
\[ \overline{PO}_{\text{stay}} = (P_s)(15) + (1.0 - P_s)(25) = \overline{PO}_{\text{descend}} = (P_s)(3) + (1.0 - P_s)(35). \]

Solving for \( P_s \), we get \( P_s = 0.45 \). Whenever the monkey estimates that \( P < 0.45 \), it should descend and drink; when it estimates that \( P > 0.45 \), it should stay in the canopy. When \( P = 0.45 \), it can do either and expect the same average payoff.

There is a quick and dirty way to estimate \( P_s \). If we define \( P \) as the estimated probability that the left-hand situation in a 2 \times 2 payoff matrix is true, and \((1 - P)\) as the estimated probability that the other situation is true, then \( P_s \) can be computed as the difference in payoffs between getting it right versus wrong in the right column divided by the sum of the differences between getting it right versus wrong in both columns. In this case, we would find (again):

\[
P_s = \frac{(35 - 25)}{(35 - 25) + (15 - 3)} = 0.45
\]

**Payoff differences versus absolute values**

The computational shortcut for \( P_s \) demonstrates an important point: where the red line is drawn on a decision maker’s meter depends on the relative differences in payoffs between right and wrong decisions in the two columns, and not on the absolute values of each payoff. Doubling the values of each cell in the matrix will not change the location of the red line. Another way to look at the red line is to consider the difference in payoffs in any column as the cost of errors when the condition for that column is the current one. As errors in the right-hand column become more costly relative to those in the left column, the red line probability moves to higher values; as the costs in the left column increase relative to those in the right column, the red line moves to lower values. This simplifies the task for a decision maker: it only needs to estimate the relative differences in payoffs between right and wrong decisions to know where to set its red line. Then all it needs to do is compare its current estimate of the probability that a predator is present to the red line.
Both male and female northern cardinals (*Cardinalis cardinalis*) can sing, although males do most of the loud advertisement singing. Suppose we are interested in estimates of the probabilities that a male, his mate, or both will sing in response to a song recorded from another male and played back just inside the territory of the pair. We find 10 different cardinal pairs, and we perform the experiment once with each. We get the following table summarizing our results by putting an “X” in a cell if the individual identified in that row sang in the trial identified in that column:

<table>
<thead>
<tr>
<th>Trial</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male sings</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Female sings</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on this (small) sample, we can estimate the probabilities that each sex will sing in reply to a playback by counting the number of times a member of that sex sang and dividing it by the total number of trials in which that sex could have sung. Letting $P(M)$ be the estimated probability that a male will sing and $P(F)$ be the estimated probability a female will sing, we get:

$$P(M) = \frac{6}{10} = 0.60,$$

$$P(F) = \frac{4}{10} = 0.40$$

**Combinatorial probabilities**

The same table allows us to compute some probabilities involving both members of a pair. For example, the probability that either the male or the female sings on an average trial, $P(M$ or $F)$ is computed by counting the number of times at least one of them sang (in our case, 8 times) and dividing by the total trials (here, 10 trials):

$$P(M$ or $F) = \frac{8}{10} = 0.80$$

Alternatively, we might want to know the probability that both members of a pair will sing when stimulated with playback, $P(M$ and $F)$. We thus count the number of trials in which both birds sang (in the table, we see this is only 2 times) and divide by the number of trials (again, 10 total trials):

$$P(M$ and $F) = \frac{2}{10} = 0.20$$

**Conditional probabilities**

The prior probabilities were based on the entire suite of trials. However, sometimes we are interested in subsets of the sample. For example, we might want to know the probability that the female will sing given that the male also sang, $P(F|M)$. Here the notation uses a “|” to separate the events of interest (females singing) and the relevant subset of the sample (males singing). We compute $P(F|M)$ by counting how many times males sang (here 6 trials), and counting in how many of those 6 trials the female also sang (2). We then compute

$$P(F|M) = \frac{2}{6} = 0.33$$
We can also compute the conditional probability that a male would sing given that a female sang. By examining the table, we see that this is:

\[ P(M|F) = \frac{2}{4} = 0.50 \]

We observe that \( P(F|M) \) and \( P(M|F) \) are not necessarily equal. In fact, they are often different values.

**Relations between probability types**

It is easy to show that these various probabilities have simple relationships to each other. One important relationship is the following:

\[ P(M \text{ or } F) = P(M) + P(F) - P(M \text{ and } F) \]

This says that the fraction of occasions on which at least one of the members of the pair sang is the fraction of trials in which males sang plus the fraction of trials in which females sang minus the fraction of trials in which both sang. The latter term corrects for the fact that the sum of the number of times that males sang and that females sang will count the number of occasions on which both occurred twice. Since we should count these events only once, we need to subtract out the number of times both occurred from either the male sum or the female sum. We saw earlier, in \( P(M \text{ or } F) = 0.8 \). We can get that same number using the equation above as:

\[ P(M \text{ or } F) = P(M) + P(F) - P(M \text{ and } F) = 0.6 + 0.4 - 0.2 = 0.8 \]

Note that if, for some reason, males and females never sang in the same trial—that is, if male and female songs were exclusive events—then \( P(M \text{ and } F) = 0 \) and \( P(M \text{ or } F) = P(M) + P(F) \). This last form of the equation is very commonly used in computing payoffs because many events are in fact exclusive. This equation is sometimes called the “**OR Rule**”: the probability that one or the other of several exclusive events will occur is simply the sum of the probabilities of the individual events.

In those cases where events *can* occur jointly, we may invoke a second equation:

\[ P(M \text{ and } F) = P(M) \cdot P(F|M) = P(F) \cdot P(M|F) \]

This also makes intuitive sense: the fraction of times that both the male and the female sing, \( P(M \text{ and } F) \), cannot be greater than the total fraction of times that the male sings, \( P(M) \). \( P(F|M) \) is the fraction of those times that a male did sing in which the female also sang. The product of the fraction of trials in which a male sang and the fraction of those trials in which the female also sang is clearly the fraction of trials that both males and females sang. The same logic can be applied if we start with the fraction of trials in which females sang, \( P(F) \), and multiply this by the conditional probability that a male will sing, given that the female did, \( P(M|F) \).

Earlier, we used the table to show that \( P(M \text{ and } F) = 0.2 \). We can get the same answer using the second equation as follows:

\[ P(M) \times P(F|M) = (0.6)(0.33) = 0.2, \text{ and} \]
\[ P(F) \times P(M|F) = (0.4)(0.50) = 0.2 \]

As with the “**OR Rule**,” the second equation has a simpler form in special cases. What if we find that \( P(M|F) = P(M) \)? This means that the probability that a male will sing is the same for the full 10 trials as it is for the 6 trials in which females also sang. Put another way, what if the presence of female song has no effect on the probability of males singing? It is easy to show that if \( P(M|F) = P(M) \) then it also has to be true that \( P(F|M) = P(F) \). When these conditions are met, we say that male and female songs are **stochastically independent**: the occurrence of one does not alter the likelihood of the other. This allows us to reduce the second equation to the following “**AND Rule**”: if two events are stochastically independent and can co-occur, then
To use these basic definitions and relations, just substitute the events in question for the responses, $M$ and $F$, in these examples.

**Utility of the rules**

These simple probability rules are used repeatedly in this book. The “backbone” for computing conditional payoffs is the relevant probability rule with each component probability weighting (discounting) the relevant payoff value. Similarly, updating using Bayes Rule (see Web Topic 8.3) also relies on these simple rules. It pays to become familiar with them!
Web Topic 8.3
Bayesian Updating

The task

A female songbird is searching for a healthy mate. She has learned that, on average, about 60% of males are healthy and 40% are sick with parasites. Her prior probabilities are thus \( P(\text{Healthy}) = 0.60 \) and \( P(\text{Sick}) = 0.40 \). The female has also learned that the speed of courtship songs is often an index of the health of the singer: healthy males sing fast songs about 80% of the time, and sing slow songs the remaining 20%. In contrast, sick males sing fast songs only 30% of the time, and sing slow songs 70% of the time. She assembles these conditional probabilities into a coding matrix that appears as follows:

<table>
<thead>
<tr>
<th>Male Health</th>
<th>Male Song</th>
<th>Healthy</th>
<th>Sick</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast</td>
<td>0.80</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>Slow</td>
<td>0.20</td>
<td>0.30</td>
<td></td>
</tr>
</tbody>
</table>

She then encounters a male who sings a Fast song. What is her best estimate of the a posteriori probability, \( P(\text{Healthy}|\text{Fast Song}) \), that this male is healthy?

The method

Bayes' Theorem provides a way to compute an updated (a posteriori) probability given (a) the prior probabilities, (b) detection of a signal or cue, say \( S_1 \), and, (c) knowledge of the conditional probabilities (coding matrix) relating signals to alternative conditions. Given two possible alternative conditions, \( A \) and \( B \), prior probabilities \( P(A) \) and \( P(B) \) respectively, and a coding matrix of the form

| Condition | Signal | A \( P(S_1|A) \) | B \( P(S_1|B) \) |
|-----------|--------|-----------------|-----------------|
| \( S_1 \) |        | \( P(S_1|A) \)  | \( P(S_1|B) \)  |
| \( S_2 \) |        | \( P(S_2|A) \)  | \( P(S_2|B) \)  |

then detection of a signal \( S_1 \) allows the computation of the a posteriori probability

\[
P(A|S_1) = \frac{P(A)P(S_1|A)}{P(A)P(S_1|A) + P(B)P(S_1|B)}
\]

Without access to additional information, the probability on the left side of this equation is the best possible estimate that can be computed. It is the Bayesian estimate.

If the signal detected had been an \( S_2 \), then the Bayesian estimate would be

\[
P(A|S_2) = \frac{P(A)P(S_2|A)}{P(A)P(S_2|A) + P(B)P(S_2|B)}
\]

When there are more than two alternative conditions, the method is the same except that there will be one term in the denominator for each possible condition.
Updating using probabilities

Applying this method to the situation facing the female bird who just heard a candidate male sing a fast song, we can substitute her prior and conditional probabilities into the first Bayesian equation to get

\[
P(\text{Healthy} \mid \text{FastSong}) = \frac{(0.60)(0.80)}{(0.60)(0.80) + (0.40)(0.30)} = 0.800
\]

If, instead, the female heard a slow song, her Bayesian estimate would be

\[
P(\text{Healthy} \mid \text{SlowSong}) = \frac{(0.60)(0.20)}{(0.60)(0.20) + (0.40)(0.70)} = 0.300
\]

Updating using frequencies

Gigerenzer and Hoffrage (1995) have pointed out that people are more likely to use frequencies (whole numbers) than probabilities (fractions) in their updating. To see how they would do this, note that the numerator on the right side of the equation (above) for \(P(\text{Healthy} \mid \text{FastSong})\) is the fraction of the population that is both healthy and singing fast songs. The denominator includes this same number and adds to it the fraction of the population that is sick but also singing fast songs. Thus the denominator of the Bayesian equation is simply the total fraction of the population of males that are singing fast songs at any given moment. The Bayesian a posteriori probability is then the fraction of males who are singing fast songs that are also healthy.

Now, simply replace the fractions in the original Bayesian equation with the actual numbers of males in each case. This is equivalent to multiplying each term in the Bayesian equation by total number of males available. For example, suppose there were 100 males that the female was likely to encounter. Of these, 60 would be healthy and 40 would be sick. Of the 60 healthy males, 80% or 48 would be singing fast songs at any given time. Similarly, of the 40 sick males, 30% or 12 would be singing fast songs at a given moment. The total number of males singing fast songs will thus be 48 + 12 = 60. The Bayesian estimated a posteriori probability that a male who sings a fast song is healthy is then just 48/60 = 0.80, the same value we got using probabilities.

Cumulative updating

How does a female update her estimate that a given male is healthy after hearing him sing several successive songs? The Bayesian process is the same except that after the second song, she should replace the prior probabilities that she used in the first computation with the new a posteriori probabilities obtained after updating from the first song. For example, we saw that if the male’s first song was fast, her estimated probability that he is healthy would increase from her initial prior value of 60% to 80%. Suppose his second song is also fast. She should compute her second update as follows:

\[
P(\text{Healthy} \mid \text{FastSong}) = \frac{(0.80)(0.80)}{(0.80)(0.80) + (0.20)(0.30)} = 0.914
\]

The conditional probabilities remain the same, but the prior probabilities change in both the numerator and the denominator. Suppose the third song that he sings is slow. She would then update again to give a new estimated
probability that he is healthy of

\[
P(\text{Healthy} \mid \text{Slow Song}) = \frac{(0.914)(0.20)}{(0.914)(0.20) + (0.086)(0.70)} = 0.752
\]

This successive process can continue as long as the same male sings songs and the female has the patience to listen and do the relevant updating. **Figure 1** shows the average trajectory of the female’s estimated probability that the male is healthy for each kind of male:

![Figure 1: Cumulative Bayesian estimates of male health given songs heard. Results are based on computer simulations of a female using the initial prior probabilities and coding matrix outlined in the text above. Any one simulation would show a jagged approach to probabilities of 1.0 (if the male were healthy) or 0 (if he were sick). Dots show mean values of 1000 simulations and error bars show range of variation around those means. In both cases, cumulative sampling eventually asymptotes to the axis representing the true state of the singing male.](image)

**Sender versus receiver errors**

The examples above assume that receivers always identify correctly whether a male’s song was fast or slow: the only “errors” in the system arise because healthy males sometimes sing slow songs and sick males sometimes sing fast songs. More realistically, the signal system will exhibit errors by both parties. Do receiver errors have a similar impact on the trajectory of successive sampling as sender errors? The answer is yes: similar increases in errors by either party have similar overall effects on successive sampling. To see this, the coding matrix for senders can be combined with that for correct identification of which signal was sent by females into an overall coding matrix (see Web Topic 8.8 for matrix details). We plot the average trajectories for 1000 computer simulations but with differing errors for senders and receivers in **Figure 2**.
Figure 2: Effects of sender versus receiver errors on updated probabilities. Each point is the mean of 1000 random simulations of a receiver using cumulative Bayesian updating to estimate the probability that a singing healthy male is healthy. Filled circles indicate sender and receiver coding matrices that are both 90% consistent (e.g., healthy males sing fast songs 90% of the time and only err 10% of the time; receivers correctly identify a song as fast 90% of the time and only misclassify a song 10% of the time). Open circles indicate males that sing the appropriate song for their health only 70% of the time but receivers that correctly classify songs by speed 90% of the time. Open triangles indicate males that sing the appropriate song for their health 90% of the time but females that classify song speeds correctly only 70% of the time. Filled squares indicate males that sing appropriate songs only 70% of the time and females that classify songs by speed only 70% of the time. As both parties make more errors, trajectories rise to asymptote much more slowly. An increase in error by one party while the other remains unchanged has the same effect regardless of which party experiences the change in error rates.

References Cited


Web Topic 8.4
Signal Detection Theory

Discrete versus overlapping signals

When a receiver detects a signal stimulus, the first task is to assign it to one of several possible categories. Once this assignment has been made, a receiver can use the prior probabilities and the relevant signal coding matrix to update the probabilities of alternative conditions being true, compare expected values of alternative actions, and make a decision on how to respond to this signal.

If the patterns in the signal are completely non-overlapping with those of alternative signals, we say that the signals are discrete. We assumed that signals were in fact discrete in outlining the red line decision process in Web Topic 8.1. But what if signals are not discrete, either because senders emit signals with at least partially overlapping patterns, or because initially discrete signals become distorted and more overlapping during propagation between the sender and the receiver? Can a receiver faced with overlapping signals still define an optimal red line to use in decision making?

An example

The answer is yes, and the method that describes this process is called signal detection theory. To see how a red line can be defined with overlapping signals, consider a female bird trying to assess the health of a potential mate by listening to his courtship song. Suppose that in this species, sick and parasitized males tend to sing slower songs, and healthy males tend to sing faster songs. However, the signals are not discrete and there is considerable overlap in song speeds between males in different states of health:

This plot shows the conditional probability that a male will sing a song at a given speed \( w \) depending on his health. There are two distributions shown: one for sick males, \( P(\text{w|Sick}) \), and one for healthy males, \( P(\text{w|Well}) \). We assume here that the two distributions are normal (bell-shaped), but that is for computational convenience and the general conclusions below do not depend on that assumption.

Red lines and types of errors

Suppose the female draws a red line on this plot: whenever she hears a song at a speed lower than the red line value,
she will reject the male; whenever she hears a song at a speed higher than the red line value, she will accept that male as a mate. Where is the optimal place to draw this line?

Looking at the same graph with a red line on it, we can see that the red line divides each of the two song speed distributions into two parts. The area under the sick male distribution bounded on the right by the red line and on the bottom by the X axis is the total probability that the female will correctly reject a male when he is in fact sick. We shall denote this probability by \( P(\text{Correct Rejection}) \). Similarly, the area under the well male distribution that is bounded on the left by the red line and on the bottom by the X axis is the total probability that a female will correctly accept a male when he is healthy. We shall denote this by \( P(\text{Hit}) \).

With the line in this location, the female cannot avoid making two kinds of errors. The area in the dark blue region to the right of the red line defines the overall probability that the female will erroneously accept a sick male as a mate. This type of error is denoted as \( P(\text{False Alarm}) \). The area of the dark red region to the left of the red line defines the overall probability that the female will erroneously reject a well male and is denoted by \( P(\text{Miss}) \).

It should be obvious by looking at this graph that moving the red line to the right will reduce \( P(\text{false alarm}) \) but it will increase \( P(\text{Miss}) \). Similarly, moving the red line to the left will reduce \( P(\text{Miss}) \) but increase \( P(\text{False Alarm}) \). Since the female cannot reduce the total probability of making some errors, the only way to find an optimal location for the red line is by minimizing the costs of the errors. For example, if false alarms are more costly than misses, then the optimal location for the red line will be at faster song speeds; if misses are more costly than false alarms, then she should set the red line at a lower song speed. To find the optimal location, we therefore need to consider the relative payoffs of each outcome, and her estimated probabilities that a given male is sick or healthy after hearing him sing.

### Fitting the red line to payoffs and probabilities

Suppose that the payoffs to a female of accepting or rejecting well versus sick males can be summarized in the following payoff matrix:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Action</th>
<th>Well Male</th>
<th>Sick Male</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Action</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reject Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accept Male</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Suppose that on average, a fraction $P$ of the males in the population are well and $(1-P)$ are sick. When the female hears a male sing at song speed $w$, she will update her estimate that he is healthy from $P$ to $P(\text{well}|w)$ and that he is sick from $(1-P)$ to $P(\text{sick}|w)$. She can now combine these updated probabilities with the relevant payoffs to compute expected values (average payoffs) for each action. The expected value for accepting this male as a mate will be

$$\bar{PO} (\text{Accept}) = P(\text{well}|w)(R_{11}) + P(\text{sick}|w)(R_{12})$$

the expected value for rejecting this male will be

$$\bar{PO} (\text{Reject}) = P(\text{well}|w)(R_{21}) + P(\text{sick}|w)(R_{22})$$

The optimal redline will occur at that $w$ for which the expected value of accepting a male is equal to that for rejecting him; at higher song speeds, the female should accept males, and at slower song speeds, she should reject males. Setting the two expected values equal to each other and rearranging, we get that the critical song speed, $w_c$, is that for which

$$\frac{P(\text{well}|w_c)}{P(\text{sick}|w_c)} = \frac{(R_{22} - R_{12})}{(R_{11} - R_{21})}$$

We can simplify this further by assuming that the female used Bayesian methods (Web Topic 8.3) to update the probabilities that the male was well after hearing him sing. Specifically, she could have updated using the following formula:

$$P(\text{well}|w) = \frac{P \times P(w|\text{well})}{P \times P(w|\text{well}) + (1-P) \times P(w|\text{sick})}$$

Plugging the right hand side of the Bayesian equation into the left side of the previous equation and rearranging, we get that the critical song speed, $w_c$, is the one for which

$$\frac{P(w|\text{well})}{P(w|\text{sick})} = \frac{(1-P) \times (R_{22} - R_{12})}{P \times (R_{11} - R_{21})}$$

The left side of this equation is simply the ratio of the Y axis values for the well versus sick distributions at $w_c$. It is called the likelihood ratio and is usually denoted by $b$. If we increase $w_c$, the likelihood ratio becomes larger since we move more into the well distribution and out of the sick distribution:
The right side of this equation includes the ratio of the prior probabilities (odds ratio) and the ratio of the differences in payoffs between right and wrong choices in the two conditions (payoff ratio). The entire right side is called the operating level in signal detection theory. All of these numbers are fixed before the male begins to sing or the female begins to make a decision. One can also think of the operating level as the ratio of the costs of the two types of errors noted earlier, each discounted by the prior probability that it will occur. As sick males become more common ((1−P) increases), or the cost of false alarms increases (R12 versus R22), the numerator of the operating ratio increases relative to the denominator, and the appropriate value of wc on the left side of the equation has to increase. If healthy males become more common (P increases), and/or the relative cost of misses increases (R21 versus R11), the right side of the equation decreases, and the optimal location for the red line, wc, moves to lower song speeds.

Discrete versus overlapping signals

The strategies for drawing red lines on meters when signals are discrete (Web Topic 8.1), and on pattern axes when signals overlap (this Web Topic), use the same ingredients. Both approaches depend on the differences in the payoffs of right versus wrong decisions, and not on the absolute values of individual payoffs. Both approaches require access to the signal coding scheme and the prior probabilities. Both invoke Bayesian updating upon receipt of a signal to define the optimal probability estimates before computing expected values of alternative actions. And both permit shortcuts to decision making if prior probabilities and payoff differences remain sufficiently stable for reasonable periods. In the case of overlapping signals, a receiver need only compare perceived signal properties to threshold values to make a quick decision. They do not even have to compute a Bayesian update since the ingredients for that update are incorporated into the determination of the optimal red line.

Further reading:


Web Topic 8.5
Prospect Theory

Background

While both animals and people often make decisions that are consistent with comparisons between expected values of alternative actions, there are also many exceptions that do not fit the classical model of decision making (Real 1996; Kahneman and Tversky 2000). Several common deviations from classical predictions recur commonly enough in human economics and animal decision making that they have received special attention. They have all been called “paradoxes” because they represent behaviors contrary to what a rational person (or animal) should do. Two of the most widely cited examples are:

The St. Petersburg Paradox

Consider a game in which a coin is tossed. If it comes up heads, you win $2 and the game is over. If it comes up tails, you get a second toss. If this second toss comes up heads, you get $4 and the game ends, and if it is tails, you get another toss, and so on. At each stage, the payoff doubles the previous value. What is the expected value of this game? The probability of getting heads on the first toss is 0.5 and the payoff is $2. The discounted payoff if the game ends at this first stage is thus (0.5)($2) = $1. The probability of getting tails on the first toss and heads on the second toss is (0.5)(0.5) = 0.25, and the payoff at this stage is $4. The discounted payoff for a game that ends at the second stage is thus (0.25)($4) = $1. At each possible end point in this game, the discounted payoff will be again $1 since the probability of getting to the next stage is halved at each step, whereas the payoff is doubled. Since the game could go on forever, the expected value for the game is $1 + $1 + $1 +……= infinity. The classical prediction is that someone given a chance to play this game for a fee should agree to pay any amount since the average outcome is an infinitely large number. However, in practice, people are only willing to pay small amounts to play this game, and are thus acting in a risk averse manner. This is known in the economics literature as the St. Petersburg Paradox.

The Allais Paradox

The St. Petersburg Paradox suggests that people generally avoid risky situations. However, it turns out that they are not consistent in this regard. For example, suppose that you are asked to make two decisions. For the first decision, you are invited to pick one of two lotteries to enter. A ticket to Lottery A has a 100% chance of paying you $30, and a ticket to Lottery B has an 80% chance of paying you $40 and nothing otherwise. If you are like most people, you will pick Lottery A over Lottery B. In this decision, you will favor the less risky option (100% vs. 80%).

In the second decision, you are invited to enter Lottery C, in which a ticket has a 25% chance of winning $30 and otherwise pays nothing, or Lottery D, in which a ticket has a 20% chance of paying $40. If you are like most people, you will this time select Lottery D. You will thus choose the more risky option (20% vs. 25%). Since the first choice in each lottery is 1.25 times more likely to pay out winnings than the second (1.00/0.8 = 0.25/0.20 = 1.25), the relative odds of winning are the same in the two decisions. The payoffs are also the same. Despite these similarities, people are routinely risk averse in the first decision and risk prone in the second.

The Four-Way Table

Inspired by the Allais Paradox, researchers again presented subjects with a choice between a non-risky lottery (Lottery A) in which there was a 100% probability of winning $X, and a risky lottery (Lottery B) in which there was a probability $P less than 100% of winning $Y and otherwise winning nothing. They then varied both $P and $Y systematically to produce different combinations of probability and payoff. For each choice situation, the researchers asked the subjects how large $X would have to be before the subjects felt equally inclined to choose Lottery A or Lottery B. The researchers were thus asking the subjects to indicate their subjective evaluations of risky gambles relative to sure bets. If the values of $X equaled the expected values of Lottery B, then the subjects would be considered risk insensitive:
the fact that the second lottery was risky had no effect on their decision and they would be acting rationally. If the value of $X$ was less than the expected value of Lottery B, then they would be risk averse, since they would prefer a guaranteed lower amount than a larger average amount in a gamble. If the value of $X$ were larger than the expected value of Lottery B, then they would treat this choice in a risk prone manner since they overvalued the more risky option.

The results of such studies are very interesting and can be summarized in the following table (from Tversky and Kahneman 1992):

<table>
<thead>
<tr>
<th>Probabilities</th>
<th>Gains</th>
<th>Losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Lottery B: 5% chance of winning $100; otherwise no gain or loss Expected value = (0.05)($100) = $5</td>
<td>Lottery B: 5% chance of losing $100; otherwise no gain or loss Expected value = (0.05)(−$100) = −$5</td>
</tr>
<tr>
<td></td>
<td>Lottery A Equivalent ($X$): $14</td>
<td>Lottery A Equivalent ($X$): −$8</td>
</tr>
<tr>
<td></td>
<td>Bias: Risk Prone</td>
<td>Bias: Risk Averse</td>
</tr>
<tr>
<td>High</td>
<td>Lottery B: 95% chance of winning $100; otherwise no gain or loss Expected value = (0.95)($100) = $95</td>
<td>Lottery B: 95% chance of losing $100; otherwise no gain or loss Expected value = (0.95)(−$100) = −$95</td>
</tr>
<tr>
<td></td>
<td>Lottery A Equivalent ($X$): $78</td>
<td>Lottery A Equivalent ($X$): −$84</td>
</tr>
<tr>
<td></td>
<td>Bias: Risk Averse</td>
<td>Bias: Risk Prone</td>
</tr>
</tbody>
</table>

These examples suggest that people are risk prone when faced with gains of low probability or losses of high probability; they are risk averse when faced with gains of high probability, and losses of low probability. Clearly, these inconsistencies depend both on the value of the payoffs and on the absolute values of their probabilities.

**The theory**

Prospect Theory was proposed by psychologist Daniel Kahneman and cognitive scientist Amos Tversky in 1979 (Kahneman and Tversky 1979) and later extended as cumulative prospect theory (Tversky and Kahneman 1992) to reconcile the actual decision making of people with the classical economic theories based on comparisons of expected values. This theory basically replaces both the probabilities and the payoffs in classical theory with nonlinear transformations of each. Specifically:

**Rescaling payoffs**

The St. Petersburg Paradox can be understood if people (and animals) do not value a given payoff in an absolute sense, but instead compare it to their current state and needs. A hungry animal might value a given item of food much more highly than a well-fed animal. Bernoulli had suggested in 1738 that payoffs should be transformed into a new variable, which he called utility, that better reflected the value of a given payoff to a given decision maker. The transformation of payoffs into utility can follow any of several types of functions. If an animal is already well off, access to further payoffs is likely to have a decelerating (concave) relationship with utility:
Bernoulli pointed out that if one replaces each of the $1 payoffs in the St. Petersburg with utility values based on a decelerating curve like the one above, the expected value of the game is now at most several dollars, and thus fits the amount of money that most people will pay to participate.

Conversely, if a decision maker were desperate, a small payoff might not help it much, but a larger payoff could greatly increase its condition. The appropriate transformation would then be an accelerating (convex) function like the following:

Propect theory combines these two possible transformations into a single sigmoid curve centered around a reference point which we denote as a payoff of 0:
A decision maker starts with some existing state defined by its position on the X axis of this graph. The change in utility that is generated by access to a new payoff is computed by moving the appropriate direction and distance along the X axis from the starting point and then comparing the new utility value at that new location to the initial one. We can see that if the decision maker starts with a payoff above the reference level, each incremental positive payoff results in a decelerating change in its utility. If it starts out below the reference point, any positive increase in payoffs results in an accelerating change in utility. The process of imposing a reference point at the payoff where accelerated curves shift to decelerated ones is called **framing**. If the reference point is in fact the status quo, then any additional payoff that moves a decision maker to a higher utility level is considered a **gain**, and any payoff that leaves the decision maker at a lower utility is considered a **loss**.

It is easy to show that risky outcomes produce smaller **average increases** in utility than non-risky ones when a decision maker begins to the right of the reference point (Smallwood 1996). Such individuals should thus be **risk averse**. Conversely, if one begins to the left of the reference point, risky outcomes produce larger average increases in utility than non-risky ones when the decision maker begins to the left of the reference point. Such individuals should be **risk prone**. The fact that people in general are more strongly risk prone than risk averse has led to a slight reshaping of the typical framing curve so that it looks as follows:

Note that the part of the curve to the right of the reference point asymptotes much sooner and to a lower absolute value.
than does the curve to the left of the reference point. This reflects a greater aversion to loss than to gains in most people. In prospect theory, this asymmetrical curve is called the subjective value function to distinguish it from the earlier symmetrical utility function.

Rescaling probabilities

While rescaling payoffs into subjective values explained some of the biases and deviations in human decision making, it was not enough. Tversky and Kahneman thus proposed rescaling probabilities as well (Tversky and Kahneman 1992; Kahneman and Tversky 2000). There are two possible reference points when dealing with probabilities: certainty ($P = 1.00$) and impossibility ($P = 0.0$). As we saw in the Allais Paradox and the Fourway Table, people tend to overweight low probabilities (which favors risk aversion for losses and risk prone behavior for gains), and underweight high probabilities (which favors risk aversion for gains and risk prone behavior for losses). Research also suggests that any transformation of probabilities should be asymmetrical: an increase in probability from 0.20 to 0.25 is perceived as much less important than a change from 0.95 to 1.00. According to these observed biases, a function that transforms probabilities into a new variable, called weights in prospect theory, is shown below:

Were weights simply proportional to probabilities, the function would lie along the dashed line. The weighting function rises above this line to the left of the graph (reflecting people’s overweighing of low frequencies), and falls well below the line on the right (reflecting underweighting of higher probabilities). The function is also obligingly asymmetrical with a more gentle concave curvature at low probabilities but a very acute convex curvature at higher values.

Combining Values and Weights

To compute an overall expected value for an action using prospect theory, each payoff that could result from that action is transformed into appropriate subjective values and multiplied by the relevant weight (as determined by the prior graph and the probability of that payoff being realized). The discounted subjective values are then summed for all possible outcomes of that action to give a net subjective expectation. Subjective expectations for different alternative actions are then compared to make a decision.

Biology and prospect theory
Prospect theory has been quite successful in predicting human decision making. As a result of his work on this and related theories, Daniel Kahneman received the Nobel Prize for Economics in 2002. Many of the biases that prospect theory seeks to explain in humans also show up in animals (Real 1996). Risk sensitivity is very common in animal decision making, and biases outlined in the Allais Paradox and the Fourway Table have been described in animals (see Chapter 8). The theory thus seems to have generality.

One concern is that prospect theory has largely been derived by fitting arbitrary functions to observed behaviors: it is thus descriptive as opposed to being derived from and clearly linked to other fundamental principles. The subjective value and weighting functions have been tweaked and shifted until they fit the behaviors, and as a result, they provide reasonable predictive power. What they lack is explanatory power. Why is the subjective value function asymmetrical? Why are people and animals more loss aversive than gains sensitive? These are questions that one would like to have anchored in other known processes and principles.

Some of the components of prospect theory can be rooted in biology (Trepel et al. 2005). For example, there are clear physiological reasons why an animal’s conversion of resources into survival and reproduction is likely to involve curved and not linear functions. Curved functions will automatically lead to optimal decision making that is risk sensitive (Smallwood 1996). Risk sensitivity can also be explained, in part, by known nonlinear scaling during sensory detection and perception (see Web Topic 8.6). However, many of the other characteristics of subjective valuation and probability weighting have no clear ties to known biological or physical laws and remain to be explained.

Literature cited


Web Topic 8.6
Weber’s Law and Risk

The problem

Animals (and people) often have to choose between an action that has a sure and known consequence and an alternative action that can lead to any of several alternative consequences. The second option is said to be riskier than the first. Suppose the consequences in each case involve access to some resource such as food. If the differences between the consequences is in the amount of food provided, as opposed to the delay in receiving food after acting, animals (and people) tend to be risk averse: even if the expected value for the riskier action is somewhat larger than the sure bet, they will choose the sure bet. If however, the consequences provide equal amounts of food, but they differ in the delay between the decision and receiving the food, animals (and people) are often risk prone: that is, they favor the riskier option (Kacelnik and Bateson 1996). This is a curious difference that demands an explanation.

Errors in payoff estimation

Optimal decision making requires comparisons between expected values of alternative actions. Expected values depend on the probabilities that different consequences of an action will occur, and the values of the payoffs for each consequence. Because animals cannot know the exact value of a given payoff until it is experienced, decision making relies on estimates of payoffs that are subject to error. These estimates are generated by pooling the information provided by ambient cues, signals from other animals, and prior experience. The result is a probability distribution of different possible values for a payoff. In many cases, this distribution will be bell-shaped (e.g., a normal distribution):

One can identify several key characteristics of such a probability distribution. The possible payoff value that is most likely is called the mode. In this example, that occurs at a payoff value of 30. The average or mean payoff value can be computed by discounting each possible payoff value by its probability and adding these all together (because the possible values are infinite, one should actually use an integral and not a sum here). The median is the payoff at which half of the overall probability is accounted for by payoffs less than it, and half by values greater than it. This particular
probability distribution is symmetrical: as a result, the mean, median, and mode will occur at the same payoff value (30). Finally, we want a measure of the variation we might see if we randomly sampled this distribution a 100 times. It seems intuitive that the wider the bell-shape of the distribution, the larger this measure of variation. A useful measure of variation when a distribution is normal is the **standard deviation**. This is defined as the distance one needs to move away from the mean in either direction to account for 34% of the most likely payoff values. In this example, the standard deviation is 5.

Not all probability distributions are so nicely symmetrical. In nature, distributions are often **positively skewed** such as this example:

![Graph showing skew distribution](image)

With a positive skew in the distribution, the mean and median will occur at a higher payoff value than the mode. While less common in nature, distributions can also be skewed in a negative direction with the mean occurring at a smaller payoff value than the mode.

**Weber’s Law**

Weber’s Law builds on the recognition that animals (and people) estimate and measure quantities with some error. Suppose that a jay is trying to decide which of two peanuts is the heavier one before carrying it off to cache it for the winter. It lifts and shakes each peanut and then makes a decision on which to store. We can experimentally give the jay peanuts that are closer and closer together in weight. Eventually we will get to a point where the jay can still identify the heavier peanut, but giving it peanuts any more similar causes the jay to choose randomly. We have identified the **just noticeable difference** (JND) in peanut weight for these jays. Weber (1834) performed similar tests on people and observed that JND’s got larger as the average measurements on the compared items got larger. In fact, the ratio between the JND and the average magnitude of the two measurements tended to be a constant across a wide range of magnitudes. As a human example, suppose you can just barely identify the heavier of 95 g and 105 g weights. The average weight is 100 g and the JND is 10 g. The ratio between the JND and the average is 10%. According to Weber’s Law, you would also need a 10% difference to identify two weights that averaged 1 kg. However now, the JND would be 10% of 1 kg or 100 g. This is a much larger minimum difference than for the 100 g weights. Put simply, Weber discovered that the perceptual error in measurement increases proportionally with the magnitude of the
measurement. This finding has since been confirmed in many species and in each of the sensory modalities.

Some 26 years later, Fechner (1860) proposed that the fixed ratio between JND and average measurement value was likely a consequence of animals trying to achieve a large dynamic range in their sensory systems. An animal with a large dynamic range can measure very small magnitudes and very large ones. However, the cost of such a broad dynamic range is greater error in measuring larger magnitudes. Fechner suggested that if sensory organs and brains perceived stimulus magnitudes on a logarithmic scale, large dynamic range and the larger measurement errors could both be explained. Stevens (1957) challenged Fechner’s logarithmic scaling and proposed a power function alternative. While the dispute over what kind of scaling is actually used in animal sensory organs and brains continues to this day (Shettleworth 2001; Copelli et al. 2002; Johnson et al. 2002; Dehaene 2003), the original proposition that most sensory systems obey Weber’s Law remains widely accepted.

**Weber’s Law and risk**

All of this means that there are actually two levels at which chance can affect expected values during decisions. The first, which we have dealt with in prior discussions, concerns which of several alternative consequences will occur when an animal chooses a risky action. However, now we see that whether the decision maker selects a “sure bet” option or a risky option, the actual payoff experienced always depends on a random draw from some probability distribution. The difference between sure bets and risky options is that in the former, the draw will be made from a single distribution, whereas the draw for a risky option could be taken from any of several alternative distributions. Another way to look at risky options is to pool the probability distributions for each possible consequence into one single probability distribution. If there are two equally likely consequences for a risky action, the pooled distribution is the sum of the probabilities of the two distributions for each possible payoff value divided by two. If alternative consequences are not equally likely, then the pooled probability for any payoff will be more similar to that of the more likely consequence than to the less likely one.

What does Weber’s Law have to do with decision making? Perhaps quite a bit (Gibson et al. 1988; Reboreda and Kacelnik 1991; Bateson and Kacelnik 1995; Kacelnik and Bateson 1996; Kacelnik and Abreu 1998). Consider a decision maker who has to choose between action A, which will always result in a payoff drawn from distribution X, or action B, which will draw a random payoff from distribution Y 50% of the time or distribution Z the other 50% of the time. Suppose the modes for these three alternative distributions have values such that mode(Y) < mode(X) < mode(Z). According to Weber’s Law, the error in measuring payoffs from distribution Z will be greater than that for the two other distributions because the most likely payoffs in distribution Z are larger than for the other two distributions. The error for distribution Y will be less than for the other distributions by the same logic. Since the total probability in a distribution has to add to one, increasing the error in a distribution will decrease the probability at the mode, and decreasing the error will increase the probability at the mode. This is shown graphically below:
We can see that the probability of drawing the mode value of a distribution decreases with the magnitude of the mode and the resulting error in that distribution.

Now, let us just consider the two possible outcomes for action $B$, the risky action. If the decision maker elects to perform action $B$, it will draw the payoff from either the $Y$ or the $Z$ distribution:

In our particular example, an animal electing to perform action $B$ is equally likely to draw a payoff from distribution $Y$ as it is to draw the payoff from distribution $Z$. We can thus combine these two distributions with equal weighting to get the
pooled payoff distribution for action $B$:

There are several ways that a decision maker could use this information (Kacelnik and Abreu 1998). The means for these two distributions are approximately equal (payoff = 50), so the fact that many animals faced with such a decision favor action $A$ when the issue is the amount of food acquired and action $B$ when the difference is in the delay in reward delivery suggests that they do not rely only on the mean values. The mode for action $B$ (payoff = 10) falls at a
much lower payoff than that for action A (payoff = 50). This would explain the data because animals generally seek to maximize their food intake, but to minimize their delay in getting that food. The same is true for the medians (median for action A = 50 and for action B = 42).

If decisions are preceded by randomly drawing a single sample from each of the two distributions and comparing the results, predictions are a bit more complicated but still possible. Sometimes, by chance, the decision maker will draw a higher payoff from the distribution for action B than for A; other times, it will draw the reverse. The question is then what fraction of the time will it draw a better value for the risky option than for the less risky one. The methods for computing this fraction are given in Kacelnik and Abreu (1998). In the example that we have been considering, action B would provide a larger payoff than action A in 47% of the draws. This means that action A will provide a larger draw more often than will action B. Again, if the choice involves different amounts of food, we would expect such decision makers to choose action A because they maximize food intake and are more likely to draw a larger payoff from the X distribution than from the pooled Y and Z distributions. If the choice involves different delays in obtaining that food, then they would select action B because they are more likely to draw the smaller delay from the pooled Y and Z distributions.

**Supporting data**

Risk sensitive decision making was at first assumed to depend on the absolute variation among the payoffs of a risky choice. This meant that one could use the standard deviation of the relevant distribution, or its square (the variance), as an index of risk. Given two alternative actions each with multiple possible payoffs, the one with the higher variance in payoffs would be avoided by risk averse animals and favored by risk prone ones. However, if Weber’s law is playing an important role in decision making, the best predictor of risk sensitivity would not be the absolute measures of variability but instead the ratio between the error and the average values being measured. One commonly used ratio that fits this description is the **coefficient of variation** (CV). This is computed as the ratio between the standard deviation and the mean of a distribution. If Weber’s Law does contribute to risk sensitivity as we have described earlier, it should be the case that animals (and people) consider a choice with a higher CV as more risky than one with a lower CV, and they should be indifferent to a choice in which both alternatives had the same CV, even though one might have higher absolute variability as measured by variance or standard deviations. This is a prediction that can be studied.

Recent work suggests that, as predicted, the CV of the relevant distributions is a better predictor of risk sensitivity than either standard deviations or variances in a wide variety of animals (Weber et al. 2004; Shafir et al. 2005). It must be remembered, however, that a good fit of data to a particular model is a necessary but not sufficient condition for believing that this model is the true cause of those results. One must also consider alternative hypotheses that can explain the same data, and then identify critical experiments or observations that will discriminate between them (Platt 1964). In this regard, several authors have suggested that similar data could arise from learning processes that do not need to invoke Weber’s Law per se (Lockhead 2004; Weber et al. 2004). As noted by several authors, Weber’s Law, components of Prospect Theory (see Web Module 1.5), Jensen’s inequality, and processes such as associative learning may all play roles in the observed risk sensitivity of animals and people.

**Literature cited**


Web Topic 8.7
Brains and Decision Making

Overview

One of the most exciting developments in the last decade is the increasing ability to monitor the neurobiology of decision makers. In animals, this has traditionally involved inserting electrodes into selected brain areas and monitoring relative activity as the animal makes a decision. Recent technical advances such as two-photon imaging and optogenetics have pushed the envelope even further by identifying events at the individual neuron and sub-neuron levels (Homma et al. 2009; Knopfel et al. 2010). Popular animal systems include roundworms, sea slugs, leeches, honeybees, zebrafish, zebra finches, mice, rats, and monkeys. In people, non-invasive functional magnetic resonance imaging (fMRI) uses the increased metabolic activity of working neurons to track the steps as a person makes different kinds of decisions. These technologies have encouraged neurobiologists to determine whether animals or people really have the necessary machinery for Bayesian updating and optimal decision making, or instead are just relying on some very clever heuristics. They also have looked for possible causes for the biases seen in animal and human decision making. Suitable areas of the brain for these functions have now been located in animals and humans, and this has encouraged a melding of economics and neurobiology into a field called neuroeconomics (Glimcher and Rustichini 2004).

Basic structure of the human brain

The human brain is divided into a series of lobes and internal regions (Figure 1):

Figure 1: (A) The four lobes of the human cerebral cortex and some of their functions. Decision making involves the most anterior (near the face) parts of the frontal lobes, parts of the parietal lobes, and deep inside, parts of the temporal lobes. (B) Top view of human cerebral cortex. Note that the brain is largely divided into left and right hemispheres. Decisions based on more confident probability and payoff estimates tend to activate relevant areas in the left hemisphere; less certain decisions may
activate the corresponding regions but in the right hemisphere.

The large and prominently fissured mass on the top is called the cerebrum or cortex. It is divided at most points into right and left hemispheres. Each hemisphere hosts four regions called lobes. While each lobe is in fact multifunctional, one can roughly assign reasoning and voluntary movements to the frontal lobe, audition and some higher level visual processing to the temporal lobe, touch and skin sensations to the parietal lobe, and vision to the occipital lobe. Deep inside the folds in each hemisphere where the temporal and frontal lobes meet are the insula; these deal with visceral functions and taste. Below the cortex is a complex of regions known as the subcortex. This includes the striatum (also known as the basal ganglia), which consists of two parallel structures that begin in the frontal lobe and arc backwards into each temporal lobe, down, and then forwards again. (Figure 2).

---

**Figure 2:** The basal ganglia (collectively called the striatum) are located in the center of the brain just below the outer cortex. The major components are the caudate nucleus, the putamen, and the globus pallidus. The ventral striatum (an important area that stores utility values for rewards of actions) consists of ventral and medial parts of the caudate nucleus and the putamen.

This area starts, regulates, and stops voluntary motor actions. Interdigitated with the striatum is another set of subcortical structures called the limbic system (Figure 3):
These include the **cingulate cortex** which is a large band above the striatum, and the **amygdala** and **hippocampus** zones in the lower parts of each temporal lobe. The limbic system controls emotion (cingulate and amygdala) and regulates what gets stored as memories in the brain (hippocampus). The cerebrum connects to the spinal cord successively through the **diencephalon**, the **midbrain**, and the **brainstem**, all of which handle switching and routing functions for nerve traffic coming into and out of the brain. The lower part of the diencephalon, the **hypothalamus**, regulates a number of autonomic functions such as body temperature, hunger, thirst, reproduction, and circadian rhythms. The brainstem hosts two systems that modulate activity throughout the brain using specific transmitter substances. One, the **dopamine system**, plays an important role in assigning reward values to recent stimuli. The second, the **norepinephrine system**, modulates mental arousal and vigilance. The **cerebellum**, which controls posture, coordination, and balance, is nestled between the cortex and the midbrain at the rear of the brain.

**Rational decision making in human brains**
Economists and psychologists had speculated that the brain might have two separate decision-making systems: one, that was fast and heuristic, and a second, that was slower but more rational (Kahneman and Tversky 2000; Camerer et al. 2005). The reality has turned out to be a bit more complicated (McClure et al. 2004; Glimcher et al. 2005; Sugrue et al. 2005; Trepel et al. 2005; Sanfey et al. 2006). Some parts of the brain, such as the prefrontal areas of the frontal lobes, seem to be active during any decision process. There are three (at least) sub-regions within the prefrontal cortex that are activated during decision making (Figure 4). The ventromedial zone is independently sensitive to changes in outcome probabilities and payoffs, but also contributes to their combination as expected values (Knutson et al. 2005; Daw et al. 2006; Sanfey et al. 2006). The adjacent orbitofrontal zone seems involved with contrasts between alternative possible payoffs, but may focus more on losses (and aversive actions) than on gains (and appetitive actions) (O’Doherty et al. 2003; Ursu and Carter 2005; Daw et al. 2006). The dorsolateral zone also appears to track current estimates of expected values and is tightly linked to a final decision zone in the posterior part of the parietal lobe (Kim and Shadlen 1999; Trepel et al. 2005; Sanfey et al. 2006).

Activity in this latter area remains sensitive to changes in probabilities and payoffs of alternative consequences suggesting that final commitments to action do not take place earlier in the prefrontal regions (Glimcher 2003; Glimcher et al. 2005; Sugrue et al. 2005). Once this parietal region is activated, the next step is performance of an action. Many of these steps may be lateralized: when probabilities and payoffs are relatively certain, the relevant parts of the left hemisphere handle the decision making; when alternatives are equally likely or probabilities are uncertain, the relevant regions in the right hemisphere dominate the decision process (Kim et al. 2004; Knutson et al. 2005). These and similar studies have verified the existence of explicit brain regions that can track probabilities, payoffs, and expected values for multiple alternatives and compare them for rational decision making.

Figure 4: Prefrontal lobe regions of human brain activated during decision making. See text for specific roles in this process.

Biased decisions in human brains
What happens when decisions are less than rational? Instead of invocation of a second and separate decision system, recent research suggests that biased decisions are generated when other brain centers such as the limbic system and striatum modulate the rational process (McClure et al. 2004; Yarkoni et al. 2005; Sanfey et al. 2006; Tom et al. 2006). The ventral striatum appears to be a general repository of positive (gain) payoff information (O'Doherty et al. 2004; Knutson et al. 2005; Daw et al. 2006). The amygdala has been proposed as a general repository of negative (loss) payoff information (Dalglish 2004). Neither site appears to store absolute payoff values but instead converts estimates into “utilities” as a function of the current physiological state of the animal, levels of risk, the degree to which contexts limit choice, and historical associations with similar situations. The striatum and amygdala have tight links to the ventromedial and orbitofrontal prefrontal zones respectively where their weighted utility estimates are then played against the more direct estimates of rational decision making. The degree to which the decision is rational appears to depend on the strength of striatal and amygdala inputs relative to the ongoing rational process (McClure et al. 2004; De Martino et al. 2006). Note that the processing of losses and gains in separate brain regions could explain the observed higher biases for losses than for gains if the influence of the amygdala were generally greater than that of the ventral striatum.

Recent fMRI studies indicate that the amygdala is a primary source of risk sensitivity and framing biases (De Martino et al. 2006). A second source of decision bias is the presence of somatic markers (Bechara and Damasio 2005). Somatic markers are combinations of autonomic responses such as accelerated heart rate, perspiration, heat and cold flashes, or general muscle tension that are triggered (usually by the hypothalamus) when certain kinds of signals or cues are perceived by the decision maker. Some somatic markers are instinctive, whereas others are acquired from prior experience. They provide one possible mechanism for invoking the past in a linear operator process. When activated, a somatic marker acts as an additional cue that the ventromedial prefrontal cortex needs to consider during its melding of direct estimates and input from the striatum and amygdala.

Much of the remaining decision machinery in the mammalian brain seems to be devoted to updating and error correction following a decision. Once an action is complete, the prefrontal cortex, amygdala, and striatum all receive input that allows them to compare the actual versus previously anticipated payoffs. When there is a large difference between these values, the anterior cingulate, which is an active observer of all decision making, and the orbitofrontal cortex alert the brain to this situation; a particularly strong difference between expected and observed payoffs may also activate the insula (O'Doherty et al. 2003; Dalglish 2004). If the rewards of the action exceed expectations, the dopamine system in the brainstem increases its activity, and if rewards are less than expected, then the dopamine system reduces its activity (Pessiglione et al. 2006). Where observed and expected rewards are similar, the dopamine system activity remains unchanged. Because the dopamine system projects throughout the brain, this provides a global broadcast of the effectiveness of the latest decision (Schultz 1998). The amygdala, striatum, and the dorsolateral cortex then play key roles in updating stored payoffs based on this recent experience. In this context, the amygdala is thought to rescale both gain and loss payoffs into weighted utilities and then feed these biases to the orbitofrontal cortex (Dalglish 2004; De Martino et al. 2006; Paton et al. 2006). The dorsal striatum is also involved in updating once actual payoffs can be evaluated, and may play a role in establishing heuristic short cuts for future encounters (O'Doherty et al. 2004). Finally, the hippocampus and other nearby regions of the temporal lobe oversee the updating or addition of memory templates for any recent cues or signals that facilitated the decision (Greene et al. 2006; Moscovitch et al. 2006; Svoboda et al. 2006).

Other vertebrate brains and decision making

The basic brain structures and functions described above appear to be shared among rats, monkeys, and humans. What about other vertebrates? Until recently, the regions of birds' brains followed a completely different nomenclature. Careful comparisons have now revealed that there are very strong parallels in avian and mammalian brains, and most of the structures identified above for mammals are now believed to have counterparts in the brains of birds (Jarvis et al. 2005). Since birds and mammals show similar evidence of rational decision making and similar biases when irrational, it seems likely that neurobiologists will eventually demonstrate similar processes in both taxa. Although amphibians have a much smaller cerebrum than birds or mammals, they also share many of the subcortical structures described earlier.
including basal ganglia, amygdala, striatum, hippocampus, and hypothalamus (Striedter 1997; Endepols et al. 2005; Medina et al. 2005). Several of these structures even appear to be present with similar functions in fish (Broglio et al. 2005; Portavella and Vargas 2005). This suggests that the basic processes of decision making evolved early in the vertebrate line and have only been elaborated by subsequent evolution (see also the multiple chapters relating brain regions to specific ecological and behavioral tasks in Dukas and Radcliffe 2009).

Invertebrate brains and decision making

Invertebrates have quite different brain structures from vertebrates. However, research on slugs and leeches suggests that successive stimuli lead to cumulative updating, and that decisions depend on mutual levels of activity in multiple nerve cells as opposed to simple association and switching circuits (Esch and Kristan 2002; Esch et al. 2002; Jing and Gillette 2003; Briggman et al. 2005). The brains of most insects consist of sensory processing ganglia that provide input to mushroom bodies where associative learning and decision making take place (Fahrbach 2006; Menzel et al. 2006). While there has yet been little neurobiological work on decision making in insects, genetic and biochemical studies suggest that decisions again involve multiple but interacting regions within the mushroom bodies (Heberlein et al. 2004; Abramson et al. 2005).

![Figure 5: The brain of a honeybee. The large masses (ME and LO) on each side do most of the processing of visual stimuli. Olfactory stimuli are processed in the paired lobes on the bottom side of the brain (AL). Decision making and memory appear to be the main functions of the large mushroom bodies (MC and LC) positioned at the top and between the various sensory lobes. (After Menzel 1983.)](image)

Literature cited


Measures of Discrete Signal Effectiveness

Introduction

The most efficient way to summarize discrete coding schemes is with matrices. Typically, these are two dimensional matrices with inputs (such as ambient conditions) listed along one axis, outputs (such as signals emitted) listed along the perpendicular axis, and cell values containing some measure of how often a given input and a given output co-occur. Different matrices can be used to characterize sender coding, signal propagation effects, and receiver assignments of propagated signals to expected categories. Or one could combine all of these effects into a single overall matrix. Below, we outline how one might collect data for a sender’s matrix, combine this with similar data on propagation and receiver matrices to compute an overall matrix, and use matrix algebra to compute various measures of signal effectiveness.

Format conventions

A coding matrix summarizes the conditional probabilities that a given input will result in a given output. Thus they are a special case of transition (also called stochastic) matrices which are widely used in statistics, ecology, and Markov chain analyses. To keep the geometry of our matrix axes similar to treatments of continuous coding schemes (in which inputs are assigned to the horizontal axis and outputs to the vertical axis), we shall use “left stochastic matrices” in which we assign inputs to matrix columns and outputs to matrix rows. The columns in such a matrix should each add to 1.0. This is the opposite configuration from the “right stochastic matrices” used in many mathematical and ecological studies.

Matrix manipulation: Matlab

We shall demonstrate computations below using MATLAB. This is a commercially available program that is specifically designed for this kind of mathematics. It is now available in both professional and student versions. MATLAB is produced by The Mathworks (http://www.mathworks.com/) and is available through a site license at many universities and colleges. All of the examples and routines outlined below are presented in MATLAB formats, and the functions we have written can be copied from this text and pasted into MATLAB M-files or used directly at the MATLAB prompt (>>). If performed at the prompt, all command lines should be followed by hitting the return key. The routines may also be adapted fairly easily to other environments such as Wolfram Research’s Mathematica (http://www.wolfram.com/).

Obtaining a sender’s coding matrix

Suppose we are studying a species of African monkey that can emit any of three different call types when it spots predators: S1, S2, S3. The three predators are leopards, eagles, and other predators like snakes. We are interested in measuring how regularly the monkeys assign the same signal type to a given predator including emitting no signal when no predator is present. We pick a sample period that is shorter than the interval between successive appearances of predators. We then observe these monkeys for a large number of sample periods and record the presence or absence of a predator in each period and note any alarm calls given. Because an initial alarm call often triggers mimicked calling by others in the troop, regardless of whether they saw a predator or not, we only record the first alarm call given in a calling bout. Suppose we thus accumulate 1,205 sample periods. We can assemble the raw data into a contingency table in which each cell indicates the number of times that a given condition (predator type or no predator present) co-occurred with a given signal (including no signal emitted). This might look as follows:
We can convert these raw data into a table of probabilities that we shall call the AND table: this summarizes the joint probability, based on our samples, that a given condition and a given signal will co-occur. The AND table is created by dividing all cell values by the grand total (here 1,205). In Matlab, we would first create the raw matrix, \( W \), as:

\[
W = \begin{bmatrix}
49 & 9 & 11 & 27 \\
4 & 81 & 7 & 9 \\
2 & 1 & 89 & 54 \\
8 & 8 & 36 & 810
\end{bmatrix}
\]

Then the AND matrix is simply

\[
\text{Wand} = W / 1205
\]

This gives us:

\[
\begin{array}{cccc}
\text{Signal} & \text{Leopard} & \text{Eagle} & \text{Other} & \text{No Predator} \\
\hline
\text{Alarm S1} & 0.0407 & 0.0075 & 0.0091 & 0.0224 \\
\text{Alarm S2} & 0.0003 & 0.0672 & 0.0058 & 0.0075 \\
\text{Alarm S3} & 0.0017 & 0.0008 & 0.0739 & 0.0448 \\
\text{No Call} & 0.0066 & 0.0066 & 0.0299 & 0.6722 \\
\end{array}
\]

We next compute the marginal subtotals for the AND matrix: the sums of the rows indicate the overall fractions of sample periods in which each signal option was given in our sample, and the sums of the columns provide the fractions of sample periods that each predator situation occurred. In Matlab, we can denote the row sums as the vertical vector \( B \), and the column sums as the row vector \( P \). Thus:

\[
B = \text{sum}(\text{Wand},2)
\]

\[
P = \text{sum}(\text{Wand})
\]

If we add those subtotals (in blue) to the AND matrix, we would get:

\[
\begin{array}{cccccc}
\text{Signal} & \text{Leopard} & \text{Eagle} & \text{Other} & \text{No Predator} & \text{Subtotal B} \\
\hline
\text{Alarm S1} & 0.0407 & 0.0075 & 0.0091 & 0.0224 & 0.0797 \\
\text{Alarm S2} & 0.0003 & 0.0672 & 0.0058 & 0.0075 & 0.0838 \\
\text{Alarm S3} & 0.0017 & 0.0008 & 0.0739 & 0.0448 & 0.1212 \\
\text{No Call} & 0.0066 & 0.0066 & 0.0299 & 0.6722 & 0.7154 \\
\text{Subtotal P} & 0.0523 & 0.0822 & 0.1187 & 0.7469 & 1.0000
\end{array}
\]
Note that the overall sum of all cell values in an AND matrix should equal 1.0.

To obtain the sender’s coding matrix, we would divide each cell in the AND matrix by its column subtotal ($P$). In Matlab, this takes a few steps:

```matlab
>> PP = [P;P;P;P]
>> S = W_and ./ PP % (note that this is the ./ operator, not the / operator)
```

Converting the resulting probabilities into percentages, this gives us the sender coding matrix $S$:

**SENDER CODING MATRIX ($S$):**

<table>
<thead>
<tr>
<th>Signal</th>
<th>Leopard</th>
<th>Eagle</th>
<th>Other</th>
<th>No Predator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alarm S1</td>
<td>78%</td>
<td>9%</td>
<td>8%</td>
<td>3%</td>
</tr>
<tr>
<td>Alarm S2</td>
<td>6%</td>
<td>82%</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>Alarm S3</td>
<td>3%</td>
<td>1%</td>
<td>62%</td>
<td>6%</td>
</tr>
<tr>
<td>No Call</td>
<td>13%</td>
<td>8%</td>
<td>25%</td>
<td>90%</td>
</tr>
</tbody>
</table>

We next consider the case of a monkey receiver who is at the outer limit of the troop dispersion while feeding. Thus it will experience any alarm signals only after they have propagated some distance from the sender. Careful measurements have allowed our researchers to compile the following propagation matrix, $T$, which lists the possible emitted signals by sender monkeys as the inputs and an acoustical classification of these sounds (by the researchers) after they have propagated this distance as the outputs. Note that the occasional addition of wind and insect noise has created one more propagated signal than is actually emitted by senders:

**PROPAGATION MATRIX ($T$):**

<table>
<thead>
<tr>
<th>Propagated Signal</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>No Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1–P</td>
<td>85%</td>
<td>3%</td>
<td>17%</td>
<td>3%</td>
</tr>
<tr>
<td>S2–P</td>
<td>6%</td>
<td>80%</td>
<td>12%</td>
<td>1%</td>
</tr>
<tr>
<td>S3–P</td>
<td>5%</td>
<td>4%</td>
<td>67%</td>
<td>8%</td>
</tr>
<tr>
<td>S4–P</td>
<td>0%</td>
<td>11%</td>
<td>0%</td>
<td>5%</td>
</tr>
<tr>
<td>No Call–P</td>
<td>4%</td>
<td>2%</td>
<td>4%</td>
<td>83%</td>
</tr>
</tbody>
</table>

The distant receiver monkey has to try to match what it hears with one of the four possible signal options that it knows senders might adopt. Based on a variety of lab and field tests, our researchers have assembled the following assignment matrix ($R$) used by a typical receiver monkey at that distance from the sender:
RECEIVER ASSIGNMENT MATRIX (R):

<table>
<thead>
<tr>
<th>Assigned Signal</th>
<th>Propagated Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1–P</td>
</tr>
<tr>
<td>S1</td>
<td>64%</td>
</tr>
<tr>
<td>S2</td>
<td>9%</td>
</tr>
<tr>
<td>S3</td>
<td>2%</td>
</tr>
<tr>
<td>No Call</td>
<td>25%</td>
</tr>
</tbody>
</table>

We shall need some estimates of the prior probabilities that each predator condition is likely to occur. The best data we have available for this are the column subtotals for the AND matrix, \( P \). Writing these probabilities as percentages, we get for the vector \( P \):

PRIOR PROBABILITIES (P):

<table>
<thead>
<tr>
<th>Condition</th>
<th>Leopard</th>
<th>Eagle</th>
<th>Other</th>
<th>No Predator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>5%</td>
<td>8%</td>
<td>12%</td>
<td>75%</td>
</tr>
</tbody>
</table>

**Chaining**

One thing we might want to do is combine the separate sender, propagation, and receiver matrices into an overall cumulative matrix. Note that in addition to summarizing cumulative errors, this table will reveal the degree to which errors in one stage are corrected by errors in a later stage. As the overall matrix will reveal, such error correction is overwhelmed by cumulative errors and the cell values for the overall matrix are always no larger than the smallest equivalent cell among the contributing matrices.

To generate the overall matrix, we load each component matrix, (S for sender, T for transmission, and R for receiver):

\[
\begin{align*}
\text{S} &= \begin{bmatrix}
0.7800 & 0.0900 & 0.0800 & 0.0300 \\
0.0600 & 0.8200 & 0.0500 & 0.0100 \\
0.0300 & 0.0100 & 0.6200 & 0.0600 \\
0.1300 & 0.0800 & 0.2500 & 0.9000
\end{bmatrix} \\
\text{T} &= \begin{bmatrix}
0.8500 & 0.0300 & 0.1700 & 0.0300 \\
0.0600 & 0.8000 & 0.1200 & 0.0100 \\
0.0500 & 0.0400 & 0.6700 & 0.0800 \\
0 & 0.1100 & 0 & 0.0500 \\
0.0400 & 0.0200 & 0.0400 & 0.8300
\end{bmatrix} \\
\text{R} &= \begin{bmatrix}
0.6400 & 0.0900 & 0.1500 & 0 & 0.0300 \\
0.0900 & 0.8000 & 0.1200 & 0.6400 & 0.0100 \\
0.0200 & 0.0100 & 0.5500 & 0.0600 & 0.0800 \\
0.2500 & 0.1000 & 0.2800 & 0.3000 & 0.8800
\end{bmatrix}
\end{align*}
\]

We can combine the effects of propagation and sender error by computing the simple matrix product \( T \cdot S \). To compute
the entire chain, we add the effects of the receiver assignments to get: \( R \times T \times S \). The only trick to remember is to place each successive effect ahead of the one being modified in the product (e.g. pre-multiply). The overall (\( O \)) result of combining the sender, transmission, and receiver matrices is then:

\[
\begin{align*}
\gg O &= R \times T \times S \\
O &= \\
&= \begin{bmatrix}
0.4552 & 0.1372 & 0.2010 & 0.0823 \\
0.1588 & 0.6074 & 0.1606 & 0.0772 \\
0.0661 & 0.0491 & 0.2676 & 0.1271 \\
0.3270 & 0.2114 & 0.3949 & 0.7248 \\
\end{bmatrix}
\end{align*}
\]

Note that the dimensions of multiplied matrices have to be compatible. This means that the number of columns in the matrix to the left in any product must equal the number of rows of the matrix to the right. Here, \( S \) is a 4 (row) x 4 (column) matrix that is pre-multiplied by \( T \) which is a 5 x 4 matrix. The result will be a \((5 \times 4)(4 \times 4) = (5 \times 4)\) matrix. We then pre-multiply our \( T \times S \) result by \( R \) which is a 4 x 5 matrix. The overall result is a \((4 \times 5)(5 \times 4) = (4 \times 4)\) matrix. The values in this \((R \times T \times S)\) table provide the overall conditional probabilities that a receiver will assign what it received to one of the expected signals (rows) when a given predator situation was true (columns). Note also that some rounding errors have occurred at this MATLAB default resolution (all columns in these tables should add to 1.0). These can be corrected by dividing each number in a column by the column total to get:

**OVERALL CODING MATRIX:**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Leopard</th>
<th>Eagle</th>
<th>Other</th>
<th>No Predator</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assigned Signal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td>45%</td>
<td>14%</td>
<td>19%</td>
<td>8%</td>
</tr>
<tr>
<td>S2</td>
<td>16%</td>
<td>60%</td>
<td>17%</td>
<td>8%</td>
</tr>
<tr>
<td>S3</td>
<td>7%</td>
<td>5%</td>
<td>26%</td>
<td>12%</td>
</tr>
<tr>
<td>No Call</td>
<td>32%</td>
<td>21%</td>
<td>38%</td>
<td>72%</td>
</tr>
</tbody>
</table>

**Forward measures of signal effectiveness**

**Forward measures** of signal effectiveness rate the regularity with which a given output is produced when a given input is true. This can be contrasted with **backward measures** which rate how regularly the correct input is inferred when a given output is received.

There are two kinds of forward measures for signal effectiveness: those that focus only on the coding matrix, and those that require both the coding matrix and the prior probabilities of the matrix inputs. The latter essentially combine the coding matrix and the priors to regenerate the relevant AND matrix. By itself, this would argue for just using AND matrices and not extracting the coding matrices from them. The reason for going to coding matrices is that these are less likely to vary with context than are the priors. If the coding matrix for senders is fairly fixed, a receiver monkey can also use it effectively in a variety of situations by simply updating its estimates of the prior probabilities of the various predators for each situation. Ideally then, one would want some measures that focus only on the coding matrix, and others that characterize a particular situation and thus combine the coding matrix with relevant priors. Luckily, both kinds of measures are available.

**Forward measures based only on the coding matrix: determinants**

The determinant of a square matrix (one with identical numbers of rows and columns) is a measure of how
heterogeneously values are distributed among the matrix cells. For a coding matrix, the most heterogeneous distribution possible is **perfect coding** in which there is a single cell in each row and column with a conditional probability of 1.0, and zeros in all other cells. The determinant of such a perfect coding matrix is 1.0 (or using percentages, 100%). At the other extreme is a completely homogeneous matrix in which all cell values are identical. This is the case when all outputs are equally likely for a given input. The determinant of such a matrix is 0 (or 0%). Levels of cell value heterogeneity between these two extremes will yield intermediate determinant values. However, note that it does not take much homogenizing to make the determinant value small: the more similar any two columns or rows in a matrix, the closer the determinant will be to 0.

What if the coding matrix is not square as will be the case for some animal signal systems? A solution has been proposed by Yanai et al. (2006) who suggest multiplying such a matrix by itself, (actually premultiplying it by its transpose, the same matrix with rows and columns reversed), to generate a new square matrix. The square root of the determinant of this new square matrix is called a “generalized determinant.” This method will give the same value when applied to an initially square matrix as one would get by directly taking the determinant of that matrix. Thus generalized determinants can place all signal matrices on the same 0–100% scale with perfect coding at 100% and random assignment at 0%. In Matlab, the relevant step for extracting the generalized determinant from a matrix \( S \) is:

\[
>> deta = \sqrt{\text{det}(S^*S)}
\]

where \( S^* \) is the transpose of \( S \) (the same matrix but with rows and columns reversed). Thus if \( S \) has \( n \) rows and \( p \) columns, \( S^*S \) has dimensions \( (p \times n) \times (n \times p) = (p \times p) \). This new matrix tabulates the sums of the squares of the inputs. It is similar to the first step in converting a rectangular \( (n \times p) \) data matrix with samples as the rows and measures as the columns into the relevant variance-covariance and correlation matrices in statistics: one “squares” the \( n \times p \) data matrix into a \( p \times p \) sums of squares matrix and then uses this to compute variances, covariances, and correlations between measures.

While one can compute a generalized determinant for any coding matrix (including transmission and receiver matrices as well as sender matrices) without reference to prior probabilities, this tacitly assumes that none of the inputs has zero prior probability (Kåhre 2002): if a prior is zero, then it really does not matter what the values are in the coding matrix, as any set of conditional probabilities will yield the same outcomes. Differences in the determinants of two coding matrices for both of which the same input has a prior of zero may be meaningless. We thus should restrict determinant measures to coding matrices in which all inputs have non-zero prior probability.

**Forward measures requiring both priors and the coding matrix**

**Average Consistency**

The simplest forward measure that is specific to a given set of input prior probabilities is consistency. One first identifies the cell in each column of the measured matrix that contains the maximal value for that column. The row containing that maximal cell value is the dominant output for that input. For our sender matrix \( S \), we can color the maximal cells in blue:

<table>
<thead>
<tr>
<th><strong>Sender Coding Matrix (S):</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Signal</strong></td>
</tr>
<tr>
<td>Alarm S1</td>
</tr>
<tr>
<td>Alarm S2</td>
</tr>
<tr>
<td>Alarm S3</td>
</tr>
<tr>
<td>No Call</td>
</tr>
</tbody>
</table>

The average **consistency** of this matrix is computed by weighting each input’s maximal cell value by the prior
probability that that input will occur, and then adding these products across all inputs. This is done simply in MATLAB using:

\[
\text{>> AC} = \text{max}(S) \times P' \% \text{ (note transpose of } P \text{ using prime operator)}
\]

For our sample sender’s matrix \(S\) above, the average consistency is 85.4%.

Perfect coding will result in an average consistency of 100%; random assignment of outputs to inputs will produce an average consistency equal to the reciprocal of the number of possible outputs. Note that average consistency is oblivious to the presence of more than one maximal value in the same row of the matrix: consistency only characterizes deviations from perfect coding along the vertical axis; it ignores any deviations along the horizontal axis. As a result, average consistency values are invariably larger than the determinant for the same matrix.

**Index of Association**

This measure was proposed for rating the heterogeneity of contingency tables by Goodman and Kruskal (1954). Because the AND matrix is simply the original contingency table divided by the total number of samples, the index of association can also be computed directly from the AND matrix. It finds the maximum row subtotal in that matrix and compares it to the maxima in each of the AND matrix columns. In words, the index evaluates how much knowing the current input helps predict the output when compared to just knowing which output is most common overall. The index varies from 1.0 (100%) for perfect coding to 0 (0%) for a table with uniform values throughout. For our sender's coding matrix \(S\) above, and the priors in vector \(P\), the index of association (called “lambda” by Goodman and Kruskal) is 48.2%. A MATLAB routine for this index is:

```matlab
function L=lambda(S,P)
% Computes Goodman-Kruskal index of association for coding matrix S
% with inputs as columns and outputs as rows, and prior probability
% vector P. Values range 0-1 with 100% for perfect coding and 0 when
% all matrix cell values are equal.
% classifications, Part I. Journal of the American Statistical
% Association 49: 732-764.
[a b]=size(S);
D=ones(a,b);
for i=1:a
    D(i,:)=P;
end
ND=D.*S;
B=S*P';
PB=max(B);
PNB=sum(max(ND));
L=(PNB–PB)/(1–PB);
```

**Other Forward Measures**

A number of other forward measures have been proposed. Many, such as Theil's index of inequality (Theil 1970), use logarithms of probabilities as part of their computation. Since animal coding matrices (particularly those approaching perfect coding) contain zeros in some cell values, these measures may not be computable without making some substitutions or approximations (as is done for mutual information, of which many of these measures are close relatives).

**A posteriori probabilities: Bayes’ theorem**

All backward measures of signal effectiveness are based on the a posteriori probabilities that a given input is true having received a given input and assuming a given coding matrix and set of prior probabilities. While animals may invoke a variety of shortcuts or approximations to estimate a posteriori probabilities, none can do better than by invoking Bayes’ Theorem. Therefore, most backward measures of signal effectiveness first compute a table of a posteriori probabilities using Bayes’ theorem. If one knew how an animal not using Bayesian methods updated
Below, we provide a MATLAB routine for computing the matrix of a posteriori probabilities predicted by Bayes’ Theorem, given a coding matrix \( S \), and the set of prior probabilities in the horizontal vector \( P \). Note that the output matrix from our Bayes M-file continues to observe the “left stochastic” convention that inputs are assigned to columns and outputs to rows. In the case of the a posteriori probabilities, this means that what were outputs in the coding matrix (e.g. signals) are now assigned to columns and the former inputs (e.g. conditions) are now assigned to the rows. This is because the cell values in the matrix give the conditional a posteriori probabilities that a given condition is true after having received a given signal and assuming that the initial coding matrix and priors are valid. The relevant MATLAB routine is:

```matlab
function B=Bayes (S,P)
% S is the coding matrix with conditions as columns and signals as rows
% P gives prior probabilities as row vector with columns as conditions
% D.*S computes the “Condition AND Signal“ matrix from S and P
% S*P' computes the row totals in the AND matrix and thus the total
% fraction of time that each signal is given across all conditions
% The last half of the expression divides each AND cell value by
% the corresponding row total; this is the actual Bayes calculation
% Values in the output matrix cells are a posteriori probabilities of
% each condition being true (rows) after having received the signal assigned
% to that column. Note reversal of axis assignments from S matrix.
% Inclusion of signals that are never used results in 0 Bayesian estimates.

[a b]=size(S);
D=ones(a,b);
DD=D;
for i=1:a
    D(i,:)=P;
end
AND=D.*S; %Compute joint (AND) matrix of conditions and signals
SS=S*P'; %compute total fraction of time each signal is given
for i=1:a
    if (SS(i)>0)
        SP(i)=1/SS(i);
    else
        SP(i)=0;
    end
end
for i=1:b
    DD(:,i)=SP;
end
B=AND.*DD; %compute a posteriori prob of conditions given signal
B=B'; %Reverse axes so that original inputs and outputs reversed
% and columns add to 1.0.
```

Let us apply this routine to compute the a posteriori probabilities for the overall matrix \( O \) and the prior probabilities \( P \) when there are no transmission alterations to the emitted signals and receivers identify emitted signals perfectly. This yields the following a posteriori matrix with the emitted signal alternatives as the columns and the inferred predator conditions as the rows:

```matlab
>> Bayes(O,P)
ans =
    0.1903    0.0584    0.0245    0.0262
    0.0918    0.3571    0.0291    0.0271
    0.2017    0.1593    0.2385    0.0759
    0.5162    0.4253    0.7078    0.8708
```
Putting these into the appropriate format and correcting for rounding errors,

**A POSTERIORI PROBABILITIES:**

<table>
<thead>
<tr>
<th>Inferred Condition</th>
<th>Alarm S1</th>
<th>Alarm S2</th>
<th>Alarm S3</th>
<th>No Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leopard</td>
<td>19%</td>
<td>6%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Eagle</td>
<td>9%</td>
<td>36%</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>Other</td>
<td>20%</td>
<td>16%</td>
<td>24%</td>
<td>8%</td>
</tr>
<tr>
<td>No Predator</td>
<td>52%</td>
<td>42%</td>
<td>71%</td>
<td>87%</td>
</tr>
</tbody>
</table>

Note that the a posteriori probabilities for each of the predators remain low even after receiving signals; however, these values should be compared to the pre-signal priors of 5%, 8%, and 12% for leopards, eagles, and other predators respectively. Clearly, receipt of signals increases the receiver’s estimated probabilities that the respective predator is present, and reduces the prior 75% probability that no predator is present.

**Backward measures of signal effectiveness**

There are two useful sets of backward measures for discrete coding schemes. Both result in absolute measures of signal effectiveness and but can be converted to relative measures comparing the effectiveness of the signal scheme to one with perfect coding.

**Reliabilities**

The first two measures concern signal scheme reliability. Suppose a given condition (one possible input in the original coding matrix) is true. Reliability is here defined as the probability estimated by a receiver that this condition is true after having processed a received signal and updated its estimates. If the relevant coding scheme is perfect, the receiver’s estimate that that condition is true will be 100%; if a coding scheme has the same values in all cells, and thus provides no information, the receiver’s estimate after receipt of a signal will be no different than before and thus equal to the prior probability for that condition. **Average reliability** weights the reliability of the signal scheme for each condition by that condition’s prior probability and adds up these products. **Relative reliability** is the average difference between the updated probability for each condition and its prior value divided by the maximal difference that would be obtained if coding were perfect. Thus if the reliability for a condition were R, and the prior probability for that condition were P, relative reliability=(R–P)/(1–P). Relative reliability thus varies between 0 and 1.0.

Computations of reliabilities in MATLAB are easy: one simply chains the various stages in the communication process. Thus if sender and receiver are far apart and thus we need to consider the sender’s matrix S, a transmission matrix T, and a receiver’s assignment of transmitted signals to expected categories matrix, R, the overall coding matrix O can be created in MATLAB as:

```
>>O=R*T*S
```

We thus have the first three steps in the communication process accounted for. We next need to characterize how the receiver uses assigned signals to update its probability estimates for each condition. As an upper limit to updating rates, we use the Bayes routine (above), the overall coding matrix O, and the prior probabilities vector P to create the final updating transition matrix BA as:

```
>>BA=Bayes(O,P)
```

Pre-multiplying O with BA will give us a square matrix G with the actual conditions as columns, the possible inferred conditions as rows, and cell values equal to the updated probabilities estimated by the receiver that each of the
possible conditions is true when the condition listed in the columns of the matrix is in fact true:

$$G = BA^\top O$$

The main diagonal of this matrix indicates how likely the receiver thinks the current condition is true; in a way, these values indicate how closely the receiver is to getting it right. The off-diagonal values indicate the probabilities that the receiver has estimated that conditions that are not currently true might be. The $G$ matrix given our overall coding matrix above and priors $P$ is:

**OVERALL RELIABILITY:**

<table>
<thead>
<tr>
<th>Inferred Condition</th>
<th>Actual Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Leopard</td>
</tr>
<tr>
<td>Leopard</td>
<td>11%</td>
</tr>
<tr>
<td>Eagle</td>
<td>11%</td>
</tr>
<tr>
<td>Other</td>
<td>15%</td>
</tr>
<tr>
<td>No Predator</td>
<td>63%</td>
</tr>
</tbody>
</table>

The average reliability is then the product of the prior probabilities $P$ and the values along the main diagonal of $G$:

$$R = P \cdot \text{diag}(G)$$

In the example, the average reliability $R$ is 63.5%. This value is high because of the high diagonal cell and high prior values for the no predator condition.

To compute the average relative reliability, $RR$, we need to compare each value along the main diagonal of $G$ with the corresponding prior values in $P$, and divide this difference by the maximum possible were the coding scheme perfect ($1 - P$). These ratios are then weighted by the prior probabilities to given an average relative reliability for the entire coding scheme. Assuming that no value of $P = 1.0$, we can use:

$$[a \ b] = \text{size}(P)$$

$$ON = \text{ones}(a, b)$$

$$W = (GG^\top - P) / (ON - P)$$

$$RR = W \cdot P'$$

In our example above, the relative reliability ($RR$) is 14%.

**Mutual information**

Mutual information was developed by Shannon (Shannon and Weaver 1949) and has been widely used by many disciplines in science and statistics. The uncertainty about which condition is true is typically given as a current probability estimate. In information theory, probabilities are rescaled as the number of binary questions (bits) that would need to be answered to remove all uncertainty. Mutual information is the difference between the uncertainty about which condition is true before signals minus the residual uncertainty after a receiver has detected and classified a signal and updated its probability estimates. It is thus a backward measure of signal effectiveness.

As with reliability, we can generate an absolute and relative average measure of coding scheme effectiveness. The absolute measure is simply the average reduced uncertainty (in bits) after receiving a signal from the scheme, and the relative measure is the fraction of the initial uncertainty that is resolved on average by receipt of a scheme signal.
We first compute the absolute mutual information using the same coding matrices presented earlier (sender matrix $S$, transmission matrix $T$, and receiver signal assignment matrix $R$). Again, we use the chain procedure to compute an overall coding matrix $O = R^T S$ and a prior probability vector $P$.

The conversion of all probabilities to bits uses logarithms to the base 2. Because 0 probabilities may occur, we need a routine that tells MATLAB what to do when such a value occurs. The following routine will do that:

```matlab
function Y=mylog2(X)
%Computes log to base 2 if number>0 and returns 0 otherwise
[a b]=size(X);
Y=X;
for i=1:a
    for j=1:b
        Y(i,j)=0;
        if (X(i,j)>0)
            Y(i,j)=log2(X(i,j));
        end
    end
end
```

The average uncertainty in bits for any vector or matrix $V$ can then be computed using:

```matlab
function H=JH(V)
%Computes entropy from V which is either a vector or a matrix
H=-sum(sum(V.*mylog2(V)));
```

To compute the mutual information provided by a coding matrix, here denoted by $S$, (but note that it could be the overall matrix $O$ if propagation distortions and receiver error are relevant), given priors in vector $P$, we can use the following:

```matlab
function [HT RH]=MI(S,P)
% This function computes the mutual information in bits provided by a
% signal matrix S and a vector of prior probabilities P. The result is
% output as the variable H. If the output is defined as a 2 element vector,
% the second element is the fraction of the original uncertainty resolved
% on average by this coding matrix. This assumes Bayesian updating.
% Uses functions mylog2 and JH.
[a b]=size(S);
D=ones(a,b);
for i=1:a
    D(i,:)=P;
end
ND=D.*S; % computes AND matrix of joint prob of inputs and outputs
XP=P; %Get priors ready for entropy calc
if (sum(XP)~=1)
    XP=XP/sum(XP); % Make sure priors add to 1.0
end
for i=1:size(XP) % Make sure priors >0
    if (XP(i)==0)
        XP(i)=1;
    end
end
HC=JH(XP); %Compute max uncertainty at start
HT=HC+JH(RC)-JH(ND); %Compute mutual information provided
RH=0;
if HC~=0
    RH=HT/HC; %Compute fraction of original information resolved
end
```

**Single routine for all discrete measures**

The following MATLAB routine, `Rel`, computes and prints out all of the above measures for a given coding matrix $S$ and
prior probability vector $P$. If no output vector is assigned, one only gets the list and results. If the function output is assigned to a vector such as $[M\ VV]$, $M$ will contain the list and values and $VV$ will be a vector with only the values.

function $[M\ VV] = \text{Rel}(S, P)$

% $S$ is a coding matrix with inputs as columns and outputs as rows.
% Column totals should add to 1.0. $P$ is a horizontal vector containing
% the prior probabilities of each column in $S$ being true.
% Average consistency (ACC) is defined as the sum of max cell values for each
% column in the coding matrix discounted by the prior probability of that
% input being true. It is a forward measure that varies between 0 when no
% information about outputs is provided by inputs to 1.0 for perfect coding.
% Another forward measure is the generalized determinant of the coding
% matrix. This also varies between 0 (no information provided) and 1.0 for
% perfect coding. A third forward measure is the Goodman-Kruskal index of
% association, lambda. This is the average reduction in uncertainty (on a
% scale of 0 to 1.0) that knowledge of the inputs provides in predicting the
% outputs when compared to a guess based on the relative frequencies with which
% each output in $S$ is given.
% The remaining measures are backwards indices in that they require the
% computation of an updated (Bayesian) estimate of input probabilities
% given receipt of a signal, a set of priors, and access to the relevant
% coding matrices. Average reliability is the mean probability across inputs
% that a receiver using this coding system will assign to conditions when they
% are in fact true. Perfect coding will yield a value of 1.0 and chance
% coding (all cell values in the coding matrix are identical, will yield a
% weighted average of the priors (e.g. there is no change from prior values
% if attending to signals). Relative reliability is the fraction of the
% improvement above using priors to estimate inputs provided by signals
% when compared to not using signals and only relying on priors. It varies
% from 0 (no improvement) to 1.0 (perfect coding).
% Entropy (bits) is another backwards measure using the Bayesian estimates of
% probabilities. It computes the difference in uncertainty before receipt of
% a signal minus the residual uncertainty after signal receipt. Relative
% entropy is the fraction of initial uncertainty removed on average by use
% of this signal set. These routines require access to the Bayes, AbsRel,
% and JH routines.
% Jack Bradbury, October 2009.

[a b]=size(S);
D=ones(a,b);
for i=1:a
    D(i,:)=P;
end
ND=D.*S; %computes AND matrix of joint prob of each input and output
RC=S*P'; %computes sums of rows of AND matrix
%Compute measures
ACC=max(S)*P'; %Compute average consistency
L=lambda(S,P);
deta=sqrt(det(S'*S)); %compute generalized determinant of matrix
[R RR G]=AbsRel(S,P);
XP=P; %Get priors ready for entropy calc
if (sum(XP)~=1)
    XP=XP/sum(XP); % Make sure priors add to 1.0
end
for i=1:size(XP) % Make sure priors >0
    if (XP(i)==0)
        XP(i)=1;
    end
end
HC=JH(XP); %Compute max uncertainty at start
HT=HC+JH(RC)-JH(ND); %Compute mutual information provided
RH=0;
if HC~=0
    RH=HT/HC; %Compute fraction of original information resolved
end
titles=char('Consistency:','Determinant:','Lambda:','Reliability:','Rel
Reliability:','Ht (bits):','Fraction Hmax:');
MV=num2str([ACC;deta;L;R;RR;HT;RH],'%6f');
Literature cited


Web Topic 8.9
Mutual Information Measures of Signal Effectiveness

Introduction

Mutual information is often used as a backward measure of signal effectiveness: that is, it computes the change in a receiver’s uncertainty (entropy) about which of several alternative conditions is true after receipt of a signal. Uncertainty and changes in uncertainty are measured in bits: the number of binary questions that would need to be answered to identify which of several alternatives is true. The uncertainty in bits can be computed as the logarithm to the base 2 of the probability that a given alternative is true. We are most interested in the effectiveness of signal sets, and thus compute the weighted average of the binary questions required to clarify which alternative in a set is currently true.

Logic of method

The computation of the mutual information, $\overline{H_T}$, provided by a discrete signal set requires access to the relevant coding matrix which tabulates the conditional probabilities $P(S_j | C_i)$ that a given output, $S_j$, will be produced when a given input, $C_i$, is true. In many mathematical treatments, inputs of transition matrices are assigned to rows and outputs to columns, but inputs are then assigned to the horizontal axis in graphs of continuous processes. For consistency, we here adopt the convention of assigning alternative inputs to the horizontal axis of both coding schemes, and outputs to the vertical axis. Thus in our discrete coding matrices, inputs are assigned to columns and outputs are assigned to rows. Since the contents of the matrices are probabilities, and every input should result in one of the listed outputs, columns in these coding matrices should add to 1.0 (or 100%).

Mutual information also requires access to a vector listing the priori probability before signaling, $P(C_i)$, that each input alternative is likely to occur.

Given access to the coding matrix and the prior probabilities, there are two ways to compute the average mutual information provided by a signal set:

as

$$\overline{H_T} = \sum_i \sum_j P\left(C_i \text{ and } S_j\right) \log_2 \frac{P(C_i | S_j)}{P(C_i)}$$

or as

$$\overline{H_T} = \overline{H(C)} - \overline{H(C \mid S)}$$

where

$$\overline{H(C)} = \sum_i P(C_i) \log_2 P(C_i)$$

and
The probabilities that a given input is true when a particular output has been received, \( P(C_i | S_j) \), are computed by combining the coding matrix with the prior probabilities using Baye’s theorem.

**Method and example**

Because the computations for mutual information can become complicated with multiple inputs and outputs, we outline below a simplifying recipe that requires generation of three successive tables. The first is the coding matrix tabulating the conditional probabilities that a given output will be generated when a given input is true. The second tabulates the joint probabilities, \( P(C_i \text{ and } S_j) \), that any given input and output will co-occur. The third table lists the a posteriori probabilities, \( P(C_i | S_j) \) based on Baye’s Theorem, that a particular input is true given the coding matrix and the prior probabilities.

We shall use an example in which a female bird seeks to choose a mate and wishes to discriminate between healthy and sick candidates. Males can sing either fast or slow songs. In the first considered population, both sick and healthy males can sing songs at both speeds, but healthy males sing fast songs more often than do sick males. Given that healthy and sick males are equally common in the population, and given the coding matrix listed in the first table, what is the average mutual information provided by song speed?

1. **Coding Matrix and Prior Probabilities:** First, set up the coding matrix by observing the fraction of time that each of sick and healthy males produce fast and slow songs. This table contains the \( P(S_j | C_i) \) values. We also add the prior probability vector below this table:

<table>
<thead>
<tr>
<th>Male Condition</th>
<th>Song Speed</th>
<th>Healthy</th>
<th>Sick</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>0.70</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>0.30</td>
<td>0.60</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

2. **Joint Probability Table:** We can use the basic probability rule that a joint probability equals the product of the prior probability one component will occur with the conditional probability that the other will occur when the first is true. Algebraically, \( P(C_i \text{ and } S_j) = P(C_i) \cdot P(S_j | C_i) \). We thus multiply the prior probability in each column by the respective cell values in the same column of the coding matrix. Thus:

<table>
<thead>
<tr>
<th>Male Condition</th>
<th>Song Speed</th>
<th>Healthy</th>
<th>Sick</th>
<th>( P(S_j) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>0.70 x 0.50 = 0.35</td>
<td>0.40 x 0.50 = 0.20</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>0.30 x 0.50 = 0.15</td>
<td>0.60 x 0.50 = 0.30</td>
<td>0.45</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that we can add up the subtotal across rows to obtain the overall fraction of time, \( P(S_j) \), that a given output, here song speed type, is produced overall.

3. **A Posteriori Probability Table:** We can now easily compute the Bayesian a posteriori probabilities that a female could compute after combining receipt of a given signal with her knowledge of the coding matrix and the prior
probabilities of each male condition. Cells in this table are computed by dividing the cell values in the joint probability table by their corresponding row in that table:

<table>
<thead>
<tr>
<th>Song Speed</th>
<th>Male Condition</th>
<th>Healthy</th>
<th>Sick</th>
<th>P(S_j)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>0.35/0.55 = 0.636</td>
<td>0.20/0.55 = 0.364</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>0.15/0.45 = 0.333</td>
<td>0.30/0.45 = 0.667</td>
<td>1.00</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td></td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

This table now gives the probability that a given male condition (column) is true given receipt of a particular signal (row). Here the rows add to 1.0 because the sum of the fraction of healthy and sick males is all the males. Note that the a posteriori probability that a male is healthy after receiving a fast song is 0.636; this can be compared to the prior probability that a male was healthy of 0.50. Receipt of a fast signal moves the probability meter in the female’s head from 0.50 to 0.6365. Receipt of a slow song moves the needle down from the prior value of 0.50 to the a posteriori value of 0.333. Clearly some information has been obtained. The next steps compute how much information was provided in bits.

4. Using the first method listed at the beginning of this online, we can now compute the average mutual information provided by this signal set. This uses the formula:

\[
\overline{H}_T = \sum_i \sum_j P(C_i \text{ and } S_j) \log_2 \frac{P(C_i | S_j)}{P(C_i)}
\]

This uses the cell values in the second table, those in the third table, and the prior probabilities to compute the average information provided:

\[
\overline{H}_T = 0.35 \log_2 \left( \frac{0.636}{0.50} \right) + 0.20 \log_2 \left( \frac{0.364}{0.50} \right) + 0.15 \log_2 \left( \frac{0.333}{0.5} \right) + 0.30 \log_2 \left( \frac{0.667}{0.50} \right) = 0.067
\]

bits.

5. Using the second method, the relevant formula is:

\[
\overline{H}_T = \overline{H}(C) - \overline{H}(C | S)
\]

where

\[
\overline{H}(C) = \sum_i P(C_i) \log_2 P(C_i)
\]

and
We first compute $H(C) = 0.50 \log_2(0.50) + 0.50 \log_2(0.50) = 1.0$ bit. We next compute

$H(C|S) = -(0.35 \log_2 0.636 + 0.20 \log_2 0.364 + 0.15 \log_2 0.333 + 0.30 \log_2 0.667) = 0.933$ bits.

Thus $H_T = H(C) - H(C|S) = 1.0 - 0.933 = 0.067$ bits.

Note that $H(C)$ is the total uncertainty that a female faces about a male’s health before he sings. $H(C|S)$, (called the equivocation), is the residual uncertainty she faces about whether he is healthy after hearing him sing. It is the reduction in uncertainty after receiving the signal that is used here as a measure of signal effectiveness. Note also that it does not depend on any payoffs that might accrue to either party as a result of the signal exchange.

One can compute a relative amount of information received by dividing the actual amount by the maximum that could have been received. Complete resolution of uncertainty in this case would require $H(C) = 1.0$ bit. The average female obtains 0.067 bits by attending to song speed. Thus she reduces her uncertainty by $0.067/1.0 = 6.7\%$.

**Extreme examples**

1. **Random coding matrix:** What happens if signals are given randomly? What is mutual information in that case? The corresponding coding matrix will show equal values in all cells for a given column:

<table>
<thead>
<tr>
<th>Male Condition</th>
<th>Song Speed</th>
<th>Healthy</th>
<th>Sick</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

The joint probability matrix is then:

<table>
<thead>
<tr>
<th>Male Condition</th>
<th>Song Speed</th>
<th>Healthy</th>
<th>Sick</th>
<th>$P(S_j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>0.50 x 0.50 = 0.25</td>
<td>0.50 x 0.50 = 0.25</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>0.50 x 0.50 = 0.25</td>
<td>0.50 x 0.50 = 0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and the a posteriori probabilities are:

<table>
<thead>
<tr>
<th>Male Condition</th>
<th>Song Speed</th>
<th>Healthy</th>
<th>Sick</th>
<th>$P(S_j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>0.25/0.50 = 0.50</td>
<td>0.25/0.50 = 0.50</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>0.25/0.50 = 0.50</td>
<td>0.25/0.50 = 0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In this case, we can see that listening to song speed has no effect on a female’s estimate that a male is healthy: her
probability meter for a given male starts at 0.50 and remains at 0.50 after he sings. We would thus expect that no information has been exchanged. This is in fact what the subsequent computations show:

Using the second method, we again have that $\overline{H}(C) = 1.0$ bit, and now the equivocation, $\overline{H}(C|S) = - (0.275 \log_2 0.5 + 0.275 \log_2 0.5 + 0.275 \log_2 0.5 + 0.275 \log_2 0.5) = 1.0$ bit. As expected, $\overline{H}_T = \overline{H}(C) - \overline{H}(C|S) = 1.0 - 1.0 = 0.0$ bits. The relative gain in information is also 0%.

2. **Perfect Coding**: Consider the opposite extreme. What if healthy males always sing fast songs and sick males always sing slow songs. This is a case of perfect coding and the corresponding coding matrix will be:

<table>
<thead>
<tr>
<th>Male Condition</th>
<th>Song Speed</th>
<th>Healthy</th>
<th>Sick</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td></td>
<td>0.50</td>
<td>0.50</td>
</tr>
</tbody>
</table>

The joint probability matrix is then:

<table>
<thead>
<tr>
<th>Male Condition</th>
<th>Song Speed</th>
<th>Healthy</th>
<th>Sick</th>
<th>$P(S_j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>1.0 x 0.50 = 0.50</td>
<td>0.00 x 0.50 = 0.00</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>0.00 x 0.50 = 0.00</td>
<td>1.00 x 0.50 = 0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td></td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

and the a posteriori probabilities are:

<table>
<thead>
<tr>
<th>Male Condition</th>
<th>Song Speed</th>
<th>Healthy</th>
<th>Sick</th>
<th>$P(S_j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fast</td>
<td>0.50/0.50 = 1.00</td>
<td>0.00/0.50 = 0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>0.00/0.50 = 0.00</td>
<td>0.50/0.50 = 1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Prior Probabilities</td>
<td></td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

These signals move a female’s probability meter all the way to 1.0 (if the male sings a fast song) or all the way down to 0 (if he sings a slow song. Song speed resolves ALL uncertainty and there should be no equivocation after receipt of the signal. This is in fact what the subsequent computations show:

Using the second method, we again have that $\overline{H}(C) = 1.0$ bit, and now the equivocation, $\overline{H}(C|S) = - (0.0 + 0.5 \log_2 1.0 + 0.0 \log_2 0.0 + 0.0 \log_2 0.0 + 0.5 \log_2 1.0) = 0.0$ bits (since $\log_2 0.0 = 0.0$ by convention). As expected, $\overline{H}_T = \overline{H}(C) - \overline{H}(C|S) = 1.0 - 0.0 = 1.0$ bits. Since the amount of information gained is equal to the original uncertainty, the relative information gained in this case is 100%.

### Successive sampling and larger coding matrices

If a female samples the same male multiple times successively, she can gain further information with each successive song that he sings. The computations are similar except that the a posteriori probabilities computed for each alternative condition become the prior probabilities for the next song he sings. Although tedious, one can compute the final a posteriori probabilities after a series of male songs and compare them to the initial prior values to come up with a value for the average mutual information provided by that song series. For computing sequential sampling, or for matrices with large numbers of inputs and outputs, it is much easier to use matrix methods. Details on how to compute mutual information and other signal effectiveness measures using matrix methods are summarized in Web Topic 8.8.
Further reading


Introduction

The major problem with using receiver responses as an index of signal effectiveness is that responses confound the effects of the amount of information provided by a signal, the receiver’s estimates of prior probabilities, and the relative payoffs of alternative actions. A female faced with choosing between two displaying males may fail to discriminate between their displays a) because the differences are too small for her to detect, or b) it does not pay for her to expend the effort to compare them.

Signal detection theory provides tools for separating the roles of the amount of information in signals from the value of that information. It allows one to compute an index, called receiver sensitivity and denoted by \( d' \), that can be used as another measure of the effectiveness of a signal set. The following discussion assumes that the reader is familiar with the general approach of signal detection theory as summarized in Web Topic 8.4.

Logic of method

Consider a hypothetical example in which females seek to identify a healthy male instead of a sick male as a mate. All males sing songs and song speed varies continuously among males. However, the distribution of song speeds for healthy males has a higher mean value than that for sick males. Let male song speed be denoted by \( w \). The task for each female is to define a “red line” at some critical value \( w_c \) such that any male whose song speed exceeds \( w_c \) will be considered an acceptable mate and any male with a slower speed will be rejected. The optimal value of \( w_c \) will depend upon a given female’s estimates of the prior probabilities of sick and healthy males, and the relative payoffs to her of correct versus wrong decisions. Thus different females, or the same female at different times, may set different values of \( w_c \).

If the distributions of song speed for sick and healthy males are at all overlapping, a female invoking her particular \( w_c \) will make some correct choices and some errors. Let \( P_{hit} \) denote the fraction of time that a given female correctly selects a healthy male for a mate and \( P_{false\ alarm} \) denote the fraction of the time that she mistakenly selects a sick male for a mate because his song rate is greater than \( w_c \).

Now consider three different populations of males that vary in the degree to which the distributions of song rate for sick and healthy males overlap. In each example, a graph on the left will plot song speed (\( w \)) on the horizontal axis and the probability that a given type of male (sick or healthy) will sing that song speed on the vertical axis. In all cases, we shall assume the distributions are roughly bell-shaped with the same variances. The mean song speed for each distribution is the one under the peak value of the bell-curve’s vertical axis. Consider first a population (A) in which there is little or no difference in the mean values of sick and healthy male song speeds: the two distributions are completely overlapping. This is shown on the left graph below:
Suppose we select pairs of healthy and sick males at random from this population and record their songs. We then play the two songs back to a test female from that population and see which speaker she approaches. We do this multiple times with different pairs of randomly sampled males to get an estimate of how often she correctly selects the healthy males ($P_{hit}$) and how often she incorrectly selects sick ones ($P_{false\ alarm}$). The values of these two measures will depend on that female’s red line value of $w_c$. We plot the two values on the graph on the right and label it $w_{c1}$. We then select a second test female who is of a different age, has different nutritional condition, or because of different prior probabilities is likely to have a different red line value of $w_c$, and repeat our experiment. We then plot her values of $P_{hit}$ and $P_{false\ alarm}$ on the graph and label them. Adding values for more females allows us to see the relationship between $P_{hit}$ and $P_{false\ alarm}$ as the relative payoffs to females of right versus wrong decisions changes. The graph on the right is called a **receiver operating curve** or **ROC** graph.

If the distributions of song speed for sick and healthy males are completely overlapping, it will be impossible for females to make accurate discriminations between them using song speed: there is no correlation between song speed and health, and it should be obvious that attending to song speed provides no information to females. In this case, the ROC graph is a straight line as shown in this example: $P_{hit}$ and $P_{false\ alarm}$ remain proportional to each other at a fixed rate. Increasing one results in an equivalent increase in the other.

Next, consider a population (B) in which song speed is somewhat correlated with male health. This implies that the two distributions are not entirely overlapping, and there is thus a non-zero difference between their means. Let us call that difference $d'$.

If we now undertake playbacks of sick and healthy male songs to a series of females, we will get the plot of $P_{hit}$ versus $P_{false\ alarm}$ shown on the right. Now, the ROC relationship between $P_{hit}$ and $P_{false\ alarm}$ bends up towards the upper left corner of the graph. This means that for all values of $w_c$, the cost to a female in terms of numbers of false alarms is much lower for every correct choice than was the case in population A. This is because there is now a significant correlation between male song rate and male health, and the information provided by songs reduces errors in female decisions.
In population (C), the correlation between male song speed and male health is even stronger than in population (B). The difference between distribution means, \( d' \), is a much larger number and the curvature of the ROC plot towards the upper left corner of the graph is even stronger:

![Graph showing the correlation between male song speed and health in population (C)](image)

These examples suggest that one should be able to estimate the difference between the distribution means, \( d' \), by estimating the degree to which the curvature in the ROC plots deviates from the straight line expected when there is no correlation between signal and condition. And surprisingly, this measure of the amount of information can be obtained using receiver responses. Even more surprising is the observation that if both distributions are bell-shaped and have similar variances, any pair of \( P_{hit} \) and \( P_{false\ alarm} \) values will fall on only one possible ROC curve corresponding to only one \( d' \) value. This means that we could estimate \( d' \) from examining the \( P_{hit} \) and \( P_{false\ alarm} \) values of only a single female.

**Standard units**

The major point of measures of signal effectiveness is to be able to compare one signal set to another, or perhaps obtain an average value for how effective most threat signals or most alarm signals are. Clearly, one cannot compare \( d' \) values if the units for one signal set are in songs/second and another is in brightness of red plumage coloration. As long as the relevant distributions are bell-shaped (Gaussian) or can be made so with appropriate transformations, one can convert the \( w \) values in any distribution plot into \( z \) scores. This is a scaling widely used in statistics and computed as follows. If the mean of a normal distribution is \( \mu \), and its standard deviation is \( \sigma \), then the \( z \) score for \( w \) is

\[
z(w) = \frac{w - \mu}{\sigma}
\]

We can thus replot any original probability distribution of \( w \) values as a probability distribution of \( z(w) \) values. This distribution will have its maximum when \( z(w) = 0 \) (e.g. when \( w = \mu \)), and all \( z(w) \) values to the left of this peak will be negative (e.g. \( w < \mu \)), and all \( z(w) \) values to the right of the peak will be positive (\( w > \mu \)). The difference between the means of two \( z \)-scaled distributions, \( d' \), will then be given as a multiple of their common standard deviation (if it is the same for both), or as a multiple of their average standard deviation (if they are different). Because \( d' \) is measured in standard deviation units, decreasing the average standard deviation of the distributions is equivalent to increasing the distances between their means: either reduces overlap between the distributions, and thus reduces errors.

Let the means for the two probability distributions be \( \mu_1 \) for healthy males and \( \mu_2 \) for sick males. We wish to convert the \( w \) axis for each distribution into \( z(w) \) values. For the first distribution,
and for the second distribution and the same \( w \),

\[
z_2(w) = \frac{w - \mu_2}{\sigma}
\]

We note that

\[
z_2(w) - z_1(w) = \frac{\mu_1 - \mu_1}{\sigma} = d'
\]

which is the measure we seek. We can thus estimate \( d' \) if we can estimate \( z_1(w) \) and \( z_2(w) \) from observations of a female’s decisions.

**Applying the method**

Suppose we perform our playback experiments on a female using songs of sick and healthy males from the same population. We now have values for \( P_{hit} \) and \( P_{false\ alarm} \) for that female. Most statistics texts have tables in the back listing the area below a normal probability curve to the right or the left of some cutoff value of a \( z \) score. Usually the reader has a \( z \) value and wants to know the corresponding probability. In our case, we know the probability, but would like to know the corresponding \( z \) score. We thus locate the measured probabilities \( P_{hit} \) and \( P_{false\ alarm} \) in this table, and then find the corresponding values \( z_{hit} \) and \( z_{false\ alarm} \) respectively. Since these \( z \) scores are based upon the same \( w \), in this case that female’s \( wc \), we can use their difference to compute \( d' = z_{hit} - z_{false\ alarm} \). To provide a feeling for the scale of this measure, a receiver which correctly identifies both sick and healthy males 50% of the time (e.g., chance) has a \( d' = 0 \), that which is accurate 70% of the time has a \( d' = 1.04 \), that accurate 90% of the time has a \( d' = 2.56 \), and that accurate 99% of the time will have a \( d' = 4.65 \).

**Additional measures from signal detection theory**

An additional parameter of signal detection theory that can be extracted from \( P_{hit} \) and \( P_{false\ alarm} \) data is **bias**: this is the degree to which a female is conservative about accepting males, and thus avoids false alarm errors at the expense of having more miss errors. It thus reflects the value of information independently of the amount of information. The simplest measure of bias is the **criterion index** \( c \): it can be computed as \( c = -0.5 (z_{hit} + z_{false\ alarm}) \). A female that has no bias accepts equal numbers of false alarms and miss errors (e.g. \( P_{false\ alarm} = 1 - P_{hit} \)), and their bias \( c = 0 \). When females avoid false alarms, \( c > 0 \), and when they avoid misses, \( c < 0 \). For any observed combination of \( P_{hit} \) and \( P_{false\ alarm} \), \( c \) depends upon the distance between that point and the diagonal running from top left to lower right corner of the ROC plot.

It is also possible to estimate the **likelihood ratio** parameter \( \beta \), which is equal to the ratio of the likelihoods that a male is healthy to the likelihood that he is sick (see Web Topic 8.4 for derivation). It can be computed using \( \ln (\beta) = c \ d' \) if the female is making optimal decisions. Using hit rates and false alarm rates, we can rewrite this as \( \ln(\beta) = -0.5 \ [z_{hit} - z_{false\ alarm}]^2 \).
Non-normal distributions or unequal variances

If we know that the distributions of $w$ for healthy and sick males are normally distributed with equal variances, we saw that we do not have to compute an entire ROC curve to obtain estimates of $d'$, $c$, and $\beta$: instead, one pair of hit and false alarm rates will do. However, distributions may not be normal or have equal variances. The only way to detect this is to plot the ROC curve by obtaining data from multiple females or by manipulating one female's prior probabilities or payoff values. We can still compute a single $d'$, $c$, and $\beta$ from such a situation; however the analysis is more complicated than that given here. See MacMillan and Creelman (1991) for details.

Further reading


Web Topic 9.1
Equations of Change and Evolutionary Models

Introduction

A standard way to study dynamic systems is to derive a suitable equation that uses current values of one or more variables in the system (right-hand side) to predict the variable values in the next time interval (left-hand side). One then assigns initial values to the system variables, and uses the equation iteratively to track the trajectory of the system over successive time intervals. Because the output values for one application of the equation are then used as the inputs for the next iteration, the process is said to be recursive and the relevant equation is called a recursion equation. The right hand side of a full equation combines the current values of the variables with a rate of change term. In many cases, modelers make the rate of change term the left side and put the variables that affect that rate of change on the right side. This is defined as a difference equation when time intervals are discrete, and a differential equation for continuous time analysis. The variable values in the next time interval are obtained by adding the rate term to the prior values if using a difference equation, or by integrating over a time period given certain initial values if using a differential equation. A good introduction to the use of recursion and rate equations in ecological and evolutionary biology is provided by Nowak (2006).

Evolutionary modeling often uses recursion equations to track potential evolutionary trajectories. Approaches differ in whether the right hand side of the equation includes both genetic and environmental terms, and if both are present, how they are defined and weighted. They may also differ in whether they limit consideration to linear effects or instead allow for nonlinear processes (see Web Topic 2.8).

Some approaches use recursion equations to identify stable equilibria by looking for values of the system variables at which the rate of change equals zero and slight perturbations off the equilibrium will return to the equilibrium point. This is typical of evolutionary game theory modeling and some applications of quantitative genetics. In contrast, practitioners of adaptive dynamics are interested in all trajectories, whether they lead to equilibrium points or not. A number of fairly sophisticated tests can be applied to recursion equations to predict types of trajectories without having to actually track examples.

Below, we provide an introduction to several types of recursion equations used to model evolutionary processes. Each is given in the rate of change format. Only basic introductions to the methods are provided; all of them can become quite complicated as more details are added. Interested readers can consult the cited references for examples.

Quantitative Genetics

The key recursion equation in quantitative genetics is the breeder’s equation (Bulmer 1980; Falconer 1996). When only a single continuously variable phenotypic trait is considered, this is

\[ \Delta Z = h^2 S \]

where the response to selection, \( \Delta Z \), is the change in the mean phenotypic value of the trait between the parental and the offspring generations, \( h^2 \), is the narrow-sense heritability of the trait (measured as the ratio of additive genetic variation in the trait to the overall phenotypic variation in the trait), and \( S \) is the selection differential measured classically as the difference in the mean trait value for those that breed and the mean trait value in the parental generation.

This equation ignores the likely possibility that a focal trait might be affected both by direct selection on it and by indirect effects propagated to it by genetic correlations with other traits also under selection. Relevant genetic correlations would include pleiotropy and linkage disequilibrium. Assuming weak selection, resilience of the additive genetic variance in the face of selection, and roughly normal (Gaussian) distributions of trait values, a polygenic version
of the breeder’s equation is

\[ \Delta \mathbf{z} = \mathbf{G} \mathbf{P}^{-1} \mathbf{S} \]

where \( \Delta \mathbf{z} \) is now a vector containing the changes in mean trait value for each of a set of genetically correlated traits, \( \mathbf{G} \), is a square matrix containing the additive genetic variance for each trait along the main diagonal and the additive genetic covariances between traits in off-diagonal entries, \( \mathbf{P}^{-1} \) is the inverse of an equivalent matrix for phenotypic variation, and \( \mathbf{S} \) is a vector containing the direct selection differentials on each trait (Lande 1979, 1980, 1981).

The latter are typically computed as the covariance between each trait’s value and relative fitness. The equation can be further simplified by combining the product of \( \mathbf{P}^{-1} \mathbf{S} = \mathbf{\beta} \). \( \mathbf{\beta} \) can then be interpreted as a vector of partial regression coefficients of each trait on relative fitness. The values of \( \mathbf{\beta} \) can also be used to create an adaptive landscape with a particular combination of phenotypic trait values defining a location in that landscape and the height of the landscape at that point indicating relative fitness (Steppan et al. 2002; Arnold et al. 2008). For this reason, \( \mathbf{\beta} \) is referred to as a selection gradient in contrast to the vector of selection differentials in the polygenic \( \mathbf{S} \). Evolutionary trajectories that lead the mean of a population to a peak or ridge in that landscape are likely to end there since on top of a peak, \( \mathbf{\beta} = 0 \). This defines evolutionary equilibria in quantitative genetics models.

Ideally, this equation would be applied recursively to predict evolutionary trajectories of traits or set to zero to identify equilibria. One concern, however, is the degree to which \( \mathbf{G} \) (called the G-matrix) remains untouched by selection between generations (Dieckmann et al. 2006; Pigliucci 2006). Considerable effort is currently underway to measure and compare G-matrices in real systems, use simulations to see under which conditions the G-matrix might or might not be stable, and otherwise test the assumptions of the polygenic recursion equation (Steppan et al. 2002; Revell 2007; Roff 2007; Arnold et al. 2008; Calsbeek and Goodnight 2009). An alternative is to relax the assumptions used in the classic polygenic equation, and use a series of concurrent recursion equations to track both phenotypic and genotypic changes across generations (Barton and Turelli 1991; Bürger 1991; Christiansen 2000). This takes extensive computation, but increasingly robust methods are now available including a set of functions for use in the commercial program Mathematica (Kirkpatrick et al. 2002).

**Adaptive Dynamics**

Recursion equations have long been used in ecology to describe not only the equilibrium points but also the dynamics of competition, predator/prey cycles, speciation, and community stability (Ellner 2006; Pastor 2008). It was only natural that the initial efforts to identify equilibria in behavioral strategies (e.g., evolutionarily stable strategies [Maynard Smith 1982]) would be extended to track the dynamics of such systems (Metz et al. 1996; Hofbauer and Sigmund 1998; Nowak 2006). This has in turn encouraged researchers to seek general recursion equations that could be applied to all types of evolutionary processes. One candidate is the canonical equation of adaptive dynamics (Dieckmann and Law 1996). This differential equation assumes that evolution is ultimately limited by low rates and low magnitudes of mutations. This assumption ensures ample time for concomitant ecological dynamics to play out before significant mutational change occurs. It thus divides evolution into two successive time scales and simplifies the mathematics. The canonical equation for a single continuous trait divides the rate of change in the average trait value (\( \mathbf{z} \)) into two multiplicative components: an overall evolutionary rate coefficient, \( k \), and a selection derivative, \( D \):

\[
\frac{dz}{dt} = kD
\]

where \( k \) equals a scaled product of equilibrium population size, mean mutation rate/birth, and the variance in mutation rate for that trait, and \( D \) equals the change in fitness resulting from deviations of the trait value away from the current
population mean. The latter is usually written as a partial derivative of fitness with respect to trait values and can be envisioned graphically as the slope of an adaptive surface in the vicinity of the current population mean. **D** is thus similar to the **β** of quantitative genetics. If **D** = 0, the population is at a singular point which may or may not be an equilibrium. In fact, at least eight theoretical, and in practice six realizable, types of singular points are possible for even a single continuous trait (Geritz et al. 1997; Apaloo et al. 2009). An evolutionarily stable strategy (ESS) is only one of these: it is defined as an outcome in which the most common strategy has the trait values defined by the singular point and no alternative strategy when rare can invade that population (Maynard Smith 1982). Another possible outcome is a convergently stable strategy (Eshel 1983). This is a singular point at which any mutant whose strategy is more similar to the singular point than the currently common strategies in a population can invade that population. Adaptive dynamics analyses show (surprisingly) that an ESS need not be convergently stable, and a convergently stable singular point need not be an ESS (Eshel 1983; Eshel and Feldman 1984; Geritz et al. 1997; Geritz et al. 1998). Another surprising outcome is possible bifurcation or branching of a population near a singular point into two coexisting strategies in the form of a stable polymorphism or even speciation. Simply looking for stable equilibria would usually miss most of these additional evolutionary possibilities.

As with the breeder’s equation, the canonical equation of adaptive dynamics can be generalized to accommodate simultaneous tracking of multiple traits. In the multiple trait version, \( \frac{dz}{dt} \) is now the rate of change in a vector (\( z \)) of values for the set of traits, the mutation rate in \( k \) depends on the values of the entire current trait vector (\( z \)) and the variance in the single-trait equation is here replaced by a variance/covariance matrix describing mutational correlations between traits. **D** is now a multidimensional measure of the fitness gradient around the current population mean (Dieckmann et al. 2006; Durinx et al. 2008; Apaloo and Butler 2009; Leimar 2009). Methods have been developed that allow one to predict patterns of stability or instability simply by examining the form of **D** (Leimar 2009).

This approach has been criticized because the rates of mutation assumed in the models are usually far smaller than is actually found in real systems (Abrams 2001, 2005; Barton and Polechova 2005; Abrams 2009). In addition, studies of quantitative trait loci indicate that some important traits depend on only a few genes and at least some of these appear to have arisen through mutations of large effect (Lynch and Walsh 1998; Roff 2007; Kelly 2009). However, the fact that applications of adaptive dynamics often give predictions similar to those of quantitative genetics suggests that its assumptions may not be that limiting (Leimar 2009).

### The Price Equation

The Price equation (Price 1970, 1972) has quite general applicability. It is used to dissect the difference in average values of some property in two populations into additive components. The focal property considered can be anything: alleles, phenotypic traits, learned behaviors, types of poetry, etc. In evolutionary applications, the relevant populations are a parental generation and an offspring generation, and the property is usually some index of allele frequencies or phenotypic traits. The general form of the equation is

\[
\Delta \bar{z} = \text{Cov}(w,z) + E(w\Delta z)
\]

where \( z \) is the value of the property of interest in a parental individual, \( w \) is that individual’s relative fitness (e.g., its absolute fitness divided by the mean fitness of the parental population), \( \Delta z \) is the difference between a parent’s property value and that of its offspring, and \( \Delta \bar{z} \) is the difference in mean property value between the offspring and parental populations. The \( \text{Cov}(w,z) \) term is the covariance between relative fitness and property values in the parental population, and is thus equivalent to the selection component of the breeder’s equation and adaptive dynamics.

The second term, \( E(w\Delta z) \), is called the transmission component, and is the weighted average (based on relative fitnesses) of the difference between a parent's property value and that of its offspring. If the property is allele frequencies, and parents produce gametes that accurately reflect their own allelic frequencies, random variations away from parental patterns summed across all parents in the population will usually cancel out and the second term will be negligible (Frank 1997; Grafen 2000, 2006, 2007b). Where this is true, the Price equation and phenotypic approaches such as evolutionary game theory converge on a common strategy for predicting evolution: in both cases, selection can
then be described as an optimization process in which relative fitness is the payoff being maximized (Grafen 2000, 2002, 2007b, 2008). The transmission component will not be negligible if a species has sex ratio or other allelic distorters during gamete production. It then must be included in modeling. It can also serve as an "error term" that provides better matches between model predictions and real systems when the underlying genetics are complicated. For example, by defining the relevant property as some higher power version of allelic frequencies, the transmission term can be used to include epistasis, nonrandom mating, and other processes into models (Frank and Slatkin 1990; Frank 1995, 1997); most quantitative genetic and adaptive dynamics models avoid these complications. Finally, because the Price equation is linear and thus additive, either the selection term or the transmission term (or both) can be further partitioned into additive sub-components to build more complicated models. For example, the selection term can be divided into group and individual components for modeling group selection processes (Wilson 1975; Wade 1985; Queller 1992).

There is a cost to the great flexibility of the Price equation: its application only provides information about the mean property values in the next generation; it cannot compute the variances or covariances in the offspring generation that would be required to predict a subsequent generation (Barton and Turelli 1987). This has led a number of critics to question the utility of the Price equation in evolutionary modeling (Lewontin 1974; Gould and Lewontin 1979; Ewens 2004). However, it should be pointed out that the only reason alternative approaches such as quantitative genetics and adaptive dynamics can plot multiple generation trajectories is because they make some rather stringent assumptions about the underlying genetics (e.g., constancy of the G-matrix in quantitative genetics and low mutation rates in adaptive dynamics).

A clear example of how to use the Price equation for a single locus situation can be found in Box 3, p. 1246 of (Grafen 2007b). For a two-allele locus in a diploid animal, Grafen suggests assigning each individual $i$ a "p-score," $p_i$, which takes the value 1 if that individual has the same focal allele on both chromosomes, 0.5 if it has the focal allele on only one chromosome, and 0 if it lacks the focal allele on both chromosomes. The left side of the Price equation, $Δp$, then represents the change in the mean value of the focal allele in the population across generations. Like the other methods, the single trait version of p scores can be generalized by tracking the dynamics of several loci concurrently. Each individual has only one relative fitness value, $w_i$, no matter how many loci and associated alleles are considered. It is thus feasible to expand the computation of each individual’s p-score as a linear sum of the values assigned to alleles at each of many loci (Grafen 2008). This sum will be similar to the combined additive genetic components of traits used in quantitative genetics models. Because the Price equation is linear and additive throughout, it is possible to accommodate uncertainty into appropriate terms by weighting alternative values by the probabilities each will occur, and adding these products together to obtain an expected value (Grafen 2000). Extensions of the Price equation approach have also been developed to accommodate cooperative behavior among relatives, different classes of individuals (age, sex, or status) within evolving populations, social interactions in networks, and evolutionary competition between groups (Frank 1997, 1998; Grafen 2006, 2007c, a; Grafen and Archetti 2008; Frank 2009; Gardner and Grafen 2009; Grafen 2009).

**Literature Cited**


Computing the Value of Information

Introduction

Animals rely on a variety of sources of information in making a decision about subsequent actions. Likely sources include genetic biases, probabilities of events or conditions occurring based on prior sampling, concurrent or recent cue information, and signals. Some sources of information are more costly to exploit than others. We expect evolutionary selection to favor receiver attention to combinations of sources where benefits outweigh costs on average. The value of information provides a measure to compare two or more alternative combinations of information source. In biology, it is usually given as the difference in average fitness between two alternative strategies. The concept originated in economics (Savage 1954; Good 1966; Raiffa 1968; Gould 1974; Chavas and Pope 1984; Ramsey 1990), but has recently been applied to a number of problems in evolutionary biology (Stephens 1989; Bradbury and Vehrencamp 2000; Koops 2004; Lachmann and Bergstrom 2004; Szalai and Szamado 2009; Donaldson-Matasci et al. 2010; McLinn and Stephens 2010; McNamara and Dall 2010; Schmidt et al. 2010).

In the context of animal communication, one might want to compute the value of information when an animal uses a particular signal set versus when that same party does not use it. In this case, we would compare the average fitness of relying on that signal set with the average fitness derived from using some non-signal combination of information sources to make the same decisions. Alternatively, we might consider the value of information provided by one type of signal set, say continuous and graded signals, when compared with an alternative discrete set. Each strategy would provide its own average amounts of information, have its own reliabilities, and result in different performance costs. The value of information would combine the benefits and costs of each strategy and compute the difference between them. In the following sections, we first apply the value of information to compare reliance by a receiver on a discrete dyadic signal set with the alternative strategy of ignoring those signals. This will demonstrate some of the utility of the approach. We will then review more formal and complex applications and conclude with some Matlab M-files for computing the value of information with more than two alternative signals and actions.

The value of information for discrete dyadic signal sets

Consider a situation in which a receiver must make a decision whether to perform action $A_1$ or $A_2$. $A_1$ yields a higher payoff when condition $C_1$ is true, and $A_2$ has the highest payoff when condition $C_2$ is true. We shall assume that this information is already available to the receiver. The actual prior probabilities that $C_1$ and $C_2$ occur can be denoted by $p$ for $C_1$ and $(1-p)$ for $C_2$. We can summarize this information in a payoff matrix with the priors indicated in a lower row:
Different receivers may have access to different information prior to making a decision. A receiver that ignores signals and cues would rely only on its best estimates, based on prior sampling, that \( C_1 \) and \( C_2 \) occur on average. If it does this well, these will be close to the prior probabilities listed in the table. A receiver that uses its prior estimates and any recent cues might have somewhat different information on which to base its decision. And a receiver that has access to signals may have different information. Each type of receiver then makes a decision on what to do next (\( A_1 \) or \( A_2 \)). We shall denote the fraction of time it makes the correct decision, (e.g. the one that gives the highest payoff), when \( C_1 \) is true by \( \phi_1 \) and the fraction of time it makes the right decision when \( C_2 \) is true by \( \phi_2 \). We can add this to our table as follows:

<table>
<thead>
<tr>
<th>Action</th>
<th>Condition</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_1 )</td>
<td>( R_{11} )</td>
<td>( R_{12} )</td>
<td></td>
</tr>
<tr>
<td>( A_2 )</td>
<td>( R_{21} )</td>
<td>( R_{22} )</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Priors</th>
<th>( p )</th>
<th>( 1-p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>( \phi_1 )</td>
<td>( \phi_2 )</td>
</tr>
</tbody>
</table>

The average payoff to a receiver making a decision given the variables in this table is:

\[
PO = p(\phi_1 R_{11} + (1-\phi_1)R_{21}) + (1-p)(\phi_2 R_{22} + (1-\phi_2)R_{12})
\]

The values of the priors and payoffs are assumed to be the same for all receiver strategies; what differs are the values of \( \phi_1 \) and \( \phi_2 \) resulting from their particular access to information. We can compute the value of information for any two strategies by finding the difference in the average payoffs (using the above formula) for the two strategies.

Suppose we have two alternative strategies that differ in the nature and amount of information used in making the same decision. We assume the payoffs of different actions and the prior probabilities of conditions \( C_1 \) and \( C_2 \) are fixed. Suppose receiver strategy \( S_1 \) uses one set of information sources at some cost of getting that information of \( -K_1 \), (where \( K_1 \geq 0 \)) and makes correct decisions fractions \( \phi_1 \) and \( \phi_2 \) of the time. Its average payoff per decision is:

\[
PO(S_1) = p(\phi_1 R_{11} + (1-\phi_1)R_{21}) + (1-p)(\phi_2 R_{22} + (1-\phi_2)R_{12}) - K_1
\]

Other receivers use a different strategy \( S_2 \) which relies on a different set or weighting of information sources at a different sampling cost \( -K_2 \) resulting in different probabilities of correct choices \( \theta_1 \) and \( \theta_2 \). Their average payoff is:

\[
PO(S_2) = p(\theta_1 R_{11} + (1-\theta_1)R_{21}) + (1-p)(\theta_2 R_{22} + (1-\theta_2)R_{12}) - K_2
\]

The value of using different information, say for using \( S_2 \) instead of \( S_1 \), is then:

\[
VI(S_2,S_1) = PO(S_2) - PO(S_1)
= p(\theta_1 - \phi_1) (R_{11} - R_{21}) + (1-p)(\theta_2 - \phi_2) (R_{22} - R_{12}) - (K_2 - K_1)
\]

We can simplify this further by letting \( \Delta R_1 = (R_{11} - R_{21}) \) and \( \Delta R_2 = (R_{22} - R_{12}) \).

If \( VI(S_2,S_1) \) is positive, then \( S_2 \) will be favored by selection over \( S_1 \). If it is negative, then \( S_1 \) is favored over \( S_2 \). By
Signaling versus ignoring signals

Let us now apply the general result above to a specific situation. Consider two classes of receivers of the same species. They are faced with the same decision, have exactly the same prior estimates of which condition is most likely true, and would get the same payoffs for each combination of action and condition. Strategy \( S_1 \) ignores all signals. As discussed in the text, it will optimally invoke a “red line” approach in which it always does \( A_1 \) if its current estimate that \( C_1 \) is true is above a threshold value, and it always does \( A_2 \) if it is not. In the text, the adopted action is called the “default strategy.” The value at which the red line is set depends on the relative differences in payoff between correct versus wrong decisions for the two conditions (see Web Topic 8.1). Suppose that prior sampling, recent cues, and current payoff value assessments place the current estimated probability that \( C_1 \) is true above the red line. Receivers using the \( S_1 \) strategy will always do \( A_1 \). This means that it always makes a correct decision when \( C_1 \) is true, but always makes the wrong decision when \( C_2 \) is true. For this strategy, \( \phi_1 = 1 \) and \( \phi_2 = 0 \) (note that we could have assumed that \( A_2 \) was the preferred action for \( S_1 \) receivers; the conclusions below would be similar).

Now consider an alternative strategy \( S_2 \) which has access to all the same prior and cue information that an \( S_1 \) receiver uses, but in addition, \( S_1 \) receivers attend to signals produced by senders that are correlated with whether \( C_1 \) or \( C_2 \) is currently true. Again, we denote the reliabilities of these receivers’ decisions by \( \theta_1 \) and \( \theta_2 \). We shall also assume that \( S_1 \) receivers expend the same costs as \( S_1 \) receivers to monitor cues and sample past events. The critical difference is then the additional costs that an \( S_2 \) receiver expends to attend to and process signals.

The value of information for a receiver that switches from \( S_1 \) to \( S_2 \) will then be:

\[
VI(S_2, S_1) = p(\theta_1 - 1)\Delta R_1 + (1 - p)(\theta_2 - 0)\Delta R_2 - (K_2 - 0)
\]

The first term on the right hand side of this expression will be negative unless reliability using signals is 100%, in which case the term will equal zero. This makes sense as a receiver using \( S_1 \) will always make the correct choice when \( C_1 \) is true whereas one using signals, which are usually imperfect, will have a reliability less than 100%. Thus an \( S_1 \) receiver that switches to attending to signals will make more errors when \( C_1 \) is true than it did before. On the other hand, the second term in the expression is likely to be positive since an \( S_1 \) receiver never makes the right decision when \( C_2 \) is true, whereas an \( S_2 \) receiver at least gets it right \( \theta_2 \) of the time. The third term will be negative as long as there are additional costs of attending to signals. For \( S_2 \) to be favored by selection, the middle term must more than compensate for the two negative terms.

Note that increasing overall reliability, for example by improving a receiver’s sensory equipment or spending more time sampling signals, will increase both \( \theta_1 \) and \( \theta_2 \). This will increase the magnitude of the positive term while reducing the magnitude of the first negative term. Taken alone, these effects will increase the likelihood that \( S_2 \) will be favored over \( S_1 \). However, it should be obvious that some minimal level of reliability must be attained before the positive term exceeds the negative one. In addition, it is likely that the costs of participating in communication \( -K_2 \) will have to rise to accomplish any improved reliability. Whether \( S_2 \) is favored or not will then depend on which terms rise faster.

We can show both points graphically as follows. Suppose the reliabilities are the same for \( C_1 \) and \( C_2 \); e.g., let \( Q = \theta_1 \)
= \theta_2$, and let us ignore the cost term for the moment. Any benefit of using signaling over ignoring signaling can then be written as:

\[ B(S_2, S_1) = p(Q - 1)\Delta R_1 + (1 - p)(Q - 0) \Delta R_2 \]

or rewriting

\[ B(S_2, S_1) = Q(p\Delta R_1 + (1 - p)\Delta R_2) - p\Delta R_1 \]

This is an equation for a straight line relating $B(S_2,S_1)$ to $Q$ with slope of $p\Delta R_1 + (1 - p)\Delta R_2$ and intercept $-p\Delta R_1$. On a graph (red line) this is:

![Graph showing the relationship between $B(S_2,S_1)$ and $Q$, with a straight line indicating the benefit equation.](image)

This shows that average reliability must be at least equal to $Q_c$ before there can be any benefit to attending to signals.

Now we can add the absolute value of the costs, $K_2$, to the same graph. Costs are likely to rise with $Q$ in an accelerating way: it may not take much investment to improve reliability a bit, but the remaining error becomes increasingly costly to remove. Adding such costs (blue line) to the benefit plot gives:

![Graph showing the benefit and cost curves for $B(S_2,S_1)$ and $K(S_2,S_1)$, with the benefit line showing the cost of improved reliability rising faster than the benefit.](image)

The value of information to an $S_2$ receiver is the difference between the red curve and the blue curve. In this example, it is maximal at an intermediate reliability value ($Q_{opt}$) which is greater than the minimal threshold value at which the value of information becomes positive. If the blue curve were to rise faster, this would push the optimal $Q$ to lower values, and eventually the red curve would be entirely beneath the blue curve: costs of improved reliability rise faster than the benefits and selection will not favor $S_2$ receivers over $S_1$ receivers.

We can derive similar functions, and similar graphs for senders. The value of information for senders is similar to that of receivers except for the relative payoffs of right and wrong decisions and the costs of participating in communication; prior probabilities and the reliabilities for a given signal exchange would be the same. Because the payoffs differ, the slope and intercept for the benefits line may differ between senders and receivers. Similarly, the cost curve for increasing $Q$ is likely to differ for the two parties. This can create a different optimum $Q$, and thus different optimum investments for sender and receiver. See Bradbury and Vehrencamp (2000) and Koops (2004) for further details on
More complex models and alternative applications

Here, we summarize several recent publications using the value of information in behavioral ecology and ecology. Stephens (1989) resurrected an earlier paper in economics by Gould (1974) defining the value of information. Whereas our example above assumed discrete conditions, signals, and actions, Stephens lets alternative actions be continuous (at least up to an interval scale), allowing him to approximate the costs of small deviations from the optimal action for each condition with a Taylor expansion. His analysis shows that the value of information depends upon the variance in optimal actions and, verifying a claim by Gould, not on the variance of alternative conditions (unless conditions and actions are sufficiently linked that increasing condition variance automatically increases optimal action variance). His approach allows one to predict whether the value of information increases or decreases if the relative shapes and heights of the payoff function curves for different conditions are varied.

Koops (2004) adopted Stephen’s notation and approach and divided the value of information into the value of perfect information and the value of incorrect information. This allowed him to examine the ratio of additional benefits to additional costs experienced by an animal relying on signals when compared to one using a default strategy. He then computed the minimum reliability required to justify engaging in communication both with and without additional costs of signaling and reception (\(K_2\) and \(K_1\) in our model above), and assuming reasonable scalings of the latter costs with reliability, derived optimal reliability values for each party.

Lachmann and Bergstrom (2004) use the value of information to compare non-combinatorial with combinatorial signaling schemes. Their analysis demonstrates the greater vulnerability of combinatorial coding schemes to potential deception. Donaldson-Matasci et al. (2010) focus on cues and community ecology, but their conclusions are just as applicable to animal signals. They note that when the value of information is based on relative growth rates within populations across generations, increases in cue or signal reliability cause changes in the value of information that scale identically with the corresponding measures of the amount of information provided. Kåhre (2002) also argues that alternative weightings of reliabilities, properly defined, can preserve the relative scaling and properties of the unweighted reliabilities. McNamara and Dall (2010) prove that the benefit component of the value of information is never negative if the receiver is using Bayesian updating of cues and signals. McLinn and Stephens (2010) apply value of information models to experiments with jays and find that both the reliability of signals and the uncertainty of the alternative conditions play roles in determining whether receivers attend to signals or not. Finally, Schmidt et al. (2010) provide a broad view of the uses of information and its relative value at different levels in natural ecosystems.

Sample Matlab Routines for Discrete Signal Systems

Below, we provide some Matlab routines for computing the value of information for discrete signal and discrete action systems with more than two alternatives each. The contrast is between using a given signal set and not using it. The primary M-file, VI, requires access to several other M-files: Bayes, AbsRel, and getPC. The latter tries to estimate the minimum probability for each alternative condition, holding the others fixed at their prior relative probabilities, at which a receiver should switch to the corresponding optimal action. Note that this is just a guess: relative probabilities when there are more than 2 conditions can vary in complicated ways with successive sampling and updating. However, the computed value of information does not depend on these computations.

```
function B=Bayes (S,P)
  \% S is the coding matrix with conditions as columns and signals as rows
  \% P gives prior probabilities as row vector with columns as conditions
  \% D.*S computes the "Condition AND Signal" matrix from S and P
  \% S*P' computes the row totals in the AND matrix and thus the total
  \% fraction of time that each signal is given across all conditions
  \% The last half of the expression divides each AND cell value by
  \% the corresponding row total; this is the actual Bayes calculation
  \% Values in the output matrix cells are a posteriori probabilities of
  \% each condition being true (rows) after having received the signal assigned
```
% to that column. Note reversal of axis assignments from S matrix.
% Inclusion of signals that are never used results in 0 Bayesian estimates.

[a b] = size(S);
D = ones(a, b);
DD = D;
for i = 1:a
    D(i,:) = P;
end
AND = D .* S;  % Compute joint (AND) matrix of conditions and signals
SS = S * P';  % Compute total fraction of time each signal is given
for i = 1:a
    if (SS(i) > 0)
        SP(i) = 1 / SS(i);
    else
        SP(i) = 0;
    end
end
for i = 1:b
    DD(:, i) = SP;
end
B = AND .* DD;  % Compute a posteriori prob of conditions given signal
B = B';  % Reverse axes so that original inputs and outputs reversed
% and columns add to 1.0.

function [M N G] = AbsRel(S, P)
% This routine takes a coding matrix S, in which inputs are listed as
% columns and outputs are listed as rows, and a horizontal vector P
% which summarizes the prior probabilities of each input listed in S,
% and computes the a posteriori probabilities that a receiver using this
% signal set and Bayesian updating will assign to each alternative input
% option (rows) when a given input (columns) is in fact true.
% These probabilities are summarized in a square matrix G in which
% the main diagonal reports correct assignments, and off-diagonal
% values are the probabilities of erroneous assignments.
% If the output is given as the vector [M N G], a number of values is
% output. M is the weighted average (by P) of the probabilities that
% a receiver will identify the correct input after attending to
% signals and updating. It will equal 1.0 for perfect coding, and
% the reciprocal of the number of inputs for chance coding matrices
% (ones with the same value in all cells). N is a weighted average of the
% changes in probabilities from priors when using signals compared to the
% maximal change that would occur with perfect coding.
% It will vary between 0 and 1. It is assigned a value of 0 if any
% prior probability is 1.0. If no assignment is given for this routine,
% only M is generated.
% If several coding matrices are involved in a sequence, e.g. if there
% is a sender matrix SS, a transmission matrix T, and a receiver matrix
% R summarizing how receivers assign transmitted signals to expected
% categories, the input to this routine should be the chain product
% S=R*T*S, and the vector P lists the prior probabilities for the inputs
% to SS. Note that this routine requires access to the Bayes function.
% Jack Bradbury October 6, 2009

G = Bayes(S, P) * S;
GG = diag(G);
M = P * GG;
N = 0;
[a b] = size(P);
ON = ones(a, b);
if min(P) < 1.0
    W = (GG' - P) ./ (ON - P);
    N = W * P';
end

function [T MN I] = getPC(R, P, C)
% This routine takes a payoff matrix R in which alternative conditions are
% listed as columns and alternative actions as rows. Cell values are payoffs
% of performing a given action when a given condition is true. Prior
% proabilities for each condition are given in horizontal vector P.
% Routine finds minimum probability for a condition that must be true before
% it pays to adopt that action which gives the maximal payoff for that condition.
% Solution varies one probability holding relative values of others fixed
% until average payoff of the corresponding action is the maximum among the
% alternative actions. Result T is a matrix in which each row gives a
% probability vector for the alternative conditions in which the main
% diagonal gives the minimal probability for that condition to elicit that
% action. MN gives the minimal probability that must be true for each column
% before the action giving the highest PO for that column will be adopted.
% I is the row that has the maximum payoff in R. C is an argument defining
% how refined the search is.

[a b]=size(R); % get size of payoff matrix
[X,I]=max(R); % Find action (row) with max payoff for each column
T=zeros(a,b); % set up empty final T matrix
MN=zeros(1,b); % set up empty final M matrix
for i=1:b % start one of b runs varying one prob while others held constant
    G=P; % Assign prior prob to each of b cells
    G(i)=0; % Set focal cell to 0 to begin sampling
    SG=sum(G); % Compute sum of remaining cell values in vector
    G=G/SG; % Standardize remaining cell values so sum is 1.0
    j=0;
    D=1/C; % define increments to increase focal cell probability
    test=1;
    while (test~=0)
        j=j+1;
        PP=j*D; % define increment at this step
        PQ=G*(1-PP); % decrease non-focal cell values by proportional amounts needed
        % to insure that sum of b cells is 1.0 once focal cell is augmented in next
        % step
        PQ(i)=PP; % Augment focal cell to next step value
        V=R*PQ'; % Compute average payoff given this set of cell values
        [X,II]=max(V); % Keep track of which combination was maximal
        test=ne(I(i),II);
    end
    T(i,:)=PQ; % Move on to next one but store maximum row in T
end
for i=1:b
    MN(i)=T(I(i),i);
end

function [MV D VD]=VI(S,P,R)
% This function computes the value of information, V, in which average
% fitness using a signal set is compared to a default strategy not using
% that signal set. It depends on a coding matrix S with conditions as
% columns and signals as rows, prior condition probabilities as a
% row vector P, and a payoff matrix R in which alternative conditions
% are listed as columns, alternative actions as rows,
% and cells give payoffs. Ideally, payoff matrix reflects changes in payoffs
% above some standard average fitness. Thus negative values imply decreases
% in fitness and positive ones are increases due to taking that action when
% that condition is true.
% The routine computes the average payoff for each alternative action given
% priors and assigns action with maximal average payoff to default strategy.
% This is action animal always takes in absence of signals (or changes in
% priors).
% It then holds relative probabilities of all but one condition fixed, and
% varies probability of remaining one from 0 up until adopting the action
% with the highest payoff for that condition just becomes the optimal
% strategy. This is the threshold probability for a switch from default to
% that strategy.
% Routine then computes a posteriori probabilities of each condition when
% attending to signals. These are compared to threshold values and
% differences are output as matrix D. All positive values in D mean signals
% led to updated probability greater than threshold-meaning signals should
% change choice of action.
Routine then computes reliability values: probabilities that receiver will assign signals to particular conditions and then take corresponding actions.

Finally, routine computes average payoff when attending to signals. This is the sum of the products between a particular payoff and the reliability probability that that action will be instigated with a given condition is true. The difference between each of these sums and the payoff if the animal only adopted the default strategy are output in the vector VD. These values for each condition are next discounted by their prior probabilities to compute the overall average payoff. The value of information is this value minus that when no signals are used and the receiver always uses the default strategy.

Computation ignores performance costs to either party of communicating.

C=50000;%Set high limit to estimate threshold probabilities
[a b]=size(S);
[V I]=max(R); %Identify optimal actions for each condition in R
[T M J]=getPC(R,P,C); %Find minimal probabilities to adopt actions
NI=R*P'; %Compute average payoffs of all actions with no signals
[NPO K]=max(NI); %Compute PO and identify default action
B=Bayes(S,P); %Compute a posteriori probabilities of using this code
MM=zeros(a,b);
for i=1:a
    MM(i,:)=M; %Create threshold matrix
end
D=B-MM'; %Compute differences in a posteriori prob and thresholds
[A1 B1 RR]=AbsRel(S,P);
W=sum(R.*RR); %Compute products of reliability prob and payoff for each cell
VD=W-R(K,:);
IPO=W*P'; %Compute overall average payoff when using signal system
V=IPO-NPO; %Compute value of information as difference

titles=char('Default Action: ','Aver PO no Signals: ','Aver PO w/ Signals: ','Value of Info: ');
VV=num2str([K;NPO;IPO;V],'%4f');
MV=[titles VV];

**Literature Cited**


Web Topic 10.1
Estimating Evolutionary Trees

Introduction

The evolutionary history of a group of species is called its **phylogeny**, and a **phylogenetic tree** is a graphical summary of this history. The tree describes the pattern and in some cases the timing of events that occurred as the group radiated into new species. The tree also documents which organisms are more closely related, and is an important tool for naming species and genera. Although we sometimes have a fossil record to examine certain morphological traits in some animal groups, this record is fragmentary, and we cannot obtain DNA sequence data except for very recently extinct and well-preserved specimens. We therefore must estimate the phylogenetic tree using extant species. This online unit first reviews the basic logic and methods for estimating trees. It then gives the justification and protocols for the two main uses of trees in animal communication studies: a) correcting for phylogenetic inflation of samples when correlating communication systems with ecological contexts and selection pressures; and b) mapping behavioral or communication signal traits on the phylogenetic tree to infer the sequence of trait evolution. This online unit draws extensively from Chapter 4 of Freeman and Herron’s textbook on evolutionary analysis (Freeman and Herron 2007) and from Hall’s useful manual (2004), and the interested reader should consult these texts for further details.

The logic of phylogenetic inference

The basic logic of inferring evolutionary relationships is that more closely related taxa should have more traits in common. We group species by their similarities and distinguish groups by their differences. In principle, all traits that have a genetic basis and that vary among taxa could be used to assess similarities and differences. Traditionally evolutionary biologists used morphological traits such as bone structures, bristles, and mouthparts, and mode of embryonic or larval development. In the last 25 years, genetic traits such as the presence of certain genes or alleles, the sequence of nucleotides in a particular gene, or the sequence of nucleic acids in a particular protein, have been used with increasing effectiveness. But some traits, loci, DNA/RNA sources, and proteins, are more informative than others.

**HOMOLOGY**

Homologous characters are those that are similar due to descent from a common ancestor. Phylogenetic reconstruction must of course be based only on homologous traits. The subset of homologies that are most useful for estimating phylogenies are called **synapomorphies**, which are similar because they are modified from a common ancestor and different from the next more closely related taxon. Synapomorphies help us identify **monophyletic groups**, also called clades or lineages (Figure 1). Synapomorphies are also useful traits because they identify evolutionary branching points. When two populations of a species become isolated geographically and start to evolve independently, some of the homologous traits in each population will start to diverge due to mutation, selection, and drift. These changed traits are synomorphies that distinguish the two independent lineages. Furthermore, synapomorphies are nested. Over time, multiple branches occur. Each branching event adds one or more shared, derived traits, and the result is a hierarchy of synapomorphies. The clustering of synapomorphies graphed in this way is called a **cladogram**. A cladogram shows the pattern, or ordered sequence, of cumulative evolutionary change.
PROBLEMS IN RECONSTRUCTING PHYLOGENIES

To accurately reconstruct a tree, researchers must identify characters that qualify as synapomorphies. This is not always easy to do, since traits that appear to be similar in two species or taxa may have evolved independently. In this case, the traits were not derived from a common ancestor, so they are neither homologous nor true synapomorphies. One reason for these superficial similarities is **convergent evolution**, which occurs when natural selection favors similar structures as solutions to an environmental factor. For example, both octopuses and ray-finned fishes possess a camera lens eye, because both depend on excellent vision to find food and detect predators. But the eye structure has evolved independently in the mollusk and the early vertebrate. Convergences can also occur for genetic traits, especially for DNA sequences, where there are only 4 possible nucleotide bases (A, G, T, C) and by chance they can be the same in short sequences for species that are completely unrelated. Mistaking a similar trait for a synapomorphy when it is in fact a convergence constitutes a serious problem for figuring out which species are most closely related.

A second problem is **reversal**, where a trait that has evolved in a lineage is subsequently lost in a terminal species. Reversals are also common in DNA data, for the same reason mentioned above concerning the likelihood of a mutation back to an earlier nucleotide base given that there are only 4 possibilities. Convergence and reversal are lumped under the term **homoplasy**. If similar traits are not due to homology then they are due to homoplasy. Homoplasy represents noise in the datasets used to reconstruct phylogenies.

The most effective way to distinguish homology from homoplasy is to use many different traits, or very long sequences, in reconstructing evolutionary relationships. For example, ray-finned fish and other vertebrates have a bony skeleton and a variety of other traits that distinguish them from octopuses and other mollusks. One would have to assume that all of these traits had also changed in concert with the eye to draw the conclusion that fish and mollusks were closely related. It requires many fewer changes to propose that camera lens eyes evolved twice in a convergent process. This type of argument illustrates the principle of **parsimony**. Parsimony is a general logical criterion whereby simpler explanations for a phenomenon are preferred over more complex ones. When applying parsimony to phylogenetic inference, we accept the tree that involves the fewest changes. While parsimony is logically appealing, sometimes it does not work. For example, trait loss (reversal to an ancestral condition of no trait) could be relatively common for
extravagant sexually selected traits that evolved in an ancestor and proved to be very costly to bearers in some populations or species. Researchers therefore need to be very careful in selecting traits for phylogenetic reconstruction that are least subject to homoplasy and more reliable as sources of synapomorphies.

ROOTING THE TREE

A tree is said to be rooted if there is a particular node, the root, from which all other nodes can be reached by moving forward. The root is the hypothetical common ancestor of all the taxa included in the analysis. An unrooted tree specifies only the relationships among the taxa, but it cannot tell us anything about the direction of evolution and the ordering of trait changes over time. Unrooted trees are illustrated with a radial format (Figure 2) because the ancestral node has not been defined. To root a tree, an outgroup species must be included among the set of species in the ingroup, the taxon of interest. An outgroup is a taxon that is more distantly related to each of the ingroup taxa than any of the ingroup taxa are to each other. Selection of a good outgroup taxon is not always easy. It must be sufficiently distantly related to the taxa being considered, but not so distant that it doesn’t share a common ancestor with the ingroup species, i.e., it is not homologous. In practice, researchers typically rely on external sources of information from (earlier) classical studies to identify one or more candidate outgroup species.

Figure 2: An unrooted tree of immunodeficiency viruses in different host species. This tree was created using genomic sequences and a neighbor-joining algorithm. Support at each internal node was assessed using 1000 bootstrap samplings. Branch lengths are based on the number of genetic changes. FIV = feline immunodeficiency virus (blue), SIM = simian immunodeficiency virus (shades of
DIFFERENT WAYS TO PRESENT TREES

Trees can be presented in different ways, and each has its advantages and disadvantages. A cladogram, which shows only the branching order of nodes, can be drawn in either slanted or rectangular style. Figure 3 illustrates this difference. Either can be used to show the tree topology, and the points at which key adaptations arise can be denoted in a similar way. If the number of species being examined is large, the rectangular style will be easier to display in a compact amount of space.

![Figure 3: Slanted and rectangular style of trees](image)

When illustrating some type of distance or time information is important, the rectangular style must be used. Such a graph is called a phylogram. When genetic sequence data has been used to create the tree, the branch lengths can be varied to show the number of sequence changes occurring on each branch. The number of changes is considered to be an indication of either evolutionary time or strength of selection (Figure 4a). If something is known about the rate of genetic change, then the x axis can be expressed as time. For example, a molecular clock can be estimated using mutation rates at a selectively neutral gene locus and calibrated to real time if some independent fossil evidence is available. This results in a phylogram where the node positions can indicate the time of divergence between two branches, or the time of speciation events (Figure 4b).
Figure 4: Phylograms with distance/time information. (A) A molecular phylogeny of *Zenadia* doves, using two other genera of doves as outgroups. Branch lengths reflect genetic distance—shorter branches indicate species or populations that had fewer changes and are considered more closely related (Johnson et al. 2001). (B) Estimated divergence times for the apes, based on a combination of data from dozens of proteins used as molecular clocks. The heavy bars show ±1 standard error around the time estimates and lighter bars show 95% confidence intervals (Stauffer et al. 2001).

**FINDING THE BEST TREE**

When many traits are used in phylogenetic reconstruction, and the number of species being compared is moderate to
large, computer software must be used to construct the tree. PAUP, PHYLIP, MrBayes, BEAST, RAXML, and PHYLLAB are six commonly used packages that build trees using several of the methods mentioned below, and others are more specific to certain methods (for details see Hall 2004, and web resources at end of this unit). The four main methods are neighbor joining, maximum parsimony, maximum likelihood and Bayesian Markov chain Monte Carlo methods. Neighbor joining is an algorithmic method that computes a single best tree, whereas the others are tree-searching methods. In the latter case, many possible trees are identified, and then one must decide which tree is best. Alternative methods involve computing a probability of likelihood that alternative trees are supported by the data.

**Neighbor joining:** This is a distance method that converts discrete character data, such as the presence or absence of a morphological trait or the identity of a nucleotide at a homologous location in a gene, into a distance value (or in other words, the inverse of a similarity value). For example, two species are separated by a genetic distance of 10% if an average of 10 nucleotides have change per 100 bases. Many discrete characters can be combined into an overall average distance value. A matrix of pairwise distance estimates is computed between all pairs of taxa, and a clustering algorithm is used to group the most similar taxa (smallest genetic distance) together. A single tree is constructed.

**Maximum parsimony:** Parsimony can be employed with multiple traits, by summing all of the changes across all of the characters and selecting the tree with the lowest sum, but this strategy may not provide the most reasonable answer. This method typically finds several to many trees with the same, or only slightly different, numbers of changes. With several more or less equally parsimonious trees, one must then use other methods to select among them, or combine them.

**Maximum likelihood:** The essence of the maximum likelihood method with genetic sequence data is to determine how likely one would obtain a particular set of DNA sequences, given a mathematical formula that describes the probability that different types of nucleotide substitution will occur, and given a particular phylogenetic tree with known branch lengths. A computer can evaluate this question for all possible tree patterns, and the one with the highest likelihood is chosen. RAXML is the best software program to use for this method.

**Bayesian analysis:** Bayesian approaches are similar to maximum likelihood, except one asks what the probability is of a particular tree being correct, given the data and a model of how the traits in question change over time. Bayesian methods usually produce a set of trees of approximately equal likelihoods. The results are easy to interpret because the frequency of a given clade in that set of trees is identical to the probability of that clade, so no bootstrapping (see below) is required to assess the confidence in the structure of the tree. BEAST is the best software program to use for this method.

Which method is best? Although it seems very efficient to use an algorithmic method that finds the single best tree, one can be misled by not considering other possible trees. The “correct” tree doesn’t exist because we are trying to deduce the order in which existing taxa diverged from a hypothetical common ancestor and the amount of change along the branches between the diverging events. We can’t be sure that a tree topology accurately reflects the historical branching order. The tree-comparing methods are becoming increasingly more popular because they provide a statistical test for estimating how much better one tree is over another. If several very good trees have been obtained with slight differences, one can combine them into a single consensus tree.

Having obtained what we believe is the best tree, we can then evaluate how well supported a particular branch of the tree is using a technique called **bootstrapping**. This question is analogous to asking about the reliability of measuring height in a group of people. After measuring some people and computing the average of the sample, one evaluates whether the sample average is an accurate representation of the true average for the entire population by computing a measure of the variance around the mean. If the variance is high, one becomes less confident that the sample average is reliable. In bootstrap analysis with trees, one compares trees with and without a particular branch. A computer creates new datasets from the existing one by repeated sampling. With many repeats, one can then compute how many times a particular branch was estimated from the resampled data. If this number is very high, 90–100%, one can be very confident that the branch is accurate. If the number is around 50%, then it is best to conclude that support for the branch is low, and collapse that portion of the tree into a **polytomy**, or a point of uncertainty, in the published tree.
Using phylogenies to answer questions

Once a well-supported phylogeny has been obtained for a taxon of interest, the phylogeny can be used to answer a variety of questions. These questions fall into two broad categories: those in which the phylogeny is used to control for taxonomic relatedness so that relationships among traits and ecological variables can be analyzed across species, and those in which the phylogeny is used to generate and test evolutionary hypotheses. We take up these two categories in turn.

COMPARATIVE METHOD STUDIES WITH PHYLOGENETIC CORRECTION

In studies using the comparative method, the investigator compares the quantitative value of two (or more) traits across a range of species, or the association of a trait with some ecological variable, looking for significant correlations. The study may address whether ambient selective pressures play an important role in observed patterns of trait diversity, or whether the relationship with ecological factors meets some theoretical expectation. A bit of history may be useful here.

The field of animal behavior in the 1950s was largely focused on relating the diversity of animal behaviors and signals to their phylogeny at the ultimate level and to their inherited physiological mechanisms at the proximate level. Behavioral traits were even proposed as useful criteria for determining or at least confirming, phylogenetic trees. The assumption that behavior was as tightly linked to phylogeny as morphology was shattered by a number of pioneering comparative studies in the 1960s. These included correlations between avian mating systems and habitats by Crook (1964), Brown (1964), Verner and Willson (1966), and Lack (1968); similar correlative studies were published on primates (Crook and Gartlan 1966; Crook and Aldrich-Blake 1968) and antelopes (Jarman 1974). Statistical tests were undertaken to identify which correlations were significant and which not. The resulting significant correlations found among so many taxa triggered efforts to derive evolutionary models that could explain the correlations: Orians’ (1969) model for the evolution of polygyny is one important example. Overviews by Emlen and Oring (1977) and Clutton-Brock and Harvey (1977) then sought to integrate process and correlations into general theories of social and mating system evolution in animals. By the late 1970s, it was largely accepted that ecology was usually a better predictor of social, mating, and communication systems than was phylogeny. However, concern arose when Felsenstein (1985) pointed out that such comparative studies among a set of species contained an inherent statistical problem: because species are part of a hierarchically structured phylogeny and related to each other to varying degrees, they are not truly statistically independent samples for correlational studies. Even if the phylogenetic effects are subtle, drawing conclusions from correlations between behavior and habitat for 50 rodents, one ungulate, two primates, and a whale might be a risky business. At best, phylogenetic relatedness adds noise to such comparisons, and at worst, sample inflation by one group of related species might lead to spurious conclusions. Clearly, the comparative method required careful corrections for phylogenetic relationships among those species sampled.

Several techniques have been proposed for how best to go about doing this (Harvey and Pagel 1991; Harvey and Purvis 1991; Garland et al. 1992; Martins and Hansen 1997; Butler and King 2004; Garland et al. 2005) but they boil down to three basic methods: independent contrasts, Monte Carlo computer simulations, and generalized least squares models. Independent contrasts, proposed by Felsenstein, involves taking the difference between the trait values for two adjacent species or groups of species on the tree; each difference is independent of other differences elsewhere on the tree, and therefore can be used as a statistically independent point for correlational analyses (Felsenstein 1985; Garland et al. 1992). Monte Carlo simulation, by randomly moving the trait values around on a given tree, provides an empirical null distribution and defines a critical 5% acceptance value against which the observed correlation can be compared (Martins and Garland 1991; Garland et al. 1993). Phylogenetic generalized least squares regression (PGLS) is currently the favored method. The investigator develops a model of the rate of evolutionary change, such as the basic random Brownian model or the Ornstein-Ulenbeck model that assumes selective optima, and uses this expectation along with an independently derived phylogeny (with known branch lengths) to figure out how much of the variance and covariance among traits can be expected just from phylogenetic history. The residuals are then used in a regression analysis between the traits of interest. This method allows for the exploration of multivariate models. Both
MATLAB and R are good software platforms with a variety of existing routines for exploring alternative models (Lavin et al. 2008; PHYSIG, see web resources at end of this unit).

A good communication-related example of the use of PGLS to examine the possible environmental selective pressures affecting signaling traits is the comparative study of song complexity in mockingbirds and their relatives (Mimidae). Ten quantitative measurements were made on song clips of 29 species and reduced with factor analysis to three orthogonal song characters: heterospecific mimicry, short-term note diversity, and song complexity. From the known geographic ranges of each species, the variance and predictability of rainfall and temperature were obtained; habitat type and the migration behavior of each species were also considered. The environmental and song variables were analyzed with multiple regression models while controlling for phylogenetic relatedness. All three of the song variables, but especially mimicry, were significantly associated with more variable and unpredictable climates. Mimicry was especially strongly associated with facultative migration. Mimicry and short-term diversity had strong phylogenetic components. Possible functional explanations for these relationships included stronger selection in more variable and unpredictable climates leading to more elaborated signals of quality, or to signals of intelligence in the mate attraction context (Botero et al. 2009).

Phylogenetic corrections are now expected of any multispecies comparative study. In a recent review of 194 comparative studies that also reported the raw (uncorrected) result, the phylogenetic correction changed the conclusion in only a few cases (Garamszegi and Møller 2010). Moreover, the direction of the change after correction either improved or diminished the correlations in roughly equal numbers of cases. This review article also points out that most studies neglect the effects of within species variation. Heterogeneous sampling within species can potentially be as important as phylogenetic effects in comparative analyses. Some additional comparative analyses of communication-related traits include the following: Aparicio et al. 2003; Hagman and Forsman 2003; Kraaijeveld 2003; Galeotti et al. 2003; Bleiweiss 2004; Stuart-Fox and Ord 2004; Emlen et al. 2005; Olson and Owens 2005; Seddon 2005; Soler et al. 2005; Richards 2006; Berg et al. 2006; Bokony et al. 2007; Del Castillo and Gwynne 2007; Mank 2007; Doucet et al. 2007.

**PHYLOGENY-DRIVEN HYPOTHESES**

Several kinds of questions are driven by explicit knowledge of the phylogenetic tree. Evolutionary biologists of course want to know which species form meaningful clades so they can name them correctly. This entails determining whether taxa classified and named using classical morphological characters still hold up once genetically-based trees have been made (de Queiroz and Cantino 2001). Why certain species are found in certain parts of the world and how geographic distributions have changed through time can be examined with phylogenetic analyses (Pagel 1999; Raxworthy et al. 2002). Co-speciation between parasites and their hosts is another interesting question, in which independent trees are built for both groups, and one asks whether they speciated together (Clark et al. 2000). Behavioral biologists would like to know whether and which behavioral traits are useful taxonomic characters (de Queiroz and Wimberger 1993; Slikas 1998; Blomberg et al. 2003). Animal communication researchers often seek the ancestral form of signaling traits and the progression of signal elaboration across species (Phelps and Ryan 2000). Finally, we need very accurate trees to ask whether a signal trait or a receiver trait evolved first.

For questions that involve mapping traits onto a tree and assessing their evolutionary pattern across clades, one needs a quantitative measure of the degree of homology (or homoplasy). Several such measures have been developed. A simple and very useful measure, especially for discrete traits, is the **rescaled consistency index**, which quantifies the minimum amount of change a character may show on any tree divided by the observed number of changes, scaled by the maximum possible amount of homoplasy (Kluge and Farris 1969; Farris 1989a, b). The index has a value of 0 when the character is completely homoplous, and 1 when the character is perfectly homologous. Another type of measure is called the **test for serial independence** (Abouheif 1999). It measures the degree of non-randomness in a sequence of trait values for the terminal taxa in a tree. The variation in the differences between successive taxa is divided by the sum of squares of the observations. Completely random sequences have a value of 2 (i.e., homoplous), a value less than 2 indicates homology, and a value greater than 2 indicates non-random alternation. This test is independent of branch lengths and any type of model of evolutionary change, which has been considered a problem. Other
randomization tests have been devised that allow such refinements. Here, the trait values for the terminal taxa are permuted many times to obtain a random expectation, and observed values are compared. Variants of these models can take into account different branch lengths on the tree, and different evolutionary models for the tempo of evolutionary change, either accelerating or decelerating (Blomberg et al. 2003). Another strategy is to compute a measure of similarity between all pairs of terminal taxa using one or more traits, and then correlate these values with the genetic distances between the same pairs of taxa (Slabbekoorn et al. 1999). This technique has the usual problem with spurious and confounding correlations. For example, a correlation between genetic distance and a display trait character could be a spurious result of a true correlation between the display trait and habitat features, because habitat is correlated with genetic distance (Rheindt et al. 2004). Regardless of which method is used, when there is a significant amount of homology for a trait mapped onto a tree, we conclude that there is a strong phylogenetic signal (not to be confused with a communication signal).

Below are four interesting examples of studies in which communication signals have been mapped onto trees to evaluate the pattern of evolution of the traits. The first example (Figure 5) examined song characteristics of oropendolas, neotropical songbirds with black and yellow plumage in the Icterid family. The well-supported tree was derived from DNA sequence data, and branch lengths reflect molecular changes. Genera include Psarocolius, Gymnostinops, and Ocyalus. Of 32 acoustic features examined, the 22 shown here revealed unambiguous evolutionary changes (i.e., significant homology). Many of these traits evolved once and were subsequently retained in the clade. The two lower taxa have complex song repertoires, while the remaining species all have one song type (11) and perform the bow display (9). The trill (6) evolved once and now defines a subclade comprising P. atrovirens and the three P. angustifrons subspecies. The harsh broadband crash note (7) showed a reversal, arising in the upper clade and then was lost in two of the P. a. atrovirens subspecies. Multiple evolutionary changes occurred in the branch for G. montezuma, a highly polygynous species whose song is long and intense with large frequency shifts and note overlap. The variable characters based on frequency properties of songs (26, 28–31) were uniquely derived in terminal taxa; this implies rapid evolutionary change of these features that are the most vulnerable to degradation during transmission in different habitats.
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Introductory rattle</td>
<td>(11) Song versatility</td>
<td>(23) Rattle rate</td>
</tr>
<tr>
<td>(3) Broadband rattle</td>
<td>(12) Intersong interval</td>
<td>(24) Trill note rate</td>
</tr>
<tr>
<td>(4) Rattle-whistle</td>
<td>(15) Song duration</td>
<td>(25) Highest peak frequency</td>
</tr>
<tr>
<td>(5) Click</td>
<td>(16) Note percentage</td>
<td>(28) Frequency range</td>
</tr>
<tr>
<td>(6) Trill</td>
<td>(18) Note overlap</td>
<td>(29) Frequency shift rate</td>
</tr>
<tr>
<td>(7) Crash</td>
<td>(20–21) Pause duration</td>
<td>(30) Max frequency shift</td>
</tr>
<tr>
<td>(9) Bow</td>
<td>(22) Pause rate</td>
<td>(31) Frequency slope</td>
</tr>
</tbody>
</table>

**Figure 5: Phylogenetic analysis of the evolution of song features in oropendolas.** Song traits shown enumerated in the table are mapped onto the phylogeny, with green indicating traits that were gained and red indicating traits that were lost. Spectrograms of male songs are shown on the right. (After Price and Lanyon 2002.)

Another interesting phylogenetic analysis examined the acoustic structure of the male call in two related groups of Australian psyllids (Hemiptera: Psylloidae, family Triozidae), plant-sucking insects related to whiteflies and aphids (**Figure 6**). One group, contained in the genus *Schedotrioza*, feeds on *Eucalyptus*; species are allopatric by microhabitat segregation because each feeds on a different host plant species. The other group is more heterogeneous morphologically and feeds on various host plants in the family Casuarinaceae; species in this group all feed on the same set of host plant species, often on the same individual plant, and are therefore usually sympatric with each other. The sounds are produced by stridulation using forewing vibrations, which are primarily transmitted through the plant substrate. In many species, females answer male calls with their own unique stridulatory sounds, and courting pairs engage in antiphonal duetting. The phylogeny shows the *Schedotrioza* to be a monophyletic and more recently derived genus, with a distinctive male call structure defining the clade. Furthermore, the correlation between genetic distance and acoustic distance was positive within the group, as it was when all species were combined, indicating a strong phylogenetic signal in the call structure. However, within the Casuarinaceae-feeding species, the genetic–acoustic distance correlation was negative, which may reflect increased selection for acoustic divergence in closely related, sympatric taxa. The phylogenetic tree indicates that this is probably not a monophyletic group (Percy et al. 2006).
Figure 6: Phylogenetic analysis of acoustic stridulatory signals in Australian psyllids. (A) Single
The third study examined plumage color patterns in the New World orioles (genus *Icterus*). Each species exhibits a unique pattern of black, orange or yellow, and white color patches, yet they vary within similar themes. There are two common themes found in several sets of species, the Baltimore type with black heads and white edging on most wing coverts, and the Altamira type with carotenoid-colored heads and white patches on the outer primaries. This study attempted to determine whether these plumage types reflected common ancestry (synapomorphy) or independently derived convergence or reversal (homoplasy). The molecular phylogeny was based on DNA sequence data from two mitochondrial genes, and revealed three clades within the genus. Forty-four feather patch areas were scored as being white, black, or carotenoid. Each one was separately mapped onto the phylogeny and various measures of homology (consistency index CI and percent phylogenetic signal strength) were computed for each. **Figure 7** shows the mapping of one patch—the coloration of the lesser wing coverts—which had one of the lowest CI values and appeared repeatedly in all three of the main clades, yet there was still a significant phylogenetic signal in this character. The table summarizes these two measures for all 44 patches. Most patch areas showed repeated reversals and convergences. The Baltimore plumage type included species from clades C and A, and the Altamira type included species from all three clades. This finding indicates that distantly related species frequently converge on a similar overall pattern. Despite the generally high level of homoplasy for individual patches, there were few uniquely derived character states. The researchers concluded that there was a surprising amount of plumage lability at one level, but a strong degree of plumage conservatism at another level (Omland and Lanyon 2000).
The final example of a phylogenetic approach to signal evolution deals with the waveform of electric pulses in African Mormyroid elephantfishes (Sullivan et al. 2000). The electric discharge is used for both electrolocation and for communicating with conspecifics. The pulses are discharged continuously and have a consistent species-specific shape. In this phylogenetic example, the morphology of the electric signal-producing organ has also been determined for each species and the innervation pattern is associated with the waveform shape, so the precise evolutionary pattern of changes in the signal generation mechanism can also be deduced. The tree shown in Figure 8 comprises 41 species found in west-central Africa. The sister group to the Mormyrid elephantfishes is the Gymnarchids, which has a primitive electric organ.

The electric organ in this taxon consists of disk-like cells called electrocytes arranged in a series of columns in the narrow region of the tail. The flat surface of the disks is oriented perpendicular to the skin surface. A nerve stalk innervates each disk, and disks in a column are simultaneously stimulated. Electrocytes can be classified into six types based on the presence or absence of penetrating stalks and the side of innervation. The type S cell is stalkless (innervated by a simple branching nerve as shown in the figure) and produces a monophasic pulse. This type is found in the Gymnarchus outgroup. There are two major categories of stalked electrocytes. In the first, stalks arising from the posterior face of the electrocyte disk are non-penetrating and receive innervation on the posterior side (type NPp). All fish with this type produce simple biphasic electric organ discharge waveforms. In the second category, the stalk penetrates through the electrocyte. In the simplest of these, the stalk penetrates through the electrocyte to the opposite side, so that it receives innervation from the anterior side (type Pa). Others have stalks arising from the anterior face that penetrate through to receive innervation from the posterior face (type Pp). These two types produce biphasic pulses. In

<table>
<thead>
<tr>
<th>Character (Overall index)</th>
<th>CI (0.30)</th>
<th>Signal (%)</th>
<th>Character (Overall index)</th>
<th>CI</th>
<th>Signal (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary coverts, ventral</td>
<td>0.25</td>
<td>2.1</td>
<td>Throat</td>
<td>0.25</td>
<td>0.6</td>
</tr>
<tr>
<td>Secondary coverts, dorsal</td>
<td>0.30</td>
<td>2.6</td>
<td>Breast</td>
<td>0.25</td>
<td>10.6</td>
</tr>
<tr>
<td>Secondary coverts, ventral edges</td>
<td>0.25</td>
<td>12.6</td>
<td>Belly</td>
<td>0.33</td>
<td>0.0</td>
</tr>
<tr>
<td>Secondary coverts, dorsal edges</td>
<td>0.29</td>
<td>1.0</td>
<td>Vent</td>
<td>1.00</td>
<td>na</td>
</tr>
<tr>
<td>Primaries, edges</td>
<td>0.33</td>
<td>0.0</td>
<td>Undertail coverts</td>
<td>0.40</td>
<td>0.3</td>
</tr>
<tr>
<td>Secondaries (all), edges</td>
<td>0.22</td>
<td>8.5</td>
<td>Edge breast patch</td>
<td>0.33</td>
<td>2.3</td>
</tr>
<tr>
<td>Dorsal secondaries, edges</td>
<td>0.33</td>
<td>0.0</td>
<td>Throat feathers (shape)</td>
<td>1.00</td>
<td>na</td>
</tr>
<tr>
<td>Tertials, tips</td>
<td>0.20</td>
<td>0.0</td>
<td>Breast patch border</td>
<td>1.00</td>
<td>na</td>
</tr>
<tr>
<td>Tertials, edges</td>
<td>0.25</td>
<td>0.0</td>
<td>Thigh</td>
<td>1.00</td>
<td>na</td>
</tr>
<tr>
<td>Ventral primaries, proximal</td>
<td>0.33</td>
<td>1.1</td>
<td>Anterior malar</td>
<td>0.33</td>
<td>6.1</td>
</tr>
<tr>
<td>Middle coverts</td>
<td>0.29</td>
<td>1.1</td>
<td>Posterior malar</td>
<td>0.14</td>
<td>4.0</td>
</tr>
<tr>
<td>Lesser coverts</td>
<td>0.14</td>
<td>44.3</td>
<td>Posterior eyestripe</td>
<td>0.29</td>
<td>0.5</td>
</tr>
<tr>
<td>Marginal coverts</td>
<td>0.50</td>
<td>5.1</td>
<td>Anterior ear coverts</td>
<td>0.33</td>
<td>0.0</td>
</tr>
<tr>
<td>Underwing coverts</td>
<td>1.00</td>
<td>na</td>
<td>Posterior ear coverts</td>
<td>0.17</td>
<td>0.2</td>
</tr>
<tr>
<td>Scapulars</td>
<td>0.50</td>
<td>0.5</td>
<td>Anterior supercilium</td>
<td>0.25</td>
<td>0.0</td>
</tr>
<tr>
<td>Upper back</td>
<td>0.33</td>
<td>0.0</td>
<td>Posterior supercilium</td>
<td>0.17</td>
<td>0.7</td>
</tr>
<tr>
<td>Middle back</td>
<td>0.60</td>
<td>0.5</td>
<td>Eye-ring</td>
<td>0.50</td>
<td>2.1</td>
</tr>
<tr>
<td>Upper rump</td>
<td>0.33</td>
<td>5.3</td>
<td>Crown and nape</td>
<td>0.14</td>
<td>3.4</td>
</tr>
<tr>
<td>Lower rump</td>
<td>0.50</td>
<td>2.1</td>
<td>Forehead</td>
<td>0.25</td>
<td>0.0</td>
</tr>
<tr>
<td>Upper tail coverts</td>
<td>0.33</td>
<td>0.0</td>
<td>Base rectrices</td>
<td>0.14</td>
<td>1.1</td>
</tr>
<tr>
<td>Anterior flank</td>
<td>0.25</td>
<td>0.3</td>
<td>Outer rectrices</td>
<td>0.60</td>
<td>5.5</td>
</tr>
<tr>
<td>Posterior flank</td>
<td>0.25</td>
<td>2.1</td>
<td>Tips rectrices</td>
<td>0.13</td>
<td>5.8</td>
</tr>
</tbody>
</table>
Brienomyrus, the stalk penetrates through the electrocyte a second time to receive innervation on the face from which it originated (called a doubly penetrating stalk with posterior innervation, type DPp). This arrangement produces a triphasic pulse. Species in one genus, Pollimyrus, have both doubly penetrating and non-penetrating stalks (type DPNP). The figure shows the point at which each of these electrocyte innervation strategies is believed to have arisen. What is especially interesting in this phylogenetic analysis is that it is very easy for reversals to occur, i.e., for a multiple penetrating stalk to revert to a simpler form with fewer penetrations, because these represent more primitive developmental stages.
Figure 8: Phylogeny of electric organ evolution in Mormyroid elephantfishes. The robust phylogenetic tree is based on genetic sequences from 4 gene loci. The six electric organ morphological types and typical waveform EOD that they might generate are illustrated on the left, and arrows point to the hypothesized position in the tree where they first arose. Numbers refer to hypothesized character state changes. Note the frequent occurrence of reversals, from the typical Pa organ type shown in red to the simpler, ancestral NPp type shown in black. (From Sullivan et al. 2000.)
Web resources

PAUP: http://paup.csit.fsu.edu/

PHYLIP: http://www.phylip.com/

MrBayes: http://mrbayes.csit.fsu.edu/

BEAST: http://beast.bio.ed.ac.uk/Main_Page

RAXML: http://www.embl.de/~seqanal/courses/commonCourseContent/usingRaxml.html and http://phylobench.vital-it.ch/raxml-bb/

R packages: http://www.r-phylo.org/wiki/Main_Page

PHYSIG: http://www.biology.ucr.edu/people/faculty/Garland/PHYSIG.html

MESQUITE: http://mesquiteproject.org/mesquite/mesquite.html

Literature cited


Web Topic 10.2
Neural Network Models and Feature Detectors

Introduction

Sensory systems evolve to enable organisms (plant and animal) to gather meaningful information from their environment and then integrate that information to make adaptive decisions. The field of sensory ecology studies how sensory systems are adapted to detect the objects important to a given species’ survival and reproduction, such as food, mates, predators, and shelters. All modalities can be studied from this perspective, and a typical goal is to describe how sensory receptors and processing centers are designed to maximize performance on a given task. For example, the peak pigment sensitivities of marine fishes are optimally designed to maximize luminance and chromatic detection in different habitats (Cummings 2004), the basilar papilla of frogs’ ears is tuned to be most sensitive to the dominant frequency of conspecific male calls (Capranica et al. 1973; Capranica and Moffat 1983; Wilczynski et al. 1992), certain chemical sensors on male moth antennae are exquisitely tuned to detect the pheromone blend of conspecific females, and touch receptors may be highly developed to rapidly detect prey (Catania and Kaas 1997; Catania and Remple 2005).

Feature Detectors

Very specific detection systems are often called feature detectors. Feature detectors are refinements in the receiving apparatus that make the detection of cues and signals easier against background noise. Such receiver detection systems are more than a peripheral filter; they typically involve higher levels of neuronal integration as well (Martin 1994; Ison and Quiroga 2008). Lateral inhibition between adjacent receptor cells plays a key role in feature detector wiring. For example, the retina employs horizontal and amacrine cells to enhance the detection of edges, horizontal or vertical bars, spots, and certain types of movement (Yang and Wu 1991; Cook and McReynolds 1998; Zhang and Wu 2009; Lipin et al. 2010). Similar systems have been described for auditory and olfactory receivers. Owls have feature detectors for locating the position of prey by the sounds they make (Konishi 1993), and species-specific templates for the songs of passerine birds ensure that only conspecific notes and syntax patterns are learned (Lewicki and Konishi 1995; Huetz et al. 2004; Theunissen et al. 2004; Lehongre and Del Negro 2011). The vertebrate olfactory system also uses lateral inhibition to enhance the recognition of certain odorants (Lledo et al. 2004).

Feature detectors are hard-wired in the sense that no learning is involved, although the neuronal pathways may need to be activated with general sensory stimulation. Specific responses to the detection of a feature detector are also often hard-wired. Hard-wiring ensures that the signal or cue is recognized the first time it is encountered and that the appropriate response (approach or evasion) is made. Feature detectors can evolve as part of the receiver tuning process of signal evolution described in main text Figure 10.2 to enhance the detection of important social signals. Here they serve in the recognition of sign stimuli, or releasers for appropriate responsive behaviors. A classic example of a sign stimulus and its corresponding feature detector is the red spot on the bill of many gulls that releases begging behavior by chicks (Tinbergen and Perdeck 1950). Feature detectors can also evolve to facilitate approach toward cues from prey or avoidance of cues from enemies. Once such feature detectors for environmental cues have evolved, they can be exploited by senders that match the key cue characteristics via the receiver-precursor scenario of signal evolution outlined in main text Figure 10.16.

Neural Network Models

Artificial neural network models can be used to explore some of the properties of sensory recognition systems in animals. Neural network models attempt to mimic the basic features of sensory systems and have been used to understand both the key mechanisms underlying sensory recognition and the hidden preferences that may emerge as a byproduct of these mechanisms (Gurney 2007; Phelps 2007). An example of a simple feed-forward model is shown in Figure 1. It consists of three layers. The input layer is a 6 x 6 array of sensory receptor cells analogous to a retina (or
olfactory epithelium or basilar papilla). The second layer is called the hidden layer, in this case comprised of 36 cells analogous to ganglion cells. Each cell in this layer is connected to every cell in the input layer above it and each input layer cell is connected to every hidden layer cell. These connections are associated with quantitative weightings that regulate the strength of the signal passing between cells, analogous to a synaptic nerve connection. The total stimulus to a hidden layer cell is therefore the weighted sum of the input from all of the sensory cells in the layer above. The third layer consists of a single output cell connected to all of the hidden layer cells. Input to this output cell is also a weighted sum of the connection strengths from the hidden layer.

**Figure 1: A feedforward neural network model for visual mate detection.** In the input layer, the blue structure represents the generic characteristics of a conspecific bird with a long tail. In the hidden layer, each cell receives input from every cell in the input layer, shown by red lines. For simplicity, the connections for only a few cells are shown. The output layer receives input from each cell in the hidden layer (encompassed within the orange lines).

When the input layer is stimulated, each of the receptor cells receives a score of zero or one. The output from these cells is then multiplied by the weighting factors to determine the total input to the hidden layer cells. The summed output from hidden layer cells is similarly multiplied by the weighting factors associated with this layer to determine the summed stimulation of the output cell. The output cell is the decision-maker — if the summed input to this cell is above a certain threshold, it responds with a “yes” and the network is said to “recognize” the stimulus pattern. The weighting factors at both levels “evolve” during training sessions in a process that mimics either learning or natural selection to achieve a certain level of discrimination. The input layer is presented with the target object in various locations and orientations and “taught” that this is the correct image using a back propagation procedure or using a genetic algorithm to assign fitness scores. The input layer is also presented with non-target images, which may be a different pattern or random patterns, and taught to reject such images. Networks with weighting patterns that make more correct decisions have higher fitness. Weights undergo random mutational changes at each trial (generation) and the networks with better discrimination abilities are retained for the next trial. Each iterated run produces a unique successful network with a different final set of weighting factors that work (enable the network to make an accurate discrimination); thus there are many solutions to the same problem.

Novel stimulus patterns are then presented to the trained networks without modifying the weightings. A network model is considered robust if it can generalize, i.e., recognize stimuli that are similar to the training target. In other words, it should be able to evolve internal representations that correspond to generic features of the target. Occasionally a network is found that responds more strongly to an exaggerated version of the training pattern. For example, a classic
model by Enquist and Arak (1993), using a network structure much like the one depicted in Figure 1, trained the retina to accept an image of a long-tailed bird (with a two-square tail) and to reject a heterospecific bird with a shorter tail (a one-square tail). When the net was presented with a bird with a very long three-square tail, it responded more strongly than it did to the two-square-tail training target. At the time, this study was touted as evidence of potential sensory biases that could explain the elaboration of sexually selected male traits. Such sensory biases could also provide an explanation for the classical ethologist’s observation of **supernormal stimuli** (Tinbergen 1948). These are the preferences sometimes shown for cues or signal features that are enlarged or exaggerated beyond the normal range (McClintock and Uetz 1996; Burley and Symanski 1998; Drăgănoiu et al. 2002; Ghirlanda and Enquist 2003). Another emergent property of neural net models is an occasional strong response to a novel pattern. No pattern recognition system performs perfectly since backgrounds vary and the pattern may appear in different orientations or at different distances. However, a variety of neural connection schemes may satisfactorily distinguish the pattern from its background on the majority of occasions. Some schemes by chance may be more sensitive to a previously unencountered pattern. This phenomenon has been termed **hidden preferences** or **latent biases** (Enquist and Arak 1993). Other network models found similar biases towards **symmetrical** patterns (Enquist and Arak 1994; Johnstone 1994). Supernormal stimuli and hidden preferences might drive the evolution of new signal forms as proposed by the sensory exploitation model.

Some criticisms have been leveled at the conclusions drawn from these simple neural network models. First, the results may depend on the details of the model. A re-examination of the Enquist and Arak (1993) model using a somewhat different training method and stronger selection regime found that the response to the extra-long tail was not greater than the response to the training tail (Kamo et al. 1998). Kamo et al. attempted to pinpoint the reason for this discrepancy by exploring effects of different training periods and backgrounds. Their network generalized well to similar images, but responses to exaggerated images with longer tails or wings fluctuated wildly over time. They suggested that Enquist and Arak’s network was insufficiently trained. Second, such simple networks are not designed to perceive particular shapes or spatial patterns, but only to respond to the average level of stimulation to a given sensory cell (Cook 1995; Dawkins and Guilford 1995). A true feature detector that perceives spatial pattern shapes must consist of a more complex hidden layer with a two-dimensional array of cells and a third series of weighting factors among the cells in this layer. Like horizontal and amacrine cells of the retina, these connections produce patterns of reinforcement and inhibition between adjacent cells. Neural network models of this type can evolve edge, spot and bar detectors similar to those described by neurophysiologists (Rubner and Schulten 1990).

Fuller (Fuller 2009) employed a more complex network to investigate the feasibility of the sensory bias model of mating preference evolution, whereby males develop traits that match female sensory biases for food items. The model species for this simulation study was the bluefin killifish **Lucania goodei**, in which males are polymorphic for blue, red, or yellow anal fins. Potential food items could be blue, red, or yellow worms, and the question addressed was whether selection for a colored worm preference could have pleiotropic effects on male trait evolution, given that both foraging and mating use the same visual sensory system. The network was modeled after a color-sensitive retina with five visual pigments (UV, violet, blue, yellow, and red) as found in the killifish. The input layer consisted of a 12 by 12 array with a mosaic arrangement of 144 cone cells, some of which were linked into double-cone units. The next layer was modeled after bipolar cells that connected to a subset of same-pigment cone cells, as well as some mosaic bipolar cells receiving input from the double-cone units. Another layer contained horizontal cells that made lateral connections among cone cells. The next layer represented ganglion cells that received input from all bipolar cells, and another layer represented processing neurons in the brain. Finally, these processing neurons connected to two output neurons, one for mate recognition and one for food recognition. Each cell in the network also connected to a single bias neuron that weighted its threshold for firing. Mutations could occur on both the weightings and the pigment wavelength peaks. Selection on populations of 100 networks was strong, with half of the most poorly performing networks eliminated in each generation, while the remaining networks were allowed to “breed” in a pairwise fashion with high heritability of the “parental” network values. In one series of simulations, the networks were trained to prefer one color of worms over the other colors, and then they were assayed for their mate choice preferences. A small but significant preference for mates having the same colored fins was obtained, but there was a great deal of variation and many cases of preference for alternate colors. This result provided modest support for the correlated responses predicted by the
sensory bias hypothesis. However, in other simulations that selected simultaneously on mate and food preference colors, there was no tendency for matching colors to evolve preferentially, showing that foraging and mate choice preferences are not constrained to evolve in concert. Even though such complex network models are still great simplifications of natural evolutionary processes, this study nevertheless demonstrated the restrictive conditions under which correlated responses could evolve.

There are many other very interesting applications of neural network modeling that address communication-related questions, as well as other types of ecological questions. For example, (Hurd et al. 1995) found support for an antithetical (opposite or polarized) form of aggressive versus submissive threat signals, Enquist and Arak (1994) documented the latent attractiveness of symmetrical patterns, Phelps and Ryan (2000) found strong biases for ancestral acoustic signal forms in frogs, and Brodin and Haas (2006) studied the process of sexual imprinting in sympatric speciation. The interested reader can find many useful articles in a Theme Issue of the March 27, 2007 Philosophical Transactions of the Royal Society B (volume 362, issue 1479).

**Literature cited**


Web Topic 10.3
Rich Media Examples of Ritualization

Introduction

As discussed in Chapter 10, displays evolve from a wide variety of sender precursors. These precursors are usually cues that provide useful information to receivers that attend to and associate them with something the sender might do next. If the sender benefits from transmitting this information, selection can favor ritualization of the cue into a signal. Signal precursors include basic survival behaviors such as foraging, grooming, fleeing, and fighting, as well as physiological processes involved in growth, maintenance, and reproduction. Numerous examples were given in the text, and here we provide video and audio examples of some of them. This Web Topic unit is divided into sections based on type of precursor activity: comfort and hygienic activities, locomotory and feeding appendage movements, intention movements, motivational conflict, physiological processes (autonomic and endocrine systems), antipredator defenses, and a few interesting oddities.

Comfort and hygienic movement precursors

Many species of animals have ritualized body hygiene actions into signals. One of the first taxa in which such ritualization was recognized contains the ducks and geese (family Anatidae). In ducks, nearly all of their courtship and mating displays can be traced to precursors with more mundane functions. Many duck displays appear to be ritualized versions of “comfort” movements: behaviors such as shaking to remove water and arrange the feathers, preening, and using the head and beak to waterproof the plumage with oils produced by the uropygial gland. In addition, ducks have ritualized many other activities; other precursors of signals include drinking, tipping up to feed, leaping into flight, swimming on the surface, and diving. Many species are quite vocal and have added conspicuously stereotyped movements to sound production during courtship. In blue-winged ducks such as shovellers, the main male courtship display is a re-ritualized version of another signal. Ducks thus provide a rich panoply of ritualized signals.

Ducks are a very ancient group of birds and their higher-level taxonomy remains disputed. However, certain subfamilies are clearly distinct units with most members performing homologous courtship and mating displays. Some species might show all the displays known in its subfamily, whereas others might have exaggerated one or more displays and dropped the others. It is relatively easy to identify the same ritualized display in related species: they may differ slightly, but they are largely conserved during speciation. For this reason, duck displays were one of the first behavioral traits to be used for taxonomic classification (Johnsgard 1965; Lorenz 1971). Below, we provide a sampling of courtship displays in two different subfamilies of ducks. This provides a useful perspective on how ritualized displays can be conserved within a taxon, but still allows for species differences.

Dabbling ducks (Anatinae): Most of the species in this group belong to the genus Anas. The ancestors of the Anatinae must have had a large display repertoire since some species show most of the displays known in the subfamily, whereas other species show only a subset, with different species showing different subsets. Mallard ducks (Anas platyrhynchos) and speckled teal (Anas flavirostris) exhibit most of the shared repertoire. This includes ritualized drinking, an action shown by both sexes but often by males that is used in a wide variety of social contexts (see Figure 1.4 in Chapter 1). Ritualized and stereotyped female displays include inciting (in which the female accompanied by one male points her bill repetitively at another male, often inducing the first male to chase or attack the second); nod swimming (in which the female encourages males to display by stretching her neck forwards and swimming rapidly through the males in large circles); and head pumping (in which the female pumps her head up and down in synchrony with a male just prior to copulation). Male displays include preliminary shakes (ritualized shaking that may include wing flaps to announce that other displays will follow); grunt whistles (a rapid forward arching of the body, emission of a whistle, and the sweeping up of a stream of water droplets towards the rear with the bill; a major courtship display); head up–tail up (sudden stretching of head and neck upwards while also sticking the tail up vertically; another major courtship display); burping (a high pitched sound emitted while conspicuously stretching the
neck and upwards or in a rotary fashion); **down–up** (rocking forward onto the breast and back; a signal most often directed at competing males); **head pumping** (male part of synchronized movement prior to copulation; male often dips bill at low point); **bridle** (opposite of **down–up** in that male rocks back on tail while lifting breast out of water; usually performed after a copulation); **nod–swimming** (same behavior as female but usually performed after a copulation).

Many of the clips below are digital versions of Super-8 movies taken from a water-level blind at the Bronx Zoo when one of the authors was a graduate student. Although the quality is low, the clips show displaying birds close up and are labeled with the name of the display and the species. A few more recent clips of mallards are included. Unfortunately, none of these clips include any corresponding sounds. In many of the clips, note the **ritualized drinking** interspersed between other displays:

- **Female inciting**: Female inciting is present in most Anatidae, not just the dabbling ducks. In this clip, examples in order of presentation are: Brazilian teal (Amazonetta brasiliensis), mallards, and common eiders (Somateria mollissima).
- **Female nod–swim**: The clip shows two nod–swims by a female mallard in the midst of many males.
- **Male preliminary shakes**: Again, a display more widely distributed in the Anatidae. In order of presentation: Bahama pintails (Anas bahamensis), garganey teal (Anas querquedula), marbled teal (Marmaronetta angustirostris), speckled (also called sharp-winged) teal (Anas flavirostris), and common eider.
- **Male grunt–whistle**: This striking display is limited to *Anas* species. Examples in order of presentation: green-winged teal (Anas crecca), mallards, and speckled (sharp-winged) teal.
- **Male burping**: Again, there is no sound accompanying these clips. In order of presentation, examples are: Baikal teal (Anas formosa), green-winged teal, Bahama pintail, red-billed duck (Anas erythrorhyncha), and garganey teal.
- **Male Down–up and Head up–Tail up**: In mallards, these tend to be given as separate displays. The **down–up** is often directed at other nearby males. In green-winged teal, a **down–up** is often, but not always followed by a **head up–tail up**. And in the Bahama pintail, which has lost the **grunt–whistle** display entirely, the main male courtship display is an obligate **down–up/head up–tail up** combination. Examples in order of presentation are green-winged teal, mallard, and Bahama pintails.
- **Head pumping by both sexes**: When a male and female are both interested in mating, they begin a synchronized head pumping that often culminates in copulation. Males often begin head pumping first, and females only join in if they are currently interested in mating:
  - **Normal male–female examples**: This clip shows a garganey teal male unsuccessfully trying to get a female to mate, followed by several mallard interactions, the last of which lead to a copulation.
  - **Male head pumping as courtship display**: Male blue-winged ducks such as shovellers (Anas clypeata) have co-opted the synchronized **head pumping** used by other species prior to copulation into their main courtship display; males swim around **head pumping** all the time. They have lost all of the other traditional dabbling duck displays such as **grunt whistles**, **down–ups**, and **head up–tail ups** but have added one novel display: ritualized **mock feeding**.
- **Male bridle**: This is usually performed by male dabbling ducks after copulations. Examples here are mallard and spectacled teal.
- **Male nod swim**: This is also performed by males after a copulation. Examples here are of mallards.

**Stiff-tailed ducks** (Oxyurinae): This small group of species has a pooled display repertoire that is quite different and surely evolved independently from that of dabbling ducks. As their name implies, males have a small rosette of stiff tail feathers that can be erected vertically when males are displaying. Males concurrently stretch their necks and raise their head into a stiff position and then perform some action ritualized from other types of movement: preening the breast, shaking, or swimming. Examples:
**Common ruddy duck** (*Oxyura jamaicensis*): The main male display is called the **bubble display**. Once the tail and head are vertically stretched upwards, the male tips his bill down and raps his breast repeatedly producing a froth of bubbles. It then sits stiffly in the vertical position before repeating the actions.

- **White-headed ruddy duck** (*Oxyura leucocephala*): This male performs a ritualized swimming next to the female with its bill aimed at the female. It then shifts into the head and tail erect position of most stiff-tailed ducks but holds this position instead of performing further actions. ([http://www.arkive.org/white-headed-duck/oxyura-leucocephala/video-09a.html](http://www.arkive.org/white-headed-duck/oxyura-leucocephala/video-09a.html))

- **Argentine ruddy duck** (*Oxyura vittata*): Males of this species (which are notorious for having one of the longest penises of any animal), display by first assuming the vertical head and tail position common to congeners. Then, instead of rotating just the head to strike its breast repeatedly like the common ruddy, these males rock their entire body forwards and backwards in a violent motion. The amplitude of the rocking gets stronger as the display progresses.

**Locomotory and feeding precursors**

**Drumming displays in woodpeckers**: The primary signal for both territorial defense and mate attraction in most woodpeckers is a mechanical drumming sound. Woodpeckers tap on branches as they foraging and excavate holes for nesting and roosting. In the process of foraging, they encounter hollow branches that produce a loud resonating sound when tapped. The advertisement sound is a ritualized series of taps, either rapid drumming or a stereotyped pattern of tapping on various resonant structures. Examples of both types are included here:

- **Yellow-bellied sapsucker** (*Sphyrapicus varius*):
  - This clip shows the remarkable chiseling skill of a foraging bird as he carves off squares of bark to expose the tree’s sap. This is the precursor activity ([http://www.arkive.org/yellow-bellied-sapsucker/sphyrapicus-varius/video-00.html](http://www.arkive.org/yellow-bellied-sapsucker/sphyrapicus-varius/video-00.html)).
  - This is an audio track of the sapsucker’s ritualized and patterned drumming sound when signaling ([http://macaulaylibrary.org/audio/flashPlayer.do?id=87149](http://macaulaylibrary.org/audio/flashPlayer.do?id=87149)).

- **Downy woodpecker** (*Picoides pubescens*): This is a sound clip of a downy woodpecker’s drumming display ([http://macaulaylibrary.org/audio/flashPlayer.do?id=44905](http://macaulaylibrary.org/audio/flashPlayer.do?id=44905)).

**Flight sounds in insects**: The wings of insects often produce a buzzing or humming sound during flight. In a few cases, this sound plays a role in mate attraction.

- **Yellow fever mosquito** (*Aedes aegypti*): The flight tones of these mosquitoes are positively correlated with body size; the female fundamental frequency is around 450 Hz whereas the male’s frequency is 600–700 Hz. An interacting receptive female and male often shift their flight frequencies to match that of the other, but it is usually the second harmonic of the female and the third harmonic of the male that show the convergence at around 1200 Hz (Cator et al. 2009; Cator et al. 2010). This video shows an example of a converging pair ([http://www.sciencemag.org/content/323/5917/1077/suppl/DC1](http://www.sciencemag.org/content/323/5917/1077/suppl/DC1)).

**Stridulation and other mechanical sounds from wing and leg movements**: We provide numerous examples of such sound signals that are ritualized versions of leg and wing movements in Web Topic 2.6.

**Intention movements**

As discussed in Chapter 10, the cues generated by animals preparing to adopt some action are prime precursors for ritualization into signals. Below we provide examples of several common categories of ritualized intention movements:

**Intention to attack threats often display weapons and readiness to bite or lunge but may also be derived from non-aggressive precursors:**
• **Male white rhinoceros** (*Ceratotherium simum*): Note curled tails and stiff-legged approach with heads down before actual fighting ([http://www.arkive.org/white-rhinoceros/ceratotherium-simum/video-12b.html](http://www.arkive.org/white-rhinoceros/ceratotherium-simum/video-12b.html))

• **Leopard cat** (*Prionailurus bengalensis*): This clip shows a classical aggressive threat with teeth bared and ears back ([http://www.arkive.org/leopard-cat/prionailurus-bengalensis/video-13.html](http://www.arkive.org/leopard-cat/prionailurus-bengalensis/video-13.html))

• **Gelada baboon** (*Theropithecus gelada*): In baboons, yawning, raising the eyebrows, and blinking have each been ritualized into threat displays. Olive and chacma baboons perform stylized yawns as threats; gelada baboons pull their lips back from their teeth. Partially closing the eyes to show the upper eyelids is also a threat. Some species raise their eyebrows or have special coloration of the upper lids to accentuate this display. This clip shows several gelada baboon males threatening using both raised lip and eye-flash displays. ([http://www.arkive.org/gelada/theropithecus-gelada/video-12b.html](http://www.arkive.org/gelada/theropithecus-gelada/video-12b.html))

• **Male impalas** (*Aepyceros melampus*): Males mark trees, open mouths, and present horns before engaging in sparring ([http://www.arkive.org/impala/aepyceros-melampus/video-12.html](http://www.arkive.org/impala/aepyceros-melampus/video-12.html))

• **Male hooded seals** (*Cystophora cristata*): Males of this species inflate a black sac on the forehead and a red nasal sac while vocalizing as threats to other males before beginning combat ([http://www.arkive.org/hooded-seal/cystophora-cristata/video-12.html](http://www.arkive.org/hooded-seal/cystophora-cristata/video-12.html))

• **Male northern elephant seals** (*Mirounga angustirostris*): These males display their body sizes by rearing up and vocalizing before fighting. ([http://www.arkive.org/northern-elephant-seal/mirounga-angustirostris/video-12.html](http://www.arkive.org/northern-elephant-seal/mirounga-angustirostris/video-12.html))

• **Ritualized fighting in male adders** (*Vipera berus*): As with the seals, these snakes display their height and strength before actually engaging in combat. ([http://www.arkive.org/adder/vipera-berus/video-11.html](http://www.arkive.org/adder/vipera-berus/video-11.html))

Submission and retreat signals are often opposite in form to attack signals:

• **Crocodile** (*Crocodylus porosus*): Whereas males lower their heads before attacking, an individual that is losing a fight raises its head in submission. ([http://www.arkive.org/saltwater-crocodile/crocodylus-porosus/video-12.html](http://www.arkive.org/saltwater-crocodile/crocodylus-porosus/video-12.html))

• **Barbary macaques** (*Macaca sylvanus*): Lip-smacking and teeth chattering are common submission signals in macaque monkeys. The teeth chattering shown here in Barbary macaques is used as a submissive signal, but is also used to indicate appeasement, affiliation, and reassurance. ([http://www.arkive.org/barbary-macaque/macaca-sylvanus/video-12a.html](http://www.arkive.org/barbary-macaque/macaca-sylvanus/video-12a.html))

• **Rhesus macaques** (*Macaca mulatta*): The following clip shows a wide mixture of both aggressive and submissive facial gestures in another species of macaque monkey. As with the Barbary macaque, lip smacking and teeth chattering are largely submissive signals. So is an open-mouth with the teeth and gums exposed, and flattening the ears against the head. An open mouth with the teeth covered is however a threat. Staring at an opponent, flashing the colored upper eyelids, and raising the eyebrows is also a threat. ([http://www.arkive.org/rhesus-macaque/macaca-mulatta/video-12a.html](http://www.arkive.org/rhesus-macaque/macaca-mulatta/video-12a.html))

• **Chameleons** (*Furcifer pardalis*): Aggressive and submissive displays are shown. Note the lateral compression broadside display and bright colors in the dominant animal and the evasive behaviors and pale colors in the subordinate. ([http://www.arkive.org/panther-chameleon/furcifer-pardalis/video-12b.html](http://www.arkive.org/panther-chameleon/furcifer-pardalis/video-12b.html))

• **American bison** (*Bison bison*): Courtship and aggressive behavior in bison; note the wallowing, calling, parallel walk, and head down aggressive threats, and the backing away submissive display by the loser of the fight. ([http://www.arkive.org/american-bison/bison-bison/video-bi09a.html](http://www.arkive.org/american-bison/bison-bison/video-bi09a.html))

Other intention signals:

• **Gerenuk** (*Litocranius walleri*): The intention to mate display in many African antelopes involves the male tapping the inside of the female’s rear legs with his foreleg. Here we see this behavior in a pair of gerenuks (see also Figure 7.2, Chapter 7). ([http://www.arkive.org/gerenuk/litocranius-walleri/video-09.html](http://www.arkive.org/gerenuk/litocranius-walleri/video-09.html))

• **Wolf spider** (*Pardosa nigriceps*): Courtship is dangerous for male spiders as females are prone to eat them if they do not recognize the male as a conspecific. Males carry their sperm in their palps, which in this species are
Motivational conflict precursors

In Chapter 10 (pp. 382–4), we discuss various mechanisms by which motivational conflict in animals might serve as a precursor to ritualized signals. Below, we provide a video clip of a ritualized signal that is thought to have arisen from each of the motivational routes mentioned in the text:

Ambivalence behaviors:

- **Three-spined stickleback** (*Gasterosteus aculeatus*): The zigzag courtship dance of the male stickleback is considered a classical example of a signal arising from motivational conflict. This video shows the entire nest-building and courtship sequence of which the zigzag dance is a part. ([http://www.youtube.com/watch?v=cBX8hWuiHTk](http://www.youtube.com/watch?v=cBX8hWuiHTk))

Broadside threats:

- **Sable antelope** (*Hippotragus niger*): Two males approach each other and the dominant performs a broadside threat that causes the subordinate to leave. ([http://vimeo.com/22088612](http://vimeo.com/22088612))

Displacement acts:

- **Laughing gulls** (*Leucophaeus atricilla*): In this clip, two individuals are contesting dominance status. Until the end, neither attacks the other, but instead bends over and pecks at the ground as if feeding. This is interpreted as a displacement act that has become ritualized into a signal of threat. ([http://www.youtube.com/watch?v=9v4cmeoHcqE](http://www.youtube.com/watch?v=9v4cmeoHcqE))

Redirected acts:

- **Grass pulling in gulls** may be a redirected act of aggression during fights. The movement is similar to the behavior of grasping the opponent’s bill or wing and pulling. This clip is from the movie **Signals for Survival** (Cornell Lab of Ornithology).

Motivation-structure rules in birds and mammals: As discussed in Chapter 10, the combination of multiple motivations and corresponding physiological states of actions appropriate to those motivations can combine in complicated ways to limit the kinds of sounds an animal might produce (see Figure 10.14). These favor higher frequencies as fear increases, and higher bandwidths as aggressiveness increases. Below, we provide some examples that appear to fit these motivation-structure rules:

- **Bonobos** (*Pan paniscus*): Threat calls of this species tend to be low frequencies whereas fearful screams are high frequency. ([http://www.arkive.org/bonobo/pan-paniscus/video-13.html](http://www.arkive.org/bonobo/pan-paniscus/video-13.html))

- **Red deer** (*Cervus elaphus*): Males with different ages, sizes, and motivations: produce roars at different dominant frequencies. Here are two examples. ([http://www.arkive.org/red-deer/cervus-elaphus/video-13.html](http://www.arkive.org/red-deer/cervus-elaphus/video-13.html))
- **Domestic dog** (*Canis lupus*): Growls with different bandwidths and dominant frequencies produce different responses in receivers. ([http://www.youtube.com/watch?v=fJl9QkNllrc](http://www.youtube.com/watch?v=fJl9QkNllrc))

- **Wart hogs** (*Phacochoerus africanus*): Relaxed wart hogs tend to make low frequency grunts. When fighting, males alternate low frequency threats before a grapple, but then grunts and high frequency squeals during one. ([http://www.arkive.org/common-warthog/phacochoerus-africanus/video-12.html](http://www.arkive.org/common-warthog/phacochoerus-africanus/video-12.html))

- **Carolina wren** (*Thryothorus ludovicianus*): Calls of this species fit nicely within a motivation-structure rule scheme. In the table below, aggression increases as one moves to the right, and fear as one moves down. Click on the call name to hear the sound:

<table>
<thead>
<tr>
<th>Increasing aggressive motivation →</th>
<th>A) Pi-zeet</th>
<th>B) Rasp</th>
<th>C) Growl</th>
</tr>
</thead>
<tbody>
<tr>
<td>D) Chert</td>
<td>E) Cheer and Dit</td>
<td>F) Chatter</td>
<td></td>
</tr>
<tr>
<td>G) Pees</td>
<td>H) Nyerk</td>
<td>I) Scees</td>
<td></td>
</tr>
<tr>
<td>Increasing fearful motivation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The calls in the first column (A–G) are tonal and relatively high in frequency. Calls become more harsh and broadband across the rows from left to right. The aggressive endpoint is the growl (C) often given before an attack, and the fear endpoint is the repeated pee, an appeasement call. The pi-zeet (A) is much like a contact call given to locate family members. The rasp (B) is given during mobbing and intraspecific fights. The chert is associated with movement and is highly variable in pitch, repetition rate, and repetition pattern, being lower and harsher when moving toward an opponent and higher and more tonal in submissive contexts. The cheer (E) is a male-only call sometimes given instead of a song during countersinging and sometimes during mobbing; females and juvenile males give the dit call during fights and before roosting. The chatter is a female-only aggressive call that sometimes overlaps her mate’s song and is directed at neighboring females. The nyerk (H) is also a female-only aggressive call often preceding the chatter. Scees (I) are distress calls given by birds that are losing a fight. (After Hagerty & Morton (1995); Owings & Morton (1998); recordings courtesy of Eugene Morton.)

Physiological processes

Piloerection: Autonomic processes such as erection or puffing of hair and feathers for thermoregulation have been ritualized by many birds and mammals, giving them voluntary control for display:

- **Belted kingfisher** (*Megaceryle alcyon*): Many birds have feathered head crests that they can raise or lower at will. Examples include jays, parrots, hoopoes, and kingfishers. While this clip is largely focused on the fishing skills of a belted kingfisher, it also shows the bird repeatedly raising and lowering its crest. ([http://www.arkive.org/belted-kingfisher/megaceryle-alcyon/video-00.html](http://www.arkive.org/belted-kingfisher/megaceryle-alcyon/video-00.html))

- **Superb lyrebird** (*Menura novaehollandiae*): Clip begins with a male’s enormous tail plumage collapsed. Once it begins displaying, it erects the ornate tail feathers into a particular configuration while mimicking the calls of many other local species of birds. ([http://www.arkive.org/superb-lyrebird/menura-novaehollandiae/video-13.html](http://www.arkive.org/superb-lyrebird/menura-novaehollandiae/video-13.html))

- **Coyote** (*Canis latrans*): Many mammals alternately sleek or fluff the fur on their bodies during dominance interactions and overt aggressive interactions. Fluffing of the fur over the whole body is usually associated with fear and defensive motivations (e.g., cats). In this clip, a group of coyotes fights over a carcass. Note how the degree of pelage fluffing or sleeking varies with the relative level of dominance. ([http://www.arkive.org/coyote/canis-latrans/video-08d.html](http://www.arkive.org/coyote/canis-latrans/video-08d.html))

Blood flow: Vasodilation of relevant tissues is a common autonomic mechanism for thermoregulation or preparing tissues for flight or other action. A number of animal species expose patches of bare skin that are highly vascularized. Increasing blood flow into this skin colors it red. Such patches may be inflationable in various ways to produce large red display structures. Whereas many animals show red patches all the time, humans and a few other species can blush transiently and rapidly. Some animal examples with long-term red skin:

- **Magnificent frigatebird** (*Fregata magnificens*): Males have a large and inflatable throat sac that is highly vascularized and red. This clip shows the sac deflated when fighting with other males, and then inflated during display. ([http://www.arkive.org/magnificent-frigatebird/fregata-magnificens/video-09a.html](http://www.arkive.org/magnificent-frigatebird/fregata-magnificens/video-09a.html))

- **Southern ground hornbill** (*Bucorvus cafer*): Both sexes in this species have red vascularized tissues around the eyes and as part of an inflatable throat sac. ([http://www.arkive.org/southern-ground-hornbill/bucorvus-cafer/video-09a.html](http://www.arkive.org/southern-ground-hornbill/bucorvus-cafer/video-09a.html))

Hormone-derived signals

- **Domestic pigs** (*Sus scrofa*): During courtship, males produce a frothy drool containing an androgen-derived pheromone; the male then nuzzles the female around the mouth and snout so she receives the pheromone, and eventually she stands still for mating. ([http://www.youtube.com/watch?v=4tmmv7M0IAk&feature=related](http://www.youtube.com/watch?v=4tmmv7M0IAk&feature=related))

Anti-predator defenses as signal precursors
- **Cotton leafworm moth** (*Spodoptera litura*): Both sexes of this noctuid moth have ears tuned to detect the sonar calls of predatory bats. Males co-opt the tuned auditory sensitivity in females by producing courtship calls in the same ultrasonic range as the bat calls. ([http://www.youtube.com/watch?v=E7kJYU_25To](http://www.youtube.com/watch?v=E7kJYU_25To))

- **Tiger moths** (Arctiidae): Tiger moths not only detect the echolocation calls of predatory bats, but also produce ultrasonic clicks that can "jam" a bat's sonar. Male tiger moths have co-opted this behavior by producing ultrasonic courtship sounds derived from their jamming clicks. This link takes you to the lab web site of Dr. William Conner. Select “Moth Movies,” and then view clips of *Cycnia tenera, Empyreuma affinis, Syntomeida epilais*, and *Euchaetes bolteri*. ([http://www.wfu.edu/~conner/](http://www.wfu.edu/~conner/))

**Literature cited**


Web Topic 10.4
Emotion, Drive, and Motivation

Introduction

Motivation theories attempt to explain how and why animals perform different behaviors at different times. The key principle underlying this book is that animals must continually make decisions about what activity to pursue given the variable nature of the social and ecological environment in time and space. This problem is particularly relevant for mobile animals, and in fact a centralized brain and conscious awareness of the environment are argued to have evolved as a consequence of the “liability” of mobility and the need to make constant decisions (Merker 2005). In this Web Topic unit we briefly review the development of motivation theories. Early models were not construed in ways that were readily testable by physiologists, but subsequent models steadily improved to enable such tests. Emotional constructs became embedded in this motivational framework. In some cases, emotions and motivation are signaled to receivers, or receivers may detect inadvertent cues from behaving individuals that reveal information about their motivational or emotional state. We therefore conclude this unit by discussing the relevance of these theories to the evolution of honest signaling.

“Drive” models of motivation

The first generation of motivational models viewed motivation as part of a homeostatic drive mechanism that energizes animals, pushing them to rectify physiological imbalances and thus satisfy their internal needs. Homeostasis, the tendency of a system to remain at a stable equilibrium, requires a regulatory system that maintains the animal’s physiological state at an optimal set point. If the physiological state moves off the set point, an error-detection mechanism triggers corrective responses to bring the animal back to the set point, like a thermostat (Cannon 1932; Hull 1943). The need to correct the imbalance generates the drive energy. This drive causes animals to respond to those stimuli that will satisfy the need and reduce the imbalance. Appropriate stimuli for the specific drive are learned and refined with experience, according to the response-reinforcement learning paradigm. In this view, response motor patterns that best meet the animal’s needs in a particular stimulus situation are selectively strengthened through reward reinforcement. Hunger and thirst drives seemed to fit this model well. If a deficit in nutrition or water is detected in the body, physiological mechanisms such as low blood glucose levels or body fluid levels initiate and direct food- or water-seeking behaviors. Once found, the animal consumes or drinks until the system set point is restored and the drive is reduced. Evolutionarily built-in goals, such as reproducing (sex drive) or chasing off a threatening rival (aggression drive), were similarly viewed as motivating set points from this homeostatic perspective. These ideas prompted researchers to search for the physiological set points and the deficit signals in the brain (Pfaff 1982). However, numerous problems emerged. First, set points appeared to be highly flexible and variable and thus not very stable (Bolles 1980). This observation lead to the alternative concept of settling points, a stable state caused by a balance of opposing forces, which gives the illusion of a set point but without the homeostatic mechanisms (Wirtshafter and Davis 1977). Second, an animal’s behavior could be strongly affected by anticipation of a goal, by social effects, and by availability of resources; in other words, other cognitive and external factors needed to be considered (Weingarten 1983; Schulkin 2003). Third, the homeostatic mechanism of drive reduction upon fulfilling the physiological need could not be demonstrated, even for hunger and thirst drives. For example, intravenous feeding or introducing food directly into the stomach does not stop an animal or human from consuming more food (Miller and Kessen 1952; Turner et al. 1975).

Another early model was the hydraulic model of motivational drive by the ethologist Konrad Lorenz (Lorenz and Leyhausen 1973), which attempted to explain why the level of motivation builds up as a function of the time since the last reward. This model uses the metaphor of the flush toilet (Figure 1). Drive grows internally like the build-up of pressure in a fluid reservoir, which at some point bursts through an outlet. Proposed internal sources of the build-up could be physiological depletion cues or secreted hormones related to hunger, thirst, aggression, and sex. The model also incorporates the presence of external stimuli that can increase the probability of triggering the motivated behavior.
An external stimulus is more likely to trigger the behavior if the drive pressure is high than when it is low. If the drive pressure is very high, it might burst through the outflow even with no external stimulus, called a **vacuum response**. This model was thus put forward as a way to explain the spontaneous occurrence of behavioral acts by animals that have been prevented from performing their normal behaviors. Although this model was not readily adopted by neuroscientists because it didn’t offer useful details of neural mechanisms, it did deal with the interaction between internal motivating factors and external stimuli in controlling behavior, and offered an explanation of the phenomenon of spontaneous acts by thwarted animals. Hogan (1997, 2005) developed a similar model of motivation around the concept of **energy** as a way to overcome some of the shortcomings of Lorenz’s model. He envisioned an energy variable that has multiple sources of internal and external input energy, and multiple means of dissipating the energy. This model still requires a threshold variable that specifies when the behavior is likely to occur.

The drive models explain some properties of motivated behaviors, such as the increase in goal-seeking behavior as a function of time since the last reward, the ability of animals to learn appropriate responses in a given stimulus situation, and the occasional examples of spillover behaviors, but they fail to provide a physiologically-based answer to the mysterious source of the driving energy, and don’t address the frequent observation of behavioral flexibility of responses by animals. Although they posit separate drive systems for different types of motivated behaviors, these models also don’t address the common observation of interactions between different drives, such as motivational conflict described in Chapter 10 of the main text (Berridge 2004; Anselme 2010).

**Figure 1: Hydraulic drive model of motivation.** Fluid enters a reservoir through the tap, which represents a continuing flow of energy from endogenous sources. The height of the fluid in the reservoir indicates the build-up of drive energy. The fluid is held in the reservoir by a spring-loaded valve, which represents the inhibitory pressure from higher-level brain areas. The weighted scale pan represents the strength of external releasing factors, which can increase the likelihood of valve release.
in addition to the pressure from the fluid in the reservoir. When the valve opens, energy flows through
the outlet into a slanted trough. The holes on the bottom of the trough coordinate muscle action
pattern responses, where the yellow numbers indicate the rank of the response strength. (After
Lorenz 1950.)

“Incentive” models of motivation

In contrast to the energy-driven and drive-reduction models for behavioral decision making, incentive motivation
models envision motivation as a mental state responsible for changing an animal’s receptiveness to specific
environmental stimuli (Berridge 2004; Anselme 2010). The first step in the development of these ideas was the
discovery that hungry rats would suddenly change their behavior when offered tasty (sweet) but nutritionless rewards
(Pfaffmann 1960). Pleasurable sensations (called hedonic rewards) appeared to be motivating factors by themselves,
even in the absence of a drive-reduction effect. Pleasurable sexual sensation, even the anticipation of it, was also found
to be a powerful motivating force (Sheffield 1966). Thus affective reactions (i.e., emotions) such as liking, disliking,
fear, and anger became more explicitly incorporated into this second generation of motivational models. Bolles
proposed that animals were motivated by the learned expectation of hedonic reward, not by internal deficit states
(Bolles 1972). Bindra then clearly laid out the alternate learning paradigm implicit in this view (1974, 1978). Rather than
learn a specific motor pattern in response to a given rewarding (or aversive) stimulus, an animal’s response is
dependent on what it perceives at the time of stimulus presentation. Via learning and experience, animals develop
cognitive representations of correlated sets of environmental stimuli that characterize particular goals or incentive
objects. The suite of stimulus characteristics representing these goals or objects then becomes strongly associated
with the hedonic reward they provide. The objects thereby acquire an incentive salience value. Response acts are
flexibly determined at the moment as a function of the animal’s current internal condition, motivational state, and
sensory inflow from both within the body and from the environment. In contrast to the drive models, a physiological
deficit does not drive the seeking behavior directly, but can magnify the hedonic impact and incentive salience of the
object or conditioned stimulus associated with the goal (Toates 1986). Motivational states thus stimulate perception and
focus attention on key eliciting stimuli in a given context so that the animal performs the appropriate set of behavioral
acts to achieve its goals. These kinds of models account for observed flexibility by animals far better than the
response-reinforcement associations postulated by the drive models. For example, animals often perform different
responses to the same stimulus situation that nevertheless achieve the same goal. An animal may take advantage of a
feeding or drinking opportunity, even though it is not particularly hungry or thirsty, because it is uncertain when such
opportunities will occur next. And animals perform nuanced behaviors in response to complex social stimuli such as a
rival or potential mate.

Incentive models predict that neural pathways for motivational resolution in the brain’s limbic area must project to higher
brain areas responsible for sensory-perception integration of external stimuli, and not directly to motor areas as
predicted by the drive models. Brain circuits and nuclei that fired in response to hedonic stimuli were soon discovered
(Stellar 1982), and a new field called affective neuroscience emerged (Davidson and Sutton 1995; Panksepp 1998;
Rolls 1999; Davidson et al. 2000; LeDoux and Phelps 2000). The search for drive-dedicated neurons initially found no
evidence that specific motivations or emotions were restricted to specific neurons. Rather, stimulation of key regions of
the forebrain basal ganglia and limbic system nuclei (hypothalamus, hippocampus, amygdala, ventral tegmentum,
nucleus accumbens, and ventral pallidum) evokes a range of motivated behaviors and both positive and negative
emotions. These brain regions are connected by circuits and feedback loops of neurons with different types of
neurotransmitters that have excitatory, inhibitory, and disinhibitory effects (Figure 2). The mesolimbic dopamine system
seems to be responsible for the arousal of non-specific “wanting” (Gray et al. 1999). Dopaminergic neurons originate in
the ventral tegmentum of the basal forebrain and send projections to the hippocampus, amygdala, nucleus
accumbens, and prefrontal cortex. Stimulation of the nucleus accumbens in particular initiates seeking or exploratory
behavior (Reynolds and Berridge 2002). The prefrontal cortex is the site of executive function, where attention is
focused on features of preferred stimuli and decision making is accomplished. Activation of the dopamine system then
stimulates the release of the neurotransmitter acetylcholine by cholinergic neurons throughout the brain; acetylcholine
facilitates the integration of sensory input in the cortex, and also operates at neuromuscular junctions in the brain stem
to cause approach and retreat actions. GABAergic neurons send inhibitory signals among the ventral tegmentum, ventral pallidum, and specific areas within the core shell of the nucleus accumbens, and appear to play a major role in activating “liking” responses and positive emotional expressions (Berridge 2003). Complex feedback loops with other types of neurons may then disinhibit these inhibitory effects. Although these circuits and brain regions do not encode specific motivational systems, some evidence is accumulating for the existence of dedicated neuropeptide transmitters for some motivational functions. The hypothalamus appears to possess neurons containing dedicated hunger peptides (e.g., leptin, neuropeptide Y, cholecystokinin, ghrelin) and specific receptors for these peptides. Thirst may be regulated by brain receptors for another neuropeptide called angiotensin II. In the section below on emotions, we summarize the brain circuits involved in other motivational systems. The point here is that different forebrain circuits combine with cortical visual, auditory, and olfactory input in the prefrontal cortex to enable animals to take appropriate goal-directed actions, as predicted by the incentive models.

Figure 2: Brain areas involved in motivation and emotion. Sagittal section through a rat brain showing key nuclei in the limbic system. Main neuronal circuits for mesolimbic dopamine reward system, showing dopamine neurons ascending from ventral tegmentum (red), glutaminergic pathways (blue) linking prefrontal cortex, hippocampus, and basal ganglia, and GABAergic inhibitory neurons (green). Green boxes show key nodes in this system. Other nuclei and brain areas discussed in this web topic unit are also shown. Abbreviations: PAG = periaqueductal gray area; DR = dorsal raphe nucleus; BST = bed nucleus of the stria terminalis. (After Kelley and Berridge 2002; Berridge 2004.)

The next-generation incentive model, called the uncertainty processing theory (UPT), attempts to address three problems with the earlier incentive models: 1) how motivational state actually affects behavior and cognitive processing, 2) how motivational specificity can be achieved, and 3) how interactions and conflicts among different motivational systems are resolved. The UPT argues that motivation is the brain’s solution to the problem of environmental uncertainty about psychologically significant events (Anselme 2010). Motivation is proposed to operate like an information processing center that enables animals to acquire knowledge about the contexts in which important events are likely to occur, highlight such events or their stimulus features as incentive objects, and reduce uncertainty about them through the recruitment of anticipatory and attentional cognitive activity. The model, shown in Figure 3, is derived directly from
current knowledge of motivational processing in the brain.

Figure 3. The uncertainty processing theory of motivation. The flow diagram shows neurotransmitter pathways to the anticipation processing center (yellow) and the attention processing center (light blue); type of neurotransmitter indicated in blue. An event important to the animal’s fitness is indicated by the red box. The animal’s psychological state of uncertainty about this event is represented by the orange box. The probability $P_E$ of an event occurring ranges from 0 to 1; uncertainty is highest at $P_E = 0.5$; certainty is high if the event is either very likely to occur ($P_E = 1$) or very unlikely to occur ($P_E = 0$). Prior experience and knowledge of this event (purple) reinforces the incentive salience value (positive or negative valence) for this event. Excitatory and inhibitory connections among the basal ganglia, together with decisions in the prefrontal cortex, determine the type of motivation elicited by the event and initiate seeking behavior. Approach or avoidance depends on the direction of reinforcement in earlier encounters with the event. (After Anselme 2010.)

High uncertainty about especially salient events such as food items, mates, rivals, or predators, could imperil reproduction and survival, so striving to reduce uncertainty about both liked and disliked events is strongly favored by natural selection. Uncertainty is proposed to be the factor that releases motivated behaviors in changing environments. As mentioned earlier, mesolimbic dopamine plays a key role in causing reward expectation and wanting. The UPT goes one step further by suggesting that dopamine encodes uncertainty for a wide range of positive and negative events. Key evidence for this assertion derives from a study showing that activation of dopamine neurons peaks when the probability of receiving a reward is 0.5, and gradually diminishes as reward probability becomes more predictable (i.e., approaches 0 or 1) (Fiorillo et al. 2003). Moreover, dopamine neurons are more responsive to the anticipation of rewards rather than to the receipt of a reward, and to novel attention-grabbing stimuli (Schultz 1998, 2002). Dopamine projections to the prefrontal cortex, basolateral amygdala, and hippocampus are involved in anticipation of a reward. These areas in turn interact using glutamate and converge their projections onto the nucleus accumbens, which processes anticipation and facilitates flexible approach responses via indirect connections to motor neurons (Ikemoto and Panksepp 1999). In the next step, motivational specificity occurs via a selective attention process. Cholinergic neurons (acetylcholine) originating in the basal forebrain project to cortical regions of the brain and are involved in discrimination and processing of sensory information from all modalities. When an animal in a given physiological state
(e.g., deprived) has the opportunity to establish physical contact with a useful object for that context (e.g., food), the object acquires psychological significance as a consequence of its hedonic value, which can then be represented mentally. The learned association between physiological state and the object’s value establishes a goal and modulates the seeking behavior. The response may be approach or withdrawal depending on whether the direction of reinforcement during prior encounters was positive or negative.

The UPT model offers a possible explanation for the behaviors animals perform when confronted with two conflicting events, as we described in the section on motivational conflict in Chapter 10 of the main text. Animals and humans presented with two types of salient stimuli show ambivalence, delayed reaction times, and increased error rates; if the responses are truly incompatible, one type of response eventually prevails while the other is inhibited (Figure 4A). This could be caused by the limited ability of the brain to attend to two different events at the same time, given the common pathway for anticipation and selective attention. The activation of the dopamine-cholinergic pathway for one event means that it is less available to deal with another event; the attention and response thresholds for this event are therefore raised (Figure 4B; see also main text Figure 10.7). If two events are both strong stimuli, they may each mutually inhibit the other as a consequence of attentional interference (Anselme 2010).

Figure 4: Attentional interference during motivational conflict. Two motivational causal factors are shown in both graphs with numbers indicating the different factors (e.g., thirst and hunger, or fear and anger). The attentional threshold is shown by the bar labeled A(1 or 2) and the higher response threshold by R(1 or 2); the bracketed M region indicates the motivational strength for the respective factor. (A) When only one motivational causal factor is stimulated to a level above its attentional and response thresholds, the attention and response thresholds for the other causal factor are raised (dashed arrows). The animal focuses its attention only on the strongly motivated stimulus and performs the appropriate response behavior, while attention and behaviors appropriate for the other causal factors are inhibited. (B) If two causal factors are both simultaneously stimulated, each raises the thresholds for the other one, but if the intensities of both stimuli are above their thresholds, the consequence is attentional interference (tan arrows), which in this case is stronger for M1 compared to M2. (After Anselme 2010.)

Whether or not dopamine turns out to encode event uncertainty, it clearly does instigate an urge to adopt resource-seeking behaviors such as approach, exploration, and investigation. In humans, dopamine injection causes euphoria. Which leads us to a discussion of the role of emotions in motivating behavior.

**Emotions**

Emotions are an important component of motivational systems. As discussed above, wanting is a dopamine-dependent process that makes animals more receptive to certain stimuli and transforms them into desired goals—the incentive salience component of motivation. The wanting system probably evolved early in vertebrate evolution to mediate the innate pursuit of key objects such as food and mates and the avoidance of stimuli associated with
predators. Although desired goals are usually liked goals (and dangerous objects are fear-inducing objects), liking and disliking represent distinct hedonic (emotional) processes compared to wanting, and liking and wanting are facilitated by different neurotransmitters and neural circuits. Expressions of hedonic pleasure or displeasure and associated neurological changes occur upon consumption of a reward (e.g., tasty food versus bitter food). Such emotions operate as reinforcers during the process of learning the incentive value of various environmental stimuli (Cardinal et al. 2002; Berridge and Robinson 2003). Within the construct of motivation, the feelings generated by hard-wired emotional systems facilitate the acquisition of environmental knowledge and skill. In the broader view, different emotions stimulate organisms to respond quickly to external events in ways that optimize fitness (Darwin 1872).

Emotion is an umbrella term that includes affect (feelings), along with cognitive, behavioral, communicative, and physiological manifestations. Emotions are triggered by external events and often lead to rapid onset and short duration effects, in contrast to moods, which are brought on by less obvious antecedents and have longer lasting effects. Most mammals possess a core set of emotions—seeking, fear, anger, play, panic, lust, and care (Panksepp 2005). Birds and other vertebrates also show many of these emotional behaviors. Evidence for these core emotional systems includes the ability to elicit specific response behaviors by stimulating key subcortical brain nuclei (Panksepp 2003a), and in human fMRI studies the activation of different brain regions by individuals experiencing different emotions (Damasio et al. 2000; Phan et al. 2002). In humans, emotional responses occur on two levels: a subconscious level that involves subcortical pathways, the autonomic system, and involuntary behaviors and facial expressions; and a conscious level that directs the behaviors to appropriate targets and modulates some of the involuntary expressions. We don’t know the extent of animals’ cognitive experiences during emotional episodes but we do see evidence of the physiological effects, communicative expressions, and directed responses. Rats and monkeys will eagerly press levers to obtain an injected dose of heroin, a plant-derived opiate that mimics the pleasurable effects of endogenous opioids found in all mammalian species. Below is a brief overview of the core emotional systems. Each one is based on a different brain circuit that links basal ganglia, thalamus, and a distinct portion of the frontal lobe (Alexander and Crutcher 1990).

The **seeking system** is the generalized activation system recruited by all vertebrates to organize the pursuit of food, mates, and shelters. The energizing (wanting) component is driven by the dopaminergic-cholinergic pathway described earlier. Activation of the ventral tegmental area and nucleus accumbens plays a critical role in causing both approach and aversive action responses (Ikemoto and Panksepp 1999; Ikemoto 2007; Zellner and Ranaldi 2010). The hedonic (liking) component is caused by opioid neurons in the nucleus accumbens shell, ventral pallidum and brainstem. These neurons synthesize endogenous opioid peptide transmitters such as endorphins and enkephalins, as well as the specialized G protein receptors for these neurotransmitters. Stimulation or activation of these regions is responsible for the pleasurable sensations that occur once a tasty food item has been consumed (Berridge 2003). In rats, primates, and human infants, the sensation of sweet taste (from taste bud sensors) triggers a characteristic tongue-protruding response, while bitter taste triggers gaping (or the disgust facial expression in older humans). The wanting and liking circuits of the seeking system can be separately knocked out without destroying the other component. Together, the pleasurable experience of an object reinforces the subsequent motivated pursuit of similar objects.

The **fear system** prepares an animal to take appropriate rapid actions in dangerous contexts. It has a dedicated circuit of excitatory (acetylcholinergic and epinephrinergic) neurons. The neural response is initiated in the lateral and central amygdala upon sensory detection of learned or innate cue features. The amygdala plays a key role in distinguishing between aversive and positive stimuli (Etkin et al. 2006; Tye and Janak 2007; Shabel and Janak 2009; Bermudez and Schultz 2010; Morrison and Salzman 2010; Tye et al. 2011). If the stimulus is deemed dangerous, a cascade of involuntary responses is mediated by nerve projections to the anterior and medial hypothalamus, on to the periaqueductal gray area (PAG) of the midbrain, and then to the lower brain stem and spinal cord (Panksepp 2004). This circuitry stimulates the sympathetic nervous system (a component of the autonomic nervous system) and triggers the release of the hormone epinephrine (also called adrenaline) from the adrenal gland, which targets various organs and causes an increase in heart rate, blood pressure and perspiration while inhibiting other non-essential body functions. The behavioral response to stimulation of these specific subcortical brain areas is freezing, startling, or fleeing; cognition may be involved in assessing which of several alternative escape behaviors is best given the current
conditions. This emotion is sometimes accompanied by a scream vocalization, and in humans by a fearful facial expression.

The context and function of the remaining core emotions is the mediation of social interactions. The anger system shares some circuits with the fear system. The orbital frontal cortex first processes sensory input from social cues and signals, typically olfactory in the case of rodents and visual and vocal in the case of primates. Neural pathways then descend to the medial amygdala, which, as in the fear system, evaluates whether the stimulus is aversive or positive. Pathways then extend to several areas in the hypothalamus, thalamus, lateral septum, and bed nucleus of the stria terminalis. These centers send projections to the periaqueductal gray area and the lower brain stem, as in the fear system. The hypothalamus also connects directly to the pituitary gland, which releases various hormones directly into the body's circulatory system that target the adrenal gland. Noradrenaline (norepinephrine) plays an important role in aggression as both a neurotransmitter in the brain and as a circulatory hormone produced by the adrenal gland. As in the fear response, these hormones increase heart rate and blood supply to the muscles and prepare an animal to fight if necessary. Various aggressive signals, facial expressions, and postures are involuntarily produced from the output of these subcortical ganglia. In addition, neural pathways from the subcortical ganglia ascend through the thalamus into the cortex to permit some voluntary control and assessment of the anger-evoking stimulus. The dopamine motivational circuit must be intact in order for animals to display the full aggressive behavioral repertoire toward a rival individual. Finally, steroid hormones such as testosterone and estrogen are also necessary to sustain aggressive behavior. Testosterone seems to primarily have an organizational effect on the brain, especially in males, by making aggression-inducing stimuli more salient; testosterone affects responsiveness in the lateral septum, amygdala, and dorsal raphe nucleus. Testosterone also promotes the development of aggressive behavioral skills through play behavior in young mammals (Nelson and Trainor 2007).

The play system also seems to have its own set of circuits, although they overlap with circuits for other emotions. Play behavior is the first type of non-mother-directed social interaction in a young mammal's life. It provides a crucial opportunity for learning adult social skills, practicing aggressive behaviors, establishing one's position in the dominance hierarchy of peers, and discriminating the sexes, and natural selection has imbued it with a pleasurable motivating reward system. The opioid and dopamine reward circuits are recruited to motivate play behavior. Cholinergic, noradrenergic, and opioid neuron circuits underlie the attentional processes needed for focusing on the rapidly performed actions and learning from errors. The cortex is not involved in play initiation, but it does affect play performance, in the sense that decorticated animals are hyperactive and very aggressive. Built-in mechanisms thus inhibit lethal acts during play. Playing animals also spontaneously give characteristic playful vocalizations. As mentioned above, testosterone is involved: males are more likely to initiate play and respond to play initiation signals by other males (Siviy and Panksepp 1987a, b; Panksepp et al. 1994; Vanderschuren et al. 1997; Knutson et al. 1998).

The panic system operates primarily in young mammals that have been separated from the parent, and elicits vocalizations of pain, distress, and crying. This behavior appears to be initiated in a region of the cortex called the anterior cingulate. Stimulation of this area leads to activation of the bed nucleus of the stria terminalis, the ventral septal and dorsal preoptic areas, and then on to the dorsomedial thalamus and the periaqueductal gray area of the brain stem. This circuit is called the thalamo-cingulate limbic pathway (Herman and Panksepp 1981; Panksepp 2003b; Newman 2007; Panksepp and Watt 2011). No learning is involved in this pathway, as it occurs in very young infants after being isolated. The effects can be modulated and diminished by the application of various opioid peptides (e.g., oxytocin) and the monoamine neurotransmitter serotonin.

Feelings of social attachment and bonding in mammals (care and lust systems) are facilitated by release of oxytocin and vasopressin, often called love hormones (Nelson and Panksepp 1998). These peptides are synthesized by specialized cells in the paraventricular nucleus of the hypothalamus. Some of these cells extend projections into other parts of the brain, such as the prefrontal cortex, basal ganglia (amygdala, ventromedial hypothalamus, septum, nucleus accumbens), and brain stem (Morgane et al. 2005), where they operate as neurotransmitters. Receptors for these neurotransmitters are also synthesized in these brain regions. Other cells in the hypothalamus send projections to the nearby pituitary gland, which stores oxytocin and vasopressin and releases these chemicals into the blood stream, where they operate as hormones. They target various organs involved in reproduction, including mammary glands and
uterus in females and gonads in both sexes. Since these peptides are too large to pass the blood–brain barrier, their joint physiological and psychological effects are believed to be coordinated by synchronous release into the brain and circulatory system. Oxytocin is particularly important in female mammals. It increases immediately after birth and regulates aspects of bonding with infants (Nowak et al. 2007). A large surge of oxytocin occurs during sexual behavior and orgasm. In monogamous mammals such as the prairie vole (*Microtus ochrogaster*), administration of central oxytocin induces pair bond formation and greater social contact; this species has many more oxytocin receptors than closely related rodents with polygamous, non-pair-bond social systems (Williams et al. 1994). Vasopressin is chemically similar to oxytocin and plays a role in facilitating paternal care in male mammals. Some other neurotransmitters also facilitate positive social interactions. Endogenous opioids such as norepinephrine and epinephrine are rewarding and can induce odor and place preferences; they are also released during bouts of affiliative interaction such as suckling, physical contact, allogrooming, and social play. Opioids are postulated to encourage animals to engage in affiliative social behaviors by inducing a euphoric state (Nelson and Panksepp 1998). The dopamine motivation and reward system also plays an important role in reinforcing sexual behavior in male mammals (Agmo and Berenfeld 1990).

An important message to be extracted from these summaries of the neurophysiological bases of emotional systems is that each is associated with diagnostic expressions—in other words, communication signals. Food-seeking is associated with expressions of liking, such as smiles and tongue extrusion, and expressions of disliking such as gaping and disgust facial expressions. Fear is associated with screams and fearful expressions; anger with aggressive postures, staring, weapon presentation, and mouth expressions; and panic with cries and expressions of pain and sadness. Care and lust systems are associated with physical contact gestures and smiles; and the play system is associated with invitation postures, play faces, and laughter. For the most part, these expressions occur involuntarily based on subcortical neural pathways to various motor systems; they can occur in animals with a non-functional cortex. They are therefore honest indicators of the emotional feelings the sender is experiencing. The signals have evolved for this purpose because of direct social benefits to senders and costs of cheating. Nevertheless, they can often be modulated and controlled voluntarily to some degree because of cortical loops in all of the systems. This issue is discussed in detail for human emotional expressions in Chapter 16.

A final point is the inevitable urge by biologists and psychologists to categorize emotions, especially in humans, because we have many more described emotions than animals do. Early attempts by several researchers converged on a two-dimensional model, in which emotions are placed on a grid or circle with two orthogonal axes, one representing level of pleasantness (e.g., positive versus negative feelings, or valence), and another axis representing arousal level (high versus low) (Russell 1980; Watson and Tellegen 1985; Thayer 1986; Russell et al. 1989; Larsen and Diener 1992; Yik et al. 1999). **Figure 5** shows a melded version of these models.
This type of model was subsequently expanded in several ways. One idea was to greatly increase the number of divisions around the circle to 28, comprised of 14 bipolar pairs of emotions (calm–tension, certainty–uncertainty, compassion–anger, fun–boredom, pleasantness–unpleasantness, happiness–sadness, pleasure–pain, satisfaction–frustration, desire–reject, love–hate, courage–fear, strength–tiredness, enthusiasm–apathy, arrogance–humiliation). Opposing emotions in each pair are situated on opposite sides of the circle, and emotions are ordered around the circle so that similar emotions are adjacent; the earlier concept of two orthogonal axes, pleasantness and arousal, was also retained (Diaz et al. 2001). Another strategy was to add a third dimension. Figure 6 shows an example with somewhat similar ordering of categories around a circle, but with a third dimension depicting intensity (Plutchik 2001).
Figure 6: Three-dimensional circumplex model. The terms in black lettering around the circle represent the primary emotions. There are four pairs of primary dyads, with the contrasting type on the opposite side of the circle. The dyads are arranged so that similar emotions are adjacent, as in a color wheel. Colored regions represent gradations between the primary emotions. Emotional intensity is represented by the radius of concentric circles, folded to create a three-dimensional structure where intensity is the third dimension, as shown in the inset. This model does not reflect the orthogonal pleasantness and arousal axes of the earlier models. (After Plutchik 2001.)

The most recent models have used multivariate statistical analysis of a large number of variables describing different emotions, including appraisal, psychophysiological changes, motor expressions, action tendencies, subjective experiences, and level of emotional regulation and control to statistically sort out the number of dimensions needed to encompass all of the emotional terms. This type of analysis finds that three or four dimensions are required: hedonic valence, arousal, degree of control or power, and unpredictability (Laukka et al. 2005; Fontaine et al. 2007). One such model is illustrated in main text Figure 16.26. Although affective neurobiologists do not find these psychological models helpful, it could turn out to be the case that the emotional axes do correspond to levels of analysis of emotional contexts in the brain. Certainly the amygdala performs an initial analysis of valence (positive or negative), and dopamine and serotonin circuits then determine the motivation or arousal level. Emotions that are close to each other
on the wheel may use more overlapping components of the neural circuits, and combination emotions may vary and balance the ratios of different neurotransmitters. As new techniques are developed for monitoring subtle neuronal and chemical changes in the brain in behaving animals, the differences and similarities among emotions may be better elucidated (Panksepp 2003a; Tye et al. 2011).

Literature cited


Web Topic 10.5  
Evolutionary Game Theory  

Introduction  

In Chapters 9 and 10 of the text, we introduced the basic logic of evolutionary game theory. However, we did not provide any guidance on how to find evolutionary stable strategies (ESSs). It turns out that appropriate methods depend a lot on how the game is structured. In this Web Topic, we provide more details on the underlying logic of evolutionary game theory, outline a general schema for classifying evolutionary games, and then provide some examples of finding ESSs for several game categories. Because this field has grown very rapidly since Maynard Smith’s influential book in 1982, this Web Topic can at best serve as an introduction. Our goal here is to provide just enough background to make the discussions in the text understandable. At the end of the module, we provide references that, with varying degrees of accessibility, allow readers to go beyond the introductory material presented here.  

Basic concepts and definitions  

- **Games versus simple optimality**: When an animal adopts an optimal behavior to cope with some inanimate context, the latter will not change in response. For example, if an animal’s optimal behavior on a hot day is to move into the shade, the weather does not respond by snowing to spite the animal. It could snow, but not because the animal moved into the shade. However, this is not the case if the context in which the animal must find an optimal behavior involves other living creatures. Unlike the weather, living creatures are likely to change their physiology and behaviors according to what our focal animal decides to do. Usually, both parties seek to maximize their own payoffs despite the actions of the other party. The interaction of two or more animate parties each trying to select the optimal behavior given the behaviors of the others constitutes a game. Finding an optimal strategy in a game is clearly different from finding an optimal strategy in an inanimate context.  

- **Game equilibria**: Each player in a game should adopt the best response to the behavior selected by its opponent. At first glance, this might seem likely to lead to an endless chain of moves and countermoves. However, if the players each hit on a best response to the other’s best response, the process will stop. At such a Nash equilibrium, it does not pay for either player to shift strategies (Nash 1950, 1951). Note that the players may have access to the same alternative strategies (a shared role) or different strategies (different roles). In either case, it is the combination of strategies adopted by the parties that defines the equilibrium. Both pure and mixed strategies (a fixed ratio of pure strategies) can lead to Nash equilibria. A Nash equilibrium can be strict, in that only one strategy for each party is its best response, or non-strict in that one or both parties can play any of several equivalent strategies and still maintain the equilibrium. Neither player need maximize their payoffs at a Nash equilibrium; in fact, either or both might do better at some other combination of player strategies. However, once both players hit on each other’s best response, it will not pay for either to shift to another strategy. Depending on the game, there can be no, one, or many possible Nash equilibria. In the latter case, which one is observed may depend only on chance and history.  

- **Evolutionary games**: While classical economics focused on the optimal interactions of a particular pair of game opponents, evolutionary game theory shifts the emphasis from specific individuals to the relative success of alternative strategies in a population of animals. If animals that adopt a given strategy have higher fitness than those in the same role but who adopt alternative strategies, the former will spread in the population over evolutionary time. In a large and well-mixed population, animals playing one strategy encounter other animals playing the same or different strategies in proportion to their relative abundances. The average payoff for a given strategy is then the sum of the payoffs between it and all strategies currently in the population, each discounted by the current abundance of that strategy. Because this average payoff varies depending on the current mix of strategies in the population, evolutionary games are always frequency-dependent. Although played out by
individuals within a generation (the “inner game” of Vincent and Brown [1988]), evolutionary games are basically multi-generational competitions between strategies (the “outer game”).

- **Evolutionarily stable strategies**: An evolutionarily stable strategy (ESS) is one that, when common, cannot be invaded by any rare alternative strategy (Maynard Smith and Price 1973; Maynard Smith 1982). This means that an ESS strategy is the best response both to itself and to any other rare strategy. If two ESSs are each the other’s best response, they generate a Nash equilibrium. The two ESSs will be the same strategy if the interacting animals have the same role. In this case, the Nash equilibrium can be strict in that only one pure strategy will be the ESS for all parties. Alternatively the equilibrium may be non-strict in that the ESS is some mix of strategies, any of which when given as part of the mixture or singly will provide equally good responses. If the interacting parties have different roles (e.g., males versus females or species X versus species Y) and the available strategies differ for the two roles, there will be two ESSs defining the equilibrium: one for each role. In this case, each ESS must be a pure strategy, and the corresponding Nash equilibrium will be strict.

- **Roles and equilibria**: Where two or more roles are present in a game, each role is likely to have its own strategy set. Where this is the case, any equilibria are called separating equilibria. Alternatively, if multiple roles adopt the same strategy set, a stable outcome is said to be a pooling equilibrium. It is also possible that some roles pool strategies whereas other roles each have their own strategy sets. Any equilibria in this situation are said to be semi-pooled. While these distinctions are not needed if there is only one role, or even for simple games with a few roles and a few strategy options per role, they do become useful in describing outcomes of games in which strategy sets are continuous and players may differ widely in available strategies. For example, in games describing optimal levels of male courtship display, males may differ in health or condition (role). If they each adopt different (honest) levels of display, any ESS would be a separating equilibrium; if those with less than average health display at one level and those of higher than average display at another, any ESS would be a pooling equilibrium. A semi-pooled equilibrium could occur if the less healthy males all adopted one display level, but above average males each used a display corresponding to its health.

- **Equilibrium dynamics**: The original definition of an ESS assumed haploid genetics, infinite population sizes, and sufficient “mixing” so that the relative abundances of the alternative strategies were identical everywhere. The definition also assumed that only one type of invading mutant was likely to occur at any time. This was in many ways a “static” definition in that it ignored whether and how strategies that originally appeared as rare mutants might increase in frequency until they arrived as dominant components of an ESS. The assumption that the population was infinite eliminated the effects of drift and thus any worries about what would happen if a population drifted off of an ESS equilibrium. Would it come back over time, or just wander away? And was an equilibrium that was resistant to invasion by a single mutant strategy necessarily resistant to multiple types of invaders at the same time?

Such questions inspired subsequent research on the adaptive dynamics of evolutionary games instead of just focusing on the end-points (Taylor and Jonker 1978; Eshel 1983; Eshel and Feldman 1984; Hofbauer and Sigmund 1988; Vincent and Brown 1988; Abrams et al. 1993; Dieckmann and Law 1996; Metz et al. 1996; Eshel et al. 1997; Geritz et al. 1997; Geritz et al. 1998; Hofbauer and Sigmund 1998; Weibull 1995; Cressman 1998; Vincent and Brown 2005; Nowak 2006; McGill and Brown 2007; Dercole and Rinaldi 2008; Apaloo et al. 2009; Apaloo and Butler 2009; Cressman 2009). These studies made it clear that the conditions for an ESS were insufficient to account for the many types of equilibria that might arise in evolving systems. They were not even sufficient for the persistence of ESS strategies in nature. In particular, two additional strategy properties are needed to fully describe all likely trajectories and equilibria.

To visualize the significance of the three strategy properties, it is useful to think of adaptive landscapes (see Chapter 9). Normally, one envisions populations as moving up the slopes of an adaptive landscape towards regions of higher average fitness as the relative frequencies of alternative strategies shift due to selection and mutation. An equilibrium occurs when the population arrives at a location where the slope of the landscape is zero. Such an equilibrium can occur on a peak (a fitness maximum), in a pit (a fitness minimum), or along a
valley (a saddle point where fitness increases along one axis but decreases along an orthogonal axis). Evolutionary games are more complicated than simple optimization processes because the shape of their adaptive landscapes can change as the frequencies of alternative strategies change (Eshel 1983; Abrams et al. 1993; McGill and Brown 2007). What starts as a population’s trajectory uphill can gradually change into a flat or even downhill trajectory as successive changes in population composition change the landscape. This greatly complicates the possible trajectories and affects the types of equilibria that might be encountered.

The original definition of an ESS dodged the possibility that the underlying landscape might change by assuming that only a single type of mutant could appear at a time, and when one did, it was so rare that its presence did not change strategy frequencies or the adaptive landscape. If either condition failed to be met, the landscape could change, and then the strategy might no longer be uninvadable. Adaptive dynamic analyses indeed identified systems in which an ESS might not be sufficient for long-term stability in real populations. However, invoking a second and independent property of strategies, convergence stability, solves this problem. An equilibrium strategy is convergence stable if a mutant strategy closer to it in the adaptive landscape than the current mean of the nearby population can always invade (Eshel 1983; Abrams et al. 1993; Rand et al. 1994; Eshel et al. 1997; Geritz et al. 1998; Leimar 2005, 2009). Successive mutants would then move the population closer and closer until it reaches the equilibrium. Equilibrium strategies that are convergence stable are called “attractors.” How does this help? Unlike the assumption made to define ESSs, natural populations are finite and experience drift in strategy composition. Suppose a population at an ESS drifts to neighboring but different strategy values. Without convergence stability, the population will be under no selective pressure to move back. Convergence stability can both move a population to a nearby ESS in the first place and keep it there afterwards.

An additional relevant property is neighborhood invadability (Geritz et al. 1998; Apaloo et al. 2009; Apaloo and Butler 2009; Cressman 2009). This is the ability of a strategy to invade neighboring strategies that are common even if it is rare. This property allows a mutant strategy that could, if common, be an ESS to gain a foothold in a population currently using other strategies. Note that as this mutant invades, it will change the relative frequencies of strategies in the population, and this could alter the adaptive landscape. Just because a strategy can invade when rare does not guarantee that it will be favored when more common. A strategy that is a good neighborhood invader in one context may or may not also be an ESS in another context.

If strategies can be either ESS or not, convergence stable or not, and neighborhood invaders or not, there are eight possible combinations (Geritz et al. 1997; Geritz et al. 1998; McGill and Brown 2007; Apaloo and Butler 2009; Cressman 2009). Only six of these are biologically realistic, and when one controls for some dependencies between options, two combinations stand out as most relevant. Perhaps the most important combination is a strategy that is an ESS, convergence stable, and a good neighborhood invader. Such a strategy will move up to a peak in the adaptive landscape where it will become a stable equilibrium. It is now widely believed that the persistent strategies one sees in nature usually represent this combination. The second relevant combination is a strategy that is convergence stable but not an ESS. These usually occur as minima or saddle points in the adaptive landscape and can, at least theoretically, give rise to evolutionary branching in which the population splits into two groups with divergent trajectories. These cases may be one source for stable behavioral polymorphisms within a species or even speciation.

- **Analytical formats**: A variety of techniques have been developed for finding ESSs. Two of the more traditional approaches were already described on p. 330 and Figure 9.4 in the text. Below, we briefly review these again and contrast them with the adaptive dynamics approach:

- **Extended form games**: consist of a branching tree whose junctions represent possible alternative actions by nature or the players (Gintis 2000; Cressman 2003; Hurd and Enquist 2005). The tree represents all possible paths that could be followed during the course of the game. Payoffs to each player are then provided at the end of each final branch. Extensive form games retain the greatest detail in a game’s structure. If a game requires sequential moves by players or nature, the cumulative process can be tracked in detail. It is also quite easy to examine the dynamics of sub-games (parts of the tree) in this format.
- **Normal form**: is the other traditional approach (Maynard Smith 1982). Here, the alternative strategies are listed in a matrix as rows for one player and columns for the other, and the payoff values for each combination of strategies are placed in the respective cells of the matrix. Collapsing an extensive form game into normal form eliminates most tracking and sub-game information. However, as we shall show below, one can often find ESSs solely with the normal form matrix; the details in the extensive form version of these games, while perhaps interesting, are often not essential to identifying stable equilibria. On the other hand, if the game consists of many successive and conditional choices of strategy, extensive form may be the only way ESSs can be identified.

- **Adaptive dynamics**: uses many of the tools and generalizations from dynamic systems analysis (Strogatz 1994). Instead of focusing on end points, adaptive dynamics analysis derives an equation for each strategy that predicts the change in that strategy’s abundance when it is played amidst the current abundances of each strategy in the population, the current size of the population, and the relevant fitness payoffs of each strategy combination (Taylor and Jonker 1978; Hofbauer and Sigmund 1998; Weibull 1998; Vincent and Brown 2005; Nowak 2006). If changes occur at intervals (e.g., no overlap of generations), the output of the relevant difference equation is the abundance of the strategy in the next time interval. If change is relatively continuous (e.g., overlap of generations and large populations), the output of the differential equation is the rate of change in the abundance of the strategy. If the population size is fixed or assumed to be infinite with good mixing, population size can be deleted from the arguments of the equation and the absolute abundances of strategies can be replaced by relative frequencies (i.e., fractions). There should be one such equation for each alternative strategy. Where the equations focus on relative strategy frequencies, one can visualize the resulting trajectories by placing them in a simplex (see Nowak 2006). This is an equilateral graph object with dimensions one less than the number of strategies being considered: two strategies have a single straight line for a simplex, and a three-strategy simplex is an equilateral triangle. Each point in the simplex represents a possible population mixture of the strategies: the vertices represent a single pure strategy, edges represent mixtures of two strategies, faces represent mixtures of all but one strategy, and the interior points are mixtures containing all strategies. One can begin analysis at a given point in the simplex and use the relevant equations iteratively to find successive points in the evolutionary trajectory. Equilibria occur when the trajectory reaches a point at which the equations all predict zero change in strategies in the next time interval. The type of equilibrium (fitness peak, fitness minimum, or saddle point) and whether the equilibrium is convergence stable or not can be determined by examining higher order derivatives of the equations. Where the mathematics are too complicated for analytical results, one can use simulations beginning from many different initial points to characterize likely trajectories, find equilibria, and identify their stabilities.

A variety of dynamic equations have been used in this manner. One, the *replicator equation*, focuses on relative frequencies and ignores mutation and population size (Taylor and Jonker 1978; Hofbauer and Sigmund 1988, 1998; Nowak 2006). It equates per capita changes in strategy frequencies with the difference between the fitness of each strategy and the population average fitness. Though seemingly simple in general form, it has proven to have considerable flexibility and utility in many specific applications. The *canonical equation of adaptive dynamics* is very similar but replaces differences in fitness between an animal adopting a given strategy and the current population mean with derivatives of fitness versus strategy in its arguments (Dieckmann and Law 1996). Other approaches such as *G-function* equations predict fitness and use derivatives of the G-function to monitor change. These equations often include strategy abundances, resource availability, and other factors as arguments. This facilitates the application of evolutionary game theory to multiple species contexts in ecology (Vincent and Brown 1988; Vincent and Brown 2005; McGill and Brown 2007).

### A general schema for classifying evolutionary games

There are six independent criteria that one can invoke to classify any evolutionary game. Different combinations of values for these criteria define different game categories. Each category tends to have its own shortcut methods for identifying ESSs, its own set of assumptions, and its own types of ESSs. The criteria are:
**Types of strategy set**: The first criterion for classifying games involves the nature of the strategy set. The strategy set includes a list of alternative behaviors, physiologies, or anatomical structures that each individual in that role could adopt. These alternative strategies can either be discrete (e.g., produce a red tail or a green tail), or continuous (e.g., amplitude modulate a 1 kHz sine wave at some rate between 50 and 453 Hz). In either case, the strategy set may also permit mixtures of the alternative behaviors or structures (e.g., sing a 60 Hz song 20% of the time and a 130 Hz song the remaining 80%). As noted earlier, a strategy that consists of a single behavior or structure is called a pure strategy, whereas a stable mixture is often called a “mixed ESS.” Note that a mixed ESS can take one of two forms: a genetic polymorphism occurs when each member of the same role performs only one strategy, but different members of that role perform different strategies. For example, suppose 40% of male moths in a given species emit organic acid $A$ as a pheromone, whereas the remaining 60% emit organic acid $B$. No male can emit both. Females, which represent an opposing role, perceive the male population as a 2:3 mixture of the two pheromones. The alternative form is a behavioral polymorphism. Here each individual performs a mixture of alternative strategies. Continuing our moth example, suppose all males emit pheromone $A$ 40% of the time and pheromone $B$ 60% of the time. Females still encounter male pheromones with a 2:3 ratio, but the mechanisms by which the mixture is achieved are different. Note that some authors restrict the term “mixture” to behavioral polymorphisms. While it usually does not matter to opponents whether a mixed ESS is generated through a genetic or behavioral polymorphism, evolutionary trajectories often differ for the two mechanisms and this affects the likelihood that a given role will ever achieve a stable mix (Bergstrom and Godfrey-Smith 1998). For example, inheritance patterns might allow behavioral polymorphisms but prohibit genetic ones. Strategy sets will thus differ in the degree to which one or the other form of mixed ESS is an option.

**Interacting player number**: Evolutionary game theory initially focused on pairwise interactions. Even though there might be many animals in a population adopting multiple strategies, any given interaction was assumed to involve only two players at a time. Most of the examples we outline below make this assumption. However, there has been considerable effort in recent years to examine games in which three or more players interact at the same time and no payoffs are distributed until each individual has played its strategy. This provides an opportunity for players to cooperate, punish each other, or otherwise alter the strategies they might have adopted if only two-individual interactions were allowed. Trajectories for multi-person games can be quite complicated (Broom et al. 1997; Bach et al. 2006).

**Role symmetry**: A third criterion for game classification is role symmetry. The opponent players in a symmetrical game have access to identical strategy sets, have equal chances of winning when playing a given strategy against another given strategy, and have identical payoffs when playing a given strategy against another given strategy. The players are thus entirely interchangeable; or put another way, there is really only one role. Two identically aged male baboons exchanging vocalizations over who gets a favored sleeping site are likely involved in a symmetrical game. In an asymmetrical game, there is more than one role, and each may have access to different alternative strategies, different probabilities of winning with a given strategy, different payoffs when they win with a given strategy, or some combination of these conditions. Asymmetric games are typical of interactions between male versus female, young versus old animals, or big versus small players. A vocalizing contest between a large adult male baboon and a small juvenile over a sleeping site is unlikely to be a symmetrical game as the probabilities of winning a fight are surely unequal, each individual is likely to have available strategies that the other cannot adopt, and the payoffs of winning may have different fitness consequences for each party. A single exchange of animal communication is always an asymmetrical game because sender and receiver do not have identical strategies. If there are multiple exchanges, and the parties alternate roles as sender and receiver, then the overall sequence can be symmetric.

**Contests versus scrambles**: All evolutionary games are frequency dependent because the average payoff of a given strategy depends upon the current abundances of alternative strategies in the population. However, a second level of frequency dependence may or may not be present in an evolutionary game. If the payoffs for a single interaction depend only on which strategies each party adopted, we say the game is a contest. The payoffs of this single interaction are thus independent of the abundances of alternative strategies in the population. There is then no second level frequency dependence. However, if the payoffs in a single interaction
vary depending upon the relative abundances of alternative strategies, the game is called a scramble (also called playing against the field). Some examples will help clarify this distinction.

Consider first a population of animals that regularly hunt food that occurs only in patches or items so small that they are indivisible. Only one animal gets to eat such a find. When several animals find the same patch, each can adopt any of several fighting tactics to win control over it. Suppose that the probability that an animal using Tactic A will win when competing with an animal using Tactic B is fixed, as is the amount of food obtained if one wins. We can also include some fixed costs to each party depending on which tactics each adopts. Although there are a number of variables, and one can include some stochasticity in who wins, the expected payoff to each party for a particular combination of tactics is independent of how many other animals in the population have adopted those or alternative tactics. This is thus a contest. It is still a frequency dependent game in the first sense because the average payoff over of any given tactic over many interactions depends on the frequency with which an animal playing that tactic encounters opponents playing the same or alternative tactics. It is not however frequency dependent in the second sense defined earlier.

Now consider male birds establishing nesting territories. There are two alternative strategies: settle in the thick woods or settle in the sparse woods. Thick woods provide more food and better cover per unit territory area and are thus preferred. However, as more and more males settle in the thick woods, they compress the territories of existing residents and make the average territory less and less suitable. Eventually, being the first settler in the sparse woods results in a territory that is as good or better than that obtained by being one more settler in the thick woods. The payoff of each strategy thus depends on how many males have already settled in each habitat. Because both the average payoff of adopting a strategy (frequency dependence in the first sense) and the payoffs of single interactions (frequency dependence in the second sense) vary with the relative abundances of the alternative strategies, this game is a scramble (Parker 1984; Krivan et al. 2008; Cressman and Krivan 2010).

- **Temporal pattern**: The fifth criterion has to do with the degree to which successive interactions should be treated as independent games or not. All animals face a sequence of decisions during their lives. Even during a single fight, courtship, or parent–offspring interaction, there may be several successive points at which each player must make a strategic choice. It is critical to know whether the outcome of an earlier choice of strategy does or does not affect the suitability of alternative strategies at a later decision point. If the earlier decision has no bearing on the later one, then we can treat each decision in the sequence as if it were an independent game. If the outcomes of early decisions do constrain later choices, then we must think of the entire sequence as the game, and each decision as a bout within that game. Such sequences are called dynamic games. An ESS is then an optimal choreography of successive strategy choices. This choreography will be fixed if the opponent’s choreography can be anticipated; alternatively, it may be conditional: e.g., if the opponent does A in the first bout, the best response in the second bout is then strategy 1; but if the opponent adopts B, then the focal animal should do strategy 2. In some dynamic games, one can identify an evolutionarily stable policy (Houston and McNamara, 1987; Mangel 1990), which is a guiding rule of thumb to find the optimal strategy at each point given current conditions. As examples, consider two male crickets that periodically try to take over the same mating burrow. If they have no signals that allow for individual recognition, and composition of the local population is sufficiently fluid, neither knows on successive occasions whether its opponent is a new individual or one it defeated (or was defeated by) on a previous occasion. Each contest may then be treated as an independent “single bout” game, and the optimal strategy for both crickets might be to threaten the other at the same level on each occasion. If, however, they can recognize and remember each other, then a prior winner would benefit from escalated aggression in subsequent interactions whereas a prior loser would do best to avoid confrontation. The latter series of interactions can only be analyzed as a dynamic game. An appropriate ESS policy might then adjust the aggressive levels of signals in successive bouts according to the fraction of prior wins (or losses) with a known opponent.

- **Spatial pattern**: Traditional evolutionary game theory assumed large populations and perfect mixing of animals adopting alternative strategies. As a result, the relative frequency of encountering an opponent playing a given strategy equaled the relative abundance of that frequency in the population. However, if offspring do not
disperse far from parents and especially if there is an overlap of generations, or if animals adopting a given strategy preferentially settle near others with the same or complimentary strategies, the distributions of strategies can become quite heterogeneous spatially. The degree of spatial heterogeneity in the game in thus a sixth criterion. In spatially heterogeneous populations, the average payoff of adopting a given strategy can depend significantly on where the animal adopting it is located. We discuss spatial games in more detail in Chapter 15 of the text. A good introduction to this type of game model can be found in Nowak (2006).

**Combinations of criteria and finding ESSs:** Games can thus be classified according to their values for each of these six criteria. All combinations of criterion values are possible: one game might be modeled as a discrete symmetric contest in which successive pairwise interactions are independent; another might be formatted as a dynamic and continuous asymmetric scramble involving 4 players per interaction. For each combination of criterion values, there are specific types of ESSs possible and specific methods for finding these ESSs. In addition, the recent advent of adaptive dynamics has made it clear that the number of possible trajectories and outcomes (above and beyond whether an ESS is present), also depend in complicated ways upon which combination of game properties are present. Many of these complications are well known from general treatments of dynamic analysis (Strogatz 1994). A few general guidelines may prove useful (Vincent and Brown 1988; Hofbauer and Sigmund 1998; Weibull 1998; Nowak 2006):

- Outcomes are more constrained and idiosyncratic for games with no overlap of generations (e.g., if difference equations are used for modeling) than if there is full and temporally random overlap (differential equations).
  - The number of possible outcomes and trajectories in discrete games generally increases with the number of alternative strategies allowed. This is corresponds to a change in dimensionality in dynamic analysis.
  - The number of possible outcomes and trajectories can increase if the number of players/interaction is increased (e.g., multiplayer games).
  - The number of possible outcomes and trajectories are greater when the relevant fitness terms in modeled equations are nonlinear than when linear. For example, both linear and nonlinear systems can produce oscillating cycles in which the population mix of strategies keeps changing overtime but along a repeating trajectory. Stable oscillations in a linear system are neutral cycles around an equilibrium point (see the rock–paper–scissors game below). These are neutral in the sense that a nearby trajectory is not attracted to a given cycle, and the distance of the orbit around the equilibrium point is determined solely by initial conditions. Nonlinear systems can generate stable limit cycles in which sufficiently nearby trajectories are attracted and pulled into the limit cycle loop. The location of a limit cycle is set by the game dynamics and not necessarily by initial conditions. Most contests are linear games and most scrambles are nonlinear.

### Finding ESSs in discrete single-bout symmetric contests

The games studied in this category are usually assumed to occur in very large (infinite) populations with complete mixing. We first outline how one might identify ESSs in these games generally, and then examine several important examples.

- **General method for discrete symmetric contests:** Discrete symmetric contests with no temporal or spatial patterns are the most easily analyzed games. This is because there is only one role, and the question is which strategy should be invoked when members of this role play the game against each other. The simplest case involves two player interactions where each player has access to the same two strategies. Because the game is symmetric, the two available strategies, S1 and S2, are the same for the two players. In such a 2 x 2 discrete symmetric game, there are four possible combinations of plays and thus four possible payoffs to each player. ESSs are most easily found using a normal form matrix as shown in Figure 1. One player is arbitrarily assigned the focal status, and the matrix then lists the payoffs it would receive given each strategy it can play and each strategy its opponent can play. For a symmetric game, it does not matter which player is selected as the focal individual.
Figure 1: Payoff matrix for 2 x 2 discrete symmetric contest. Because the game is symmetric, the two players are interchangeable and we only need to list one payoff for each pair of strategies played. The convention is to list the payoffs to the player on the left. Payoffs are given in the same currency for all cells in the matrix and denoted here by $PO_{ij}$. The first subscript ($i$) is the strategy played by the focal player and the second ($j$) is that of its opponent.

To find the ESSs in this game, we can use the “dot method.” First, we look at the left column in this matrix. In this column, the opponent plays S1. What is the best response of the focal player to this move? Assuming that the payoffs (denoted $PO_{ij}$ where $i$ is the focal player’s strategy and $j$ is the opponent’s) are in units of fitness and that natural selection will tend to maximize fitness, the best response by the focal role is S1 if $PO_{11} > PO_{21}$, and S2 if $PO_{11} < PO_{21}$. We put a dot in the larger cell in this column. If the two payoffs are equal (e.g., there is a tie), then it does not matter which strategy is used as a response and we put dots in neither cell. We then look at the second column in which the opponent will play S2. In this case, we compare $PO_{12}$ to $PO_{22}$ and place a dot in that cell with the larger value. If there are no ties in either column, the outcome has to be one of the four cases shown in Figure 2:
Figure 2: Possible ESSs in 2 x 2 discrete symmetric contest. For each column in the payoff matrix, a dot is placed in that cell conferring the best payoff to the focal player. There are four possible outcomes. **Case I**: S1 is the best response regardless of what the opponent plays. S1 is thus a pure ESS. **Case II**: S2 is the best response regardless of the opponent’s play. It is then a pure ESS. **Case III**: The best response is the strategy opposite to that played by the opponent. If mixtures are allowed, the ESS is a fixed mixture of S1 and S2. **Case IV**: The best response is to match the play by the opponent. Which strategy is the ESS depends upon the initial frequencies of the two strategies in the population.

In **Case I**, the best response is always S1, regardless of which strategy is employed by the opponent. S1 is then the ESS and because it consists of a single strategy (as opposed to a mixture of strategies), it is a pure ESS. In **Case II**, the best response is always S2. We can then call S2 the pure ESS. If there is a tie in one of the two columns, then that column exerts no selective force on the choice of strategy by the focal player. The row with a dot in the untied column is then the ESS. If both columns are tied, then there are no selective forces on strategies and hence no ESSs.

**Case III** is a bit more interesting. In this case, the best response is the strategy not used by the opponent. Imagine a population in which all players only performed S1. A mutant animal which hit on S2 would have higher fitness than the majority playing S1 and S2 would increase in frequency with successive generations. Similarly, if we started with a population in which all played S2, a mutant performing S1 would be able to invade and this strategy would become increasingly common over time. If S2 is better when S1 is common, but S1 is better when S2 is common, it seems obvious that there ought to be some intermediate mixture of S1 and S2 at which neither has an advantage over the other. Any time the population drifts away from the optimal mixture, the rarer strategy will have an advantage and will increase to bring the population back to the ESS. The system is thus convergence stable even in the face of some drift in frequencies. Note however that a mixed ESS cannot be a fitness maximum since playing either pure strategy or the mixed ESS, once the population has reached the ESS,
will yield the same payoff.

What would the ESS mixture be? Suppose we let $f$ be the fraction of the population exhibiting S1 at any time. A focal player adopting S1 will encounter opponents playing S1 a fraction $f$ of the time and on each occasion will receive a payoff of $PO_{11}$. It will also encounter opponents playing S2 a fraction $(1-f)$ of the time and on each occasion will receive a payoff equal to $PO_{12}$. Recalling the methods for calculating average payoffs detailed in Web Topic 8.1, we can compute the average payoff to a focal player always playing strategy 1 as

$$PO_{S1} = f \cdot PO_{11} + (1-f) \cdot PO_{12}$$

Similarly, if the focal player were always to play S2, its average payoff would be

$$PO_{S2} = f \cdot PO_{21} + (1-f) \cdot PO_{22}$$

We have argued that there must be some value of $f$, denoted by $f_e$, at which equilibrium occurs and thus it does not matter which strategy you play. This means that

$$PO_{S1} = PO_{S2}$$

when $f = f_e$. Setting the above equations equal to each other and solving for $f_e$, we get that

$$f_e = \frac{PO_{12} - PO_{22}}{(PO_{12} - PO_{22}) + (PO_{21} - PO_{11})}$$

In words, the fraction of the time we should see S1 at equilibrium is equal to the absolute difference between the values in the right hand column of the matrix (the one occurring with frequency $[1-f]$), divided by the sum of the absolute differences between the values in each column. Clearly, this must be some fraction between 0 and 1.

As we noted earlier, a mixed ESS might be achieved by a genetic polymorphism, in which $f_e$ of the players always play S1 and $(1-f_e)$ of the players always play S2, or by a behavioral polymorphism in which each player performs S1 a fraction $f_e$ of the time and performs S2 the remaining $(1-f_e)$ of the time. As far as the analysis is concerned, the two solutions are equivalent. In real life, not all genetic systems will allow the establishment of a genetic polymorphism at any given frequency $f_e$. Genetic impediments are less likely for behavioral polymorphisms.

Case IV is the reverse of Case III: the optimal choice is to match your opponent’s choice of strategy. If we imagine an initial population all performing S1, we can see that S2 is always at a disadvantage and thus cannot invade. If we begin with a population all playing S2, S1 is now the less favored option. Again, it seems likely that there ought to be some intermediate equilibrium at which payoffs of playing either are equal. In fact, the calculation of $f_e$ for Case IV by setting

$$PO_{S1} = PO_{S2}$$

gives exactly the same result as for Case III. However, in this case the equilibrium point is not stable. Suppose one begins with a population at this equilibrium mixture. Barring drift, it would pay no player to shift strategies and thus it is a Nash equilibrium. However, if the frequency of players adopting S1 drifts upwards so that $f > f_e$, S1 is now the better strategy and subsequent evolution will drive the system towards 100% of the animals using
If \( f \) had drifted lower so that \( f < f_e \), S2 would be favored and S1 would disappear (e.g., \( f = 0 \)).

**Case IV** thus has two ESSs: either pure S1 or pure S2. Which is observed in the short term depends upon whether the initial frequency of S1 is greater or less than \( f_e \). Over the long term, most finite populations will experience enough drift for moderate numbers of the initially rarer strategy to accumulate. Although one strategy may be most common and thus the ESS at the start, drift may eventually push \( f \) past \( f_e \) causing the system to flip to the other ESS. The ESS most likely to be found over the long term is the strategy which is least common at the equilibrium point. The logic is as follows. Suppose \( f_e \) is less than 0.5: that is, the frequency of S1 at the equilibrium is much less than that of S2 (e.g., \( f_e << [1- f_e] \)). If we begin with a population at the S2 ESS (\( f \approx 0 \)), S1 players will appear only as rare mutants (genetic heritability) or innovators (cultural heritability). If \( f_e \) is a small enough value, only a few such mutants are needed before \( f \) exceeds \( f_e \). Once this occurs, S1 becomes the favored strategy, and the system will move to an S1 ESS. By the same token, if we begin with a S1 ESS (\( f \approx 1 \)), it will take a very large number of S2 mutants, since \( f_e \) is small, before \( f \) can drift below \( f_e \) and cause a switch to S2. The strategy which is least common at the equilibrium is thus the one which is most likely to invade by chance if the other is the ESS, and the one least likely to suffer invasion when it is the ESS. As we saw above, the value of \( f_e \) is determined entirely by the relative payoffs in the matrix. A simpler shortcut to find the more likely ESS is to compare the absolute values of the differences between payoffs in each column of the payoff matrix. The column with the greatest absolute difference (called the **risk dominant strategy**), is the one that is most likely to arise over evolutionary time.

Note that a 2 x 2 discrete single-bout symmetric contest with no ties **always** has at least one ESS (as long as mixed ESSs are allowed). As discussed later, this may not be true for discrete symmetric contests with more than two strategies.

**Well-known examples of 2 x 2 discrete single-bout symmetric contests:** Just as there are three general outcomes in the 2 x 2 games reviewed above (a single pure ESS, a mixed stable ESS, or two alternative pure ESSs), there are three corresponding games that have had significant impact in evolutionary biology and behavioral ecology. Each of these has been given various names and incarnations, but the predictions from each have proved remarkably robust, even when scaled up to more complicated versions.

- **Take games**: Take games have two basic strategies: passive minds its own business, and cheat exploits others. As an example, consider foraging terns. Passive individuals ignore others and have an average catch of \( P \) fish per time interval. Cheats also fish on their own but when they spot another tern with a captured fish, they dive at it and steal the fish. When a passive and a cheat hunt in the same area, the passive tern loses \( B \) fish to the nearby cheat. Its average payoff is then \( P-B \). The cheat gains \( B \) fish from the passive, but at a loss of \(-C\) fish that it might have caught had it kept fishing on its own. Its payoff is then \( P+B-C \). When two cheats meet, half the time one gets to steal \( B \) fish from the other at a cost \(-C\), and the other half of the time, it has \( B \) fish stolen from it. The average payoff is then \( 0.5(P+B-C) + 0.5(P-B) = P-0.5C \). The payoff matrix for this evolutionary game is then:
Figure 3: Payoff matrix for a take game involving foraging terns. See text for justification of payoff values.

If \( B > C \), then \( P + B - C > P \), and cheat is the better strategy in a population of passives (left hand column). We can thus put a dot in the lower left cell of the matrix. Cheat is also the best response in a population of cheats (right hand column) if \( P - 0.5C > P - B \). This will be the case if \( B > 0.5C \). If \( B > C \) (left hand column condition), then it is also true that \( B > 0.5C \) (right hand condition). In short, if \( B > C \), then cheat is a pure ESS (Case II).

Note that at the ESS, everyone does worse (e.g., \( P - 0.5C \)) than they would have done if they had all remained passives (e.g., \( P \)). A payoff which is higher if everyone avoids cheating is called a Pareto optimum. However, there is always a temptation to cheat, and once some cheating starts, eventually all players have to become cheaters just to keep up. Some factor outside of the game would be required to keep the players at the pareto optimum.

Take games appear in a wide variety of incarnations, but they invariably give the same result: passive independence is vulnerable to invasion by selfish individuals with the result that everyone ends up worse in the end. Arms races are a classic example of a take game. A second famous example is the prisoner's dilemma in which two prisoners would do best if neither provides evidence on the other, but the temptation to get more lenient treatment causes one to defect, and the ESS is then for both to defect with much worse consequences than if they had both kept silent. We discuss this game further in Web Topic 13.1. Herd animals that try to put themselves closer to the herd center to reduce their own predator vulnerability (the selfish herd) are playing a take game (Hamilton 1971). In evolutionary ecology, the red queen hypothesis (Hamilton 1980; Bell 1982; Salathe et al. 2008) can also be viewed as a take game. Theft and unregulated capitalism can both lead to take games in human societies: one purpose of laws and regulations is to try to keep the population at a pareto optimum. The contexts vary: the outcome is routinely the same.

Give games: The opposite case is called a give game (also called a coordination game in the economic literature). Instead of food acquired, we here use a general currency, fitness units, for payoffs. There are two strategies: passive players, as before, mind their own business, whereas donors can give up to \( B \) fitness units to half of the other players that they encounter at a cost of \(-C\) fitness units to themselves. When donors encounter other donors, half the time they gain \( B \) fitness units from the other player’s donation, and half the time they lose \(-C\) fitness units by donating to the other player. As long as \( B > C \), exchanges between donors can be favored by selection. However, there is no way donor mutants could invade a population of passives as their fitness would be \( P - 0.5C \), whereas the average fitness of passives...
when the mutants were rare would be $P$. Even if donors were common, being a passive would have a higher payoff ($P+0.5B$) than being a donor ($P+0.5B-0.5C$). Passive would remain a pure ESS. One way out of this dilemma is for donors to be selective to whom they donate. This might involve individual recognition and memory of prior interactions, kin recognition, or some other factor that allowed them to identify who was and was not a donor. If they then reduced their rate or amount of donation to passives, so that the average donation to a passive was $b < B$ and the cost to the donor was $k < C$, could being a donor then be an ESS? And if so, how selective would they have to be? The relevant normal form matrix is shown in Figure 4:

![Figure 4: Payoff matrix for a give game involving selective donation.](image)

It is clear that as long as there is some residual cost to donating fitness units to passives, the passive strategy is the best response in a population of passives ($P > P - k$). We thus put a dot in the top cell in the left hand column. If $(B-C) > 2b$, then donor is the best response in a population of donors. Thus if donors can be sufficiently selective, either passive or donor can be a pure ESS depending upon initial conditions. Donor can be the more likely ESS if $(B-C) > 2(b+k)$.

This points up a general outcome for give games: charitable cooperation is at best a Case IV situation; there are invariably two possible ESSs, one of which is not to cooperate, and no guarantees that the charitable cooperation outcome will be the one seen in nature. Put another way, it really does not pay to be a donor unless there are already enough other donors around. This raises the question of whether charitable cooperation could ever invade a population of passives. A variety of mechanisms have been suggested including initial cooperation due to kin selection or heterogeneous settlement patterns to get past this hurdle. The bottom line is that selfish arms races are the common outcome of take games, but only under certain circumstances can give games end up at a cooperative ESS. We discuss give games in more detail in [Web Topic 13.1](#).

- **Stable mixture games:** Stable mixtures are the expected outcome whenever the best response is the opposite of what one’s opponent plays. Perhaps the best known example is the *hawk and dove game* (Maynard Smith 1982). The biological context is a conflict between two individuals over some indivisible commodity such as a single food item, a mate, or a roost. Winning the commodity increases the fitness of the winner. The naive expectation might be that the two animals should always fight over the commodity. The observation in nature is that animals often use relatively low risk display signals to decide who should get the commodity; one animal then leaves without a fight. It is not immediately clear why the two animals do not always fight over the commodity. Hawk versus dove is a simplified game invoked to explain these observations. It assumes that all contestants are equal (making the game symmetric), and allows for only two discrete strategies: fight (hawk) and display peaceably (dove). Only two players are allowed to encounter the same commodity at a time and payoffs of a single interaction are independent of how many
players are hawks or doves. The game is thus a contest. Players have no memory and may never meet again; thus this is not a dynamic game.

We again use fitness units as the relevant payoff currency. The commodity is worth $V > 0$ fitness units and the fitness cost of losing an escalated fight is $-D$. Two hawks always engage in an escalated fight during which one wins and the other loses. Since they have equal chances of winning or losing, the average payoff of such a fight is $0.5(V-D)$. If a hawk and a dove jointly encounter the commodity, the dove flees with payoff 0 and the hawk gets it with payoff $V$. If two doves meet, they coo at each other with negligible costs until one (randomly determined) gives up. The average payoff for two doves is thus $0.5V + 0.5(0) = 0.5V$. The payoff matrix is shown in Figure 5.

![Figure 5: Payoff matrix for the simplest version of the Hawk–Dove game. See text for payoff explanations.](image)

The dot in the right hand column must go in the upper right cell since $V > 0.5V$ for any positive $V$. The dot in the left hand column would go in the upper left cell if $V > D$. Hawk would then be a pure ESS (not shown in figure) and the animals would always fight over the commodity. But if the costs of injurious fighting exceed the value of the commodity, then $V < D$ and the dot would go in the lower left cell. This is then a Case III game and one would expect to find either a genetic or behavioral polymorphism in which hawks represented a fraction:

$$f_e = \frac{V - 0.5V}{(V - 0.5V) + (0 - 0.5(V - D))} = \frac{V}{D}$$

One can make this game much more complicated (see several chapters in the text), but the basic message is that if the costs of aggressive actions are high enough, it may pay both parties to resolve the dispute by some form of display at least some of the time.

**Discrete single-bout symmetric contests with three strategies:** While all of the two-strategy games discussed in the prior section have at least one ESS, this is not necessarily the case for otherwise similar games with three or more strategies. One very illustrative example is the rock–paper–scissors game (Maynard Smith 1982; Weissing 1991). This game has three strategies with the properties that rock beats scissors, scissors beats paper, and paper beats rock. The classical game has the payoff matrix shown in Figure 6A:
Figure 6: Payoff matrices for rock–paper–scissors games. (A) Classical payoffs normalized so that each combination of plays results in a gain or loss of the same amount. (B) Payoffs for the generalized version of game in which different payoffs can occur for different combinations of plays. Note that games with rewards (payoffs >0) or penalties (payoffs < 0) for ties (cells along the main diagonal) can be renormalized in either version without changing the dynamics or equilibria of the game by adjusting column values so that the main diagonal only has zeros.

Visual inspection of the matrix in Figure 6A suggests that no pure strategy can be an ESS in this game: each strategy can be invaded by one of the others. However, one might imagine that there could be some sort of equilibrium mixture. In fact there is, but the dynamics of this game may or may not lead a population to that equilibrium. The equilibrium for the Figure 6A version occurs when each strategy constitutes 1/3 of the population; the equilibrium might occur at some other mixture for the generalized game of Figure 6B. It can be shown that the dynamics of these games can be predicted by calculating the sign of the determinant of their payoff matrices (Weissing 1991; Hofbauer and Sigmund 1998; Nowak 2006). There are three possible cases that are best visualized using the simplex of the game (Figure 7). Since there are three discrete strategies, this will be an equilateral triangle, and since no pure strategy can be an ESS, any interesting trajectories will in the triangle’s interior.
Figure 7: Evolutionary dynamics of rock–paper–scissors game plotted on the relevant strategy simplex. Red dot shows equilibrium point. (A) If the determinant of the payoff matrix is zero, the population will follow a neutral (fixed) limit cycle around the single equilibrium point. The distance of the cycle from the equilibrium will depend on the starting point of the population. Two possible trajectories are shown. (B) If the determinant of the payoff matrix is positive, the population will follow a decreasing spiral that asymptotically approaches the equilibrium point. (C) If the determinant is negative, the dynamics will follow an increasing spiral that asymptotically approaches the edges of the simplex.

As shown in Figure 7, there are three possible outcomes. The classical rock–paper–scissors game of Figure 6A has a determinant equal to zero: the population will cycle around the equilibrium point without getting either closer or further away on successive loops (Figure 7A). The starting mixture of strategies in the population determines the location of the loop within the simplex. If the determinant is positive, the population mixture also cycles around the equilibrium, but on each pass it gets closer to the equilibrium. It thus approaches the equilibrium asymptotically (Figure 7B). If the determinant is negative, the population is also a spiral but in this case, each loop takes the population further from the equilibrium. What happens once it is asymptotically close to the simplex boundaries depends on the values of the matrix and the degree to which evolution in this population is subject to drift. It is instructive to note that replicator and standard ESS analyses will often give similar predictions for this game as long as generations are overlapping (e.g., allowing differential equations of change) and payoffs are not too heterogeneous. However, once these conditions are not met, the replicator dynamics and ESS solutions can make rather different predictions: what may be stable for one generational schedules may not be in the other (Weissing 1991).

The point of this example is that once there are three or more strategies in an evolutionary game, evolutionary trajectories can become quite complicated. There may be NO ESS and any equilibria may only function as reference points for cyclic changes in the current mixture of strategies. Such rock–paper–scissors-like cycles have actually been found in microbes, sessile animals, and the mating and communication strategies of some lizards (Alonzo and Sinervo 2001; Frean and Abraham 2001; Sinervo et al. 2006; Sinervo et al. 2007).

- **Discrete symmetric contests with more than three strategies**: As noted earlier, discrete games with more than three strategies can exhibit additional trajectories and outcomes not seen in lower dimensioned games. It is impossible in most cases to identify ESSs in this type of game with more than three strategies, and even in some with three strategies, using dot methods or simple inspection. Occasionally, one can obtain insights by examining sub-games within the larger matrix. However, usually, one must resort to finding equilibrium points that meet Lyapunov conditions, or attempt to linearize the curvature of the adaptive surfaces near equilibrium points.
and use higher order derivatives (e.g., a Jacobian matrix) to assess equilibrium stability (both at the point and locally) (Hofbauer and Sigmund 1998; Cressssman 2003; Leimar 2009).

Finding ESSs in discrete single-bout asymmetric contests

- **General methods for asymmetric discrete contests**: We now consider discrete contests in which there are two separate roles. The appropriate payoff matrix will need to show two payoffs for each combination of strategies (one for each player). The convention (in behavioral ecology) is to divide the matrix cell for each pair of strategies diagonally in half and to enter the payoff to the player on the left in the lower left corner of the cell, and that to the player at the top in the upper right hand corner (Figure 8).

![Figure 8: Payoff matrix for discrete 2x2 asymmetric contest.](image)

In this example, there are two players who represent different roles (A and B) and each role has access to two alternative strategies. Because the strategies available to the roles in an asymmetric game are usually different, we assign strategies one and two to role A and strategies three and four to role B. Any combination of strategies by the two players may result in different payoffs for the two players. As before, we use subscripts to indicate the strategy used by player one followed by the strategy used by player two. In parentheses, we indicate to whom that particular payoff is given.

A 2 x 2 asymmetric discrete contest may have no ESS, one ESS, or several ESSs. It will not have mixed strategies similar to those seen for symmetric contests if players are fixed in a given role and cannot, by definition, perform one of their own strategies part of the time and one of the strategies of another role the rest of the time. If animals are not fixed in a given role, as might occur if they make mistakes about their relative roles, then mixed asymmetric ESSs are possible.

To identify the pure ESSs (if any) in a 2 x 2 discrete asymmetric contest, we use the “arrow” method. This indicates the best responses by placing arrows parallel to each of the four sides of the payoff matrix. For example, we might first consider what the player in role A of Figure 8 should do if its role B opponent plays strategy three. If $PO_{13}(A) \succ PO_{23}(A)$, then the role A player should adopt strategy one over strategy two. We would then place an arrow parallel to the left side of the payoff matrix pointing up. If $PO_{13}(A) \prec PO_{23}(A)$, then we make that arrow point downwards. We then consider what the role A player should do when the role B player opts for strategy 4. If $PO_{14}(A) \succ PO_{24}(A)$, then we place an arrow parallel to the right side of the matrix pointing up; if the converse is true, then we point the arrow down. We next consider the optimal strategies for the role B player. If player A elects to use strategy one, what should the role B player do? If $PO_{13}(B) \succ PO_{14}(B)$, then the
role B player should adopt strategy three. We then place an arrow parallel to the top of the matrix pointing to the left. If the converse were true, we would make that arrow point to the right. Finally, we identify the best response for the role B player when its opponent adopts strategy two. This results in an arrow parallel to the bottom of the matrix pointing either to the right or to the left.

Any point (marked with an asterisk) at which two arrowheads meet is an ESS for that game. For a simple 2 x 2 discrete asymmetric contest, there are 16 possible arrangements of the arrows and 3 classes of outcomes (Figure 9):

- **Asymmetric Hawk and Dove**: An example will help demonstrate this approach. Consider a version of Hawk–Dove game in which the two interactants have different probabilities of winning an escalated fight over a contested commodity (Hammerstein 1981). As in the symmetric example above, let the value of the commodity be $V$ fitness units, and the injury cost of losing an escalated fight be $-D$. The only difference from the earlier version is that the probabilities of winning an escalated fight are no longer 0.5 for each party: instead, the **dominant** animal has a probability of $P_D > 0.5$ of winning such a fight, whereas the **subordinate** has a chance of $P_S < 0.5$ of winning. We shall assume that if the two parties simply display at each other, they have an equal chance of giving up first. The resulting payoff matrix is shown in Figure 10:
As long as $V > 0$, we can draw two arrows immediately: the arrow on the right hand side should point upwards (since $V > 0.5V$) and that below the matrix should point to the left (for the same reason). The questions are then which directions the top and left hand arrows will point. If $P_{SV} - P_{DD} > 0$, the top arrow will point to the left; if $P_{SV} - P_{DD} < 0$, it will point to the right. If $P_{DV} - P_{PS} > 0$, the left hand arrow will point upwards; if $P_{DV} - P_{PS} < 0$ it will point downwards. There are three possible cases which are shown in Figure 11:

![Figure 10: Payoff matrix for asymmetric Hawk–Dove game.](image1)

**Figure 10: Payoff matrix for asymmetric Hawk–Dove game.** The only difference between this and prior symmetric version is an unequal probability of the two contestants winning an escalated fight.

![Figure 11: Three possible outcomes of asymmetric Hawk–Dove game.](image2)

**Figure 11: Three possible outcomes of asymmetric Hawk–Dove game.** ESSs are shown as asterisks and relevant arrows are drawn. Conditions guaranteeing this distribution of arrows are provided in similar format below each example.

Since the three conditions shown in Figure 11 all scale according to the same parameters, another way to present the outcomes is shown in Figure 12:
Figure 12: Relationships of outcomes in asymmetric Hawk–Dove game to asymmetry in winning escalated fights \((P_D)\) and relative cost of losing an escalated fight \(\frac{D}{V+D}\).

Figure 12 presents quite intuitive results. At low relative costs of losing an escalated fight \(\frac{D}{V+D}\) and small asymmetries in the chance of winning \(P_D\), the ESS is for both the dominant and the subordinate to play Hawk and escalate (Case I). The subordinate will lose more often than the dominant, but the chances of winning are still good and the cost of losing negligible. However, as the probability \(P_D\) that the dominant wins and/or the costs of losing a fight \(\frac{D}{V+D}\) increase, the ESS shifts to one in which the dominant routinely plays Hawk and the subordinate plays Dove (Case II). If the cost of losing a fight is large and/or the value of the commodity small, either of two ESSs can evolve: one in which the dominant plays Hawk and the subordinate retreats, or one in which the subordinate plays Hawk and the dominant retreats. In this context, it is not worth fighting over the commodity and even a somewhat non-intuitive convention, such as the subordinate challenging the dominant and the latter retreating, resolves the issue quickly and simply.

Finding ESSs in continuous contests and scrambles

Any of the examples in the previous section can be made into continuous games by allowing players to vary their strategies in graded ways. Thus Hawk might grade into Dove by varying the amount of energy put into an attack. The strategy set is then the amount of energy invested in escalations. Similarly, a give or take game can be made continuous by varying the amount of resource given or taken per interaction. The question is then what is the ESS amount to give or take. Even the behavioral polymorphism ESSs of discrete games are continuous strategies: what is the optimal fraction of time that a player should spend on each strategy?

**General method:** Games with continuously variable strategy sets invariably have continuous functions relating fitness to an actor’s strategy and the current population mix. The obvious (and traditional) way to find ESSs in continuous games is to compute the first derivative of this function and set it equal to zero (to identify any equilibria), and then ensure that the second derivatives of the function at each equilibrium are negative (implying a fitness maximum) (Maynard Smith 1982; Parker 1984). These will then be ESS values of the strategies. As mentioned earlier, more recent approaches also require examination of the behavior of the function at points in the vicinity of the equilibrium to determine whether the latter is convergence stable or not. The use of Lyapunov functions or linearization and subsequent eigen-analysis are now widely used for this task (Hofbauer and Sigmund 1998; Weibull 1998; Cressman 2003). Note that just because a strategy set and the corresponding fitness functions are continuous does not mean that the ESS cannot be discrete: as we shall see in the example below and in certain other contexts, continuous games can have separating, pooling, or semi-pooling equilibria depending on the game dynamics. In the latter two cases, the equilibrium calls for a segmentation of the continuous strategy sets, and even players with different attributes (roles) can be assigned to the same segment (Parker 1982; Bergstrom and Lachmann 1998; Lachmann and Bergstrom 1998; Bergstrom et al. 2002).

Scramble games almost always have continuous fitness functions even if the alternative strategies are discrete (e.g., settling in habitat A versus habitat B). Where these continuous functions for a scramble game can be defined, the same utilization of derivatives of those functions can be undertaken to find ESSs and assess convergence stability. It should be noted however that while fitness is usually a linear function of strategy frequencies for contests, it will be nonlinear for scrambles. This means that the dynamics and possible outcomes for scrambles can be much more complicated than those of a similarly dimensioned continuous contest.

**Two classic continuous contests:** In the analysis of the Hawk and Dove game, we argued that two Doves resolved their conflict over a commodity by displaying at each other until one gave up. We did not discuss why or when either should give up the contest. One of the classical solutions to this problem is the war of attrition. It comes in two forms: symmetric and asymmetric. Since these games are cited in Chapter 11 of the text, we briefly review them here:

- **Symmetric war of attrition** (Maynard Smith 1974; Parker 1974; Bishop and Cannings 1978; Bishop et al.
1978): Consider two opponents each of which competes for some commodity $V$ by selecting some amount of display time $t$ from a continuous distribution. Neither knows before the confrontation what display time has been chose by their opponent, and no party is allowed to gain additional information while interacting. During the confrontation, the opponent that selected the longer display time wins the commodity (hence the name “war of attrition”). If both pick the same $t$, then they settle the interaction by chance. Once the confrontation is over, each party will have paid some cost of display: this will be the same for both parties since this is a symmetric game and they each displayed for the time selected by the loser. Let this cost be $-kt$ where $k$ is the rate of loss per unit time displaying in the same currency as the commodity $V$. In a population in which all players use the same $t$, the average payoff would be $0.5V - kt$ (since all confrontations are ties).

It should be obvious that there is no pure ESS in this game: a population in which all animals used a given display time could be invaded by a mutant picking a slightly longer time. Eventually, once all animals expended $kt > 0.5V$, a mutant that selected $t = 0$ could invade. No pure strategy is uninvadable. On the other hand, there is a mixed ESS in this system in which each player selects a display time $t$ with probability $p(t)$:

$$p(t) = \frac{2}{k} e^{-\frac{kt}{V}}$$

This is a negative exponential distribution: most individuals will select $t = 0$, but other values are possible with increasingly longer values of $t$ being increasingly rare (Figure 13A).

![Figure 13: Mixed ESSs for wars of attrition games.](image)

- Asymmetric war of attrition (Parker and Rubenstein 1981; Hammerstein and Parker 1982): Now consider the same game as above except that either the value of the commodity, $V$, or the costs of display, $k$, differ between the two parties. As long as one or both parties can at least partially assess their own and their opponent’s values of the ratio, $V/k$ (e.g., based on body size, age, prior experience, etc.), the ESS is to draw a random display time from one of two distributions (Figure 13B). If a contestant thinks it has a lower $V/k$, it selects a time from the distribution of $t$ values $< S$; if it thinks it has a relatively high $V/k$ ratio, it should draw a random time from the distribution of $t$ values $> S$. Note that this ESS only exists if players
have imperfect information: were a player to know for sure who would win a protracted display contest, there is no reason to display at all. Note also that it is possible that both players might draw display times from the same distribution. The ESS value of $S$ will vary with the relative values of $V$ and $k$ for the two parties, and with the probability, $Q$, that each animal accurately assesses it and its opponent’s relative $V/k$ ratios: specifically, $S$ increases with an animal’s own $V$, decreases with its $k$, and decreases with its ability to accurately assess the situation $Q$. Put simply, if animals make few mistakes in judging their $V/k$ relative to that of the opponent, $S$ will be a small number, and individuals will only draw from the longer $t$ distribution if they are fairly sure they will win. If mistakes are common, $S$ will be a larger number and individuals are more likely to draw from the lower $t$ distribution. As assessment errors and $S$ increase, the likely variation in contest times is also predicted to increase. This latter prediction is often met in real situations: where displaying contestants are similar in condition and endurance, one often sees much more variability in contest durations than if they are highly asymmetrical.

**Some additional considerations**

Most of the evolutionary games that arise in animal communication contexts can be analyzed with the methods outlined above, although implementation can become algebraically complex. However, there are several communication contexts in which the classical expectations are constrained or additional factors relevant and this can change the analysis. We outline several of these contexts below:

- **Sequential games**: A number of the evolutionary games that arise in animal communication are sequential games: instead of each party deciding ahead of time which strategy to play and then playing them simultaneously, the parties make their plays successively. Optimal strategy selection for all but the first play is thus conditional on what the other player just did. The sequential assessment game for conflicts discussed in Chapter 11 is a good example. If there are only two stages or bouts, ESSs may be identifiable by examining normal form matrices. However, even in this case, extensive form analyses may better clarify assumptions and game logic. For example, Hurd (1995) used extensive form analysis of sender/receiver interactions to show how sender costs might stabilize communication when the two parties had disparate interests. Hurd and Enquist (2005) then extended this approach to classify communication games. For complex sequences, some form of dynamic programming (also called backward induction) is commonly employed to identify ESSs (Mangel and Clark 1988; Houston and McNamara 1999). Even more complicated sequences may require simulation methods (Hamblin and Hurd 2007). Note that where the sequence is stretched out over time, for example, in a life history game, fitness changes at each bout may need to be combined multiplicatively instead of additively (Seger and Brockmann 1987; Getty 2006). Cressman (2003) provides a general review of extensive form game analysis and shows how evolutionary dynamics can be combined with this approach. Gintis (2000) examines a wide variety of evolutionary and economic games in extensive form. One construct that has proved very useful in analyzing sequential games is the notion of state: this might be a running measure of energy supplies, body integrity, information (e.g., the sequential assessment game), or some other variable that changes cumulatively during successive bouts. For a detailed explanation of how state can be incorporated into sequential evolutionary games, see Houston and McNamara (1999).

- **Finite populations**: As discussed earlier, traditional analyses of evolutionary games assume infinitely sized populations. This minimizes the stochastic effects of drift. Once focal populations are finite, drift cannot be ignored and becomes increasingly important as the population size gets smaller. A strategy that would always be a pure ESS in an infinite population can in fact be the one that goes extinct first in a small finite one. In addition, a small number of mutants in a small population can significantly change strategy frequencies and thus change the adaptive landscape for evolutionary games in that population. The assumptions used to define an ESS for infinite populations may thus no longer hold and require different definitions for stability (Fogel et al. 1998; Traulsen et al. 2005; Fudenberg et al. 2006; Traulsen et al. 2006a; Traulsen et al. 2006b; Traulsen et al. 2007; Zhou et al. 2010).

An excellent introduction to evolutionary games in finite populations is provided by Nowak (2006). Whatever
population size is being examined, he assumes that it is stable so that each birth results in a corresponding
death and vice versa (called a Moran process). Game payoffs that have only a slight effect on a player’s fitness
exert only weak selection, whereas those with a major effect exert strong selection. The possible outcomes of
evolutionary games in a finite population, even for a simple 2 x 2 discrete symmetric contest, can be quite
different from the classical case depending on both population size and the level of selection. For example,
Nowak cites a 2 x 2 matrix in which strategy A is clearly the pure ESS for infinite populations. However, B can be
an ESS in very small populations, both A and B can be ESSs in intermediately sized populations, and only for
large populations does the game converge on the infinite case with A as the only ESS. Another example involves
any Case IV game in which there are two ESSs: for infinite populations, the risk dominant strategy is the one
expected to be seen most commonly. However, in large but finite populations under weak selection, a strategy
can only take over the population if its frequency at $f_e < 1/3$ (Nowak et al. 2004; Ohtsuki et al. 2007a). This is a
more stringent condition than being risk dominant (where its strategy at $f_e$ only needs to be $< 1/2$). Altrock and
Traulsen (2009) find similar differences in the dynamics of the rock–paper–scissors game when played in finite
instead of infinite populations. Not surprisingly given these patterns, both the arms races of take games and the
potential cooperation of give games can exhibit quite different outcomes when played in finite populations
(Nowak et al. 2004).

- **Heterogeneous populations**: As with finite populations, the predicted ESSs of evolutionary games in
  heterogeneous populations may differ significantly from those of homogeneous (well-mixed) populations. The
  primary reason is that the average payoff for a given player is not based on the global frequencies of the
  alternative strategies, but instead on some biased subset of these. There are two (at least) ways that this might
  happen. First, animals might only have interactions with spatial neighbors. If a population is viscous, such that
  offspring do not settle far from parents, and all interactions are local, a focal player is unlikely to encounter
  strategies at rates equal to at their global frequencies. Alternatively, individuals may have the mobility or linkages
  needed for exposure to global strategy frequencies, but they actively limit their interactions to other players who
  adopt particular strategies (Pacheco et al. 2006).

There has been considerable research done on spatially heterogeneous evolutionary games (Nowak and May
1992; Taylor et al. 2004; Ohtsuki and Nowak 2006; Santos et al. 2006; Taylor and Nowak 2006; Grafen 2007;
Lehmann et al. 2007; Ohtsuki and Nowak 2007; Ohtsuki et al. 2007b; Grafen and Archetti 2008; Tarnita et al.
2009a; Tarnita et al. 2009b; Nowak et al. 2010). Heterogeneous populations are best modeled as networks, a
topic we discuss in some detail in Chapter 15. Note that most spatial game analyses also assume that the
population is finite. Nowak (2006), again, provides a good introduction to spatial evolutionary games in finite
populations. One interesting outcome of recent work is that a parameter that depends only on the network
structure can be used to predict the outcome of any 2 x 2 discrete symmetric contest even if the population is
heterogeneous and finite (Tarnita et al. 2009b). This means that the same payoff matrix might result in quite
different ESSs depending upon the population’s network structure.

- **Genetic constraints**: Traditional evolutionary game models assume haploid genetics and only rare mutations.
What happens to evolutionary game predictions in populations of typical diploids with realistic rates of
recombination and mutation? For pure ESSs, the outcomes are usually unchanged for diploid polygenic systems,
although the dynamics may be more complicated. Mixed ESSs, however, pose a more complicated problem. For
example, the symmetric war of attrition mixed ESS is not attainable as a genetic polymorphism; it can only be
achieved by a behavioral polymorphism (Maynard Smith 1982). Whereas the mixed ESS predicted for Hawk and
Dove games in infinite populations might be realizable as either a genetic or behavioral polymorphism, the
outcome in a finite population can be much more variable. Not surprisingly, given our earlier discussion of games
in finite populations, a genetic polymorphism, a behavioral polymorphism, or a pure strategy can be the most
likely outcome depending upon the population size, levels of selection, and levels of stochasticity (Vickery 1987;
in finite populations can vary with the population size (Schaffer 1988), and the most likely mix in a behavioral
polymorphism may be skewed away from $f_e$ if the payoff of one pure strategy against itself is much further from
the equilibrium payoff than that of the other (Ficici and Pollack 2007).

**Useful links**

Below, we have listed a few online sites that provide software for simulations, analysis of simple game models, or other tools for studying evolutionary games:

- **Dynamo** ([http://www.ssc.wisc.edu/~whs/dynamo/](http://www.ssc.wisc.edu/~whs/dynamo/)): A set of evolutionary game theory routines that can be run in the Mathematica environment by W. Sandholm and associates. Visualization tools for simplex analysis are available. Routines are limited to certain combinations of numbers of roles and strategies.

- **Gambit** ([http://www.gambit-project.org/doc/index.html](http://www.gambit-project.org/doc/index.html)): This is a free program developed by researchers at California Institute of Technology in collaboration with colleagues in the UK and New Zealand. Most versions run on all platforms. The focus is on Nash equilibria and games can be constructed in either normal or extensive form. Strategies must be discrete.

- **Social Dynamics Modeling** ([http://www.socdynamics.org/id4.html](http://www.socdynamics.org/id4.html)): This site provides a long list of social dynamic, including game, models and links to software to study those models. Different software runs in different environments. Popular environments include Matlab, Mathematica, R, Stella, C++, Java, and Excel.

- **VirtualLabs Packages** ([http://www.univie.ac.at/virtuallabs/](http://www.univie.ac.at/virtuallabs/)): This site posted by Christopher Hauert provides no downloadable software, but it does provide a wide variety of simulations and tutorials on different kinds of problems in evolutionary game theory.

- **GameTheory** ([http://home1.gte.net/vze3hvm3/java.htm](http://home1.gte.net/vze3hvm3/java.htm)): This package provides introductory examples and tools for general game theory. It is only available for Windows platforms. Non-commercial users can get full copies for free.

- **Bimatrix Game Solutions** ([http://banach.lse.ac.uk/form.html](http://banach.lse.ac.uk/form.html)): This site posted by Rahul Savani provides a GUI in which one can specify the dimensions of a discrete game and enter the relevant payoff matrices. It then finds any equilibria, average payoffs, etc.

**Recommended further reading**

Some of the advanced books on evolutionary game theory, following the tradition in economics, intersperse text with exercises whose answers may or may not be provided at the end of the book. Since these exercises are often crucial steps in the general treatment, unanswered exercises can be frustrating to readers who do not have the time or skill to solve them. Also following traditional economics, and perhaps spurred by the initially informal way evolutionary games were defined, many books are organized into series of definitions, theorems, and lemmas, each followed by detailed specification of terms and proofs. Trying to extract the general point given all the proofs can also be challenging, especially when no chapter summaries are provided. Below, we list some of the sources and add our annotations on the “type” of source in order of increasing difficulty:

- **Maynard Smith** (1982): This is the classic introduction to evolutionary game theory by one of the key players who invented the approach. It is very readable, and although dated in some ways, still constitutes one of the best starting points on the topic. Some algebraic skill is required. See also a follow-up summary by Houston and McNamara (2005).

- **Parker** (1984): Although cited much less often than Maynard Smith, G. A. Parker played an extremely important role in the early promotion and application of evolutionary game theory approaches. Again, though dated, his chapter cited here is one of the most succinct introductions available. He has also written a fascinating personal account of the history of evolutionary game theory (Parker 2010).

- **Nowak** (2006): This book is a very accessible introduction to general evolutionary processes in finite populations, the replicator equation, and the application of these ideas to evolutionary games. It also provides an introduction to spatial games. Unlike the prior two sources, this book is the first in this list to focus on evolutionary dynamics
instead of end-points. Reading the book first will help readers digest a more recent update on the same topics in Nowak et al. (2010).

- **Vincent and Brown** (2005): This broad-scope book also invokes some definitions and proofs, but all of the examples are worked out for the reader. The approach focuses on G-functions which define how a player’s fitness depends on its strategy, the current population mix of strategies, the abundances of each strategy, the population size, and amounts of available resources. Derivatives of G-functions are then used like the replicator equation to study evolutionary dynamics and plot trajectories. There is a strong emphasis on interspecific games and community ecology, but also some serious discussion of intraspecific game models. A more recent update on some of the issues raised in the book can be found in McGill and Brown (2007).

- **Hofbauer and Sigmund** (1998): This book is one of the foundations of evolutionary games based on replicator dynamics. It is heavy on theorems and lemmas, and many critical points are only mentioned as exercises (with no answers provided). Despite this, the book is a much more readable source than their earlier volume on the same topic (Hofbauer and Sigmund 1988).

- **Weibull** (1998): This was published in the same year as Hofbauer and Sigmund’s second volume. It covers much of the same ground, and is also full of theorems and lemmas, but all the examples are solved. This book is somewhat terse at points, and requires some reasonable algebraic skill, but it can provide a useful counterpoint to or additional perspective on topics in Hofbauer and Sigmund.

- **Gintis** (2000): In contrast to most of the prior texts, Gintis devotes most of this book to helping the reader learn how to use the tools of game theory, particularly extensive form games. The book is full of exercises, only some of which are solved (in the back). He has designed the text to help readers think like a game theorist and get them involved in the necessary protocols and logic. It helps if readers already know some of the relevant theory. Note that the book is not primarily aimed at evolutionary games although it includes some very useful examples.

- **Cressman** (2003): This book focuses on the benefits of using extensive form to analyze evolutionary games. A major claim is that the complicated dynamics of higher dimensional games are much more transparent and easily understood using extensive form analysis. The book includes some complicated definitions and theorem proofs, but most examples are solved in the book, not left to the reader. This explicit linking of evolutionary dynamics and extensive form analysis is a very useful perspective.

- **Other sources in brief**:
  - **Thomas** (1986): This now-dated book was aimed more at economics than biology but still has some useful insights.
  - **Vega-Redondo** (1997): This volume is also somewhat dated, but has some very good presentations on basic ESS concepts, replicator dynamics, and stochastic games.
  - **Sandholm** (2011): This recent text contains a very advanced, and mathematically intense, integration of evolutionary games in large populations with evolutionary dynamics. Convergence stability is of major interest. An innovative classification of different types of population games puts a new spin on the many different game examples that have accumulated over the last decades. Sandholm is also the supervising author for the Dynamo game software mentioned above.
  - **Dercole and Rinaldi** (2008): These authors review the logic and applications of the canonical equation for adaptive dynamics. Although the focus is more on ecological than behavioral contexts, several sections discuss evolutionary games.
  - **Dugatkin and Reeve** (1998): This multi-authored book provides an interesting, albeit somewhat dated, set of examples of applications of evolutionary game theory to animal behavior contexts. The chapter by Hammerstein is a succinct summary of the theory at that time, and the chapter by Johnstone provides a useful introduction to evolutionary game models of communication.

**Literature cited**


The Handicap Principle

Amotz Zahavi first developed the handicap model of signal evolution in 1975 as an alternative to R. A. Fisher’s (1958) runaway model of mate choice (Zahavi 1975, 2003). The question was, how can extremely extravagant sexually dimorphic traits sometimes evolve in males, when those traits clearly impose a significant cost to the bearer (i.e., the long tail of the male peacock)? Fisher’s argument was that initially, females would evolve a preference for a male trait that was associated with male fitness or quality. For example, larger males need longer tails for flying and steering, so a female preference for longer tails would result in them obtaining larger, better quality mates. Once many females acquired this preference, males could benefit from increasing their tail length beyond what was required for flying. The inefficiency cost of bearing a longer tail was compensated by the attraction of more mates. The more females preferred males with longer tails, the more males benefited by growing even longer tails, initiating a runaway process. Long tails in males and female preference for long tails would become genetically linked. The runaway process would come to a halt when the cost of further tail elongation became too great for males to bear. Fisher assumed that once all males invested in growing their tails, tail length was no longer correlated to their quality. Females maintained their preference for the tails in order to produce sons that would be successful and preferred by the next generation of females.

Zahavi countered that sexual selection via female choice should be viewed as a sender–receiver signaling interaction, where males strive to advertise their quality and females are selected to pay attention only to those signal traits that honestly reveal male quality. Rather than a runaway process that ends with no correlation between tail length and male quality, he argued that costly tails could help to reveal differences in male quality because better quality males can carry a heavier burden than lower quality males. Under sexual selection, males would thus evolve costly signaling traits as a test of their quality to discriminating females. Zahavi called this idea the handicap principle, deriving the term from the analogy of handicapping faster racehorses with heavier weights. Female receivers benefit by selecting males with handicaps because they can be assured of the quality of their mates, while male senders benefit because they can better advertise their quality and thus acquire more or better female mates. But both sexes also lose: males lose by investing time and energy in advertising, and females may receive less help from their mates and bear sons which are less fit to withstand the pressures of natural selection. But an individual without the handicapping marker does not advertise its quality, so a potential mate cannot detect it. Moreover, the particular marker must handicap the specific aspect of quality that is important to the receiver. Thus selection for the handicap creates a biological link between the information encoded in the signal and the form of the signal trait. In Zahavi’s words, “a rich person can signal the degree of his wealth by wasting money. His signal is reliable since a poorer man cannot waste as much money. A courageous man can display the degree of his courage by taking a risk, which a less courageous individual would not dare to take. However, taking a risk of bodily harm does not display wealth, and spending money does not display how brave a person is.” Figures 1 and 2 show two animal examples of possible handicap signals.
Figure 1: A foraging handicap? Both the male and female American white pelican (*Pelecanus erythrorhynchos*) develop a fibrous, dorso-laterally flattened structure on the bill (called the “maxillary horn”) during the annual courtship period. Zahavi speculates that the horn interferes with the ability of the bird to see around the tip of the bill, requiring a forager to remember where it last saw a prey fish and project its likely movements. An inexperienced bird would be less able to do this, so horn size reflects the foraging ability of a bird (Zahavi & Zahavi 1997). We don’t know whether the horn impairs vision and foraging. Like many other pelican species, the white pelican forages by swimming on the surface and jabbing its head under the water to scoop up small fish in the flexible pouch on the underside of the bill. The horn is absent in young birds and larger in paired breeding birds, presumably because they are older, so its size is correlated with age. The horn must impose some kind of cost because it is shed as soon as mating is finished. Aggressive bill jabbing and holding is often directed at this part of the opponent’s bill, so another possibility is that the condition of the horn reflects fighting ability (Knopf & Evans 2004). Although it looks like a handicap, there is no proof whatsoever that it really is a handicap signal. (Photo courtesy of Tripp Davenport.)

Figure 2: Lateral compression threat displays in lizards. As discussed in the main text, while
performing these broadside lateral displays, the male compresses his ribs, which impairs respiration. The male is literally holding his breath! The reduction in aerobic metabolism causes a compensatory increase in anaerobic metabolism, which generates lactate and diminishes locomotory capacity. Only males in good condition can afford this temporary reduction in aerobic capacity and still mobilize the energy to fight. The natural duration of time the posture is held is correlated with the male’s intrinsic endurance on a treadmill, and forced exercise reduces subsequent display duration. The display thus expends the quality attribute being advertised (Brandt 2003). This is a much more convincing example of a handicap than the pelican in Figure 1. (Photo of Sceloporus occidentalis courtesy of Yoni Brandt.)

Controversy over the handicap term

The handicap term itself is deemed to be unfortunate, even by Zahavi himself: “…the word ‘handicap’ is misleading because it has the connotation of a loss. Senders are not losing – they invest in order to gain: an individual that takes on a reasonable handicap in order to signal is like a businessman investing in advertisement” (Zahavi 2007b). Getty is moreadamant, pointing out that in the sports analogy, the function of weight handicaps is to even out the performance of different quality competitors (Getty 2006). The handicap principle, on the other hand, argues that the selective pressure exerted by receivers for increasingly extravagant display traits enhances the variation in signal trait expression among senders and makes it easier for receivers to discern true differences in sender quality.

Another controversial expression used by Zahavi to describe signal evolution is selection for “waste” and “inefficiency” (Zahavi 1991). According to Zahavi, non-signal characters are selected for efficiency, and the smaller the investment required to achieve a particular result the better. By contrast, signals must have a cost to be reliable indicators of quality, and the greater the investment, the greater the reliability. Strong selection by receivers for increasingly costly and reliable signals can lead to extravagant signal traits that seem wasteful and inefficient. Furthermore, the nature of the “wasted” commodity is a key to the information encoded in the signal. While the need for costliness may be unique to signal traits, all traits can be expected to have some cost and to have tradeoffs with other survival and reproductive fitness components. John (1997) further points out that “costliness is not synonymous with inefficiency (which alludes to net costliness instead). From the gene’s point of view, costly signals are neither wasteful nor inefficient if they maximize fitness.” Thus terms like handicap and inefficiency, while colorful descriptors of some aspects of signal evolution, have tended to exaggerate and confuse the underlying principle of the need for some form of cost trade-off that makes cheating a suboptimal strategy compared to honesty.

Controversy over verbal versus math models

Zahavi’s verbal handicap model, along with the controversial terms sometimes used to describe it, immediately stimulated several evolutionary theoreticians to attempt to model it quantitatively. They believed that mathematical models made the assumptions of a model clearer than verbal arguments and could show more rigorously whether or not the suggested conclusions were correct. These early models concluded that handicap signals could not evolve as proposed (Davis & O’Donald 1976; Maynard Smith 1976; Bell 1978; Kirkpatrick 1986). But the theoreticians misunderstood the gist of the argument, and Zahavi’s admittedly confusing graphical representation of it. These early models incorrectly assumed the costs of the signal lowered the fitness of all males in the same proportion. Parker (1979) subsequently showed that either the costs of signaling, or the potential benefits obtained when receivers respond favorably, had to be differentially scaled for different senders. These alternatives became known as the “differential costs,” or quality handicap model, and the “differential benefits,” or signal of need versions of the handicap model. When signaling costs differentially penalize either lower quality or less needy senders, reliable signals can be maintained (Pomiankowski 1987; Grafen 1990a, 1990b; Iwasa & Pomiankowski 1991; Johnstone 1995, 1997). Maynard Smith quickly accepted the basic tenet of the handicap principle (1991) and publicly apologized to Zahavi at a meeting of the Royal Society in 1992, admitting that he had been wrong.

Additive versus multiplicative models

Grafen’s differential cost model requires that higher-quality senders pay lower marginal costs for advertising in order to maintain stable honest signaling. But this requirement is correct only if the fitness benefits and costs are strictly additive, i.e., measured in the same currency (Figure 3).
Figure 3: Additive model. The green line indicates benefit to the sender for expressing a signal size of $a$, and the three blue lines show the cost of the signal for low, medium, and high quality senders. The marginal cost of a signal at any given signal size is the slope of the cost line. The optimal signal size for each quality class occurs where the marginal cost equals the marginal benefit (slopes are equal) and the net benefit ($B-C$) is maximal, as shown by the orange vertical lines.

However, additivity is not expected from most life history models, where benefit is measured in terms of mating success and cost is measured in terms of viability (or mortality rate). These components must be multiplied to obtain net fitness. A more general criterion for honest handicap signaling is that higher-quality senders must have higher marginal fitness effects of advertising. This effect could arise from higher fitness benefits as well as lower costs. The general criterion implies the existence of a ridge on the fitness surface for two correlated characters, quality (or viability) and signal intensity (Figure 4). In a multiplicative model, higher quality senders can have higher marginal viability and fitness costs and still be more efficient because of their higher marginal fitness benefits (Getty 1998, 2006). Thus Getty further argued that the additive concept of the handicap term was technically incorrect, metaphorically misleading and a poor guide for empirical research on signal evolution. He also argued that signal evolution should not be viewed as a “missing piece of Darwin’s puzzle”, but rather as an integral piece of the process of evolution by natural selection.

Figure 4: Multiplicative model. The green line is fecundity or mating advantage $F$ for a signal size of $a$. The three blue lines show the residual viability $V$ as a function of $a$ for low, medium, and high quality senders. The cost of the signal is the drop in residual viability, and the marginal viability cost is the steepness of the negative slope. Fitness is the product of $F$ and $V$ represented by the orange lines, with the optimal value indicated by orange circles. In (A), the marginal viability cost is higher for higher quality senders, which is inconsistent with the Grafen model, but could happen if there were a power-efficiency (i.e., intercept-slope) tradeoff as shown in the graph. The optimal signal size increases with sender
quality despite the violation of the Grafen requirement. In (B), the marginal viability costs are lower for higher quality senders, consistent with the Grafen requirement. The optimal signal size increases with sender quality. But at equilibrium, the high-quality senders are not “wasting” more viability, they are actually more efficient than low-quality senders. (After Nur & Hasson 1984; Getty 2006.)

Are all signals handicaps?

Zahavi views all signals, and some characters and behaviors that we would at first not even think of as signals, through the lens of the handicap principle (Zahavi & Zahavi 1997; Zahavi 2007a, b, c). If the principle is defined extremely broadly, such that any type of cost is a “handicap,” then he would often be right. But a very broad view that places an umbrella over everything prevents us from exploring some interesting variation in the process of evolving signals that encode different kinds of information in different ecological and social contexts. For example, Zahavi removes the physiological constraint distinctiveness of an unfakable index signal, compared to the cost–benefit tradeoff a quality handicap signal, by stating that all handicaps are indices with an analog relationship between trait value and sender quality (Zahavi 2007; unpublished manuscript). Zahavi dismisses the idea of arbitrary conventional signals in conflict of interest contexts that are honest as a result of a receiver retaliation rule. Status badges, for example, are designed to give the illusion of excellence. Thus the vertical breast strip in great tit makes the breast look narrower than it is, and only a high-quality individual can afford the narrowing effect of a wider band. Zahavi proposes that a reliable threat must leave the sender open to attack. Displays that transmit information about the sender’s confidence, motivation, or willingness to take risks should thus expose vulnerable body parts or assume postures that make it difficult to launch an attack. He suggests that the broadside threat display common in fish, antelope, and lizards exposes the flank of the sender and leaves any weapons on the front of the animal in an inconvenient position to attack. Threats that involve stretching some part of the body, with back arching, fins, or fur, are not likely to deceive the opponent about the sender’s body size, but place the sender in postures from which flight or attack are difficult, and provide a structure that the opponent can grab (Zahavi 1981; Zahavi & Zahavi 1997).

Zahavi further asserts that amplifiers, such as lines on the body that make assessment of body size easier, cannot be separated from body size and are handicaps that selectively penalize smaller individuals. Finally, Zahavi claims that there is essentially always a conflict of interest between any two individuals, even closely related ones, and to be on the safe side, signals should always have an honesty-guaranteeing cost. Low-cost or cost-free signals should rarely or never occur (Zahavi 1993, 2003).

Számadó takes this umbrella approach to task in his review of handicap signals entitled “The cost of honesty and the fallacy of the handicap principle” (2011). In agreement with our approach in the main text, handicap signals are restricted to signals for which a realized strategic cost can be demonstrated, and where there is a biological constraint that links the size, quality, or intensity of the display to the attribute of interest to the receiver. Proulx (2001) explicitly modeled the relationship between sender’s display costs and utility (fitness advantages) of the advertised quality to the receiver; when utility and quality are decoupled, no signaling equilibrium can be attained. Other types of signals that do not entail this realized cost should not be called handicaps. For instance, signals for which the honest sender does not pay the extra strategic cost at equilibrium, only paying if it cheats, involves a potential cost. Conventional threat signals with a receiver retaliation rule, and proximity threats involving close approach to the rival, are examples of potential cost that are only paid if a weak sender sends a stronger signal than it can support. Similarly, index signals may entail an efficacy cost, but no added strategic cost that guarantees honesty. Signals with pooling equilibria, where senders within broad ranges of qualities give signals of the same intensity, can in theory lead to partially informative low-cost signals (Bergstrom & Lachmann 1998). Finally, signals for which sender and receiver have shared or overlapping interests, such as alarm calls and identity signals, are often cost-free.

Altruism as a handicap signal

Acts such as feeding offspring that are not one’s own, perching as a sentinel while others forage, or mobbing predators, have been traditionally viewed as examples of altruism. By definition, an altruistic act benefits other individuals at some cost to the donor. Traditional explanations for these behaviors range from kin selection to various forms of reciprocity (see textbook Figure 13.1 and Web Topic 13.1). Zahavi has proposed that such behaviors are instead costly handicap signals indicating quality and propensity to cooperate (Zahavi 1977, 1990, 1995). Highly altruistic individuals gain social prestige in the opinion of others, which Zahavi argues benefits the altruist by increasing its attractiveness as a mate or collaborator. Roberts (1998) expanded on this idea that altruism could serve as a competitive display of quality. Lotem et al. (1999) and Putland (2001) emphasized the role of intersexual selection in biparental species with altruistic helping behavior, whereby females prefer as mates former helper males
that demonstrate or “show off” their parental ability. These various versions of the social prestige hypothesis fall into the class of cooperative mechanisms called “indirect positive pseudo-reciprocity” (see Bergmüller et al. 2007; Connor 2007; Bshary & Bergmüller 2007; and Web Topic 13.1) because the altruist receives an eventual benefit via its effect on third party receivers. Note that social prestige is not the same thing as dominance status. Dominance involves obtaining benefits via aggression, e.g., gaining priority of access to food and mates by intimidating rivals. Prestige and dominance could generate different hierarchies within groups, but it is likely that the two hierarchies are somewhat correlated since dominants are often in better condition and better able to afford the cost of bestowing altruism (Wright 1999, 2007; Barclay & Reeve 2012).

As with other explanations for cooperation, the evolutionary stability of social prestige mechanisms has been examined theoretically. Quantitative models demonstrate that, in principle, costly altruistic acts can become stable signals of quality. Earlier models of indirect reciprocity had demonstrated that stable cooperation could evolve in social groups if most individuals are known to each other. This condition is also necessary for the building of a reputation as a cooperator. Cooperators might then benefit over the long term because others are more likely to trust and collaborate with them (Boyd & Richerson 1989; Milinski et al. 2002; Nowak & Sigmund 2005). Such models differ from direct reciprocity and conditional altruism (tit-for-tat) models, which require the initial altruist to be repaid later by the same individual it had aided. Indirect pseudo-reciprocity allows reputation (also called “image score”) to be built by “word of mouth” or eavesdropping, and the resulting altruism is unconditional (not dependent on what the beneficiary did at a prior meeting). Lotem et al. (2003) expanded the existing indirect reciprocity model into a signaling model by first introducing individual variation in quality. When the probability of repeated requests for aid is sufficiently common, and high-quality individuals can bear the cost of an altruistic act better than low-quality individuals, then an ESS can evolve in which only high-quality individuals engage in tit-for-tat reciprocal exchanges and low-quality individuals defect. Altruistic behavior thus signals high quality. Adding a signaling benefit component to this game means that altruists gain benefits in contexts other than those in which they performed an altruistic act. When this is true, reputation or prestige causes unconditional altruism to be stable at least over a given set of parameter values. Specifically, when the cost of altruistic acts is sufficiently high, so that low quality individuals cannot obtain a net benefit from performing them, then altruism can serve as a stable handicap signal of quality. Similar quantitative models lend further support to these conclusions (Gintis et al. 2001; Pilot 2005). Thus costly signaling of quality via altruistic acts and the build-up of prestige within groups could be one of several valid explanations for the evolution of altruism within cooperating groups. This explanation avoids the restrictive requirements of kin selection (relatedness), group selection (highly structured populations and low dispersal rates), and direct reciprocity (delayed return of a similar altruistic act by the recipient), but does require individual recognition (or memory of rendezvous sites), a social network, and memory of past acts by others (Nowak 2006).

What types of evidence would be needed to demonstrate that altruism serves as a signal of social prestige, and to distinguish this social prestige mechanism for the evolution of helping behavior from alternative mechanisms? First, one needs to provide evidence of signaling to third party receivers. For instance, the helping behavior could be accompanied by conspicuous signals, a helper's behavior could differ depending on the presence or absence of an audience, or competitive helpers could be shown to vie for the attention of bystanders. If these effects cannot be shown, and instead, helper behavior functions only to promote the condition or survival of the recipient offspring, then alternative hypotheses for the helping behavior are more likely (Wright 1999, 2007). For example, if improved survival of the offspring primarily increases only the helper altruist’s inclusive fitness component, then the kin selection hypothesis (Hamilton 1964) would be the most likely explanation; if improved offspring survival leads to an increase in group size and enhanced ability to defend the territory, gain more group resources, or disperse more successfully, then the group augmentation hypothesis (Kokko et al. 2001) might be a sufficient explanation for the behavior. Second, given that some evidence of signaling has been found, additional evidence concerning the type of benefit obtained by the altruist is required, because there are two alternative signaling hypotheses—the social prestige hypothesis and the pay-to-stay hypothesis. The pay-to-stay hypothesis proposes that subordinate group members “pay rent” in order to be allowed to remain in the group and potentially gain breeding positions later in life (Gaston 1978; Kokko et al. 2002). Evidence that dominant territory owners punish or evict helpers that slack off in their helping effort, and that harder-working helpers obtain survival and/or breeding status benefits, would support this alternative pay-to-stay hypothesis. Suppose that a careful study does not support a pay-to-stay, enhanced group size, or kin assistance explanation for altruistic acts and that conspicuous signaling when audiences are present is noted. To demonstrate support for the social prestige hypothesis, one would also need to show that: (1) altruists eventually obtain a benefit by being selected by a mate or collaborator; (2) third party receivers do evaluate potential mates/collaborators on the basis of their helping effort; and (3) these third party receivers prefer the better-performing altruists. Real systems are complicated by the fact that the hypotheses for helping listed above (kin selection, group augmentation, pay-to-stay, and social prestige) are not mutually exclusive, so analyses might require a quantitative
ranking of their relative contributions to the evolution of the helping behaviors.

What evidence has been obtained for the social prestige hypothesis? The question has been primarily investigated in cooperatively breeding species, but many of these systems involve closely related individuals where the role of kin selection cannot be excluded. Luckily, a few studies have been able to focus on cases of unrelated reproductive helpers and others on the interspecific mutualistic interactions of cleaner fish and their clients. We describe some of these studies below.

Evidence of a signaling role for helping behavior has been described in the colonial and cooperatively breeding sociable weaver (Philetarius socius) (Doutrelant & Covas 2007). Here, the levels of helping behavior are enhanced by the presence of an audience, and this increases the helper’s likelihood of being observed feeding offspring. Specifically, helpers spent a longer time than parents holding the prey at the edge of the colony before feeding, and they waited for the number of third party observers to increase before feeding. However, a benefit for these behaviors has not yet been identified in this species, so social prestige and pay-to-stay hypotheses cannot be distinguished. In another colonial cooperative breeder, the pied kingfisher (Ceryle rudis), good evidence of the pay-to-stay mechanism has been found in the form of punishment to slackers and a subsequent mating advantage to more altruistic male helpers (Reyer 1984, 1990). Similar evidence also supports the pay-to-stay hypothesis in superb fairy-wrens (Malurus cyaneus) (Mulder & Langmore 1993), and the cichlid fish Neolamprologus pulcher (Balshine-Earn et al. 1998, Bergmüller & Taborsky 2005; Bergmüller et al. 2005). However, whether an additional social prestige effect is present has not been excluded.

Zahavi and co-workers (Carlisle & Zahavi 1986; Zahavi & Zahavi 1997; Kalishov et al. 2005) describe how group-living Arabian babblers (Turdoides squamiceps) compete with one another to perform feeding and sentinel acts and sometimes interfere with the altruistic endeavors of others (although these observations were mainly seen during a year with an excessive number of one-year-old helpers). Dominant individuals generally perform more provisioning, guarding, and mobbing activities than subordinate individuals. Dominants usually refuse to be fed by subordinates by walking away from them, or interfere with aid-giving gestures directed toward them by subordinates. Willing recipients of aid, on the other hand, adopt a specific crouched, begging-like posture. These observations are suggestive of a social prestige signaling component for provisioning. Wright and colleagues attempted to test the hypothesis more rigorously by looking for evidence of audience effects, as well as evidence that helpers used different rules for feeding nestlings and serving as sentinels than breeders. There was no evidence of differential competitive or conspicuousness-enhancing behaviors by helpers of different ages, sexes, and statuses, and they also used the same provisioning rules as breeders, which depended on age of nestlings and size of the group (Wright 1997, 1998a, b). Sentinel duty was most often performed by the breeding male and conformed to state-dependent models of individually selfish anti-predator strategies (Wright et al. 2001a, b, c). Thus evidence for prestige signaling in this species remains circumstantial. Dominant breeders who benefit the most from brood survival expend the most aid-giving effort (i.e., straightforward parental investment). The more likely hypotheses for helping in this species are kin selection and perhaps group augmentation. If there is some role for social prestige signaling, then the prestige and dominance hierarchies are concordant and their relative roles difficult to untangle (Wright 1999, 2007).

Another well-studied avian helper-at-the-nest species is the Australian bell miner (Manorina melanophris), which breeds in dense neighborhoods where non-reproductive helpers (primarily males) and male breeders regularly feed related and unrelated nestlings in a number of nests. Helpers and males also frequently join together to mob predators in the area. If a female loses her mate, she selects the male helper that performed the most provisioning at her recent nests as her new mate, strong evidence of a benefit to vigorously feeding helpers (Clarke 1989). Helpers also give individually distinctive calls when visiting nests, thereby advertising their aid-giving acts. However, careful observational and experimental studies to look for an audience effect on helper behavior found no evidence for this important prediction of the social prestige signaling hypothesis (McDonald et al. 2001a, b). This lack of evidence does not necessarily exclude the operation of a prestige-signaling system in this species, since the other key pre-requisites are present; helpers advertise vocally and so may not require that a female breeder audience is in view. Cooperation in this species may thus be driven by a combination of kin selection, group augmentation, and indirect pseudo-reciprocity.

The cleaner wrasse–client fish mutualistic interaction provides the strongest evidence for an altruist signaling system (Bshary & Schaeffer 2002; Grutter & Bshary 2003; Bshary & Grutter 2006; Pinto et al. 2011). The cleaner fish (Labroides dimidiatus) may cooperate and remove ectoparasites from client fish or they may cheat by feeding on client mucus. The cleaner fishes prefer mucus, so refraining from this behavior constitutes an altruistic act. Eavesdropping clients spend more time next to known cooperative cleaners than unknown cleaners, and they respond to any occurrence of cleaner cheating by avoiding them. These observations demonstrate that client fish
score and choose individual cleaners on the basis of the quality of their service. Trained cleaners also learn to feed more cooperatively (contrary to their preference) when in a “client-feedback” than in a “non-client-feedback” situation. Finally, cleaners immediately increase their level of cooperative behavior in the presence of a bystander client fish. Cleaners thus benefit in terms of many future clients if they display altruistic behavior in the presence of a bystander client.

**Chemical signals as handicaps**

Zahavi points out that chemical signals between individuals also require investment in reliability, just like other signal modalities. The cost or handicap may be the ability of the sender to bear the damage caused by the signaling chemical, or it may be the difficulty of producing a particular chemical. Carotenoids used in the production of color signals are proposed as an example of ability to bear the toxic cost of a chemical signal. Circulating carotenoids are beneficial in small amounts because they deactivate harmful oxygenated radicals (anti-oxidant). However, in larger amounts carotenoids may increase the lifetime of radicals (pro-oxidant), such that only high quality individuals may be able to bear this cost (Haila et al. 1996; McGraw et al. 2005). An example of signals that are hard to produce may be the mating pheromones of yeast cells, complex molecules such as glycoproteins that require special investment for their synthesis (Nahon et al. 1995). The alpha mating peptide of yeasts is produced from a complex glycoprotein pro-peptide. Short peptides may not display the phenotypic quality of the secreting sender because they are easy to produce, whereas large proteins with post-translational modifications may vary in relation to the phenotypic quality of the sender. Several other examples of chemical blends and pheromone concentrations that are correlated with mating success and sender status have been described (Iyengar & Eisner 1999; Endler et al. 2006), but more work needs to be done to determine the mechanism by which chemical signals provide honest information. Whether these kinds of chemical signals involve a costly handicap is not at all clear.

Zahavi goes even further to suggest that chemical signaling within multicellular bodies also may require investment in reliability, even though all cells within an organism share the same genome and have no conflict of interest (Zahavi 2008; Zahavi & Perel 2011). The reason why reliability may be needed is to avoid signaling by cell phenotypes that should not signal, or to inhibit the signaling cells from producing too much of the signal. Costly chemical handicaps could guarantee these requirements. Zahavi suggests that within-organism chemical signaling does not function to instruct the receiving cell to take certain actions, but to provide information about the quality or state of the signaling cell. The response of the receiver cell may depend on the phenotypic quality of the sender cell, and receiver cells of different qualities may respond differentially. Examples of costly or toxic chemicals used in between-cell signaling are NO, CO, and steroids. Zahavi claims that cell biologists and endocrinologists are not asking the right questions about cell–cell interactions, and offers this creative alternative view from a between-animal communication perspective.

**Literature cited**


Pinto, A., J. Oates, A. Grutter & R. Bshary. 2011. Cleaner wrasses *Labroides dimidiatus* are more cooperative in the


Web Topic 11.1
A Detailed Description of Three Conflict Resolution Models

Introduction

War of attrition games provided the earliest models for conflict resolution that involved strings of successive displays or other actions per contest (see Web Topic 10.5). Although these contests involve repeated bouts, and are thus technically sequential games, the war of attrition models got around the sequential aspects by having each player select a persistence time at the outset, and then letting them display at each other until the shorter time bid had run out. Whichever individual picked the shorter time was then the loser. To prevent cheating that would undermine any ESS, players were not allowed to vary either the intensity or the rate of repetition of their display during the contest. The symmetric war of attrition assumed all players suffered the same rate of cost accrual during the contest and valued the contested commodity equally. Under these restrictive conditions, a mixed ESS was possible favoring settling of conflicts by repeated displays. Since player equality is unlikely to be the case in reality, the asymmetric war of attrition was proposed. Here, players were allowed to differ in cost accrual rates and/or contested commodity valuation. To achieve an ESS in this situation, the asymmetric war of attrition assumed that players could glean sufficient information about their own and their opponent’s fighting abilities before the contest to assign themselves to either “loser” or “winner” roles. Each player then drew a persistence time from a random distribution appropriate to that role. As with the symmetric war of attrition, players were not allowed to vary display intensity or repetition rates. This combination of assumptions and strategies did lead to a stable ESS.

One problem with the asymmetric war of attrition is that it assumes that players can assess cues in each other before the contest that are correlated with likely fighting abilities. While body size is an obvious factor that might both affect fight outcomes and be assessable by players before a contest, there are many other likely determinants of fight outcome that are not so assessable. Examples include physical condition, energy reserves, prior experience, and motivation. In addition, fight outcomes might also depend on relative stress or damage accumulated by players during an escalation; this might be very difficult to estimate before a contest but play a major role in outcomes. Clearly, the asymmetric war of attrition cannot handle any of these cases.

In this Web Topic, we review models in which each of three different types of initially non-assessable factors are assumed to play significant roles in the outcomes of contests with repeated and sequential displays. In the first example, the sequential assessment model, all of the same assumptions are made as for the asymmetric war of attrition except that no useful assessments are possible before the contest begins. Instead, players gradually extract information about their own and their opponent’s relative fighting abilities as the contest proceeds. At some point, these estimates become sufficiently accurate that the players can confidently identify who would be a “winner” and who a “loser” if they escalated; the contest then ends with the presumed loser quitting. In a second example, the critical factor that affects contests is the amount of stored energy or some equivalent resource prior to the contest. This resource gets “used up” as the contest proceeds, and eventually, one player reaches some ceiling or threshold beyond which it cannot afford to use up the remaining resource. It quits and the contest ends with it being the loser. The third type of model focuses on the role of damage accrual and/or other externally inflicted costs that might accumulate as a contest progresses. Again, each player will have its own threshold level of such costs that it can afford to suffer before it quits. The three models are similar in that each focuses on some state variable (information, energy reserves, or damage, respectively) that changes progressively as the contest proceeds. This raises a second issue that is normally associated with sequential games: what is the optimal policy for scheduling the rate of change in that state variable at different points in the contest? We take up this question for each of the three types of models (see Web Topic 10.5 for a review of terms and logic of game theory).

The sequential assessment model (SAM)

The sequential assessment model, like the asymmetric war of attrition, assumes that each player uses the same display
repeatedly and does not vary either display intensity or instantaneous repetition rates (Enquist and Leimar 1983). Also like the asymmetric war of attrition, the goal of the interaction is to obtain an accurate estimate of the relative fighting abilities of the two parties. The state variable that changes during the course of the contest is the amount of error associated with those estimates. The sought ESS is a policy that identifies how accurate the estimates need be before one party assigns itself a “loser” role and quits.

The model assumes that the true fighting abilities of two contesting players (A and B) depend on the rate at which each accrues costs during an escalated contest, both from their own actions and from those of their opponent. If $c_A$ is the true rate at which A accrues costs when contesting with B, and $c_B$ is the equivalent rate for B, relative fighting ability from A’s point of view can be measured as

$$
\theta_A = \ln\left(\frac{c_B}{c_A}\right)
$$

When $\theta_A > 0$, B suffers higher cost accrual than A, and thus A can be considered to have higher relative fighting ability; when $\theta_A < 0$, then A suffers more than B and has the lesser relative fighting ability. It should be obvious that $\theta_A = -\theta_B$.

When the contest begins, neither party has a good estimate of the true $\theta_i$ values. They then try to get more accurate estimates by observing each other’s displays during the contest. During each sample, an estimated value of $\theta_i$ will depend both on the true value and on some random error. Suppose the random error in any one sample $i$ is $z_i^A$ and that by $B$ is $z_i^B$. The model assumes that these random errors are drawn from a normal distribution with mean zero and standard deviation $\sigma$. The best estimates at any step $n$ in the contest would then be the cumulative average of the current and all prior samples. If at contest step $n$, A’s current estimate of $\theta_A$ is $x_n^A$, and B’s estimate of $\theta_B$ is $x_n^B$, then

$$
\begin{align*}
    x_n^A &= \frac{1}{n} \sum_{i=1}^{n} \theta_A + z_i^A \\
    x_n^B &= \frac{1}{n} \sum_{i=1}^{n} \theta_B + z_i^B
\end{align*}
$$

At the beginning of the contest, each individual has only a crude estimate of the $\theta_i$ values. As the contest progresses, the effects of the random errors begin to average out to zero (e.g., the sampling error for the cumulative estimate after $n$ steps is SE = $\sigma/\sqrt{n}$), and the current estimate approaches the true relative fighting ability. The model assumes that once one animal estimates that its $\theta_i$ is negative, it will end the contest by retreating or giving up. Since the standard error of the estimates should decrease as $n$ increases, a smaller $|\theta_i|$ is necessary to trigger quitting as $n$ increases. The evolutionarily stable policy can thus be described as a giving-up line on a graph of current estimates versus the number of steps so far, $n$. When one contestant’s current estimate crosses this line, it is sufficiently certain of its lower fighting ability that it makes the decision to quit (Figure 1).
**Figure 1:** Graphical solution for the sequential assessment game. Current estimates by each party of its relative fighting ability, $x_A$ (red line) and $x_B$ (blue line), are plotted on the vertical axis for successive numbers of steps $n$ (horizontal axis). The solid horizontal line indicates equal estimated fighting abilities (e.g., $x_i = 0$). The dashed line shows the ESS policy for giving-up. Early in the fight, a player’s assessment of itself, $x$, must be quite low for it to give up because of the high level of uncertainty (e.g., large standard error of estimate). As the contest progresses and the estimates of true relative fighting ability become more accurate with repeated sampling, the giving-up line rises towards the 0 line. When an individual’s $x$ crosses the line, it quits the fight. In this example, that was the blue player. (After Enquist and Leimar 1983.)

The smaller the difference in true fighting ability, and the higher the error of assessment, the longer the fight will be on average (**Figure 2**). Contests between near equals are also more variable in duration than those between unequal competitors and the (slightly) poorer contestant may sometimes win (**Figure 3**).
Figure 2: Average duration of a fight (in number of steps) as function of the absolute value of relative fighting ability. The three curves represent increasing (black to red to blue) standard deviations in the error levels of opponent assessment: higher levels of errors result in a slower decrease in fight durations. (After Enquist and Leimar 1983.)

Figure 3: The probability of victory (vertical axis) as a function of relative fighting ability. Losing or winning becomes more certain when contestants are more disparate in ability.

As the cost of fighting increases, the giving up line moves higher towards the x=0 line, meaning fight duration is shorter...
Figure 4: Giving-up line as a function of the cost of the interaction. The red curve is the giving-up line for a high level of interaction cost. Each line below it represents half the cost of the one above. As cost increases, the giving-up line moves up and contests are shorter. A similar series of curves is generated when the value of the resource decreases. As the value increases, the giving-up line is shifted down and contests are longer. (After Enquist and Leimar 1983.)

This game is a somewhat more realistic model than the asymmetric war of attrition, because it allows contestants to control the cost (duration) of the fight as it proceeds and to decide whether to continue or quit based on information gained during the fight. On the other hand, it yields some similar predictions to those of the asymmetric war of attrition, namely that contest length is short if the individuals are very different in size or fighting ability, and highly variable but generally longer when the contestants are similar.

In the original sequential assessment model, players were allowed to shift between multiple behaviors as long as these gave the same kind and amount of information. However, real animals often seem to have a sequence of escalating behaviors that are seen in conflicts. Enquist et al. (1990) later developed a different version of the sequential assessment game with several behavioral options that could be adopted in stages. Thus when the additional information provided by any given display stage became asymptotic, players could then switch to another display that provided more or different information. Usually, better information requires riskier or more expensive displays. In fish for example, lateral display may provide only partial information on size, tail beating leads to better but riskier size estimation, and mouth wrestling or head butting provides even better information. The model predicts that the alternative stages should be ordered so as to be maximally efficient in assessing relative fighting ability. This model thus provides an optimal policy for adjusting display type at different stages in the contest. In this extended model: (1) all contests should be organized into phases consisting of one or several behavior patterns with constant intensities and rates of repetition within a phase; (2) the contest should begin with the least costly behaviors that provide some information about fighting ability asymmetry, and after repetition with diminishing returns, switch to new, more costly but more effective behaviors in subsequent stages; (3) the division into phases should be independent of relative fighting ability; and (4) contests with great asymmetry in relative fighting ability should end in an early phase, whereas matched individuals may proceed through a series of escalations reaching a final phase of more dangerous fighting. Figure 5 shows a sample trajectory with three behaviors.
Limited energy models

The sequential assessment game assumes that each time a combatant performs the same display, it provides additional information about its ability to fight should the contest escalate. This mechanism works best for assessing instantaneous attributes such as coordination or motivation. It may not be a good way to assess an opponent’s stamina and endurance, which could also play an important role in escalated contests. Like the asymmetric war of attrition and the sequential assessment game, the two models below focus on populations in which players do not share the same cost accrual rates or disputed commodity valuation. Both assume that each player has its own reserves (e.g., energy or some other limited resource) that get used up during the contest. No initial assessment of these reserves is possible, and players just display until one hits a threshold in cumulative costs beyond which it is not prepared to continue. It then quits and becomes the loser. Note that the outcomes of these contests are predetermined before the displays even begin; it is just that players cannot assess what these are. It is only after a contest has finished that this information is revealed. The first model below asks whether there can be an ESS in such contests if initial assessments are ignored and opponents just play out the endurance competition. The second model assumes that such an ESS exists and examines the optimal policy schedule for cost accumulation during such an endurance contest.

Limited energy models/War of attrition without assessment (WOAWA): The WOAWA model (Mesterton-Gibbons et al. 1996) assumes that individuals differ in the amount of energy or other limiting resource that is available for a protracted contest; distributions of maximal resources among the population’s individuals are assumed to be unimodal with a long tail at higher values. The longer a given player continues a contest, the less resource is available for other fitness-enhancing functions. Key parameters for this model are $\beta$, the rate at which a contest uses up this resource, and $\alpha$, the efficiency with which residual resource after a contest can be turned into fitness. The cost/benefit ratio, $\beta/\alpha$, is denoted by $\theta$. A second key parameter is $\kappa$, the coefficient of variation in the amount of total resource held by different contestants in the population. Analysis of the model shows that an ESS can exist when players do not assess each other prior to initiating a contest as long as the relative cost/benefit ratio $\theta$ is small enough and/or the variation among contestants, $\kappa$, large enough (Figure 6A). In addition, the ESS identifies the maximal fraction, $\upsilon$, of the total
available resource at the start of the contest that an animal should commit before quitting (Figure 6B).

**Figure 6: ESS outcomes of WOA model.** (A) Combinations of coefficients of variation in initial resources ($\kappa$) and cost/benefit ratios of contests ($\theta$) that preclude (blue) or favor (tan) a stable ESS in which contestants do not assess each other before beginning a contest, but just play out their resource until it hits a critical fraction, $\upsilon$, of their total resource available. First opponent to hit this ceiling quits and is therefore loser. High cost/benefit ratios and/or low variation among combatants in initial resource level do not support this ESS. Combinations that favor the ESS may leave losers with less variation in residual resources after the contest than winners (light tan) or winners may show less variation than losers (dark tan). (B) ESS fraction of total available resource that should be assigned to a contest $\upsilon$ as a function of population variation in initial resources available (horizontal axis) and cost/benefit ratio (different colored lines). As cost/benefit ratio increases, average fraction of resources that should be allocated to a contest decreases. (After Mesterton-Gibbons et al. 1996.)

These authors also considered which statistical model best fit the distributions of resource identified in wild populations; the Weibull distribution ([http://en.wikipedia.org/wiki/Weibull_distribution](http://en.wikipedia.org/wiki/Weibull_distribution)) appeared to give a better fit than a lognormal or gamma distribution. They also pointed out that while the asymmetric war of attrition predicts an inverse correlation between actual contest durations and the asymmetry in player fighting abilities, the WOAWA model predicts a positive correlation between contest duration and the residual resource remaining in losers of contests. This provides some interesting tests for comparing which of these two models, if either, fits a real system.

**Limited energy models/Energetic war of attrition model (EWOA):** The EWOA analysis looks for the optimal allocation of display effort during an endurance contest (Payne & Pagel 1996). Since the relative frequencies of players with different maximal endurance times are assumed to be fixed, and adoption of particular effort schedules by various players has no effect on those frequencies, this model is more of a simple optimization problem than a game. However, it provides some interesting predictions about when an endurance display system can or cannot pay for its costs.

In this model, players are competing for some commodity of value $V$, and each player has its own cost ceiling for a particular contest. This cost ceiling could be reached by performing high intensity or rapidly repeated displays throughout a short duration contest, or alternatively low intensity/infrequently repeated displays over a longer period. A focal animal’s instantaneous level of display (intensity, repetition rate, or both) at any time $t$ in the contest is denoted by $a(t)$ and the cumulative "signal" generated by this and all prior display in this contest is denoted by $s(t)$. The cumulative cost of displays at point $t$ in the contest is

$$C(t) = F(t) + T(t)$$
where $F(t)$ is the cumulative energy cost and $T(t)$ is the cumulative time lost (or fatigue acquired), both scaled in the same currency. At frequent intervals during the contest, each player compares its cumulative signal to its threshold value $X$. As long as its cumulative display is less than its $X$ and its opponent is still displaying, it also continues to display. Once $s(t) \geq X$, the player either flees (ending the contest), or escalates it into a more violent stage. Assuming no player escalates after $s(t) \geq X$, denoting the average duration of a contest against an opponent $B$ that has a threshold $X_B$ by $\tau(X_B)$, and denoting the distribution of players with different values of $X_B$ by $N(X_B)$, the average payoff of endurance contests for a focal player $A$ is:

$$E_A = \int_0^{S_{A}^{\text{max}}} \{V - C_A(\tau(X_B))\}N(X_B)dX_B + \int_{S_{A}^{\text{max}}}^{\infty} \{-C_A(\tau(X_B))\}N(X_B)dX_B$$

The first term reflects cases where $A$ won, and the second cases where $B$ won.

Holding $V$ and $N(X_B)$ constant, the authors then examined how the costs $C$ might vary with different emphases on display intensity versus contest duration for a given $X$. Replacing $C$ with expressions defining its dependence on $a(t)$, and setting the first derivative of that equation to zero and the second derivative to negative values identified the values of $a(t)$ during a contest that maximized the payoff. Three types of outcomes were identified (**Figure 7**).

**Figure 7: Possible outcomes for the EWOA analysis.** A Type I outcome arises when cumulative time costs of continued display increase rapidly with contest duration whereas cumulative energy costs rise only minimally. Then a single high-intensity display is favored (maximal $a(t)$ for the short duration of the contest). Type II contests arise when both cumulative energy and time costs increase during the contest. Type III contests arise when energetic costs rise rapidly with contest duration but time costs remain low. A minimally energetic display (low $a(t)$) is then given repeatedly for long periods as in the classical war of attrition. Whether a constant or changing display level is optimal during a contest depends on whether the cumulative time costs increase in accelerating (top dashed red line), linear (solid red line), or decelerating (lower dashed red line) manner. (Modified from Payne & Pagel 1996.)

In Type I contests, the optimal strategy is to produce a single maximum intensity signal; the duration of the contest then provides no information on player endurance. This situation can arise if the $T(t)$ costs increase non-linearly with $t$ but
energetic costs, $F(t)$, increase only slowly if at all during the contest. In Type III contests, the opposite occurs: it is best for each player to produce as costless a display as possible for as long a time as possible. In some respects, this is the classical war of attrition model. This is likely if energetic costs increase nonlinearly with contest duration, whereas time costs are minimal. Of greatest interest in this analysis are Type II contests. Here, a stable endurance game, in which each player’s stamina is honestly displayed by its maximal display duration, is only compatible with the relevant ESS if the cumulative energetic and cumulative time costs both increase significantly during the contest. In contrast with the sequential assessment game, in which each opponent should produce successive displays identically so that the average converges on their true fighting ability, the EWOA model can favor players increasing or decreasing the intensity or repetition rate of their displays as the contest proceeds: if time costs increase in an accelerating way, players should increase $a(t)$ as $t$ increases; if cumulative time costs increase in a decelerating way, players should decrease $a(t)$ as $t$ increases. This could explain changes in display intensity in natural examples for reasons other than escalating to obtain more information (as in staged sequential assessment games). Note however that for stable outcomes, all players have to adopt the same escalated or de-escalated display at the same time. How this “matching” might be achieved was not discussed in this paper, but is taken up in the next model.

### Cumulative assessment model (CAM)

Whereas the sequential assessment model focused on cumulative acquisition of information, and the two prior models focused on the cumulative energy costs of protracted contests, the cumulative assessment model combines the accumulated effects of energy consumption and acquired damage during a contest. It focuses on a contestant’s successive decisions about whether to continue or quit given the cumulative sum of these combined costs (Payne 1998). It is most relevant to species that employ ritualized fighting in which only a certain amount of direct physical damage or stress can be tolerated. It can also be applied to non-contact interactions as long as the contestants are subject to external time costs not under their control, such as predation risk or lost foraging time. At no point are rivals assessed, and instead players only self-assess their own energetic expenditures and accumulated externally caused effects like damage.

As with the EWOA model, this analysis examines how contestants who differ in some fighting ability or intrinsic quality related to fighting, $q$, should alter the intensity, here denoted by $R$, of their actions over the course of a contest to maximize their expected payoffs. $R$ can be a measure of the magnitude of a display and/or its instantaneous rate of repetition. All opponents are able to vary $R$ during the course of a contest. It is assumed that each contestant persists in the interaction until the total costs including damage inflicted by the opponent surpass some threshold, and ignores the effects of its own attack upon the opponent, other than to note whether the opponent is still fighting or has fled. Like the EWOA model described earlier, this is a model of fighting tactics (how to perform optimally against a given opponent for a certain threshold) and not a strategy of how to choose the best threshold.

In this model, the overall cost suffered by each contestant over time is a combination of energetic costs $F(t)$ and damage costs $D(t)$:

$$C(t) = F(t) + D(t)$$

Consider a focal player with quality $q$ that enters a contest with another player of quality $q'$. The rate at which energy costs accrue to the focal player at any time $t$ in the contest depends on the intensity of the actions $R(q,t)$ that it chooses to adopt, and the rate that it accrues damage costs depends on the intensity of the actions $R(q,t)$ adopted by its opponent over which the focal player has no control. The overall rate of cost accrual is then

$$\frac{dC}{dt} = aR^n(q,t) + dR(q',t)$$

where $a$ and $d$ are scaling parameters, and $n$ is an exponent that allows for the possibility that energetic and damage costs do not accumulate with the same power of $R$ over time.
The contest continues until one contestant flees at time $T(q)$ which occurs when its tolerance threshold $X(q)$ has been reached. The threshold reflects the costs an animal is willing to suffer in the contest, which in turn depend on contextual factors such as the value of the commodity being contested ($V$) and its own quality $q$. The overall expected payoff is given by

$$E(q) = \int_{0}^{\infty} \{V - C(T)(\hat{q}, q)\}p(\hat{q})d\hat{q} - \int_{q}^{\infty} C(T)(q, \hat{q}))p(\hat{q})d\hat{q}$$

where $p()$ is the probability density function of possible opponent qualities. As with the EWOA model, the first term accounts for contests that are won (opponent flees first), and the second term accounts for contests that are lost. The goal is to choose the optimal policy for each contestant as expressed by $R(q,t)$ that maximizes $E(q)$.

The author then examines a likely general case in which at least one party linearly escalates its $R$ at which point the other should also increase its $R$ and again is limited to linear increases. The question is then what the optimal intercept (initial $R$) and slope (rate of increase in $R$ over time) should be for players with different quality ($q$) values. Stable policies are only present if the exponent $n$ is greater than 1: energetic costs must rise in an accelerating manner as $R$ is increased or no policy can be stable. The predicted ESS policy intercepts and slopes for contestants with linearly increasing $R(t)$ are shown for high and low $q$ individuals in Figure 8.

**Figure 8:** Optimal policies for increasing display intensity ($R(t)$) during a contest according to CAM model. In all cases, red line shows optimal policy (particularly intercept and slope) for high quality individuals and blue line indicates optimal policy for low quality individuals. High quality individuals should always adopt a higher initial value (intercept) than low quality ones. (A) When contests are likely to be brief, high quality individuals should adopt a high initial $R$ and lower quality individuals a lower intensity. There is no time for the latter to catch up and damage the former enough to change outcome. They should just quit. (B) When contests are likely to be long, both parties should pick low initial intensities. Since high quality individuals can always afford to increase intensity at a steeper rate than low quality, the latter can never catch up. (C) For intermediate duration contests, high quality individuals again choose a high initial $R$. However, there is a chance low quality players can inflict sufficient damage on high quality opponents if they increase intensities faster and thus have a chance to win. Intermediate duration ensures that this strategy does not trap lower quality players into a long period of high costs. (After Payne 1998.)

Optimal policies differ depending on the likely duration of a contest. In all cases, higher $q$ individuals should begin display at a higher $R$ than lower $q$ individuals. Thus the relevant linear trajectories always have a higher intercept for high $q$ than for low $q$ individuals. There are three general cases:
• **Short contests**: If contests are likely to be very brief (Figure 8A), cumulative costs are unlikely to be a concern and high $q$ individuals should adopt a very high (and expensive) initial value of $R$. Because it starts at such a high $R$, these players subsequently can only increase $R(t)$ slowly. They thus have a high intercept but gentle slope. Given this strong attack with little time left to increase its own $R$ and inflict sufficient counter-damage on the opponent, a lower $q$ individual should simply flee.

• **Long contests**: If contests are likely to be very long in duration, then both parties should adopt low initial $R$ values (Figure 8B). Since the higher $q$ individuals can endure cumulative costs better than low $q$ individuals, the higher $q$ individuals can afford to increase $R(t)$ faster than can the low $q$ individuals: thus the optimal slope for the high $q$ individual line should be steeper than that for the low $q$ individuals. Low $q$ individuals can never catch up and will thus be forced to quit sooner than high $q$ individuals.

• **Intermediate duration contests**: When contests are likely to be of intermediate length, the optimal policy for high $q$ individuals is again to begin with a relatively high $R$, but because this is expensive, they can afford to increase $R(t)$ at only a moderate slope (Figure 8C). Unlike the short contest situation, there may be enough time for lower $q$ individuals to increase their $R(t)$ at a fast enough rate (steeper slope) that they can inflict enough damage on the opponent to make it quit first. Sometimes it will succeed, and sometimes it will fail, but given sufficient time, it may be worth a shot.

A critical assumption of this model is that there needs to be some component of the cumulative costs that is beyond the control of each individual. Individuals can control their energy expenditures, but they cannot control the damage or stress imposed by the rival’s actions. Other possible external sources of cost include attraction of predators, increased vulnerability to parasites, lost foraging time, impaired ability to mate guard, and lost mating opportunities. The CAM model thus differs from the EWOA model primarily in the addition of cumulative costs due to external factors out of the focal animal’s control. Note that in addition to the escalation patterns illustrated in Figure 8, the cumulative assessment model predicts that contest duration will be positively correlated with the quality (energy reserves and defensive skill) of the loser. However, controlling for loser quality, contest duration should be negatively correlated with quality of the winner, since a higher quality winner will inflict costs on the loser at a higher rate. Depending on the relative importance of energetic and damage costs, these two predictive curves could vary in their strength.

**Testing the models**

The three models discussed above that predict contest trajectories (SAM, EWOA, and CAM), are sufficiently different that they generate divergent predictions that can be tested empirically. In the following table, different measures of relative fighting ability are all collected under a general term: resource holding potential (RHP) (see Chapter 11 in the text for more detailed definitions of this term). Below, we contrast a variety of predictions of these three trajectory models (an abbreviated version of this table also appears in the text):

<table>
<thead>
<tr>
<th>Decision based on:</th>
<th>Sequential Assessment (SAM)</th>
<th>Energetic War of Attrition (EWOA)</th>
<th>Cumulative Assessment (CAM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Difference between average opponent RHP</td>
<td>Sum of own actions</td>
<td>Sum of opponent’s actions</td>
</tr>
<tr>
<td>Assumes display level matching in population:</td>
<td>Depends on version</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Assessment of opponent:</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Escalation:</td>
<td>Not within a phase, but in sequential phases</td>
<td>Escalation and de-escalation possible</td>
<td>Escalation and de-escalation possible</td>
</tr>
<tr>
<td>Contest duration most strongly correlated with:</td>
<td>RHP asymmetry between opponents (–)</td>
<td>Loser RHP (+)</td>
<td>Loser RHP (+) and winner RHP (–)</td>
</tr>
</tbody>
</table>

**Table 1. Summary of the predictions of the SAM, EWOA, and CAM models.** (Sources: Briffa & Elwood 2009; Payne 1998.)
Contest duration increases with increasing mean opponent RHP?  

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Possible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Display characteristics:</td>
<td>Non-dangerous index signals or ritualized fighting tactics</td>
<td>Energetically costly chasing or handicap signals with enforced intensity matching</td>
<td>Dangerous displays</td>
</tr>
</tbody>
</table>

It is easy to stage contests between known sized opponents and measure contest duration. This sort of data has been collected for a large number of species and used to test the contest duration predictions as a way to distinguish among the models. Taylor and Elwood (2003) importantly pointed out that these patterns can be misleading. They showed with simple simulations that if a pure self-assessment process such as the EWOA was in operation, and there was a perfect positive correlation between loser RHP and contest duration, a spurious negative correlation between RHP asymmetry and contest duration could arise. This occurs because in a population with a normal spread of body size or RHP, smaller individuals would usually have large opponents. Body size is thus negatively correlated with contestant size asymmetry, generating the spurious correlation between asymmetry and duration. Moreover, if true assessment is occurring, such that RHP asymmetry is negatively correlated with contest duration, a spurious positive correlation between loser RHP and duration, and a spurious negative correlation between winner RHP and duration, would be generated.

To make matters even more complex, if the CAM model is in operation, the same positive correlation with loser RHP, negative correlation with winner RHP, and negative correlation with RHP asymmetry will be generated. Although the CAM predicts that the loser and winner correlations should be stronger than the RHP asymmetry correlation, while the SAM predicts a stronger RHP asymmetry correlation, noisy data can obscure such subtle differences. Below, we show some of these primary versus spurious predictions of the three models:

<table>
<thead>
<tr>
<th>Model</th>
<th>Loser RHP vs. Duration</th>
<th>Winner RHP vs. Duration</th>
<th>RHP asymmetry vs. Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM</td>
<td>(1)</td>
<td>(2)</td>
<td>(1)</td>
</tr>
<tr>
<td>EWOA</td>
<td>(1)</td>
<td>(2)</td>
<td>(1)</td>
</tr>
<tr>
<td>CAM</td>
<td>(1)</td>
<td>(1)</td>
<td>(2)</td>
</tr>
</tbody>
</table>

Figure 9: Predicted relationships of contest duration as a function of winner and loser characteristics for the three fighting trajectory models. Contest duration is on the y-axis in each of these graphs. (1) indicates a primary prediction; (2) indicates a spurious correlation. (After Gammell & Hardy 2003; Taylor & Elwood 2003.)

As a consequence of these complications, it is essential in a study testing the fit to alternative models to examine the
other differences among the models, including the dynamics of escalation, the presence of matching intensities, and the type of displays or fighting tactics employed (Briffa & Elwood 2009).

Literature cited


Web Topic 11.2
Resource Value and Ownership Asymmetries in Fighting Strategy Models

Introduction

In Web Topic 11.1, we focused on game and tactical models for contests in which outcomes are decided by the patterns of change in some state of a focal contestant: specifically, the states examined were information about the focal animal’s relative fighting ability, its energy reserves, and the accumulated damage that it might have sustained during the contest. Although these models usually included a term for the value of the contested commodity, this was held constant for all players. In this Web Topic, we examine similar games but where the contested commodity has different value for the two contestants. A special case occurs when one party has ownership of the commodity and is challenged by another. Again, we begin with the classic war of attrition models and relax or modify their assumptions to see when an ESS is possible and, if relevant, what scheduling policy of costs might be optimal to achieve that ESS. A final section reviews the mechanisms by which animals might assess resource value and ownership.

War of attrition with variable resource value

As we saw in Web Topic 10.5, the symmetric war of attrition model assumes that two animals compete for an indivisible commodity of value $V$, and that their only strategic choice is the length of time they will continue in a contest. The only ESS is a mixed one based on a random selection of a persistence time from the following probability density function:

$$P(t) = \frac{k}{V} e^{-kt/V}$$

The individual with the shorter time quits, and the other individual wins the commodity. All individuals are assumed to suffer the same rate of cost accrual per unit time, $k$, and place the same value $V$ on the contested commodity. As long as $-k$ is constant for all parties, we can replace $t$ with investment cost $x = kt$.

One variant of this model, called the war of attrition with random rewards, considers the situation in which contestants may be in different commodity-requiring states on different occasions, and each individual only knows its own state and not that of its opponent (Bishop et al. 1978). For example, animals may differ in their hunger state; hungrier individuals should be willing to compete more strongly for food. The model assumes that being in a certain state is independent of the opponent’s state. The payoff to an individual in state $u$ that selects an investment cost $x$ against an opponent selecting $y$, or $E_u(x,y)$, depends on whether $x$ or $y$ is larger:

<table>
<thead>
<tr>
<th>$E_u(x,y)$</th>
<th>$x &lt; y$</th>
<th>$x = y$</th>
<th>$x &gt; y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-x$</td>
<td>$V_u/2 - x$</td>
<td>$V_u - y$</td>
<td></td>
</tr>
</tbody>
</table>

If a focal animal selected the smaller investment $x$, it loses and its payoff is then just $-x$. If both parties picked the same investment $x$, they split the commodity or have equal chances of getting it and their net payoff is $V_u/2 - x$. If the focal animal is prepared to perform the larger investment $x$, its payoff is then $V_u - y$ since it is the time selected by the loser that determines contest duration. The overall payoff for individual J of playing a given strategy $a$ against opponent $I$ is then

$$E_u(a,l) = P_a(V_u - a^*) - (1 - P_a) a$$
where $P_a$ is the probability of winning against $I$ if $J$ plays $a$; and $a^*$ is the expected investment cost if it wins, and $a$ is its cost if it loses. The ESS is a probability density function $G(x)$ with discrete sets of abutting curves depending on the number of need states. In the case of four different need states, the distribution of $G(x)$ appears as in Figure 1 below.

![Figure 1: Probability density function $G(x)$ from which focal animal $J$ should choose investment cost $x$. Four need states (e.g., hunger) are shown as separate colors with most needy state on far right. An animal in a given state should randomly select an investment cost ($x$) from the probability function defined by the line at the top of its colored segment. (After Bishop et al. 1978.)](image)

The investment cost that an animal is prepared to make to win a contest will increase the more it values or needs the contested commodity. If there are differences of state between animals, the mean payoff given the ESS above is positive and greater for those animals for which the payoff of victory is higher. This means that hungrier individuals gain by being aggressive, and have a higher net payoff after allowing for the costs of aggression than less hungry individuals.

### The sequential assessment model with asymmetric resource value

We saw in Web Topic 11.1 that the sequential assessment model assumed that contestants know little about each other’s fighting ability at the outset of the contest, and update this information as one or more types of displays are repeated over time. The ESS is based on a giving-up line that specifies a threshold estimate that one contestant is reasonably certain it will lose. This individual then quits and the contest is terminated. In the original model, contestants both knew the value of the resource ($V$) they were contesting at the outset and valued it to the same degree. As the value of the resource was increased, the giving-up line moved down and both contestants would persist for longer. This meant that contests over more valuable resources would be longer on average, everything else being equal.

A variant of this model allows the two contestants to place different values on the resource (Enquist and Leimar 1987). It is called the **sequential assessment game with random rewards**, and like the war of attrition with random rewards, assumes that animals vary in their resource valuation. The distribution of valuations in the population is assumed to be independent of the distribution of fighting abilities. The model then assumes that information about both relative fighting ability and opponent resource valuation can be acquired and updated during the contest. In practice, this might require different signals to indicate resource value and fighting ability. The model further assumes that as each
contestant improves its estimates of its opponent's valuation of the resource, it does not change its own valuation or fighting strategy as a result. The ESS then predicts separate giving-up lines for each opponent (Figure 2). Contestants with a high valuation will rely on a giving-up line that is below that of an opponent with a lower valuation. Thus contestants with higher valuations will tend to be more persistent in contests.

Figure 2: The sequential assessment model ESS giving-up lines for players with different resource valuations. (See Web Topic 11.1, Figure 4 for details on this type of graph.) Contestants with low valuation (e.g., red line) have giving-up lines higher in this graph than those with high valuation (blue line). If fighting abilities are similar, a contestant with the higher valuation is predicted to persist for longer and is therefore more likely to win against an opponent with a lower valuation. (After Enquist and Leimar 1987.)

Since relative fighting ability and resource value are drawn from independent distributions, the duration and outcome of a fight can vary in complex ways. For example, if contestant A has a greater valuation than an opponent B, the longest contests will occur when B is slightly stronger, and contests won by B will tend to be longer than those won by A. Figure 3 illustrates how variation in resource valuation against a random opponent affects the expected length of contests.
If the cost of acquiring information about relative fighting ability is decreased, either by decreasing fighting costs or decreasing the estimation error, the ESS changes so that the giving-up lines come closer together for opponents with different valuations. The consequence is that the outcome of fights will be more determined by relative fighting ability than by asymmetries in subjective resource value. Increasing cost or error will have the reverse effect of spreading out the giving-up lines. For sufficiently large cost or error, the ESS changes qualitatively, in that individuals with small valuations will decline to enter a contest. Increasing the range of variation of subjective resource valuations in the population or decreasing the range of variation of relative fighting ability has a similar effect on the ESS, making the outcome of fights more determined by asymmetry in subjective resource valuation than by asymmetry in fighting ability.

**Sequential assessment model with asymmetric resource value information**

An initial information asymmetry about resource value is most likely to occur for owner–intruder conflicts. The owner has spent some time with the resource and has more accurate information about its value than the intruder. This situation was modeled as a sequential assessment game in which the owner bases its decision to continue a contest or quit on both local resource value and information acquired about relative fighting ability, whereas the intruder bases its decision on fighting ability and an average estimate of resource valuation for the entire population (Enquist and Leimar 1987). This results in one ESS giving-up line for intruders, and a series of ESS giving-up lines for owners depending on their valuation of the resource. The respective giving-up lines are shown in Figure 4.
Figure 4: Owner–intruder game with information asymmetry. The black line is the intruder’s giving-up line, and the colored giving-up lines represent different possible valuations of the resource by the owner (higher values are lower on the plot). The shape of the intruder’s switching line is flatter, reflecting the fact that the owner’s behavior includes information about resource value. (After Enquist and Leimar 1987.)

The owner will become more persistent as its valuation of the resource increases. If the owner has a high resource valuation it will tend to persist and win against intruders, but if its valuation is less than the average for resources in this population, the intruder will win. Thus intruders will take over most poor resources but fail to acquire most high resources. A recent summary of whether this or the prior model predicts real animal contests can be found in Arnott and Elwood (2008).

Sequential assessment model with uncorrelated role asymmetry

Another variant of the sequential assessment model investigated the stability of fighting strategies based only on a role asymmetry that is uncorrelated with either fighting ability or resource valuation (Leimar and Enquist 1984). The logic is that both parties may do better to use some role asymmetry to settle a contest fairly quickly without enduring costs of a protracted contest. The invoking of the role asymmetry is thus a convention: which role tends to win and which lose may be completely arbitrary. In the context of a sequential assessment game in which both parties are trying to obtain more accurate estimates of their relative fighting ability, asymmetric roles would enter the model by specifying different giving-up lines for the two roles. The giving-up line for one of these roles would be crossed sooner on average than either would cross the shared line without invocation of role differences. The alternatives are best seen in the standard sequential assessment graph (Figure 5):
Figure 5: ESS solutions for the sequential assessment game with an uncorrelated role asymmetry. The black line indicates the giving-up line when role asymmetries are ignored. This is in fact one ESS of this type of game. The alternative ESS is to invoke the role asymmetry by assigning one role to the blue giving-up line and the other to the red giving-up line. While chance always plays a role in sequential assessment games, the role assigned to the blue giving-up line will on average quit first. (After Leimar and Enquist 1984.)

Starting from a symmetric contestant situation, if one role is gradually made more favored with respect to fighting ability or resource value, the dichotomous solution becomes more likely. Thus the bourgeois strategy, in which resource owners typically win fights against intruders because intruders recognize the role asymmetry and back down, is more stable when owners either value the resource more, or are on average better fighters. Asymmetry in average strength is likely to appear when a resource is contested several times and stronger individuals have an advantage in such contests. Stronger individuals thus accumulate as territory owners, and owners will then be more persistent than intruders. Thus when the asymmetry involves resource ownership, the model predicts that: 1) owners will win when opponents are of equal strength, 2) contests won by the owner will tend to be shorter than those won by the intruder, and 3) the longest contests will occur when the intruder is slightly larger. These predictions have been realized in numerous lab and field studies. See Appendix A1 in Kokko (2006) for an extensive list of examples.

A self-consistent model of the bourgeois strategy

Previous models of territorial conflict resolution invoking an uncorrelated convention based on ownership have two serious problems. One is that the game invariably has two ESSs: one favors respecting the bourgeois convention (owners persist but intruders give up due to the asymmetry) and the other relies on the contrary convention in which intruders persist while owners retreat. This non-intuitive anti-bourgeois ESS would result in territories constantly changing owners, and if reproduction occurs on the territories, it would be continually interrupted. This solution has also been called a paradoxical ESS, in contrast to the common sense bourgeois ESS. It is certainly not what is usually seen in nature. A second problem is that when a population approaches fixation of the bourgeois strategy, all suitable habitat becomes filled with uncontestable territory owners, and territories become a rare and highly valuable resource. This situation will select for highly aggressive floaters who have nothing to lose if they take great risks in a takeover fight; this is called a desperado strategy. A better modeling approach is to acknowledge that there will be density dependent and population-level feedbacks on the payoffs of alternative strategies: the value of territories would then increase if they become rare and the mortality costs of searching and fighting for territories would also change. Such
models are said to be self-consistent in that the payoffs can change in ways that are consistent with density
dependence and population ecology (Houston and McNamara 2006). Another way to view this is that such games
should really be treated as scrambles and not contests (see Web Topic 10.5).

Kokko et al. (2006) thus re-examined this type of game as a two-role continuous scramble. Unlike the prior models
discussed in this Web Topic, this game was treated as single-bout interactions instead of long sequences. The two
roles were territory owners and floaters without territories. Floaters interact with owners: either party can be aggressive
and persistent (“daring”) or meek and non-persistent (“careful”). Each strategy in this game is a probability pair: one is
the probability that a typical individual will be daring if an owner, and the other is the probability that it will be daring if
a floater. The authors call these pairs of probabilities the “aggressiveness” scores for each role. Note that the two
probability values in a strategy pair are assumed to be uncorrelated. The equations for single interaction payoffs
explicitly incorporated possible density dependent effects making the game a scramble. The authors also considered
cases in which relative fighting ability was versus was not a factor in interaction outcomes. The question was then what
conditions might favor an ESS in which the uncorrelated asymmetry of territory ownership resolved an interaction.

The authors examined this model using evolutionary simulations. Populations were started at some combination of
floater and owner aggressiveness values and many interactions were staged. Only owners were allowed to produce
offspring, but all offspring began life as floaters. Mutants with a pair of probabilities different from the current population
could invade and shift the current values if their levels of aggressiveness in the two roles were a better reply to the
current strategy than the current strategy was to itself. The authors could then track the dynamic trajectories of a
population over time for a given set of rates for mortality and territorial intrusion. Trajectories that ended at a fixed and
convergently stable point were then ESSs.

Four general extreme outcomes are possible in this game (Figure 6). A no respect ESS can arise in which individuals
always attack if in the floater role and always defend if in the owner role. This can be an ESS if baseline mortality for
intruders is high, costs of fighting are low, and vacant territories are scarce, so that floaters have little to lose by acting
as desperadoes. As the costs of fighting are increased, the ESS moves from a no respect pair strategy towards a
complete respect one (e.g., the bourgeois strategy) in which individuals always respect ownership if in the floater role
and always defend if owners. Depending on parameters, an intermediate partial respect ESS can arise in which
individuals always defend if owners, and attack with some probability between 0 and 1 if floaters. A fourth outcome
(called hippie world by the authors) involves no aggression by either party. This turns out never to be an ESS under
reasonable parameters. The last possible outcome is the anti-bourgeois paradoxical case in which individuals always
attack if in the floater role and always retreat if an owner. This is most likely when baseline mortality for owners is high,
that of floaters is low, and mortality risks of fighting are high.
Figure 6: Evolutionary trajectories of the self-consistent feedback model for a particular set of mortality and intrusion rates. In this example, relative fighting ability had no effect on conflict outcomes. Of the possible ESSs (no respect, partial respect, full respect, and anti-bourgeois), evolutionary trajectories for this set of mortality and intrusion rates clearly lead to only two possible ESSs: partial respect and anti-bourgeois. This two-ESS outcome is the typical outcome if conflicts are dangerous. Including relative fighting ability leads to similar outcomes, but will usually move the partial respect ESS closer to the full respect corner of the plot. (After Kokko et al. 2006.)

For most reasonable parameters, this model results in two possible ESSs: one is the anti-bourgeois strategy and the other can be the no respect, the full respect, or the partial respect ESS. This again raises the question of which ESS is most likely when two are present. It is instructive to examine the evolutionary trajectories shown in Figure 6: the “basin of attraction” of any ESS is the relative size of the plot area in which points likely lead to that ESS. In the example shown, this clearly favors the partial respect ESS. This is because the frequency-dependent changes in single interaction payoffs as evolution proceeds often affect floater and owners in different directions. Floaters often experience negative selection on aggressiveness as the population level of floater aggressiveness increases. This is what favors intermediate floater aggressiveness in the partial respect ESS. Owners, on the other hand, do not experience negative frequency-dependence and thus should always be fully aggressive at the non-paradoxical ESSs.

**Do animals assess resource value and use this in conflicts?**

The above fighting strategy models assume that animals, at a minimum, assess a resource’s value for themselves before engaging in a contest over it. Resource valuation is implicit in uncorrelated asymmetry bourgeois models since owners should not defend their territory unless they value it sufficiently. But this raises the question of whether animals can also assess their opponent’s valuation of the resource. This would only be feasible if each animal adopts some state or behavior that reflects its personal valuation of a contested site or item and that can be detected and interpreted by others. Animals might or might not benefit by advertising their personal valuation. For example, a current owner of a resource that it valued highly might benefit by declaring its intentions to defend that resource without compromise; less motivated intruders would then give up without a contest. On the other hand, intruders often cannot evaluate a resource immediately upon encountering it, and thus must assume that it is of average quality. A current owner that declares an unlimited motivation to defend that resource suggests that it might be much more valuable than average. Intruders might then be attracted to such declarations. For researchers interested in whether resource
valuation plays a role in conflict resolution, the absence of overt advertising might mean that the animal is hiding its valuation, or it could mean that resource valuation is not relevant or practiced by this animal. This poses a challenge to those who wish to test the predictions of the above models.

One set of indicators that an animal is evaluating the resource is the level of costs that they are willing to pay during an aggressive interaction. This level can be revealed during natural or induced conflicts as increases in contest duration, increases in vigor of certain activities, use of more risky displays and tactics, increases in the probability of injury, or increases in metabolic expenditure (measurable by researchers as increased oxygen consumption, lactate accumulation, or glucose depletion). One limitation to this approach is that measurements made at the end of a fight only reflect the loser’s resource valuation; they do not allow a full estimate of the costs that the winner was prepared to suffer (Arnott and Elwood 2008). An alternative technique for probing an animal’s subjective resource valuation during a contest is to interrupt the interaction at random points in time with a novel startling stimulus; this causes the attacker to break off the fight and hide; the speed with which it resumes the attack is then an indicator of its motivation (Elwood et al. 1998).

Suppose it is found that animals do evaluate the resource and this affects their behaviors. To investigate whether they can also assess their opponent’s resource value, researchers usually examine the degree to which a contestant varies its persistence, approach, and retreat behavior as a function of the opponent’s resource value. One can also search for potential cues or signals that may be associated with known variation in a sender’s resource value. Audio and video playback experiments using these potential motivational signals may be used to determine whether receivers adjust their approach, aggressive, and retreat behavior in response to these cues and signals.

Several studies examining animal resource valuation found no evidence of assessment by either party, despite a wide variation in resource value and clear benefits from higher-valued resources. For example, male scorpionflies (Harpobittacus nigriceps) provide females with an arthropod prey item as a nuptial gift during courtship. The larger the prey, the longer a male can copulate and the higher his fertilization success. Males sometimes attempt to steal another male’s prey item by engaging in aggressive contests. The duration of fights won by the usurper (because the owner gave up) was not influenced by prey size, which directly contradicts the theoretical predictions (see main text Figures 11.9B, C, and D) and suggests that the owner does not assess its own prey quality. Intruder persistence time in fights they lost also did not increase with prey quality, again indicating a lack of assessment (Thornhill 1984). Several other studies that failed to find evidence of resource assessment include contests over nest sites in male sand gobies (Pomatoschistus minutus) (Lindström 1992), contests over silk retreats and catch-nets in net-spinning caddisfly larvae (Arctopsyche ladogensis) (Englund and Olsson 1990), contests between female parasitic wasps (Pachyceropoidea vindemmae) for dipteran pupa hosts of different quality (Goubault et al. 2007), contests over prey among group-living pholcid spiders (Holocnemus pluchei) (Jakob 1994), and contests between male fallow deer (Dama dama) during the rut (Jennings et al. 2004). In most of these species, however, assessment of relative fighting ability did affect contest duration and outcomes.

On the other hand, numerous food deprivation studies have shown that hungrier contestants fight harder for a food resource, as predicted by models of subjective resource valuation (see main text Figure 11.9A, and reviews by Arnott and Elwood, [2008] and Enquist and Leimar [1987]). Internal reproductive state also affects aggressive motivation, as demonstrated in shelter-related contests among brooding versus non-brooding red swamp crayfish (Procambarus clarkii) and American lobsters (Homarus americanus) (Figler et al. 1995a,b; 1997a,b; 1998), nest site contests in male cichlid fish (Tilapia zillii) with different testis weights (Neat et al. 1998), and contests over hosts between female parasitoid wasps (Goniozus nephantinis) with different egg loads (Stokkebo and Hardy 2000). The effect of resource scarcity was demonstrated in male cricket (Acheta domestica) contests over females by subjects that had been deprived of female contact for different lengths of time (Brown et al. 2006, 2007).

Similarly, the predicted increase in contest duration when both parties value a resource highly has been found in a variety of species (see text Figure 11.9). One example involves sized-matched male red-spotted newts (Notophthalmus viridescens) fighting over females of different sizes (see text Figure 11.10). Larger females produce larger clutches and are therefore more valuable. Intruder males engage in wrestling contests with a male in possession of a female;
contests are longer when the newts fight over a larger female. This positive correlation between duration and female size indicates that the intruder is able to assess female value. Intruders appear to use tactile and visual cues during the encounter to assess female size (Verrell 1986). A similar story was described for male wrestling contests over females in the amphipod *Gammarus pulex*, where intruders tug at the female and appear to rapidly assess female size and moult stage, which is associated with their fecundity (Dick and Elwood 1990).

Asymmetric information, where owners but not intruders have information about resource value, clearly affected contest dynamics in several well-described cases. Owners are expected to increase their persistence and defensive effort in relationship to increasing value of their resource, whereas intruders lacking precise information use an average value. This results in intruders fighting too hard for poor resources and not hard enough for good resources (text Figure 11.9D). In orb web spiders (*Metellina* spp), a male guards a female and waits for her to catch a large prey item, after which he initiates courtship. Larger females are more fecund. Intruder males will fight with a guarding male, but females remain in their retreats during these male contests. Guarding male persistence was positively correlated with female size, but intruder male persistence was not, indicating that intruders lacked information on female quality (Bridge et al. 2000; Hack et al. 1997). Similarly, in female-guarding male dung flies (*Scatophaga stercoraria*), the duration of fights lost by owners was positively correlated with the value of the female, whereas duration of fights lost by intruders was not (Sigurjónsdóttir and Parker, 1981).

In territorial owner–intruder conflicts, we expect intruders to have less information about the territory’s quality than the owner, but in some cases the intruder does appear to acquire resource value information. One example is the funnel-web spider *Agelenopsis aperta*, where the contestants are females (Riechert 1978, 1979, 1984). If the intruder is more than 10% larger than the owner, it will take over the web. But an owner that is only slightly smaller than an intruder usually prevails against a larger intruder, so ownership does confer some advantage. This species has a repertoire of 33 acts performed during contests, which escalate through the following phases of increasing cost and risk: locating with possible assessment of relative body size (6 acts); signaling (14 acts), threat (3 acts), and contact (4 acts). Webs have high value to the owner both because of the investment in constructing them, and because of their location in good prey capture areas. When a web is supplemented with food, increasing its perceived value to the owner, the signaling phase of the contest and the diversity of signals employed is greatly increased. In staged owner–intruder contests, both contestants increased their contest cost with increasing web quality, indicating that the intruder also assesses web quality. However, when contests were staged between two intruders on a web whose owner was removed, there was no relationship between contest cost and web quality. Taken together, these results suggest that intruders battling an owner obtain information about web quality from the owner’s behavior. At least some of the signals appear to provide intruders with website quality information.

Hermit crabs provide a unique situation for investigating the question of self and opponent resource value assessment, because two separate resources are involved—the gastropod shell currently occupied by each opponent (Arnott and Elwood 2008; Elwood and Briffa 2001). The value of a shell is determined by the size of the shell relative to the body size of the crab. If contestants differ in size, the value of a given shell will differ. A larger crab might fight another individual to obtain a better shell, which could leave a suboptimal shell for the loser. But if the loser had a shell that was too large for it, both individuals could benefit from exchanging shells. Contests are asymmetric with two roles: attacker and defender. In a typical interaction, two crabs encounter each other and perform a cheliped display to assess relative body size. The larger individual usually assumes the attacker role and approaches and grasps the defender’s shell, causing it to withdraw inside its shell. The attacker runs its claws over the exterior of the defender’s shell and turns the shell to feel the aperture, a better way to assess the shell’s interior volume. If the attacker decides to escalate, it rotates the defender’s shell and begins a vigorous rapping of its shell against the defender’s shell; rapping occurs in repeated bouts of rapid taps and continues until either the defender is evicted or the attacker becomes fatigued and gives up. In some species (*Pagurus longicarpus*), attackers base their decision to confront another individual only on the basis of the low quality of their own shell and apparently do not assess the quality of the defender’s shell (Gherardi 2006; Tricarico and Gherardi 2007).
In the well-studied European hermit crab (*Pagurus bernhardus*), attackers gather information on the defender’s shell size after grasping and feeling it, and then make the decision whether or not to escalate (Figure 7) (Dowds and Elwood 1983; Elwood and Briffa 2001). The vigor of rapping is correlated with the attacker’s valuation of the opponent’s shell relative to its own, and successful eviction is associated with greater rapping intensity and duration (Figure 8).

Defenders in this species are apparently unable to obtain any information about the attacker’s shell quality to modify their persistence and giving-up decision. However, in at least some species (*Clibanarius* spp, *Calcinus tibicen*), the defender apparently is able to assess the attacker’s shell size, possibly using the fundamental frequency of the rapping sound. Here, the defender gives up more quickly when the attacker’s shell is of higher value from the defender’s perspective. The attacker also assesses the defender’s shell fit during the aperture investigation and preferentially attacks small individuals with overlarge shells. Resource assessment by both parties in these latter species results in interactions characterized more by negotiation than by conflict; shells are primarily exchanged between crabs only when both individuals stand to benefit (Hazlett 1987, 1989, 1996).
Figure 8: Assessment of resource value in shell-exchanging hermit crabs. (A) Logarithm of the number of rapping bouts per contest by attacking hermit crabs (*Pagurus bernhardus*) when the larger crabs possessed shells that were 50% smaller or 80% of the optimal size, and the smaller crab occupied a shell that was optimally sized for the larger crab (Briffa et al. 1998). Blue bars reflect rapping bout number when the larger crab succeeded in evicting the smaller one; brown bars indicate larger crab gave up and smaller crab was able to retain shell. Eviction required more rapping than non-eviction, and the number of raps the larger crab was willing to perform was greater when the fit to its current shell was worse. (B) Logarithm of the number of rapping bouts required for a larger crab to evict a smaller defender housed in a high quality shell as a function of the quality of the contested shell (Arnott and Elwood 2007).

The owner–intruder role asymmetry appears to be readily assessed in most contests involving territories, prior residency in an area, and mate guarding systems. The primary evidence for this assertion is that ownership frequently confers a competitive advantage on the owner (Kokko et al. 2006). In a summary of 99 studies that examined the effects of both ownership and relative fighting ability on the outcome of contests, 77 studies found a positive effect of ownership on the probability of winning. In 58 of these studies, relative fighting ability also played a role, and ownership was the sole determinant of winning in 20 cases. Relative fighting ability was the sole determinant in 21 studies, indicating that ownership was not important or not assessed. Only one study concluded that neither ownership nor relative fighting ability affected contest outcome (Leimar and Enquist 1984). The species in these studies included sea urchins, spiders, crustaceans, insects, fish, amphibians, birds, and mammals, attesting to the ubiquity of the “owners usually win” phenomenon and to the fact that high cognitive skill is not required to achieve this assessment.

**Literature cited**


Web Topic 11.3
Positive Allometry of Weapons and Ornaments

Introduction

Most non-scientists shown two breeds of cattle that differed two-fold in average length might surmise that the larger species had horns and penises that were twice as long as those of the small species, offspring and brains that were twice as heavy, a similar diet but a need to eat twice as much daily, and as a consequence, required access to twice the pasturage required by the small breed. They might even guess that the larger breed would live twice as long as the smaller one. In fact, most of these seemingly good guesses would be wrong. And while we used body length as the reference measure, most seemingly good guesses would have been as wrong had we used body mass, foot mass, blood volume, or running speed as the reference trait. The general finding is that the ratios between the magnitudes of two different organismal traits do not stay constant, (e.g., keep the same relative proportions), when compared in the same animal at different stages of development, different animals of the same age and species, or different animals at equivalent ages in different species (Huxley 1932; von Bertalanffy and Pirozynski 1952; Gould 1966). If the ratio for the two traits did stay constant when one was varied, we would call those traits isometric; if, on the other hand, the ratio varied as the magnitude of one of the traits varied, we would say the traits were allometric. The observation that many pairs of traits vary allometrically turns out to have major consequences for the organisms’ physiologies, behaviors, ecologies, and life histories (Schmidt-Nielsen 1984; Peters 1993; Brown and West 2000; Bonner 2006). These important consequences immediately raise the question of why some traits vary isometrically whereas others vary allometrically. In this Web Topic, we briefly review some basic background on biological allometry, note some of the more important general relationships, then turn to the specific role of allometry in the evolution of ornaments and weapons subject to sexual selection.

Fundamentals of allometric analysis

QUANTIFYING ALLOMETRY A measure of how two different biological traits might co-vary is called their scaling. Scalings can usually be quantified by relating the magnitude of one trait, say $y$, to the magnitude of the other, $x$, using a power law:

$$ y = k x^a $$

where $k$ is the scaling coefficient and $a$ is the scaling exponent. If the best estimate for $a = 1$, then the two traits scale isometrically and their ratio will remain constant regardless of the value of $x$. If $a > 1$, then the magnitude of $y$ is an accelerating function of $x$, and is said to be positively allometric in relation to $x$. If $0 < a < 1$, then $y$ still increases when $x$ does, but at a decelerating rate. In this case, $y$ has a negatively allometric relationship with $x$. Note that if $y$ is positively allometric when compared to $x$, then $x$ will be negatively allometric with respect to $y$.

What happens to the ratio between $y$ and $x$ when the two are related by a power law? If $y = kx^a$, then $y/x = kx^{a-1}$. If $a=1$, then $y/x = k$: this simply says that if two traits are isometric, their ratio is a constant. If $a>1$, then the ratio $y/x$ is an increasing function of $x$; if $0 < a < 1$, then $y/x = k/x^{1-a}$ which means the ratio gets smaller as $x$ gets bigger.

In practice, one collects a set of equivalent and independent sample pairs of $x$ and $y$, and then uses a statistical method to obtain a best fit to the power equation variables $k$ and $a$. It is much easier to do this if one compares the log $y$ to the log $x$ since if there is a power law fit

$$ \log y = \log k + a \log x $$

This is an equation for a straight line with intercept $\log k$ and slope $a$. Note that the only way that the slope $a$ can be greater than 1 is if the range of values of $y$ is greater than the range of values of $x$ (Eberhard et al. 1998). This will have significance later.
Several statistical methods are available to find the best-fit line to such a data set (Harvey and Mace 1982). Ordinary least square regressions assume that there is no error in the measurement of the $x$ variable. When the measurement errors for $y$ are known to be significantly larger than those for measurements of $x$, least square regressions can be used to find best-fit lines. When it is known that the measurements of $y$ and $x$ have similar errors, least square regressions will underestimate the slope of the true relation between the variables. In these cases, alternative methods, such as reduced major axis analysis, should be used to find best-fit lines. Where the sample points might not be independent (e.g., if a large subset of the species in a multi-species comparison consists of species belonging to the same genus), various methods have been developed to control for these dependencies (Harvey and Pagel 1991). Once the relevant controls and methods have been applied, the slope of the resulting (hopefully) straight line provides an overall measure of $a$, the allometric exponent, and the intercept $\log k$ reflects the average ratio of the two traits across the entire sample. Where the function is clearly not linear on a log-log plot, more than one parameter will need to be extracted to describe the relationship between $y$ and $x$.

**DIFFERENT USES OF ALLOMETRIC MEASURES** When traits are compared at different developmental stages in the same species, the resulting relationship reflects the *ontogenetic allometry* between the traits. A trait relationship based on variation among individuals of the same species and at the same life stages is called *static allometry*. Patterns based on comparisons between many different species reflect *evolutionary allometry*. Even where the same species are involved, these three different comparisons can result in quite different values of the scaling exponent $a$. Usually, allometric relationships are stronger (e.g., have steeper slopes) for ontogenetic than for static, and for static than for evolutionary contrasts.

![Figure 1: Types of allometric relationships.](image)

Where one measures two traits in two different samples, say static contrasts for adult males and adult females, or multi-species contrasts for birds versus lizards, one can then compare both the intercept values (reflecting the average trait ratios) and the line slopes (corresponding to the scaling patterns) for the two data sets. Males and females might have similar intercepts but different scaling patterns; birds and lizards might have similar slopes but different intercepts. Another common use of allometry is to correct for body size effects when comparing other traits. For example, one
might want to compare testis size in monogamous versus polygynous mammals. Absolute testis size will surely increase with body size regardless of the mating system. To control for this, one would first compute the allometric relationship between testis size and body size across the entire sample; the direction and magnitude of deviations above or below this line could then be correlated with mating system. This is called the **comparative method** (Harvey and Pagel 1991) and is discussed further in Web Topic 10.1.

**Sources of allometric relationships**

There are two general causes of the allometric relationships seen in biological systems:

- **Physical laws**: Basic geometric, physical, and chemical laws constrain all organisms. Combining a particular physical law and a particular geometry will often generate allometric relationships. For example, consider an animal that has to keep warm. The rate at which it can produce heat energy depends upon its total mass $M$ which is proportional to its volume $V$. The rate at which it loses heat depends upon its surface area which is proportional to $V^{2/3}$. The energy that it has left over after thermoregulation is then $kV - mV^{2/3}$. For fixed $k$ and $m$, this difference gets larger as $V$ increases. An animal could thus gain several benefits from being larger. These include: a) reducing its foraging effort and thus risks until it just met its thermoregulation costs; b) switching to a lower quality but more abundant food; or c) continuing its current diet and foraging effort but using the extra energy for other activities such as soliciting mates or raising more offspring. This geometrical example led early researchers to the prediction that animal metabolisms should scale as the 2/3 power of their body mass with all the attendant advantages (Rubner 1883).

In fact, nearly all organisms examined, including plants, protists, and animals, have metabolic exponents between 2/3 and 1 (McMahon 1975; Pennycuick 1992; McNab 2002; Brown et al. 2004; Glazier 2010). Resting birds and mammals tend to be on the low end of this range, whereas plants and very active animals tend to lie on the upper end. There remains considerable dispute about why this range of exponents is found. On one side is the “1/4 rule” school arguing that transport of energy, nutrients, blood, and sap, not the surface area of the organism, limits how much metabolism can be sustained; all critical functions of the organism are then predicted to scale as integer multiples of 1/4 (allowing for both positive and negative multipliers) (West et al. 1999; Savage et al. 2004; Banavar et al. 2010). Combined with the assumption that body mass sets the upper limit on energy/nutrient production, this approach predicts that metabolic rates should scale with the 3/4 power of body mass, whereas heart rate and respiration should scale as the –1/4 power of body mass. Reproductive rates and life expectancies are also predicted to scale as integer powers of 1/4. While a 3/4 value is often found for metabolism, there are many exceptions (hence the large observed range of exponent values).

An alternative view is that whether the entire body mass or some transport factor is the limiting factor in metabolism can vary depending upon the current activity of the organism (Glazier 2010). An inactive organism (such as a plant or a resting animal) does not need a metabolic rate sufficiently high to be limited by its transport system or surface area. Thus its metabolic rate should scale according to its mass (an exponent of 1). The metabolic output of a moderately active animal is much more likely to be limited by its transport system or surface area as it must unload heat and wastes generated by the activity, and replace internal oxygen from outside sources. These activities should scale closer to the 2/3 exponent predicted by the surface area model. For very high activity bursts, most animals shift into an anaerobic metabolic state in which they rely on previously stored energy and oxygen and let wastes accumulate (e.g., lactic acid). Here again, the metabolism will depend on mass only and not on the transport system; the corresponding exponent for metabolic rate will again be 1. This model thus predicts that the range of exponents seen across taxa and contexts simply reflects a shifting of the limiting constraints between energy provision (body mass) and transport functions (circulatory systems or surface area). As evidence, they note that the intercept of the log–log plot lines should be higher as activity levels rise, and therefore the observed exponents should be correlated to some degree (but in a U-shaped manner) with the intercepts. Available data do tend to support these predictions.

Once determined, the allometric scalings for metabolism can lead directly to predictions of allometric scalings for
other traits such as behavior and reproduction. For example, one might expect home range size in animals to scale similarly to their metabolism (e.g., with exponents between 2/3 and 1). In general, this is what is found, but the range of variation is much greater than that found for most physiological or anatomical traits (Hendriks 2007; Hendriks et al. 2009). The increased variability appears to be due to the independent roles of an animal’s physiology and the abundance and spatial dispersion of food; the latter often follow fractal patterns in nature, and the observed scaling of home ranges with body sizes will thus depend both on the underlying metabolic needs and the fractal scaling of the food supply (Haskett et al. 2002; Buchmann et al. 2011). Metabolic scalings are also the basis of a prediction that the optimal lifetime reproductive effort of organisms should be a constant: assuming a metabolic exponent of 3/4, each female in a stable population should optimally produce a lifetime mass of offspring that is 1.43 times her own mass (Charnov et al. 2007). If this were true, there should be an inverse relationship between the average mass of an individual offspring and the total number produced. In fact, this is what is found: scaling exponents for individual offspring masses average 1/2 for heterotherms (plants and heterothermic animals) and range between 3/4 and 1 for homeotherms; offspring numbers for the same species have scaling exponents which are ranked inversely to those for offspring masses (Hendriks and Mulder 2008).

For our purposes, it is not critical whether metabolic rates in organisms scale with 2/3, 3/4, or some other value between 2/3 and 1. As long as the metabolic scaling exponent is < 1, there will be more strategic options available to larger organisms than to smaller ones. This can have very significant consequences for social organization, mating systems, life histories, and even communication (Schmidt-Nielsen 1984; Reiss 1991; Peters 1993; Niklas 1994; Calder 1996; Brown and West 2000; Bonner 2006; Hendriks 2007; Weiner et al. 2009; Reiss and Schmid-Araya 2010).

**Selection**: Most animals must compete for access to resources, refuges, and appropriate mates. This typically generates directional selection on any trait that might give their carrier an edge in competition (Green 1992). The simplest reason why such a trait might be positively allometric with respect to body size is that, for physical reasons, large animals have more spare resources to invest in competitive traits than small animals (Petrie 1992). This of course assumes that body size is under normalizing selection or at least a lower intensity of directional selection than the competitive trait. Alternatively, if selection favors greater among-individual variation in one trait than another, this will cause the first trait to show static positive allometry when compared to the second. Sexually selected traits are often found to be more variable than non-sexually selected traits (Pomiankowski and Møller 1995; Rowe and Houle 1996), and game theoretical models for some traits, such as personality, predict stable mixtures of many different phenotypes (McNamara et al. 2009; Botero et al. 2010; Dingemanse and Wolf 2010; Wolf and Weissing 2010). A third possible factor is that positive allometry of signal traits makes it easier to discriminate between individuals with large trait values (Wallace 1987). Given that measurement error in most sensory organs and brains tends to increase with mean stimulus magnitude (Weber’s Law; see Web Topic 8.6), positive allometry of signal traits would minimize receiver error and facilitate honest signaling.

Since resources are limited for most animals, positive allometric investment in one trait will lead to reduced resources, and thus negative allometric investment, in other traits (Nijhout and Wheeler 1996; West-Eberhard 2003). The optimal distribution of available resources among competing traits will depend significantly on the relative contributions that each investment brings to overall fitness. One common tradeoff involves investment in a given trait versus that in further growth. Bonduriansky and Day (2003) have provided a useful model of this tradeoff that predicts whether a given pair of traits will have an isometric, positively allometric, or negatively allometric relationship. Given an assumption that body mass scales with an exponent < 1 as an animal ages and grows, these authors predict resulting relationships for the following contexts:

- If fitness largely depends on the absolute size of the trait → isometry
- If fitness largely depends on the size of the trait relative to body size → isometry
- Given directional selection for larger traits and stabilizing selection on relative trait size due to survival costs → isometry
- Given directional selection favoring both larger trait and body sizes but stronger selection on trait size →
Given directional selection for larger traits, survival increasing with body size but decreasing with trait size → negative allometry

Given directional selection for larger traits, viability increasing with body size but in a decelerating manner; no viability consequences of trait → positive allometry

Given stabilizing selection on the trait but directional selection on body size → sigmoid allometric curve resulting in polymorphic populations

The basic point is that the type of allometric relationship arising from a tradeoff between investing resources in one trait versus in growth depends entirely on the shapes of the fitness functions for the trait and increased body size. In general, positive allometry only arises when the marginal increase in fitness by having a larger trait is greater in a larger animal than it would be in a smaller animal. This is similar to the requirement of the handicap principle for signal honesty: signals will only be honest if the fitness consequences of a high quality individual giving a signal for high quality are greater than that same signal given by a low quality individual (see Chapter 10).

The Bonduriansky and Day model identifies selective regimes that will favor different types of allometric relationships. But what mechanisms are available to respond to this selection? The generation of static allometries depends upon differences in the previous ontogenetic allometries of the sampled individuals. Here we shall define ontogeny broadly to include continuous maturational trajectories as well as periodic renewal investments by adults (such as the annual moulting of bird plumages or the recreation of antlers by male deer). What could have caused individuals to undertake different ontogenetic trajectories earlier in their lives that generate static allometries now?

The generally accepted answer is that most organisms show some degree of phenotypic plasticity during ontogeny: the same genetic makeup can result in different phenotypes depending upon external cues and conditions such as food abundance, temperature, population density, injury, parasite loads, and the like (Schlichting and Pigliucci 1998; West-Eberhard 2003). The function describing the probabilities of exhibiting different phenotypes for a given genotype is called that genotype’s norm of reaction. When two traits are considered concurrently, one can think of the norm of reaction for their combination as a policy for defining their ontogenetic allometry: a sample policy might favor focused investment in overall growth when food is scarce during ontogeny, but more investment in specific organs, at the expense of overall growth, when food is abundant (Emlen and Nijhout 2000; Shingleton et al. 2007). The points in Figure 1 represent individuals of the same age. The distance of each point from its vertical axis intercept represents the total resource it had available to invest in traits x and y during its ontogeny. The fact that different individuals adopted different relative allocations of resource to x and y (indicated by the different slopes for the points), presumably reflects how their norm of reaction policy reacted to the total amount of resource that was or seemed likely to be available. The physiological mechanisms by which these policies are implemented may simply depend on differences in the degree to which different traits are sensitive to the same cues and conditions: if overall growth is more sensitive to ambient food levels than growth of a particular trait, the two schedules can get out of synchrony and this can change the final allometry (Shingleton et al. 2007; Tobler and Nijhout 2010). It is also possible that a relevant policy allows for shifts between several alternative schedules at switching points depending on ambient cues and conditions (Tomkins et al. 2005). At least in insects, the norm of reaction of individual genotypes is not much narrower than the average allometric relationships seen in wild populations consisting of many genotypes (Emlen and Nijhout 2000). This makes analyses easier, but is not necessarily the case for all organisms. When we plot static allometry, we may be seeing the consequences of one norm of reaction and varying cue values (e.g., some insects), or many different norms of reaction corresponding to different genotypes in the population.

There is good evidence that allometric policies are associated with significant additive genetic variance, and can respond to selection by shifting to patterns above and beyond any physically-determined allometries. Male stalk-eyed flies (Diopsidae) show strong positive allometry between the width of their eye-stalks and body size;
artificial selection in laboratory stocks can increase or decrease the allometric exponents, and natural variations in sexual selection and consequent viability are surely the reason for differences in the degree of allometry seen between species in this family (Wilkinson 1993). Similar artificial selection experiments altered the allometric relationships between horn and body size in scarab beetles (Emlen 1996). Studies of wild caught three-spined stickleback fish also suggest that allometric exponents are subject to selection independent of physical allometric effects on body size (McGuigan et al. 2010).

Allometry of ornaments and weapons

Given this prior background, we can now turn to recent studies of allometry in animal secondary sex characteristics such as weapons and ornaments. It is often observed that these traits show positive allometry relative to body size or at least have higher allometric exponents than other traits (Huxley 1932; Gould 1966, 1974; Echelle et al. 1978; Petrie 1988; Reiss 1991; Green 1992; Petrie 1992; Andersson 1994; Emlen and Nijhout 2000; Baker and Wilkinson 2001; Tomkins et al. 2005; Tomkins et al. 2010). A more recent review by Kodric-Brown et al. (2006) reported allometric exponents for weapons and ornaments in stag beetles, fiddler crabs, earwigs, and anoles ranging from 0.93–15.7, with modal values between 1.5 and 2.5. They then provided a model, described below, which they argued predicts positive allometry for all secondary sexual traits. Bonduriansky (2007) replied that neither his model nor the data supported this extreme a claim and cited a list of examples in which secondary sex characters did not show positive allometry, and others where non-sexual characters did (Bonduriansky 2007). Whereas some of Bonduriansky’s counter-examples have since been found to show positive allometry if reduced major axis analyses are used instead of ordinary least square regressions (Cuervo and Møller 2009), and traits associated with copulation such as penis length and claspers should be excluded since they routinely exhibit isometry or negative allometry (Eberhard et al. 1998; Eberhard 2009), there are still a number of species in which secondary sex traits are isometric or negatively allometric (e.g., anurans [Schulte-Hostedde et al. 2011] and feathers in some birds [Cuervo and Møller 2009]). Our reading of the literature is that most secondary sex characteristics do appear to show positive allometry, but enough do not that it would be unwise to use allometric relationships to predict trait functions or vice versa.

The Kodric-Brown et al. (2006) model builds on the earlier one by Bonduriansky and Day in that it also assumes that resources are limited, and thus investments in ornaments and weapons must trade off with investments in body growth. In their model, this tradeoff is linear: the investment available for general growth equals the total available minus that allocated to ornaments and weapons. Total fitness is equal to the sum of that contributed through body size effects and that contributed through the size of ornament/weapons relative to body size. It is assumed that either increased body size or increased relative ornament/weapon size increases fitness but with functions that are either asymptotic (body size effects) or rapidly decelerating (ornament/weapon size effects). As a result of these assumptions, combinations of body size and relative ornament/weapon size that would produce a given value of fitness plot as a hyperbola on a graph of relative ornament/weapon size versus the logarithm of body size. Different total fitnesses generate different hyperbolae. That hyperbola that just touches a species’ given tradeoff line is the optimal allocation (this is the fitness set analysis method of Levins [1968]). The authors then track the optimal allocations for a given individual over time, using the end point of its most recent growth as the starting point for the next stage. Interestingly, the resulting ontogenetic trajectories are almost always positively allometric. Also of interest, the steepness of the linear tradeoff has no effect on the allometric exponents; instead, it is the shape of the fitness hyperbolae that determines the degree of allometry. The tradeoff function, along with the disposition of the fitness hyperbolae, does affect the allometric intercept value.

Since there are exceptional species that exhibit isometry or negative allometry of sexually selected ornaments and weapons, there must be factors not included in the Kodric-Brown et al. model that need to be considered as well. Viability costs may not increase linearly with increases in ornament or weapon size, and thus the fitness generated by the weapon or ornament alone may not be a monotonically increasing function; after some threshold value, this component of fitness may decrease with increasing size due to physical constraints or nonlinear predation risks (Bonduriansky 2007). The separate contributions to fitness by body size and ornaments
or weapons may not be additive and independent: multiplicative and interaction effects are certainly possible. Alternatively, the partitioning of tradeoffs into two components, ornament/weapons versus body size, may be too simplistic and multiple components may be at play. For example, it has been shown in stalk-eyed flies that the flight imbalances caused by large male eyestalks have selected for a variety of counter-balancing adaptations (Husak et al. 2011). These, and similar compensations in other sexually selected species, minimize the viability costs usually expected for an exaggerated ornament (Husak and Swallow 2011). No single pair of traits is likely to evolve independently of other traits (Emlen and Nijhout 2000; West-Eberhard 2003). Optimization when there are multiple components can lead to quite different outcomes than if components are examined only as dyads. In addition to ontogenetic and physiological linkages, traits may be genetically correlated in ways that complicate the independent effects assumed in the Kodric-Brown et al. model. Clearly, major progress has been made in understanding the observed patterns of allometry in animal weapons and ornaments, but there are still issues to be clarified.

**Literature cited**


Web Topic 11.4
Songbird Territorial Negotiations

Introduction

Territorial songbirds use their songs to settle boundary disputes much like humans use words to negotiate business deals. Seventy-five percent of songbird species possess song-type repertoires ranging in size from a few to over 100 types, and many of the remaining species with one official song type vary their song structure in meaningful ways, such as by shifting the pitch or varying the duration with the number of repeated elements (Morton and Young 1986; Ritchison 1988; McGregor and Horn 1992; Shackleton and Ratcliffe 1994; Martin-Vivaldi et al. 1999). Most species also add or switch to call-like vocalizations, such as twitters, rattles, buzzes, or soft song, as contests escalate (Jarvi et al. 1980; Dabelsteen and Pedersen 1990; Lampe 1991; Rehsteiner et al. 1998; Poesel and Dabelsteen 2006; Anderson et al. 2008b). All songbirds learn the fine details of their species-specific song by listening to adult tutors, practicing the production of complex notes and note syntax, and crystallizing their songs at some point in early adulthood. Species differ greatly in when they learn, from whom they learn, and the degree to which they copy whole songs from other adults or innovate unique notes and song structures (Beecher and Brenowitz 2005; Brenowitz and Beecher 2005). In this Web Topics unit we discuss how birds (usually, but not always, the males) use their song types and other variable aspects of song structure and delivery pattern to make words and phrases that enable them to negotiate boundary disputes.

Variable song features and singing strategies

Male birds sing to attract mates as well as to defend territories. Females seek high-quality males that directly or indirectly maximize their fitness. As receivers, they compare males based on their signals and exert selective pressures on those aspects of singing behavior that reveal important differences in male health, genetic quality, and territory quality. Song traits preferred by females include high song rates or duty cycles, large or complex repertoires, high performance of difficult note structures, and songs of local, as opposed to foreign, population origin (Searcy and Nowicki 2005). Male receivers exert different selective pressures on singing behavior (Cardoso et al. 2007). Territorial rivals not only need information about relatively static male traits such as fighting ability and experience, but they also need to assess dynamic information about current motivation and valuation of particular patches of real estate. Thus song systems for territorial defense are designed to facilitate countersinging interactions: songs are relatively short and discrete and separated by silent gaps for listening to responses, repertoire sizes are small to intermediate, and the age window for learning may be timed so that males can acquire at least some song types that they share with their territorial neighbors. Short-term temporal patterns of song type delivery with respect to the neighbor’s singing are the key to assessment of relative motivation and negotiation of boundary locations (Todt and Naguib 2000).

Table 1 shows a list of the key singing strategies and song traits observed in songbirds. The table indicates whether the traits mainly vary between or within males; in some cases a trait may vary at both levels. While strategies such as song-type switching and matching obviously require males to possess song-type repertoires, others can be employed by species with a single song type. Also shown are the likely assignments of these song traits to categories of signals based on honesty-guaranteeing cost. Some of these features may be index signals associated with male age (repertoire size, song performance) (Mountjoy and Lemon 1995; Gil et al. 2001; Garamszegi et al. 2005; Forstmeier et al. 2006; Kiefer et al. 2006; Garamszegi et al. 2007; Nicholson et al. 2007; de Kort et al. 2009b), dominance (Christie et al. 2004; Botero et al. 2009), or local origin (song-type sharing with neighbors) (Wilson et al. 2000; Wilson and Vehrencamp 2001). Others are likely handicap signals of condition or need (song rate, song duration, amplitude, note structure and performance) (Ritchison 1988; Alatalo et al. 1990; Rehsteiner et al. 1998; Martin-Vivaldi et al. 1999; Bower 2005). Matching, overlapping, switching, and short-term diversity are probably conventional signals of aggressive or submissive motivation (Vehrencamp 2000). Switching rate was hypothesized to be a trait constrained by exhaustion, placing it in the category of a handicap or index signal of male quality. However, this idea has not been supported by recent evidence, and motivation appears to be the more likely explanation (Brumm et al. 2009). Soft song has been
studied in several species and found to be strongly correlated with subsequent aggressive behavior, especially attack behavior when a taxidermic mount is presented to the territorial male (Searcy et al. 2006; Anderson et al. 2007; Laidre and Vehrencamp 2008; Ballentine et al. 2008; Anderson et al. 2008b; Hof and Hazlett 2010; Akçay et al. 2011). Most authors argue that such songs can only be detected when close to the receiver, so delivery of this signal comes with a proximity risk and we classify this singing style as a proximity signal.

### Table 1: Variable song features of songbirds. Features may vary within and/or between males. Categories based on type of honesty-guaranteeing cost: C = conventional, I = index, H = handicap (quality or need), P = proximity.

<table>
<thead>
<tr>
<th>Song feature</th>
<th>Between male variation</th>
<th>Within male variation</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repertoire size</td>
<td>X</td>
<td></td>
<td>I</td>
</tr>
<tr>
<td>Song-type sharing with neighbors</td>
<td>X</td>
<td></td>
<td>I</td>
</tr>
<tr>
<td>Song duration</td>
<td>X</td>
<td>X</td>
<td>I, H</td>
</tr>
<tr>
<td>Singing rate</td>
<td>X</td>
<td>X</td>
<td>H</td>
</tr>
<tr>
<td>Song amplitude</td>
<td>X</td>
<td>X</td>
<td>I, H</td>
</tr>
<tr>
<td>Note structure</td>
<td>X</td>
<td>X</td>
<td>I</td>
</tr>
<tr>
<td>Song performance</td>
<td>X</td>
<td>X</td>
<td>I</td>
</tr>
<tr>
<td>Type-matching</td>
<td>X</td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Overlapping</td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Song-type diversity</td>
<td>X</td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Song-type switching</td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Soft song</td>
<td>X</td>
<td></td>
<td>P</td>
</tr>
</tbody>
</table>

Different species make use of a subset of these singing strategies, in combination with calls and/or visual displays, to form a graded system of aggressive escalation. Below we describe the escalation rules for four well-studied species.

**Great tit (Parus major)**

Great tit males have small repertoires of 1–8 song types. On average they share one song type with a given neighbor. Songs consist of two- or three-element phrases that are repeated a variable number of times to create songs (called strophes) of different duration. **Figure 1** shows an example. The same song type is sung repeatedly for long periods of time before switching to another type, called the “eventual variety” mode of singing by repertoire species. The birds appear to be open-ended learners, as they can add new song types, and may increase or change their repertoire in response to stimulation from playback or new neighbors (Franco and Slabbekoorn 2009).
Figure 1: A series of song strophes by a male great tit. The number of times an element is repeated determines the duration of the song, which can be varied strategically during countersinging contests. (Photo courtesy of Danny Gibson; spectrograms from recordings provided by Greg Budney.)

The key variable song features used during countersinging interactions are song duration, overlapping, and matching. In escalated contests, a male first challenges a target neighbor by singing a shared song type. The neighbor may then respond by also singing this song type. Matching by senders is considered an aggressive intention signal, as the sender follows it with an aggressive approach response (Krebs et al. 1981). Overlapping also occurs in the early stages of a countersinging interaction when the birds are some distance apart (Dabelsteen et al. 1996; Langemann et al. 2000). When overlapped repeatedly, males sometimes shorten their song so that they can reverse the pattern and become the overlapper. At this point, variation in song duration becomes the mechanism of negotiation. Longer songs appear to be more threatening, and males that increase their song duration relative to the rival are subsequently more likely to approach the sound source (Lambrechts and Dhondt 1987; McGregor and Horn 1992; Amy et al. 2010). In other cases, males change their song duration in the opposite direction to the rival or playback, increasing duration when the rival decreases duration and vice versa. Matching is thus a necessary first step in these duration and overlapping contests, as the two birds must be singing the same song type with the same phrase duration to adequately match each other’s song duration and delay between song starts (McGregor et al. 1992). Figure 2 shows some basic rules of escalation for the great tit.
Male great tits that are dominant in winter flocks sing longer songs on average than subordinates, and show less “positive drift” in the successive intervals between phrases in a song. Dominant males are more likely to respond to playback of long songs by increasing their own song duration, whereas subordinate males decrease their song duration (Lambrechts and Dhondt 1987). Moreover, the consistency of element frequency within strophes increases with age, and playback of more consistent songs elicits a stronger aggressive response compared to playback of less consistent songs (Rivera-Gutierrez et al. 2010, 2011). Dominant males also have wider black breast stripes (Järvi and Bakken 1984). These acoustic and plumage signal features are correlated with other measures of male quality such as survivorship and reproductive success (Lambrechts and Dhondt 1986; Rivera-Gutierrez et al. 2010).

**Black-capped chickadee (Poecile atricapillus)**

In the black-capped chickadee, a relative of the great tit, males possess a single song type consisting of a two-note whistle (*fee-bee*), but the entire song can be shifted up or down in absolute frequency to produce a continuous range of “frequency types.” Countersinging males employ two short-term singing features: overlapping and frequency matching (Figure 3).
Observations of naturally occurring interactions recorded with a microphone array system revealed that overlapping is very common during contests (at least one instance of overlapping in 80% of contests) but not associated with closer approach. Frequency matching is less common (37% of contests), generally preceded by overlapping, and somewhat
more frequent in contests where the rivals approached each other. Responses of receivers to different playback treatments generally show stronger responses and greater agitation to frequency matching than to overlapping. Thus overlapping and matching form a graded system of escalation, with matching associated with higher levels of escalation (Shackleton and Ratcliffe 1994; Otter et al. 2002; Mennill and Ratcliffe 2004; Fitzsimmons et al. 2008; Foote et al. 2008).

Chickadees form winter flocks with a dominance hierarchy, and dominance status in the prior winter’s flock affects some aspects of singing behavior and responses to playback (Mennill and Ratcliffe 2004; Ratcliffe et al. 2007). Dominance status is signaled by the brightness of the white head patch and song fine-structure: dominant males can maintain the same frequency ratio between the fee and bee notes when shifting absolute song frequency, whereas subordinate males exhibit a smaller frequency ratio when shifting to a higher absolute frequency (Christie et al. 2004; Doucet et al. 2005).

**Song sparrow (Melospiza melodia)**

Song sparrows are widespread throughout North America and subdivided into 34 morphological subspecies (Zink and Dittmann 1993). Males typically possess repertoires of about 8–10 discrete song types delivered repetitively in eventual variety mode. Song-type sharing varies among subspecies and populations. Males in some western subspecies are sedentary and share on average 20–40% of their repertoires, but for any two neighbors this value can range from 0 to 90% (Hill et al. 1999; Wilson et al. 2000). Song learning extends into the spring of a young male’s first year of life, when he has an opportunity to eavesdrop on singing neighbors, but the repertoire crystallizes after that point (Nordby et al. 2001, 2002; Burt et al. 2007; Nordby et al. 2007; Nulty et al. 2010). Migratory and eastern populations generally show very low song sharing, but a Nova Scotia population with notably high site fidelity by returning migratory males exhibits moderate song sharing levels similar to western populations (Hughes et al. 1998; Foote and Barber 2007). For the well-studied western populations with moderate song sharing, song matching is an important signal pattern (Figure 4). Song rate and switching rate may be more important in migratory eastern populations (Kramer and Lemon 1983; Kramer et al. 1985; Searcy et al. 2000; Anderson et al. 2008a). All populations appear to use soft song and wing-waving as close-distance aggressive intention signals (Anderson et al. 2008b).
For song-sharing populations, type matching is an aggressive signal that predicts subsequent approach and elicits a strong aggressive approach response in receivers (Nielsen and Vehrencamp 1995; Beecher et al. 2000a; Burt et al. 2001; Vehrencamp 2001). Singing a non-matching but shared song type in the vicinity of a known neighbor, called repertoire matching, is considered a directed challenge but is less threatening than a type match (Beecher et al. 1996). At any point during an interaction, a bird may de-escalate the encounter by singing an unshared song type, which cannot be matched by the rival (Beecher and Campbell 2005). Rules of escalation are summarized in Figure 5. Birds initially challenge a known male neighbor by singing a song they share with that individual. The rival may then escalate by singing this same song type, remain neutral but engaged by singing a different shared song, or de-
escalate by singing an unshared type. The first bird can then decide whether to continue type matching and persist with an aggressive approach, or switch to a different song type (Burt and Beecher 2008).

Some males share no song types with one or more of their neighbors, and thus they are unable to make use of these matching strategies to negotiate boundaries. Their only option is to use the songs in their repertoire that contain some similar elements to the rival's current song, especially similar introductory notes (Burt et al. 2002). Such partial matching appears to be less effective than whole song-type matching. In song-sharing populations, males that shared few or no songs with their neighbors had lower territory tenure than males that shared 2 or more song types (Beecher et al. 2000b; Wilson et al. 2000). Moreover, non-sharers had more frequent and more intense aggressive interactions with their neighbors compared to neighbors that shared several song types (Wilson and Vehrencamp 2001). Because song sparrows are age-restricted learners and cannot learn new songs after about six to nine months of age, non-sharers are generally individuals who were unable to acquire territories close to their tutors and dispersed farther away from their natal area (Nordby et al. 1999, 2002). Non-sharers may thus be less dominant birds, who must fight more intensely to gain a territory and suffer higher mortality as a result (Wilson et al. 2000). In eastern non-sharing populations, males will also sometimes reply to playback with a song type that is similar to the playback song (Anderson et al. 2005; Searcy et al. 2006). Birds that did give partial match replies approached the speaker more aggressively than those that did not match, indicating that even in these populations with little whole-song sharing, partial matching has some salience as a signal of aggressive intentions.

**Banded wren (Thryothorus pleurostictus)**

Males in this sedentary Neotropical species possess repertoires of 18–26 song types, 75% of which are typically
shared with any given neighbor (range: 50–90%) (Molles and Vehrencamp 1999). Males can switch rapidly among song types in an immediate variety mode, or they can deliver the same song type repeatedly for a variable length of time. Variable song features used by this species include switching rate, song-type diversity, song matching, overlapping, trill note consistency, and trill performance (trade-off between frequency bandwidth and trill note rate). Most song types possess a terminal trill; trill types vary in their performance score (Figure 6).

![Image of a banded wren](image)

**Figure 6: Banded wren.** Male working on the construction of his nest, a woven covered structure usually located in an ant acacia tree. Nine song types are illustrated here. Each song type has an introductory part consisting of various frequency sweeps and alternating notes, followed by a prominent and loud trill, and usually ending with a whip note. Trills vary in frequency bandwidth and note rate. (Song types after Trillo and Vehrencamp 2005. Photo courtesy of John Burt.)

Banded wrens use the same matching rules as described above for the song sparrow, using shared song types to initiate and continue an interaction, type matching for escalation, and unshared song types for de-escalation. But they
add another layer of complexity on their song system with variation in switching rate. Countersinging contests typically begin with rapid switching among a subset of shared song types, in which the rate of immediate song type matching varies depending on aggressive motivation. As males move closer to each other, one may begin to sing a song type repeatedly, and the opponent may keep switching, sing another song type repeatedly, or escalate by singing the same song type repeatedly (Figure 7). Males can de-escalate by singing unshared song types, or by overlapping as a defensive strategy (Vehrencamp et al. 2007). Thus banded wrens use shared and unshared song types, type matching, and repertoire matching in a way similar to the song sparrow, but they also can vary the rate of type matching, the delay of type matching, and the rate of switching among song types (Molles and Vehrencamp 2001; Molles 2006). Particularly when males are engaged in an extended bout of repeated type matching, they may compare each other’s trill performance and note consistency, as proposed by Logue and Forstmeier (2008). Males use song types with different structural features in different contexts. For example, they use songs with high-performance broadband trills during the dawn chorus and during countersinging interactions close to the shared boundary (Trillo and Vehrencamp 2005). Playback of songs with high values of these features often cause receivers to remain farther from the speaker, even when the playback is performed in the center of the territory (Illes et al. 2006; de Kort et al. 2009a; de Kort et al. 2009b). Overlapping playback also causes males to approach more cautiously, but overlapping by senders predicts their subsequent retreat from the speaker, and therefore has been interpreted as a defensive signal (Hall et al. 2006; Vehrencamp et al. 2007).

**Figure 7**: Countersinging interaction between two neighboring banded wren males (Male D in red, male J in green). Shared song types are shown in the white area, and unique song types for each male are shown in colored areas. Time sequence is depicted on the X axis. The birds began the interaction from the centers of their territories, each using a different subset of their shared song types (repertoire matching), but occasionally one bird type matched the other. At about song 45, male D sang a song type repetitively and the males approached their common boundary. At several points, male J delivered unshared song types. At song 65 they came even closer together and D sang another song type repeatedly, which J eventually started to type match repeatedly. The birds retreated without a fight in this instance. (After Molles 2006.)

**Summaries of playback studies on matching, overlapping, switching, and**
The signal value of different singing behaviors during male–male contests has been examined in a variety of songbird species using sound playback and observational techniques. Territorial birds are easy subjects for playback experiments because their movements are restricted to their territorial boundaries, and their individual histories (age, repertoire composition, song structure, mating success, reproductive behavior, etc.) are known if the study population has been individually color-marked. With the development of interactive playback methodologies in particular, researchers can finally examine the effects of matching, overlapping, and switching treatments on the responses of owners. In order to determine the signal value and function of some pattern of singing, three types of evidence should be examined (Vehrencamp et al. 2007): (1) the singing pattern should be given by senders in some contexts and not others, or by senders having a particular condition, status, age, or breeding stage. This information helps to establish whether the signal is given in aggressive contexts, or by more dominant birds, or by birds in better condition; (2) delivery of different singing patterns should predict or be associated with different subsequent behaviors by the sender. Signals that precede subsequent aggressive approach behavior would be classified as threat signals, whereas signals that precede retreat behavior would be classified as submissive or defensive signals; and (3) receivers of different singing patterns should respond differently. On-territory playback studies simulate the invasion of the territory by a potential rival male, and measuring the acoustic and spatial responses can help us interpret the likely function of the singing strategy. But such interpretations of responses are beset by the problem of the “peaked curve” of response strength versus signal intensity, in which low responses to a treatment are sometimes interpreted as a less threatening stimulus and sometimes interpreted as aversion to a highly threatening stimulus (see main text Figure 11.28). Details of the experimental protocol can also affect the results and interpretation—playbacks situated in the center of the territory generally elicit much stronger responses than playbacks situated on or outside the boundary. Treatments involving conventional signals may give different results from treatments involving index or handicap signals. Nevertheless, all three types of evidence should be combined to draw a consistent picture of the function of the singing pattern and its signal value. A detailed review of singing patterns by Searcy and Beecher (2009) using this approach provides a welcome assessment of the functions of song type matching, frequency matching, overlapping, low-amplitude song, song type switching, and vocal performance, and the interested reader should consult that source. In the sections below, we briefly summarize the results of playback experiments that investigated matching, overlapping, switching, and song performance, focusing in particular on playback studies that examined both the sender’s subsequent behavior after delivering certain signal variants and the receiver’s responses to these variants, i.e., points 2 and 3 above. Separate, differently designed experiments are usually required to assess these two points.

**SONG MATCHING** In songbird species with repertoires of learned song types or continuously varying song pitch, the birds in a neighborhood typically share at least some types with their neighbors. During countersinging interactions with a given neighbor, the particular song types shared with a that neighbor are used selectively, and matching rates typically increase above chance levels (Lemon 1968; Krebs et al. 1981; Schroeder and Wiley 1983; Falls 1985; Stoddard et al. 1992; Shackleton and Ratcliffe 1994; Beecher et al. 1996; Duguay and Ritchison 1998; Beecher et al. 2000a; Rogers 2004; Trillo and Vehrencamp 2005; Burt and Vehrencamp 2005; Rogers et al. 2006; van Dongen 2006). When a vocalizing individual hears another individual reply to its current song with a matching type, it can be relatively certain that the responder is attempting to grab its attention. Matching is an obvious way to acoustically point toward another individual in a network of individuals giving omnidirectional signals. Call or song matching has an affiliative function in some cases, such as coordinating nest visits between mated pairs or establishing contact in fission–fusion societies, but in male–male territorial interactions it typically serves an aggressive “challenge” function. In order for matching to evolve into a threat signal, senders must reliably follow a match with an increased display of aggression, such as approach, and receivers that are strong, dominant, and/or motivated must respond aggressively to impose a retaliation cost on potential buffers. If the threat is effective, truly weak or unmotivated receivers should retreat.

Table 2 summarizes playback studies of matching in species that have been examined from both the sender and receiver perspective. Experiments to assess sender perspective require broadcasting song types that the subject bird has in its own repertoire, to evaluate the relationship between its probability or rate of matching and its approach behavior. In all species shown in this table, if senders matched a playback stimulus, they were subsequently more
likely to approach the playback speaker, regardless of where the speaker was located. Deeper analysis of banded wren and song sparrow responses showed that matching birds nearly always approached, but that some birds approached without first matching. Experiments to assess the receiver perspective necessarily involve interactive playback experiments that immediately match or do not match the subject’s songs. Among studies using this method, some found the predicted stronger approach when subjects were matched compared to when they were not matched, but others found no difference between treatments. A key protocol feature that differentiates these studies with these two outcomes is the stimulus source and playback location: studies showing stronger approach when matched employed neighbor song stimuli from an appropriate boundary location, whereas studies showing no difference employed stranger song stimuli from the territory center. These latter studies found extremely close approach and high levels of agitation to both treatments, suggesting a ceiling effect. Given that aggressive birds sometimes approach without matching first, this undifferentiated strong response to playback in the center of the territory is not surprising. Taken together, these results suggest that matching is a conventional threat signal that operates primarily between neighbors who know each other’s repertoires and facilitates the usually respectful “dear enemy” relationship. Searcy and Beecher (2009) conclude that type matching is lower level aggressive signal that may predict an increasing level of escalation but not attack, what we have called a challenge signal.


<table>
<thead>
<tr>
<th>Species</th>
<th>Treatments or sender songs</th>
<th>Stim source</th>
<th>Pbk loc</th>
<th>Sender aggressive behavior</th>
<th>Receiver aggressive behavior</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great tit</td>
<td>TM vs. NM</td>
<td>S</td>
<td>C</td>
<td>TM&gt;NM</td>
<td>TM=NM</td>
<td>Krebs et al. 1981; McGregor et al. 1992</td>
</tr>
<tr>
<td>Black-capped chickadee</td>
<td>FM vs. NM</td>
<td>S</td>
<td>C</td>
<td>FM≥NM</td>
<td>FM=NM</td>
<td>Otter et al. 2002; Mennill and Ratcliffe 2004; Fitzsimmons et al. 2008</td>
</tr>
<tr>
<td>Eastern song sparrow</td>
<td>PM vs. NM</td>
<td>S</td>
<td>C</td>
<td>PM&gt;NM</td>
<td>PM=NM</td>
<td>Anderson et al. 2005, 2008a</td>
</tr>
<tr>
<td>NW song sparrow</td>
<td>TM vs. RM</td>
<td>N</td>
<td>B</td>
<td>TM&gt;RM</td>
<td>TM&gt;RM</td>
<td>Burt et al. 2001</td>
</tr>
<tr>
<td></td>
<td>RM vs. NM</td>
<td>N</td>
<td>B</td>
<td>RM&gt;NM</td>
<td>RM&gt;NM</td>
<td>Beecher and Campbell 2005</td>
</tr>
<tr>
<td>SW song sparrow</td>
<td>TM vs. NM</td>
<td>N/S</td>
<td>B</td>
<td>TM&gt;NM</td>
<td>TM&gt;NM</td>
<td>Nielsen and Vehrencamp 1995; Vehrencamp 2001</td>
</tr>
<tr>
<td>Banded wren</td>
<td>TM vs. RM</td>
<td>N</td>
<td>B</td>
<td>TM&gt;RM</td>
<td>TN&gt;RM</td>
<td>Molles and Vehrencamp 2001; Vehrencamp et al. 2007; de Kort et al. 2009b</td>
</tr>
<tr>
<td></td>
<td>RM vs. NM</td>
<td>N</td>
<td>B</td>
<td>RM&gt;NM</td>
<td>RM&gt;NM</td>
<td></td>
</tr>
</tbody>
</table>

**SONG-TYPE SWITCHING** Even birds with small song-type repertoires can vary the rate of switching between song types. Switching rate is measured as the number of switches per opportunity to switch (Searcy et al. 2000). Earlier
adaptive hypotheses for the function of switching among song types included reducing habituation (Falls and Dagincourt 1982), deceiving intruders about the number of birds present (Krebs 1977), and reducing fatigue of the vocal muscles (Lambrechts and Dhondt 1988). These hypotheses have since been thoroughly tested and found to be insignificant (Yasukawa 1981; Dawson and Jenkins 1983; Yasukawa and Searcy 1985; Haftorn 1995; Brumm et al. 2009). Instead, evidence is accumulating that various switching strategies function as agonistic signals between countersinging males. Switching in bout singers can be synchronized with a countersinging rival’s switch to produce an acoustic pointing signal (Kramer and Lemon 1983; Horn and Falls 1991). The rate of change to new song types can vary continuously to form a graded signal (Kramer et al. 1985; Horn and Falls 1991; Molles and Vehrencamp 1999). Switching rate appears to be a truly conventional signal, in the sense that either a low or high switching rate can represent the more threatening signal variant depending on the species. Species that normally deliver song types in bouts (eventual variety) typically increase their switching rate in more aggressive contexts, and species that normally switch after every song (immediate variety) typically decrease their switching rate in more aggressive contexts (Vehrencamp 2000; Collins 2004). The red-winged blackbird is an exception; it is an eventual variety singer that reduces switching rate when confronted with a live male rival.

Few species have been the subject of both sender and receiver perspective playback experiments for switching patterns (Table 3). Song sparrows are bout singers that increase their switching rate in more aggressive contexts, and respond more aggressively to higher switching-rate playback, but there is no evidence that a high switching rate predicts subsequent approach. Ortolan buntings are also bout singers that increase their switching rate in more aggressive contexts; but young males respond less aggressively to high switching rate playback as if they are intimidated, whereas older birds respond aggressively to both switching treatments. As mentioned earlier, banded wrens are immediate variety singers that reduce their switching rate in more aggressive contexts. They respond more aggressively to low switching-rate playback. Low switching predicts aggressive behavior in some situations (males with fledglings). In countersinging interactions, repetitive singing occurs in the later stages of escalation when birds are close, but seems to indicate unwillingness to negotiate further with song switching and matching; it is also a victory signal by the winner of a close encounter.


<table>
<thead>
<tr>
<th>Species</th>
<th>Singing mode</th>
<th>Stim source</th>
<th>Pbk loc</th>
<th>Sender aggressive behavior</th>
<th>Receiver aggressive behavior</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ortolan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>H = L (old)</td>
<td>a</td>
</tr>
</tbody>
</table>
SONG OVERLAPPING A repertoire of song types is not required for countersinging birds to use overlapping as an agonistic signal. In the first studies of song overlapping using European robins, European blackbirds, and nightingales, overlapping seemed to be an aggressive signal. These species appeared to overlap in more aggressive contexts. Robins and nightingales responded more aggressively to being overlapped, but blackbirds avoided song posts from which overlapping songs were played. More recent studies on other species have lead to a rather mixed picture about overlapping. While overlapping was initially believed to be an aggressive signal in great tits, intensive escalating playback on this species found that birds were more likely to overlap from a distance than when close. Studies on the black-capped chickadee ultimately found no particular evidence for overlapping as an aggressive or defensive strategy, but it may serve a pointing function and predict the next stage of escalation: frequency matching. Interactive studies on the corn bunting, yellowhammer, and banded wren all found an aversive response to being overlapped. This result was interpreted as overlapping being an intimidating aggressive signal. However, a sender-perspective study in the banded wren showed that overlapping by senders predicted their subsequent retreat, suggesting that overlapping instead might be a defensive threat signal, like a jamming signal in electric fish.

Searcy and Beecher (2009) concluded that there is little evidence of overlapping as an aggressive signal, largely because most positive results of increased overlapping by senders in more aggressive contexts can be accounted for by an increase in song rate and chance overlapping. Naguib and Mennill (2010) countered this view by arguing that despite the tendency for most birds to avoid overlapping or being overlapped, rare overlapping events could still be meaningful signals; moreover, many studies have found differential responses to different overlapping treatments. Eavesdropping experiments have been particularly informative. In these experiments, two speakers simulate overlapping and overlapped birds in earshot of a territorial subject. The subject often responds by showing more aggressive behavior toward the overlapping stimulus speaker. Searcy and Beecher (2011) countered again that most responses to being overlapped are interruption or shortening of the current song, which is a natural response by a vocalizing animal and not necessarily indicative of aggression. It is likely that overlapping may serve different functions in different species, aggressive in some cases, defensive in others, and not a signal in yet others. More studies are needed that examine overlapping and approach from both the sender and receiver perspective, making sure to correct for chance levels of overlapping as recommended by Searcy and Beecher (2009).

**Table 4: Summary of overlapping experiments.** Strength of sender and receiver responses to overlapping (O) and

<table>
<thead>
<tr>
<th>Species</th>
<th>Strength</th>
<th>E</th>
<th>S</th>
<th>C</th>
<th>(H &gt; L)</th>
<th>H &lt; L (young)</th>
<th>Treatments varied within-song and between-song switching rates in a factorial design.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buntinga</td>
<td></td>
<td>E</td>
<td>S</td>
<td>C</td>
<td>(H &gt; L)</td>
<td>H &lt; L (young)</td>
<td>Osiejuk et al. 2007</td>
</tr>
<tr>
<td>NW song sparrow</td>
<td></td>
<td>E</td>
<td>S</td>
<td>B</td>
<td>(H &gt; L)</td>
<td></td>
<td>Stoddard et al. 1988</td>
</tr>
<tr>
<td>SW song sparrow</td>
<td></td>
<td>E</td>
<td>N</td>
<td>B</td>
<td>H = L</td>
<td>(H &gt; L)</td>
<td>Nielsen and Vehrencamp 1995</td>
</tr>
<tr>
<td>Eastern song sparrow</td>
<td></td>
<td>E</td>
<td>S</td>
<td>C</td>
<td>H = L</td>
<td>(H &gt; L)</td>
<td>Kramer and Lemon 1983; Kramer et al. 1985; Searcy et al. 2000</td>
</tr>
<tr>
<td>Redwing blackbird</td>
<td></td>
<td>E</td>
<td>L</td>
<td>C</td>
<td>L &gt; H</td>
<td>(H &gt; L)</td>
<td>Yasukawa 1981; Searcy and Yasukawa 1990</td>
</tr>
<tr>
<td>Banded wren</td>
<td></td>
<td>I</td>
<td>N</td>
<td>B</td>
<td>L &gt; H</td>
<td>(H &gt; L)</td>
<td>Molles and Vehrencamp 1999; Molles 2006; Vehrencamp et al. 2007</td>
</tr>
<tr>
<td>Tropical mockingbird</td>
<td></td>
<td>I</td>
<td>S</td>
<td>B</td>
<td>H &gt; L</td>
<td></td>
<td>Botero and Vehrencamp 2007</td>
</tr>
</tbody>
</table>

a Receiver responses measured for older (2+ years) and younger (1 year) territory owners.
b Treatments varied within-song and between-song switching rates in a factorial design.
alternating (A) songs.

<table>
<thead>
<tr>
<th>Species</th>
<th>Stim source</th>
<th>PbK loc</th>
<th>Sender aggressive behavior</th>
<th>Receiver aggressive behavior</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nightingale</td>
<td>S</td>
<td>C</td>
<td>O &gt; A</td>
<td>O &gt; A</td>
<td>Hultsch and Todt 1982; Naguib 1999</td>
</tr>
<tr>
<td>European blackbird</td>
<td>S</td>
<td>C</td>
<td>O &gt; A</td>
<td>O &lt; A</td>
<td>Todt 1981; Wolffgramm and Todt 1982</td>
</tr>
<tr>
<td>Black-capped chickadee</td>
<td>S</td>
<td>C</td>
<td>O = A</td>
<td>O = A</td>
<td>Mennill and Ratcliffe 2004; Fitzsimmons et al. 2008; Foote et al. 2008</td>
</tr>
<tr>
<td>Great tit</td>
<td>S</td>
<td>C</td>
<td>A &gt; O</td>
<td>O = A</td>
<td>Dabelsteen et al. 1996; Langemann et al. 2000</td>
</tr>
<tr>
<td>Corn bunting</td>
<td>S</td>
<td>C</td>
<td>O &lt; A</td>
<td></td>
<td>Osiejuk et al. 2007b</td>
</tr>
<tr>
<td>Yellowhammer</td>
<td>S</td>
<td>C</td>
<td>O &lt; A</td>
<td></td>
<td>Osiejuk et al. 2004</td>
</tr>
<tr>
<td>Banded wren</td>
<td>S</td>
<td>C</td>
<td>A &gt; O</td>
<td>O &lt; A</td>
<td>Hall et al. 2006; Vehrencamp et al. 2007</td>
</tr>
</tbody>
</table>

**VOCAL PERFORMANCE**

Vocal performance includes various measures of note fine structure, such as note shape, repeated-note consistency, repeated-note drift, maximum or minimum pitch, and trill performance (bandwidth and note rate tradeoff) (see review by Podos et al. 2009). These song features usually vary within males on a slower time scale than matching, switching, and overlapping, and often differ among males as a function of age and condition (Janicke et al. 2008; Araya-Ajoy et al. 2009; Ballentine 2009; de Kort et al. 2009a). Several studies have demonstrated female preferences for high-performance songs (Vallet et al. 1998; Forstmeier et al. 2002; Ballentine et al. 2004; Cardoso et al. 2007; Holvek et al. 2008; Byers et al. 2010; Cramer et al. 2011). Thus song performance is likely to be an index signal of male age, experience, dominance, or general quality (Searcy and Beecher 2009). Nevertheless, males can choose to deliver higher performance songs in aggressive contexts, and in some cases may increase the vocal performance of a given song type in aggressive contexts (Trillo and Vehrencamp 2005; Kunc et al. 2006; Price et al. 2006; DuBois et al. 2009). Only a few playback studies have investigated the role of these features in male–male interactions. Most results indicate significant avoidance of high performance song playback, even by territorial males stimulated by playback in the center of their territory. In banded wrens, three performance level treatments were given and responses showed a peaked response for the middle level. In the nightingale, less successful males were intimated while more successful males behaved more aggressively toward the high-performance stimulus. Swamp sparrows responded more aggressively towards the higher performance stimulus.

**Table 5: Summary of song performance experiments.** Strength of aggressive response to high (H) versus low (L) performance songs. Song feature associated with older or more successful males. All playback experiments conducted in the center of the territory using stranger or modified song stimuli.
In conclusion, the variable song and singing features of bird song can signal a wide range of different types of information to rival males negotiating territorial boundaries, including: motivation to approach, attack, defend, and retreat; fighting ability, age, and experience; dominance and subordinance; current condition; and pairing status. These variable parameters can be considered “signals,” and can often be classified by their honesty-guaranteeing cost into handicap, index, conventional, and proximity signals. Species differ in their mix of variable parameters, and functions for the same signaling parameter may differ between species. Despite the tremendous progress in decoding the song systems of numerous species, much fieldwork remains to be done to verify the true function of some song signals.

**Literature cited**


Beecher, M.D. and S.E. Campbell. 2005. The role of unshared songs in singing interactions between neighbouring


Morton, E.S. and K. Young. 1986. A previously undescribed method of song matching in a species with a single song
type, the Kentucky warbler (*Oporornis formosus*). *Ethology* 73: 334–342.


Trillo, P.A. and S.L. Vehrencamp. 2005. Song types and their structural features are associated with specific contexts in


Web Topic 12.1
Quantitative Genetic Models of Sexual Selection

Introduction

Models of sexual selection focus primarily on intersexual selection and assume traditional sex roles. They aim to understand the processes by which male sexually selected traits and female preferences for those traits coevolve. The male traits are therefore cues and signals for mate attraction and courtship. Female mating preferences are behavioral, structural, or physiological traits that bias conspecific mating or fertilization success toward males with certain trait values. The myriad models make different assumptions about the nature of the male and female traits, the costs and benefits of these traits with respect to different components of fitness, and the genetic underpinnings of the coevolutionary process. Some models assume there is direct selection on female preference for certain male traits that improve female reproductive success or survival, while other models explore indirect selection on female preference caused by a genetic correlation with another trait under selection. One hypothesis for intersexual selection, the sensory bias model, was initially introduced in verbal form. In order to compare and test this model in relation to the other models (e.g., Fisherian, good genes, direct benefits, and sexual conflict models), Fuller et al. (2005) parameterized this hypothesis in the quantitative genetics framework. In the process, they clearly laid out the key similarities and differences among all five types of models. This Web Topics unit explains the multivariate quantitative genetic approach to sexual selection and summarizes their article.

Multivariate quantitative genetics framework

Quantitative traits are phenotypes determined by a large number of genes with small additive effects. Evolutionary change in the mean population value of such a trait therefore occurs gradually as selection acts on genetic mutations with small overall effects on the trait. As explained in Chapter 9 and Web Topic 9.1, the response to selection, $\Delta \bar{z}$ (measured as the change in the mean phenotypic value of a trait between the parental and the offspring generation) is equal to the product of the trait heritability, $h^2$ (measured as the ratio of the trait's additive genetic variance to its phenotypic variance in the population) and the selection differential on the trait in the parental generation, $S$ (measured as the difference between the mean value of the trait for those that breed and the overall average in the parental population):

$$\Delta \bar{z} = h^2 S$$

When there are multiple traits evolving in tandem, the equation must be expanded to take into account the genetic linkages between the traits. This is accomplished by combining the additive genetic covariances ($C$) and the additive variances ($V$) into a G-matrix. The selection differentials on the traits are combined with the overall phenotypic variation in the traits to produce a selection gradient for each trait, called $\beta$. Adding in the vector of mutational effects, $u$, the shorthand matrix formula for this expression becomes:

$$\Delta \bar{z} = G\beta + u$$

To investigate sexual selection, we need to consider the simultaneous coevolution of at least three phenotypic traits: a male display trait ($t$), a female preference for the male display trait ($p$), and a viability or residual fitness trait ($v$). The residual viability applies to both sexes and includes the fitness components of fecundity and survivability left over after excluding the sexually selected fitness components of mate number and mate quality. Finally, the selection gradient can be separated into two additive components: the part due to natural selection $\beta_N$, and the part due to sexual selection $\beta_S$. The full quantitative genetics equation for the change in each trait’s mean value in one generation is given by:
Below, we show the values these components must have to initiate each sexual selection process, as well as the values that are likely to exist at the equilibrium point. All of the tables use the following notations. Traits that have a significant amount of additive genetic variation will be indicated with $V > 0$; likewise, if they have a significant genetic covariance with another trait, this will be indicated with $C > 0$. When a trait is undergoing significant positive or negative selection, the relevant selection gradient $\beta$ will be shown as either greater or less than zero, respectively. When the gradient is zero, either a selection peak or trough has been reached and there is no net natural or sexual selection; $\beta \equiv 0$ indicates that there must be no selection on the trait. All models require some heritable genetic variation in female preference behavior ($V_p > 0$), and at equilibrium all models predict a balance between positive sexual selection and negative natural selection on the male trait. These conditions will be indicated in the figures by green text; variable states that are essential for the particular selection model are shown in red. The response and mutation vectors will not be shown in the summaries below, as they are the same in each case.

**Fisherman runaway model**

For the Fisherian model, required initial conditions are positive additive genetic variance for both the male trait and female preference. The non-random mating that results from this female preference for the male trait leads to a genetic covariance between these two traits ($C_{tp} > 0$). This happens because choosier females produce offspring that inherit the alleles for stronger female preference and the alleles for greater male trait values, resulting in gametic disequilibrium and coupling of the two traits. A change in female preference $p$ may thus bring about a change in the male trait $t$. If the covariance term is large, the female preference and male trait will rapidly increase in the runaway process, but if ($C_{tp} < V_t$), the population will return to an equilibrium point where selection on $p$ and $t$ are balanced (see main text Figure 12.6). At the equilibrium, there is no natural or sexual selection pressure on the female preference trait, and positive sexual selection for the male trait is balanced by negative natural selection on the male. These initial and equilibrium conditions are shown in Figure 1.
Figure 1. Initial and equilibrium conditions for the Fisherian model. Critical parameters required to initiate an increase in female preference are indicated in colored text. (After Fuller et al. 2005.)

**Good genes model**

The good genes model requires a positive covariance between the male display trait and genetic variation for residual fitness ($C_{tv} > 0$). This essentially means that the trait is condition-dependent, such that males with greater values of the trait have higher survival. Given that female preference also has some positive genetic variance, a positive covariance between $p$ and $v$ can be generated ($C_{pv} > 0$). Even though there is no direct selection on $p$, the preference trait can be indirectly selected for because of this covariance driven by direct selection for high viability $v$. If the male display trait is also genetically variable ($V_t > 0$), covariance between $p$ and $t$ is generated ($C_{pt} > 0$), which is not essential but can accelerate the good genes process. **Figure 2** shows initial and equilibrium stages for the good genes model. In this example, the initial genetic variation in $v$ is generated by biased mutation ($u_v$) but other mechanisms could result in a positive $V_v$. At the equilibrium, there is direct natural selection against the female mating preference because of costs female incur to search for high quality males, but this is counterbalanced by indirect viability gains through the preference–viability covariance term.
Direct benefits model

The direct fitness model is the most straightforward of the five models because no genetic correlations are required. By choosing males that provide some type of benefit that improves their fecundity or survival, or the survival of their offspring, females gain directly ($\beta_{Sp} > 0$). At the equilibrium, females will experience some natural selection cost for being choosy (Figure 3). Genetic variance for the male trait is not required, but the female preference is likely to select for such variance, and then the male trait will also undergo positive sexual selection and negative natural selection, as in the good genes model.
In the sensory bias model, females initially evolve a naturally selected preference for some type of environmental feature, such as a food characteristic. This preference is indicated in the first stage of the model by a positive $\beta_{Np}$. In the second stage, males evolve a trait that mimics this feature, and if there is genetic heritability for the male trait ($V_t > 0$), it will respond to sexual selection via the female preference for it ($\beta_{St} > 0$). In one possible type of equilibrium scenario, female preference settles at a local naturally selected stable point and is no longer under selection. As in all of the models, the male display trait reaches a stable balance between sexually selected benefits and naturally selected costs ($\beta_{St} = -\beta_{Nt}$) (Figure 4).

**Sexual conflict model**

Sexual conflict models rely on the prior existence of a female mating preference for a male trait via one of the above models. The first stage thus reflects this initial sexually-selected benefit to females, which creates an opportunity for males to take advantage of this preference. For the scenario shown in Figure 5, a female preference is assumed to
have arisen by way of the direct benefits model ($\beta_{Sp} > 0$). Males then evolve traits or mating tactics that benefit them while reducing the fitness of females. The example discussed in the main text is the transfer of manipulative chemicals in the semen that causes females to mate again more rapidly, or lay more eggs than is optimal for them. This sets up a cycle of conflict in which females try to resist male manipulation, while males counter with more forceful tactics. In one equilibrium scenario shown here, both sexes experience positive sexually selected benefits and negative natural selection via the costs incurred by the conflict.

![Figure 5. Initial and equilibrium conditions for the sexual conflict model.](image)

By placing all of the sexual selection models within this single quantitative genetics framework, strategies for distinguishing among them become obvious (Mead & Arnold 2004; Kokko et al. 2006). Clearly, looking for heritability of female mating preferences will not help distinguish the models, as they all require this component. Similarly, looking for evidence of a natural selection cost for males with larger trait values also will not be helpful, as they all predict this relationship. The sensory bias model can be distinguished from the others by looking for evidence of a natural selection benefit for female preference independent of the identity and attributes of the male, since all other models predict that the benefit to females arises from the benefits to selecting particular males (Fuller et al. 2005). The Fisherian and good genes models have many components in common, and in fact the Fisherian conditions will be met whenever the good genes model conditions are met (Kokko et al. 2002, 2003). In particular, looking for a covariance between male trait and female preference will not distinguish these two models. However, if viability benefits to males and possibly females can be demonstrated, a case can be made for the good genes model (Kokko 2001). The sexual conflict model should be distinguishable from the others by finding positive sexual selection and negative natural selection gradients for both sexes.

**Literature cited**


Web Topic 12.2
More Examples and Multimedia Clips of Courtship Sequences

Introduction

Nearly all animals engage in some form of courtship before mating, and when they do, they invariably use signals to mediate the process. As discussed in Chapter 1, signal diversity stems from a combination of factors: use of different media in which the signals are generated, transmitted, and received and thus different physical constraints; different phylogenetic histories and thus precursors; different types of information being exchanged; and different pressures and criteria for honesty guarantees. Given broad taxonomic variation in all of these factors, courtship represents the most diverse category of animal communication signals. In this module, we provide links to videos that illustrate both the scope and complexity of animal courtship displays. One cost of the emphasis on video recordings is an under-representation of courtship that is mediated by olfaction. However, this modality is relatively well covered in the text and Web Topic 6.3, and unless senders perform some overt behavior to dispense pheromones, there would be little to see in a video anyway. In addition to the examples in this module, additional courtship signals are included in Web Topics 10.3 and 13.2. Because it is instructive to see how different species in the same taxon have exploited a common ancestral repertoire and faced similar constraints, several taxa receive fairly detailed treatment below. We have tried here to select taxa for this module that have not been covered in other Web Topics.

Spiders

For a taxon of largely solitary species, spiders can have surprisingly elaborate courtship signals. In part, this may be due to the risk that females may attack and eat approaching males. Courtship is initiated when males are attracted to females that have released airborne pheromones or marked pheromones on their webs. Males then approach warily while producing seismic stridulations or drumming signals that are transmitted through the ground or along web threads. Making physical contact is extremely tricky for male spiders and usually involves repeated visual and seismic displays at a safe distance, pheromone release, and eventually quick leg touches. The final stages of courtship are even trickier for males: unlike insects where mating involves contact of the very tips of their abdomens, male spiders must gain access to a central region of the female spider’s ventral side. The genital openings of both spider sexes are located on the anterior ventral side of the rear body segment (called the opisthosoma). If spiders had to oppose their genital openings to copulate, males would be extremely vulnerable. However, before courting, males collect sperm packages from their genital pores into special storage organs on their pedipalps (the pair of small appendages protruding forward from the head). To mate, males then position themselves so that they can insert species-specific tubular structures on their pedipalps into the female’s genital pores. This gives males a bit of flexibility, and spider species differ in the relative orientations of the two sexes during copulation. Still, the risks to males remain high and in many species, females kill and eat the male after copulation if they can. Below, we examine several types of courtship in spiders:

Jumping spiders (Salticidae): This is one of the largest and most diverse families of spiders. They do not build webs but instead stalk invertebrate prey using their excellent vision and highly agile mobility. Males in particular tend to be brightly colored, with different species having quite distinct color patterning (see Figure 1.2, Chapter 1). Males court females with highly stereotyped movements of the pedipalps and first pair of legs accompanied by seismically transmitted stridulation and drumming sounds:

- **Studying Salticid courtship**: This clip from the program Nature interviews Dr. Damien Elias, a specialist in courtship displays of this group. It provides a good introduction to the complexity of these courtship signals and how one might go about studying them. ([http://www.youtube.com/watch?v=sMa9FdNM5jo](http://www.youtube.com/watch?v=sMa9FdNM5jo))
- **Jumping spider** (*Habronattus coecatus*): Close-up video of courtship in this species of jumping spider: [http://www.youtube.com/watch?v=QuWMmAgnpCI](http://www.youtube.com/watch?v=QuWMmAgnpCI)
- **Jumping spider** (*Habronattus viridipes*): Initial close-ups of female followed by courtship by male:
• Zebra spider (*Salticus scenicus*): Both male and female are wary of each other in this clip despite male displays. (http://www.arkive.org/zebra-spider/salticus-scenicus/video-09.html)

• Peacock spider (*Maratus sp.*): Male courtship displays of spectacularly colored salticid spider from Australia. (http://www.youtube.com/watch?v=ppP03ERHbUI&feature=related)

Other spiders:


• Garden spider (*Araneus diadematus*): Intruding males use specific vibrations of web threads to persuade a resident female that they are a courting conspecific. Females are still fairly aggressive. A later stage in the clip shows delicate touching with legs as in raft spider before copulation. (http://www.arkive.org/garden-spider/araneus-diadematus/video-09a.html)

• Black-widow spider (*Latrodectus mactans*): Small light male very gingerly approaches large dark female with leg touching and eventually mates with female. (http://www.arkive.org/black-widow-spider/latrodectus-mactans/video-09a.html)

• Red-rumped tarantula (*Brachypelma vagans*): Male warily approaches female’s burrow and is initially chased away; he returns and performs stereotyped drumming of his legs on the ground at her burrow entrance. She finally emerges and accepts mating. (http://www.arkive.org/mexican-redrump-tarantula/brachypelma-vagans/video-00.html)

• Nursery-web spider (*Pisaura mirabilis*): In this species, a courting male wraps a freshly killed prey such as a fly in a silk package and presents it to the female as part of the courtship sequence. Although not shown here, these females will often eat the male after mating. (http://www.arkive.org/nursery-web-spider/pisaura-mirabilis/video-09b.html)

Other invertebrates

• Leopard slug (*Limax maximus*): Individuals find each other using olfactory signals in slime trails. One then leads the other up a tree using a comination of tactile and olfactory signals where they jointly hang from a branch by a thick mucous thread and twine around each other. Since these are hermaphrodites, each extrudes its giant penis and these also twine into a spiral with a bell-shaped inflation at the tip. Sperm are transferred and the slugs either drop or break the mucous thread in this amazing sequence. (http://www.arkive.org/leopard-slug/limax-maximus/video-00.html)

• Emperor scorpion (*Pandinus imperator*): Like spiders, both sexes of scorpions have genital pores on the ventral side of their bodies, in this case between the bases of the rear legs. Unlike spiders, males do not load sperm into their pedipalps, which in scorpions form the large anterior “claws.” Instead, a male and female will face each other and, grabbing each other’s pedipalps, move back and forth as courtship proceeds. Eventually, the male will deposit a threadlike white spermatophore on the ground which he then pulls the female over so that it touches her genital pore. The spermatophore has a special trigger on its tip that ejects the sperm into the female’s pore given the right pressure and orientation of the female’s body. (http://www.arkive.org/emperor-scorpion/pandinus-imperator/video-09a.html)

• Fruit fly (*Drosophila melanogaster*): This well-studied species has a fairly elaborate courtship sequence involving pheromones, production of male near-field sounds, and tactile stimuli. An annotated overview and sound production are shown in these two clips:
  
  * Annotated example of full courtship sequence: http://www.youtube.com/watch?v=SVV-Oo1QA8M
  * Courtship song: http://www.youtube.com/watch?v=Dmge39zdJTA

Stag beetle
Lucanus cervus: Both sexes of stag beetles secrete pheromones from glands on their front legs. Males attracted to females expose the female to their own pheromone and circle the female with their large mandibles expanded. Males are very competitive and multiple males attracted to the same female will fight vigorously for access to a mating. Males will even fight over dead females. This clip shows a male circling a female and then mounting her for copulation. (http://www.arkive.org/stag-beetle/lucanus-cervus/video-09b.html)

- **Seed Beetles** (*Callosobruchus chinensis*): This clip is from the research described in Figure 12.11, and shows the female (left) resisting the male’s attempts to mount by kicking with her rear legs. (http://www.cell.com/current-biology/supplemental/S0960-9822%2809%2901703-5)

**Fish**

Fish exhibit a wide diversity of courtship signals. Pelagic and some reef fishes perform group spawning in which multiple members of both sexes simultaneously eject large numbers of gametes into the water column, usually around sunset. While one might assume that these species do not need courtship displays, they often show a variety of stereotyped movements, color changes, and chasing prior to a group spawn (e.g., tuna). Another common mating system in fish involves male defense of a patch of substrate and even the building of a nest. Males fertilize the eggs and guard them until they hatch (sticklebacks, damselfish). In nocturnal forms, such as toadfishes, males produce loud “songs” to attract females. Male pipefish and sea horses collect females’ eggs in a pouch on their body rather than in a substrate-bound nest. Other species such as guppies have internal fertilization and the typical competition between males to be selected by a receptive female has led to male courtship displays. In cichlids, pairs may form longer-term pair bonds and share in parental care; some species even have helpers at the nest. Finally, some species (again cichlids) exhibit lek behavior in which clusters of males display competitively to visiting females. In short, courtship behavior is widespread in fishes and as diverse as their mating systems.

Fish have a wide variety of modalities to incorporate into their courtship signals. Visual components include rapidly changed color patterns and movements, and many species add sounds to their courtship displays. Olfactory signals are likely ubiquitous in fish, both via secretions in their skins and via the passing of urine. One modality that is limited to fish is electrical signaling; although largely limited to freshwater species, the taxa that use electrical signals are highly diverse taxonomically and in signal patterning. Some examples:

- **Brown surgeonfish** (*Acanthurus nigrofuscus*): These common reef dwellers have aggregated into a school near the surface in late afternoon. Subsets of the school rush upwards, release gametes in a group spawn, and dive back into the school. (http://www.arkive.org/brown-surgeonfish/acanthurus-nigrofuscus/video-11a.html)

- **Three-spined stickleback** (*Gasterosteus aculeatus*): This is a well-studied species in which males build and guard a nest. The clip begins with a male preparing the nest. He then encounters a very gravid female and performs displays including leading her to the nest, which she then enters to lay the eggs. The male then enters the nest to fertilize the eggs. The details of this display are outlined in Figure 12.28. (http://www.arkive.org/three-spined-stickleback/gasterosteus-aculeatus-aculeatus/video-ac09b.html)

- **Damselfish** (*Pomacentridae*): Males of most species of damselfishes defend nesting territories on coral reefs. They are extremely aggressive and try to chase much larger fish away from their territories. Males can change body colors rapidly and markedly. In the bicolor damselfish (*Stegastes partitus*), males leave their territories shortly after dawn, exchange their usual territorial defense coloration for a different courting pattern, and patrol for possibly receptive females. When they find one, they perform a series of stereotyped dipping and tilting motions accompanied by chirping sounds in attempts to lead her back to their nest. Once at the nest, a female indicates her receptivity by changing her color pattern and the male points at the nest and quivers. The two then take turns laying eggs and fertilizing them until the female leaves. Other species of damselfish show similar combinations of movement, color pattern, and sound production during courtship. See Deloach and Humann (1999) for more details. Examples:
  - **Bicolor damselfish** (*Stegastes partitus*): This clip shows adult males in their usual territorial coloration of 2/3
of the anterior dark coloration and the posterior third light. Then it shifts to males defending artificial nest sites (plastic tubes) and one male that is courting a nearby female with the courtship color pattern of the anterior and posterior thirds dark, and a light middle third. This sequence is followed by a short shot of the nest and eggs of a dusky damselfish (*Stegastes adustus*). (Video clip courtesy of Dr. Kenneth Clifton.)

- **White-spotted damselfish** (*Dascyllus albisella*): Pair spawning in Hawaii: ([http://www.youtube.com/watch?v=vjTj33DHDxE&feature=related](http://www.youtube.com/watch?v=vjTj33DHDxE&feature=related))
- **Sergeant major** (*Abudeful saxatilis*): Territorial males usually exhibit a color pattern of black stripes against a yellow and white background. Here a territorial male tries to keep a school of wrasses from eating its egg clutch. ([http://www.youtube.com/watch?v=mNY7b6rBsg](http://www.youtube.com/watch?v=mNY7b6rBsg))
- **Sergeant major** (*Abudeful bengalensis*): In contrast to the territorial coloration above, courting male sergeant majors shift to a much darker background between the stripes especially in the head region. To see an example, use the following link: [http://www.fischverhalten.de/](http://www.fischverhalten.de/). Once at the site, select the preferred language, then click “Spawning” in the topic list on the left. When the next page opens, go to nearly the bottom of the page, where there are thumbnails for six short video vignettes (two rows with three thumbnails each). The one in the lower right, with the label beneath it (in English) "Damselfish (*Abudeful bengalensis*)" shows a female Sergeant major (smaller and grayer coloration) adding eggs to the clutch of a territorial male (larger and with much darker head coloration). When finished, the female moves to one side and the male then inseminates the new eggs.

- **Garibaldi** (*Hypsypops rubicundus*): Males of this spectacular damselfish from the kelp forests of the US West coast also defend nest sites. Their bright orange color makes male advertisement of its nesting site very conspicuous to distant females. This clip shows some courtship and other behaviors. ([http://www.youtube.com/watch?v=k_JVQGBAdNY](http://www.youtube.com/watch?v=k_JVQGBAdNY))

- **Blue-head wrasse** (*Thalassoma bifasciatum*): This is one of the most common wrasses on Caribbean coral reefs where they are one of the major predators on the eggs in male damselfish nests. During spawning, eggs and sperm are broadcast into down-currents that flow off of the reef and into the plankton. Initial phase fish of both sexes are small and slim with yellow bodies and a dark central stripe. Large females are capable of changing sex into bright blue, black, and white terminal males. Terminal males attempt to defend optimal spawning sites where they chase off other males and advertise themselves by performing a series of short vertical dashes (signal jumps). When a female approaches, the male starts swimming in tight circles and may perform some spiraling dashes upwards. Eventually, both parties dash up into the down-current, release their gametes in a burst, and dash back down to the reef. On small reefs, terminal males can often control optimal spawning sites. As reef size increases, this becomes increasingly difficult and schools of initial phase males either occupy spawning sites up current from a terminal male’s site, or by sheer force of numbers, displace the terminal male. The result is a group spawn of many initial phase males and the female. Several of these activities are visible in this clip provided by Dr. Ken Clifton.

- **Spotted handfish** (*Brachionichthys hirsutus*): Mobility is limited for both sexes in handfish. In this species, males court females by spreading their fins and tails (as shown here), fertilization is external, and females guard the eggs for the 7–8 week incubation period. ([http://www.arkive.org/spotted-handfish/brachionichthys-hirsutus/video09a.html](http://www.arkive.org/spotted-handfish/brachionichthys-hirsutus/video09a.html))

- **Multiple fish species**: This site has collected a number of stills, text, and video on the behavior of different fish taxa. Most are freshwater, but a few marine clips are included. To use the following site, select the preferred language, then click “Courtship display” in the panel on the left. ([http://www.fischverhalten.de/](http://www.fischverhalten.de/))

- **Multiple freshwater species**: A similar site with no English option: [http://www.fischreisen.de/Buntbarsche.html](http://www.fischreisen.de/Buntbarsche.html)

- **African electric fish** (*Brienomyrus brachystius*): This well annotated video shows a male and female courting and then spawning. The video was taken in the dark using infrared light (in which the fish cannot see), and the electrical discharge patterns produced by each individual are shown graphically and after conversion from electrical to audible sounds as a sound track. ([http://www.nbb.cornell.edu/neurobio/Hopkins/media/CourtshipClip/2006_07_06_2-4_courtship.swf](http://www.nbb.cornell.edu/neurobio/Hopkins/media/CourtshipClip/2006_07_06_2-4_courtship.swf))
Neotropical electric fish (*Sternopygus macrurus*): A recording of courtship electrical signals by a male. The pattern is shown as a spectrogram of the discharge frequencies and converted into sounds so that we can monitor the pattern. ([http://www.nbb.cornell.edu/neurobio/Hopkins/sternopygus/sternopygus_singing.htm](http://www.nbb.cornell.edu/neurobio/Hopkins/sternopygus/sternopygus_singing.htm))

Amphibians and reptiles

- **Great crested newt** (*Triturus cristatus*): This clip shows a male with raised crest and tail performing highly conspicuous motion displays to a female. As noted in Figure 12.33, pheromones and tactile signals also play major roles in newt courtship. After a bout of displaying, the male moves ahead of the female and deposits a white spermatophore (circled in the clip); the female then follows and picks up the spermatophore with her cloaca. She then lays fertilized eggs on the vegetation. ([http://www.arkive.org/great-crested-newt/triturus-cristatus/video-09.html](http://www.arkive.org/great-crested-newt/triturus-cristatus/video-09.html))

- **Smooth newt** (*Triturus vulgaris*): This clip shows a similar but shorter sequence in a congener. Again, note the deposition of the spermatophore and its retrieval by the female’s cloaca. ([http://www.arkive.org/smooth-newt/triturus-vulgaris/video-09a.html](http://www.arkive.org/smooth-newt/triturus-vulgaris/video-09a.html))

- **Forsten’s tortoise** (*Indotestudo forstenii*): Turtles and tortoises face special challenges during mating. Nearly all species release sex-attractant pheromones from their cloacas, and desert tortoises have additional glands on their heads that are wiped on the forelegs to disperse pheromones. The vomeronasal organ is well developed in turtles and tortoises (Mason and Parker 2010). In this clip, a male tortoise approaches the female and clearly sniffs at her cloacal region before butting her until she stops moving and lets him mount. ([http://www.arkive.org/forstens-tortoise/indotestudo-forstenii/video-09.html](http://www.arkive.org/forstens-tortoise/indotestudo-forstenii/video-09.html))

- **Banded iguana** (*Brachylophus fasciatus*): These lizards combine erection or sleeking of crests and throat dewlaps, different sequences of head-bobbing and head-shivering movements, and striking changes in coloration during courtship. All of those components can be seen in this clip of a courtship and copulation. ([http://www.arkive.org/fiji-banded-iguana.brachylophus-fasciatus/video-09a.html](http://www.arkive.org/fiji-banded-iguana.brachylophus-fasciatus/video-09a.html))

- **Nile crocodile** (*Crocodylus niloticus*): Crocodiles and alligators do not have a functional vomeronasal organ, but they do respond to cloacal scents and pheromones from glands on the throats of males. Males rub their throat glands on the snouts of females during courtship. In this clip, a male crocodile performs a surprisingly low-key courtship of a female before beginning mating. Note that he rubs his throat on her snout and back early in the process. ([http://www.arkive.org/nile-crocodile/crocodylus-niloticus/video-09a.html](http://www.arkive.org/nile-crocodile/crocodylus-niloticus/video-09a.html))

Birds with predominantly male courtship displays

We now turn to examples of courtship displays in birds. In this section, we focus on species where courtship is relatively asymmetric: males perform elaborate displays to females who compare males in an ostensibly impassive manner but then pick one for mating by soliciting to him. Where possible, we show equivalent displays in related species to show how species differences are both limited by conserved antecedents and fostered by differences in ecology, mating system, and genetic drift. Readers may also want to review the section on displays of birds of paradise in Web Topic 4.4 and ducks in Web Topic 10.2.

**Grouse** (*Tetraoninae*). The 17 species of grouse are typical heavy-bodied and largely terrestrial Galliformes. They are unusual in that adults of many species feed on plant materials that are relatively indigestible or toxic to other animals (e.g., conifer needles, sage brush, catkins, etc.); their precocial chicks are largely insectivorous. Clutches are large and in most species, females perform incubation and offspring care without help from males. There are three types of mating systems in grouse that tend to grade into each other given variation within types. Males of smaller species tend to defend large breeding territories in which one (hazel grouse) or up to three (ptarmigan) females mate with the male and nest in his territory. At the other extreme, males defend small and contiguous display territories on leks to which females come to compare males and select one for mating (sharp-tailed grouse, two species each of black grouse, prairie chickens, and sage grouse). An intermediate type of system also occurs in which each male defends a display
territory to which females come to mate, but males are out of earshot or at most, clustered slightly into exploded leks (spruce grouse, blue grouse, ruffed grouse, and capercaillie). Sexual dimorphism in body size, coloration, and elaboration of plumage and eye combs increases as displaying males become increasingly clustered in space. For more details see Hjorth (1970) or Johnsgard (1983).

Male grouse court females by spreading strikingly patterned tail feathers, drooping their folded or slightly open wings, strutting in jerky or rapid running movements around the female, and emitting stereotyped sounds. Most have a fleshy vascularized comb above each eye that can be engorged to make it stand up and be colored red (similar to chickens); some species have other tints on their combs. Many species perform a flutter-jump display in which they leap to varying heights above the ground and flap their wings as they settle back. Most also have some form of strut display performed to nearby females. The typical strut posture involves leaning forwards, fanning the tail (usually exposing brightly patterned feathers), and slightly drooping the wings. Many species have air sacs on the throat or breast that are inflated statically or dynamically during the strut to make them visible while species-specific sounds are emitted. Struts usually involve short runs, stiff-legged stamping, and stereotyped movements of the whole body or the head. Females are usually relatively passive during courtship: they watch and listen, and if multiple males are present, visit a series of males before mating. Females indicate readiness to mate by adopting a typical avian squatting posture. Mate choice on grouse leks can be highly skewed: in species like sage grouse as few as 10% of the males are chosen by females for most of the matings. These basic components are conserved throughout most of the group, with various species-specific embellishments:

- **Red grouse** (*Lagopus lagopus*): Red grouse are one of many sub-species of the pan-northern hemisphere willow ptarmigan. They exhibit the first of the types of mating system listed above. Each fall, males establish a breeding territory of one to two hectares which they then roam in winter and spring with combs erected while giving several different types of territorial calls. Males court females attracted into their territories by fanning their tail, drooping their wings, and making stereotyped circling movements around the female while wagging their heads. Both sexes make soft calls shortly before copulation. The female then nests in the male’s territory.
  - Male defending territory and advertising his presence with calls: [http://www.arkive.org/red-grouse/lagopus-lagopus/video-sc00.html](http://www.arkive.org/red-grouse/lagopus-lagopus/video-sc00.html)

- **Ruffed grouse** (*Bonasa umbellus*): This forest species of grouse is found throughout Alaska, Canada, and the northern United States. It has the intermediate type of mating system described earlier. Males exhibit small red eye combs during the breeding season. Displaying males are dispersed but often within ear-shot of each other. Each has one or more favored display logs from which it emits a stereotyped drumming: this consists of an accelerating series of low frequency booms (see Chapter 2 for mechanism). These sounds carry long distances and attract females to their display sites. Once a female is present, the male adopts a typical grouse strutting posture with the body leaning forward, the tail fanned, and the wings slightly drooped. Male ruffed grouse also erect a collar of nape feathers that form a crescent around the head. Males then move slowly towards the female while making hissing sounds and rotational movements of the head. The final approach is made in a swifter “rush” with the wings dragging on the ground.
  - Example 1 of male drumming: [http://www.youtube.com/watch?v=sW5gMV5WQP4](http://www.youtube.com/watch?v=sW5gMV5WQP4)
  - Example 2 of male drumming: [http://www.youtube.com/watch?v=t2KI-OKuqW4](http://www.youtube.com/watch?v=t2KI-OKuqW4)
  - Male in strutting posture with ruff erected: [http://www.youtube.com/watch?v=LbRbxJTAOAM&feature=related](http://www.youtube.com/watch?v=LbRbxJTAOAM&feature=related)
Blue grouse (*Dendragapus obscurus*): Blue, sooty, and dusky grouse are different subspecies, or races of a species inhabiting the conifer forests of the US Pacific north-west, northern Sierra mountains, and Rocky mountains. Like ruffed grouse, they exhibit an intermediate mating system of dispersed and promiscuous displaying males. Males exhibit low frequency hoots to attract females: coastal populations (Pacific northwest and California Sierras) tend to make loud hoots that can be heard at long distances; inland populations (Rocky Mountains) perform the behavior but can barely be heard unless very close to the bird. When other males or females are in sight, a male may perform several different patterns of flutter-jump with accompanying sound emission. Once a female is close, the male fans its tail, engorges its eye combs, droops its wings slightly, and inflating underlying air sacs, exposes two bare patches of throat skin each of which is surrounded by a rosette of white feathers. He then performs a circling rush towards the female ending with a loud hooting call. This may be repeated over smaller distances until the female solicits mating or leaves.

- Male performing long distance hoots in Sierra mountains of California: http://www.youtube.com/watch?v=LKiXaLa93gw
- Male performing soft hooting in Wyoming: http://www.youtube.com/watch?v=UrQHOG0-AL0&feature=related
- Male strutting with throat sacs visible: http://www.youtube.com/watch?v=s-MxaHrYrGg
- More male strutting; no concurrent sound recordings: http://www.youtube.com/watch?v=xO9MH4EAHdg
- Various clips of male hooting and strutting; dubbed sound track of music: http://www.youtube.com/watch?v=W2XX2PzY1Bi&NR=1

Spruce grouse (*Falcipennis canadensis*): Spruce grouse occur throughout the northern conifer forests of Canada and Alaska. They have the same intermediate mating system of dispersed displaying males seen in ruffed and blue grouse. Males have the usual colored eye combs, but lack the bare skin patches seen in blue grouse. Males can produce a very low frequency hooting to attract females. They also perform a number of variations on flutter-jumps: a typical one is to fly in a conspicuous way to a low branch where it sits with combs erect and tail partially cocked while making a few wing flicks. The courtship posture is typical of most grouse. When courting a female, the male walks towards her opening and folding alternate side of its tail with successive steps to produce a swishing sound. Males may also fan and collapse the tail to produce sounds, bob the head from side to side, or perform ritual pecking at the ground. Like ruffed and blue grouse, spruce grouse will perform a stereotyped rush at a female while hissing and swishing the tail feathers; in spruce grouse, this ends with the male squatting in a position similar to that of soliciting females. Some examples:

- Male alternatively adopting and relaxing strutting posture: http://www.youtube.com/watch?v=WMrDVBvFAd4
- In middle of clip, good example of male flutter-jump flight: http://www.youtube.com/watch?v=ZAYHH-Wkvw&feature=related
- Excellent closeup video of male courting female at grit site: http://www.youtube.com/watch?v=DFuVeHDKfx0&feature=related

Capercaillie (*Tetrao urogallus*): This Eurasian boreal forest species is the largest of the grouse. As with other species, territorial and courting males fan their tails, engage their eye combs with blood, and droop their wings. However, in this species, the head and neck are stretched up vertically while the bill is snapped accompanied by a variety of swishing sounds. This is one of the species in which males form exploded leks, but they do often position their favored display sites near to each other. In this clip, a male performs the standard display both to the camera and to another male. (http://www.arkive.org/capercaillie/tetrao-urogallus/video-13.html)

Black grouse (*Lyrurus tetrix*): Males of this species display on small leks (5–10 males). Interference between
males is common and fighting is frequent. Major courtship displays include: (a) short flutter-jumps into the air followed by a fluttering descent and a hissing sound; (b) the roo-coo in which the male tips its body forwards, spreads the tail, lowers the folded wings, and dances in circles around a female while rapidly emitting a roo-coo sound. Females on leks walk around sampling males but perform no courtship signals of their own except to crouch into a solicitation posture when ready to mate. Both displays and males fighting are shown in this clip. (http://www.arkive.org/black-grouse/tetrao-tetrix/video-09a.html)

- **Greater prairie chicken** (*Tympanuchus cupido*): Males of this North American species displays on leks of 5–20 males (leks may have been larger before habitat loss and hunting lowered populations). Eye combs in this species are yellow–orange, and both sexes have 8–10 elongated nape feathers called pinnae. Males provide long-range advertisement with flutter-jump displays and loud cackling squeals. In the main courtship strut to a nearby female, a male erects his nape pinnae straight up, stamps his feet as he moves past or around her a short distance, stops and crouches, and finally fans his tail open and closed while emitting a low frequency multiple-syllable coo and inflating and deflating the paired orange throat sacs. He then repeats this sequence. Unlike sharp-tailed grouse (see below), male conflicts usually involve different postures and displays than are used to court females.
  - Birds flying into the lek at dawn and displaying to females:
  - More male displays:
    http://www.arkive.org/greater-prairie-chicken/tympanuchus-cupido/video-09b.html

- **Sharp-tailed grouse** (*Tympanuchus phasianellus*): The sharp-tailed grouse is a close relative of prairie chickens, and their ranges often overlap in North American prairie areas. It is a lek species with 2–30 males/lek. Males advertise their presence on the lek with loud cackling calls and flutter-jump displays. Females often announce their arrival at the lek with their own cackling. As with the congener, courting males assume a bent-over posture and erect their yellow eye combs. Unlike most other grouse, the wings during strutting are not drooped but held extended laterally from the body, and unlike prairie chickens, the vertically oriented tail is a fluffy oval instead of a fan of stiff feathers. The paired throat sacs are purplish unlike the orange of the prairie chickens. Males court females by circling them with stiff-legged steps like prairie chickens, but the sounds produced are a mix of low amplitude foot-drumming and loud cackling. Intermale aggression is common on sharp-tailed grouse leks and many of the same displays directed at females when courting are also directed at neighboring males. In addition to cackling and drumming, males produce a cooing sound that is similar to the courtship booming of prairie chickens, but is here often directed at other males. As seen in the clips below, a given male switches between directing displays at females and other males repeatedly during a short interval.
  - General overview by Colorado Division of Wildlife:
    http://www.youtube.com/watch?v=GQ78xhuZYhE
  - Closeup of single male displaying with no female nearby:
    http://www.youtube.com/watch?v=0mgTkaK4ezk
  - Single male circling a female while strutting:
    http://www.youtube.com/watch?v=lscWWvieZqc

- **Greater sage grouse** (*Centrocercus urophasianus*): This is the largest North American grouse species and is an inhabitant of the sagebrush deserts of the far west. Males display on very small territories within leks that can host over 150 males. Unlike lek species with smaller numbers of males, interference between males when females are present is rare. This genus appears to have lost the flutter-jump display of ancestral grouse. Males perform a single courtship display that is highly stereotyped: the body posture is more erect than in other lek species, although the tail is fanned in typical grouse fashion. The wings are slightly drooped but held against the sides of the body. The male takes several steps forwards and then heaves the heavily muscled and air-filled breast forward twice. Each time, it rebounds and rubs stiff breast feathers against the wings producing two swishes. These are accompanied by low frequency coos and spreading and contracting of the bare green skin over the two chest sacs. The strut ends with two rapid inflations and deflations of the chest sacs, each of which produces a popping sound, and a high frequency whistle emitted between the two pops. The overall body
movement flips the male’s nape plumes upwards after which they fall back. The male then stands still for a few seconds and repeats the display. Male sage grouse usually do not herd females (e.g., move in decreasing circles around them), and often face away from them when doing their strut. This appears to be less threatening to females who often approach displaying males closely. This excellent clip with synchronized sound provides close-up views of males strutting including a final slow motion section. (http://www.youtube.com/watch?v=m0M8pZnNlnI)

- **Gunnison sage grouse** (*Centrocercus minimus*): The small relic population of this second sage grouse species is currently threatened with extinction. Males are slightly smaller than greater sage grouse males, and their nape plumes are much longer and fuller. Leks have a similar structure and overall behaviors are similar to those of the greater sage grouse. The pattern of the strut however contains more chest-sac popping, louder wing swishes, and multiple quivering of the tail feathers. The nape plumes usually flip both upwards and forwards in contrast to the lesser motion in the other species. In the clip provided, the different emphasis on the head plumes is obvious and the accompanying sounds are clearly more “bubbly” than those of the greater sage grouse. Note that the sound track in this clip is for illustration only and is not a synchronized with the video. (http://www.youtube.com/watch?v=I1DI_C7uTDw)

**Pheasants** (*Phasianinae*): Pheasants constitute another large group (50–60 species depending on the taxonomist) of Galliform birds with striking courtship displays. Like grouse, adults tend to be mostly vegetarians but are much less specialized and their diets may also include berries, fruits, tubers, fungi, and some insects. Whereas male grouse fight by battering each other with their wings, pheasants have sharp spurs on their lower legs and fight by leaping into the air and trying to stab their opponents with their spurs. Pheasants live throughout south-east Asia in habitats ranging from thick lowland forest to barren montane scrub. Mating systems include monogamous pairs that remain together all year (monals), serially polygyny in which males remain with a female up to or through incubation and then move on to find another mate (tragopans), harem polygyny in which a male guards a group of females (common pheasant and junglefowl), and various types of promiscuity including dispersed displaying males (argus pheasant) and leks (some populations of peafowl). Although there is a tendency for greater sexual dimorphism as one goes from monogamous to promiscuous mating systems, many monogamous pheasants, like monogamous pairs in many other avian taxa, are strikingly dimorphic in plumage coloration. For more details, see Beebe (1936), Delacour (1977), Savage and Ridley (1987), Johnsgard (1999), and Madge and McGowan (2002).

As with the grouse, a common ancestral set of courtship displays has been modified and elaborated, dropped, or augmented depending upon the habitat, mating system, and simple drift. The shared repertoire includes loud crowing calls by males to defend territories, challenge other males, or attract females. As with grouse and their flutter-jumps, many male pheasants perform conspicuous wing flapping and short stylized flights to advertise their presence at intermediate distances. When females are nearby, males of many species perform ritualized tidbitting. This behavior is widespread in Galliformes and consists of a hen pointing or pecking at a food item to instruct chicks on suitable foods. Males have co-opted this behavior to get the attention of females and in some species, such as jungle fowl, they actually point at or peck at food finds to the benefit of females. In other species, male tidbitting is more of a stereotyped movement with no food present. Nearly all species have bare skin around the eyes and cheeks, and in some this may extend under the bill. This skin is usually colored red or blue, although a few exceptional species have white or black skin. The skin in some species is elaborated into fleshy horns, wattles, or bibs that can be engorged with blood or otherwise expanded. Perhaps the most widespread pheasant courtship display is waltzing. This is a lateral display in which males fan the tail, expand the wings, and then tip their body to one side so that one wing touches the ground while the other points nearly straight up. The result is an elliptical sheet of feathers oriented almost vertically with the male’s head on the side nearest the female. Males may remain static or circle the female while waltzing. In many species, the plumage exposed during waltzing is strikingly patterned and colored. Different species of pheasants have different enlarged muscles under the skin that are used to erect and hold their wing, tail, or covert feathers in position during displays (Osborne 1968). Some species also have wiry feathers ending in knobs on their heads that amplify their display movements. As with non-lek grouse, many pheasants make a short rush at the female at the end of a display.
Because pheasants are highly fancied by aviculturists, and much easier to keep in captivity than grouse with their specialized diets, most of the clips below were made in captivity:

- **Koklass Pheasant** (*Pucrasia macrolopha*): The two sexes in this seasonal montane forest species are somewhat dimorphic. Pairs remain together all year (social monogamy). The male remains with the female through incubation and the family moves off together after hatching. Males are highly territorial and crow to defend territories. During the waltz, the male erects the paired pinnae on its head, rears up, lowers the wing on the side of his body away from the female, and exposes the white throat patches. He may also vocalize.
  - Male giving territorial crow: [http://www.youtube.com/watch?v=HovVFQws050](http://www.youtube.com/watch?v=HovVFQws050)
  - Male performing waltz with no female nearby towards which to aim it. This might be used here as a threat. [http://www.youtube.com/watch?v=Lo5qq_pXYm0](http://www.youtube.com/watch?v=Lo5qq_pXYm0)
  - Brief clip of male performing waltz display to female (note asymmetry of wings: [http://www.youtube.com/watch?v=Jwr1NlsOUzg](http://www.youtube.com/watch?v=Jwr1NlsOUzg))

- **Monals** (*Lophophorus*): These pheasants live in seasonal alpine meadows. They often go around in small unisexual parties and males do not appear to be territorial. Both sexes call and this may facilitate getting together during the breeding season. Although originally thought to be monogamous, they are now thought to be serially polygynous.
  - Himalayan monal (*L. impejanus*):
    - Note that the sound on this clip is dubbed and not the natural sounds of the birds. First sequence shows male crowing (no sound). Males can perform a standard waltz (not shown here) in which they stand laterally to the female, raise the chest, drop the near wing and raise the far one. In this clip, they perform a frontal courtship display which is likely a re-ritualized waltz. The male lowers its head, spreads its wings and tail, and slowly fans the tail forward and backward (shown several times here). The display is symmetrical when the female is in front of the male, but tipped in the usual waltz-manner when she is to one side. ([http://www.arkive.org/himalayan-monal/lophophorus-impejanus/video-09.html](http://www.arkive.org/himalayan-monal/lophophorus-impejanus/video-09.html))
  - Chinese monal (*L. lhuysii*):
    - The male in this clip approaches the female in an upright posture and partially fans his tail. He then performs a frontal display that is somewhat similar to that of the congener (above). However, the wings are spread more, and the body is kept low to the ground while the fanned tail is flapped as the bird rapidly runs in a loop. The display ends with a more traditional asymmetric waltz position and some ritualized tidbitting. ([http://www.arkive.org/chinese-monal/lophophorus-lhuysii/video-09.html](http://www.arkive.org/chinese-monal/lophophorus-lhuysii/video-09.html))

- **Tragopans** (*Tragopan*): The five species in this genus are all residents of semi-montane forests where they are one of the few pheasants to nest in trees instead of on the ground. The mating systems are poorly known, but serial or simultaneous polygyny best fits the available evidence. Sexual dimorphism is strong: in addition to a distinctively colored plumage of bright white spots on a red or gray-brown background, males have brightly colored fleshy horns and a large bib that can be inflated during courtship displays. Males have very loud calls that carry long distances and suggest defense of large territories. When they encounter females, they may perform a stiff-legged clucking and/or tidbit. If the female remains, the male inflates its horns and bib and begins a jerky vertical bobbing that waggles the horns and ripples the colored bib. Usually, he then moves behind a rock or fallen log so that the female can just see the waggling horns, the bobbing head, and a tiny top part of the bib. The male then adds shaking of the partially opened wings to the repetitive movements. The male may
accompany this with synchronized clicks or clucks. Suddenly, the male rears up from behind the barrier as high as it can extend its body, displaying the bib and horns in full expansion. Then it rushes at the female with the body erect and the wings drooping in a waltz-like fashion, or alternatively collapses back behind its barrier for a repeat performance. If you have never seen these displays, you must look at the following clips!

- **Temminck’s tragopan** (*T. temminckii*):
  - Male shaking out bib and doing displays to female:  
  - More male display:  

- **Satyr tragopan** (*T. satyra*):
  - Male calling, expanding head horns and chin bib, display ending with copulations:  
  - More footage of male display:  

- **Western tragopan** (*T. melanocephalus*): Brief part of clip showing male with expanded bib:  

- **Cabot’s tragopan** (*T. caboti*): Male shaking out bib and displaying briefly:  
  [http://www.youtube.com/watch?v=aE2jSoXp9bU](http://www.youtube.com/watch?v=aE2jSoXp9bU)

**Peacock pheasants** (*Polyplectron*): The 8 species of peacock pheasant inhabit either lowland or montane tropical forests. Most appear to be solitary much of the time. They are relatively monomorphic in plumage coloration with large “ocelli” designs on the wing and tail feathers. Mating systems remain unclear, but serial polygyny seems most likely. Species in this genus appear to exhibit nearly all of the common repertoire for pheasants. Males have favored display areas that they clear and use for calling. Males are often within earshot and respond to each other with counter-calling. Males approach visiting females with stiff-legged gaits, tidbitting sounds, and may even offer females food morsels in their bills. The main courtship displays are frontal and lateral waltzes with asymmetric tilting of the body. These are positioned so as to accentuate the eye-spot patterns of the plumage. Waltzing may be performed statically or while circling around the female. In this genus, females appear less passive during courtship than other taxa, and exhibit a number of head shaking, preening, and circling motions of their own prior to copulation.

- **Palawan peacock pheasant** (*P. napoleonis*):
  - General male display:  
  - Tidbitting and closeup of male display:  

- **Malaysian peacock pheasant** (*P. malacense*):  

- **Mountain peacock pheasant** (*P. inopinatum*):  

**Junglefowl** (*Gallus spp.*): The habitats for the four junglefowl species range from lowland forest edge to subalpine seasonal zones. The red jungle fowl is the ancestor of domestic chickens. Sexual dimorphism is strong: in addition to colored bare skin around the eyes and face, males have colored fleshy combs on the head and fleshy lappets below the bill. Male plumage is iridescent and brightly colored. In wild populations, jungle fowl live in small flocks that roost together at night and forage over the same home range all year (Beebe 1936; Collias and Collias 1967). Flocks tend to be relatively small (two to six individuals) in the wild, but can be considerably larger in feral and domesticated populations (Beebe 1936; Collias and Collias 1967, 1996; Collias et al. 1996). Each flock hosts one dominant male, multiple females, and varying numbers of subordinate males. At dawn, the dominant male crow and leads the group off to forage, the subordinate males usually trailing along at
the end. The dominant male defends the females with crowing and attacks when intruder males come too close during the day’s excursions. In the breeding season, males may court females with tidbitting, food provisioning, wing flapping, neck feather fluffing, headshaking, crouch-scratching, and waltzing displays (Johnsgard 1999). Because dominant males and flock females are already well acquainted, they may dispense with elaborate courtship: the female simply presents, and the male mounts. This is often the case in domestic chickens as well. However, wild females do not always mate with their dominant male, and this provides an opportunity for other males to court them which they do vigorously. When given a choice experimentally, females base their mate choice on both male display performance, particularly waltzing, and on the state and size of male secondary traits such as combs, wattles, and spurs (Zuk et al. 1995). After mating, females leave the flock to nest solitarily, and perform all parental care. Later, they rejoin the flock.

- **Red junglefowl** (*G. gallus*): The following clip shows males crowing, fighting, and neck feather fluffing. 

- **Peafowl** (*Pavo spp.*): The two species in this genus, the largest of the pheasants, inhabit forest edge and broken woodland habitats. Both are highly sexually dimorphic in coloration and plumage structure: males have extremely elaborated tail coverts which host broad plumes with multiple ocellus patterns in iridescent colors. The smaller stiff tail feathers are used to help hold up the giant covert plumes. Both sexes have a tuft of feathers on their crown. Although a harem mating system was originally suggested for both species (Beebe 1936; Delacour 1977; Johnsgard 1999), an exploded lek system has been shown in natural populations of Indian peafowl (Harikrishnan et al. 2010), and even more clustered leks near villages or in introduced feral populations (Dakin 2011; Petrie et al. 1991; Yasmin and Yahya 1996). Like junglefowl, peafowl often have communal night roosts in a high tree but then break up each dawn into males that head to display areas and females that form small foraging groups. Males advertise their presence at display sites with loud calls. When females are near, males erect the enormous train into a vertically positioned fan, softly beat the partially opened wings, and periodically quiver the array making a rattling sound. Males often aim themselves away from females forcing females to move to a new position in front of the male to observe the display. When females are ready to mate, they squat and the male rushes over, lowering the train, to copulate.

  - **Indian peafowl** (*P. cristatus*): Males calling and displaying to females: [http://www.arkive.org/indian-peafowl/pavo-cristatus/video-00.html](http://www.arkive.org/indian-peafowl/pavo-cristatus/video-00.html)

- **Great argus pheasant** (*Argusianus argus*): This species is an inhabitant of lowland tropical forests where they tend to live relatively solitary lives. Males and females are similarly colored and both have small crests and bare skin around the eyes. Males however have enormously elongated tail feathers (not coverts as in peafowl) and iridescent ocelli that are exposed when displaying. Males call to attract females to widely dispersed courtship sites. Calls can be heard up to 100 hectares around the male. Each male clears a small arena in which it performs displays to visiting females. Once a female enters the arena, the male may perform tidbitting or circle the female with a stamping gait. It then performs a classic pheasant waltz display in which one only marginally open wing is raised and the other lowered with the head facing the female. After several repetitions, the male faces the female and spreads his wing feathers into a giant fan revealing the iridescent ocelli on the plumes. He tucks his head behind one wing and begins a forward and backward pumping of the wing fan which also causes the long tail to wave forwards and backwards overhead. Females invite copulation by squatting, and the male then rushes to her to mate as in many other species.


**Some other birds with predominantly male courtship displays**
Below is a sampling of a few other interesting courtship displays:

- **Great bustard** (*Otis tarda*): Males of this open grassland Gruid often display on leks or exploded leks. Here a single male is seen displaying to a female.  
  (http://www.arkive.org/great-bustard/otis-tarda/video-09b.html)

- **Lesser florican** (*Syphetoides indicus*): Another bustard species displaying in grasslands, in this case high grass in India. Male performs vertical jumping displays to advertise its presence to females. 
  (http://www.arkive.org/lesser-florican/syphetoides-indicus/video-09.html)

- **Club-winged manakin** (*Machaeropterus deliciosus*): Most of the neotropical manakins are lek species. This clip shows a male of this species performing one of several types of displays on its favored perch.  
  (http://macaulaylibrary.org/video/flashPlayer.do?id=58492)

- **Wire-tailed manakin** (*Pipra filicauda*): Here, a male of another manakin species performs several different displays to a female sitting on the male’s display perch. 
  (http://www.arkive.org/wire-tailed-manakin/pipra-filicauda/video-09b.html)

- **Three wattled bellbird** (*Procnias tricarunculatus*): Bellbirds belong to a group of neotropical birds in which exploded leks are common. Here is a clip of a male calling. You can hear other calling males just within earshot. 
  (http://www.arkive.org/three-wattled-bellbird/procnias-tricarunculatus/video-00.html)

**Birds with mutual courtship displays**

- **Blue footed booby** (*Sula nebouxii*): Mutual display by courting male and female: 
  (http://www.arkive.org/blue-footed-booby/sula-nebouxii/video-09a.html)

- **Japanese cranes** (*Grus japonensis*): Cranes and storks often court and form pair bonds at communal “club” sites where many different individuals come to find a mate.  
  (http://www.arkive.org/japanese-crane/grus-japonensis/video-09f.html)

- **Fisher's lovebirds** (*Agapornis fischeri*): Parrots typically form long-term pair bonds. However, at each nesting attempt, they share in the preparation of the nest, and allopreen and courtship feed each other. Here, some of these activities are shown for a small African species. 
  (http://www.arkive.org/fischers-lovebird/agapornis-fischeri/video-09.html)

- **Long-tailed manakin** (*Chiroxiphia linearis*): Not all mutual display is between a male and a female. Pairs of males in the manakin genus *Chiroxiphia* share a single display territory in which they emit a “toledo” call synchronously to advertise their presence, and then do a joint dance to visiting females in which they repeatedly switch perches. Usually, a dominant member of the pair gets any matings, but the subordinate moves up to dominant status once the former one dies. 
  - Pair of males make toledo call and then move to display perch to do dance: 
    http://www.youtube.com/watch?v=UJSj6il1m0M0
  - Another pair doing dance: 
    http://www.youtube.com/watch?v=MlezPZgOH0U

- **Lance-tailed manakin** (*Chiroxiphia lanceolata*): Another species with a similar joint male display dancing for a visiting female: 
  (http://www.youtube.com/watch?v=2QQq5P3PLCw)

**Courtship displays using objects**

As we have seen, males of some species of fish, and some birds (not shown here), build nests in which they entice females to lay eggs after mating with the male. There are some other taxa, however, that build structures solely to attract females for mating; the females will then leave to lay the eggs or raise the offspring elsewhere. Since males that build such display structures are competing for matings, sexual selection has led to their building very elaborate
display structures. Perhaps the most amazing examples are bower birds. These construct various alleys and tents and decorate them with brightly colored or rare items. Females visit the males who show off their decorations. Studies using robotic females have shown that, in these species, the courtship between the sexes is very complicated: males cannot move on to more advanced stages unless and until females give them appropriate signals to go ahead (Patricelli et al. 2006). Males are very competitive and often steal from or even destroy a neighboring male’s bower. Interestingly, species with the dullest male plumage tend to construct the most elaborate bowers and the most diverse decorations.

- **Flame bowerbird** (*Sericulus aureus*): Young male constructing alley-bower: [video](http://www.youtube.com/watch?v=hNtB6UMAh6M&feature=related)
- **Satin bowerbird** (*Ptilonorhynchus violaceus*): In addition to its decorated alley bower, males of this species are superb mimics of the calls of other species.
  - Gerald Borgia web page: Dr. Borgia is a world expert on bowerbirds. This web page provides an introduction to the group, and links to an excellent set of video clips of males displaying successfully and unsuccessfully to females, engaging in destruction of neighbors’ bowers, etc. [webpage](http://www.life.umd.edu/biology/borgialab/)
- **Vogelkop bowerbird** (*Amblyornis inornatus*): These drab males build elaborate display huts and decorate them with a wide variety of colorful or shiny objects.
  - Nice clip on house, decorations, and behavior: [video](http://www.youtube.com/watch?v=Ktip0DqlT4k)
  - Discovery Channel clip: [video](http://dsc.discovery.com/videos/life-vogelkop-bowerbirds-display-treasures.html)

**Mammals**

Mammal mating systems include relatively solitary species (usually with large male territories containing separate home ranges of multiple females), harems in which a given male oversees the same females for an extended period, monogamous pairs that remain together all year, mixed sex groups with established dominance hierarchies, and various promiscuous systems ranging from well-defined leks to chaotic mating swarms. Males and females that do not live together often require several stages of courtship before either male or female will allow the other to get close enough to mate. Members of stable groups often know each other and require only simple coordination behaviors before copulation; females in harems may have no choice of mate, and as soon as receptive will be approached by the harem male. Females in lek species visit multiple males before attempting to select one based on competitive displays; as with grouse, mammalian lek species differ depending on whether females are relatively free to choose (e.g., hammerheaded bats) or subject to frequent male interference (e.g., ungulate leks). Below, we provide a small sampling of the courtship behaviors associated with different mammalian mating systems:

- **North American otter** (*Lontra canadensis*): North American otters are relatively solitary much of the time but have sufficiently overlapping home ranges that olfactory cues left by oestrus females attract neighboring males. Not surprisingly, both parties are fairly aggressive and wary during pre-copulatory interactions. [webpage](http://www.arkive.org/north-american-otter/lontra-canadensis/video-09.html)
- **Ring-tailed lemur** (*Lemur catta*): Ring-tailed lemurs live in stable mixed-sex groups. Because females are dominant to males, they have a fair amount of freedom to choose a mate. Oestrus females are detectable through olfactory cues: this clips shows examples of females actively soliciting a particular mate and one case of a female rejecting one. [webpage](http://www.arkive.org/ring-tailed-lemur/lemur-catta/video-09a.html)
- **Flying foxes** (*Pteropus spp.*): Both sexes of these large fruit bats live in communal roosts during the day. In the breeding season, males establish and defend territories on a few branches within the colony roosting tree. They mark the surface of the branches with secretions from shoulder glands. Females initiate mating by entering a
male’s territory, and can halt mating by flying off to another site. However, central males in the colony tend to accumulate multiple females who remain within their territories throughout the breeding season. Males are very solicitous and perform extended genital licking on females prior to mating. Eventually, receptive females assume a suitable posture and the male mounts from the rear, using an exceptionally long penis to get past the female’s inter-femoral tail membrane. Several examples:

- **Little red fruit bat** (*P. scapulatus*): Good example of genital licking and final copulation: [http://www.arkive.org/little-red-flying-fox/pteropus-scapulatus/video-09b.html](http://www.arkive.org/little-red-flying-fox/pteropus-scapulatus/video-09b.html)
- **Black flying fox** (*P. alecto*): Females can be wary when males first initiate courtship licking as in this example. [http://www.arkive.org/black-flying-fox/pteropus-alecto/video-09b.html](http://www.arkive.org/black-flying-fox/pteropus-alecto/video-09b.html)
- **Mandrill** (*Mandrillus sphinx*): Mandrills are another primate that lives in stable mixed-sex groups. However, the dominant male in a mandrill group rules overall and a female in oestrus has no opportunity to exercise choice or engage in preparatory courtship. ([http://www.arkive.org/mandrill/mandrillus-sphinx/video-09a.html](http://www.arkive.org/mandrill/mandrillus-sphinx/video-09a.html))
- **Northern elephant seal** (*Mirounga angustirostris*): Males arrive at breeding beaches first, spread out and establish dominance hierarchies. Females arrive, give birth, and come into oestrus shortly before or at weaning of the current young. Males are much larger than females making it difficult for females to reject an interested male. However, they are able to incite new fighting among nearby males by moving to certain locations, and then mate with the winner of that outcome. This clip shows a typical copulation in this species. Note marked sexual dimorphism in size. ([http://www.arkive.org/northern-elephant-seal/mirounga-angustirostris/video-09b.html](http://www.arkive.org/northern-elephant-seal/mirounga-angustirostris/video-09b.html))
- **Impala** (*Aepyceros melampus*): Impala live in fairly stable harem herds. The harem male and females are well acquainted and courtship is often fairly minimal. ([http://www.arkive.org/impala/aepyceros-melampus/video-09a.html](http://www.arkive.org/impala/aepyceros-melampus/video-09a.html))
- **Topi** (*Damaliscus korrigum*): Topi males either defend foraging territories within which they have mating access to oestrus females currently feeding there, or (even in the same population) males display on a lek. In the latter case, each male defends a small display territory by marking it with urine and feces, standing tall in a conspicuous site, walking stiffly with the ears cocked and tail straight out, or actively chasing out intruders. Both sexes get down on their front knees and rub their horns, orbital glands, and foreheads in the ground, preferably muddy ground. This apparently mixes glandular secretion with mud on the horns to facilitate enhanced pheromone dispersion. Males will also meet at a common boundary and perform ritual fights in which they fall on their front knees and clash horns. When females enter a male’s display territory, he lowers his ears, and approaches her with a high stepping courtship walk, usually followed by lifting the ears and sniffing of the female’s genitalia. Females have some choice of mates (by entering a particular territory), but topi, like other lekking ungulates, do show high levels of male interference. This video was taken by Jack Bradbury with input and guidance from Mike Rainy in Kenya.

- **Whales and dolphins** (Cetacea): One might think that being relatively weightless in water makes mammalian mating easier. However, there is a significant challenge to establishing suitable alignments between male and female without the earth’s surface as a substrate and given surging seas. In addition, competing males can appear from any direction in water resulting in high levels of interference in many cetaceans. In most species, the male swims upside down under the female, both keep moving, and the male stabs upwards with his long penis until he finds the female’s genital opening. Sexual play is common in both sexes, particularly younger dolphins, although adult male dolphins continue to poke and prod others with their penises in many different contexts. This may provide the dexterity needed for actual mating, but also seems to have other signal functions. Acoustic signals and touch clearly play significant roles during courtship. Males of some species also perform leaps and other gymnastics when courting females, and in a few species, males mate-guard females after copulation. Some cetacean mating examples:
• **Spinner dolphin** (*Stenella longirostris*): Copulation must take place in crowded social contexts. Here, two males both try to assume the belly-up position prior to mating.  

• **Atlantic spotted dolphin** (*Stenella frontalis*): An initial pair is soon harassed by multiple males:  

• **Dusky dolphin** (*Lagenorhynchus obscurus*): As with the spinner dolphins, male and female keep moving very rapidly and try to achieve copulation without interference of other nearby males:  

• **Southern right whale** (*Eubalaena australis*): Groups of male southern right whales often aggregate around a receptive female, making it difficult for any particular male to copulate. The alignment challenge is also considerable with an animal this big moving in three-dimensions. This clip shows both problems, although not as many males are present in this example as have been reported.  
  [http://www.arkive.org/southern-right-whale/eubalaena-australis/video-09d.html](http://www.arkive.org/southern-right-whale/eubalaena-australis/video-09d.html)

**References**


Web Topic 13.1
Direct Benefit Models of Cooperation

Introduction

Why humans should or do cooperate has been of concern to philosophers and economists for millennia. Humans often engage in extensive cooperation with unrelated individuals showing that kin selection and greenbeard biases are not necessary to compensate for the costs of being a cooperator. This immediately raises the question of whether the same mechanisms might explain the many cooperative interactions seen in other species; at minimum, kin selection cannot explain cooperative mutualisms between different species. In this Web Topic, we introduce the evolutionary game models advanced to explain cooperation without genetic compensations, and note particular cases that may be relevant to animal communication. Because the literature on this topic, especially that dealing with human cooperation, is so extensive, we can barely do justice to it here. However, we provide some further reading sources at the end of the module that provide gateways to this larger literature. Readers of this module should be familiar with the classification of evolutionary game models and the logic of “take” and “give” games reviewed in Web Topic 10.5.

Defining and classifying cooperation

A necessary condition for cooperation is that at least one individual undertakes an action at some cost to itself and other individuals benefit from the first individual’s investment. The first individual’s action could have been prompted by likely compensating benefits to itself, or its efforts might have been “purloined” as the result of theft, deceit, or coercion (Connor 1995). Setting coercive manipulations aside for the moment, there are a number of ways that a cooperative interaction could generate net benefits for both parties. In this module, we will also ignore genetic benefits as a result of kin selection or greenbeard biases: our focus here is on cooperation among unrelated animals, or at least, on contexts in which relatedness makes no significant contribution to the cooperation economics. These are usually called the “direct benefits of cooperation” (Bergmüller et al. 2007; Leimar and Hammerstein 2010).

Consider an animal B that is pursuing some direct benefit to itself. There are three ways that animal A might interact with B with net positive benefits to both parties and without either party invoking coercion or deceit:

- **By-Product Mutualism**: In the simplest case, animal A puts itself sufficiently nearby or ensures that it is sufficiently connected to animal B that animal B’s efforts not only benefit animal B but also provide some spill-over benefits to animal A. This is called by-product mutualism (West-Eberhard 1975; Brown 1983). In these cases, A benefits by B’s actions without altering the costs or benefits experienced by B. By-product mutualism can be one-sided or two-sided: in the latter case, each of A and B pursues its own goals, but they coordinate their actions so that each party receives some by-product benefit through the actions of the other. This two-sided option is called coordinated by-product mutualism.

- **Pseudo-reciprocity**: Again suppose that animal B is pursuing its own selfish activities, but animal A, instead of remaining passively on the sidelines, expends some cost to help animal B achieve B’s selfish goals. Again, animal A receives by-product benefits from B’s actions, but perhaps because of its investment in B’s efforts, it might receive more or different benefits. This is called pseudo-reciprocity (Connor 1995; Leimar and Connor 2003; Connor 2007). Since A can vary the size and nature of its investment, and its strategy can affect the payoffs for B, this interaction can be formulated as a game and ESSs sought (Leimar and Connor 2003). Note that given its definition, pseudo-reciprocity includes most animal communication as a special case: the receiver is faced with a decision about future states or actions and the sender provides information that facilitates that decision in a way that benefits both parties. Note also that, as with the prior case, both parties might coordinate pseudo-reciprocity with the other (mutual pseudo-reciprocity).

- **Reciprocity**: In the third case, animal B again pursues its own interests, and animal A suffers some costs to help animal B in that pursuit. However, in this case, there is no immediate spill-over benefit to animal A. The only way A can recoup its cost and achieve a positive payoff from its actions is if A and B subsequently exchange roles...
such that B pays costs to provide benefits to A. The key difference between pseudo-reciprocity and reciprocity is that benefits accrued by the recipient in the former spill over on the donor without the recipient changing its role, whereas roles must be exchanged before both parties achieve net positive payoffs in the reciprocity case (Trivers 1971).

![Figure 1. Criterion tree for classifying types of cooperation among unrelated animals.](image)

Answering four basic questions can assign most biological examples to a specific category. Using the individuals mentioned in the text, the first question concerns whether animal A did or did not invest in animal B (which will pursue its own interests in any case). If animal A did invest, the second question concerns whether this investment was somehow forced or tricked out of animal A. If not, the third question is whether animal B responded by also investing in the interaction (above and beyond its ongoing self-investment). Whether this leads to pseudo-reciprocity or reciprocity, the final question is whether animal A’s compensation comes from animal B (direct exchange) or some third party (indirect exchange). (Modified from Bergmüller et al. 2007; Connor 2007.)

Combining these three alternatives with possible manipulation, several authors have proposed criterion trees for assigning real examples to categories. One example is shown in Figure 1. Where the party invested in by animal A is the same one who delivers the consequences to animal A, the exchange is said to be direct; if the feedback on animal A comes from a third party, the exchange is said to be indirect. Manipulation, pseudo-reciprocity, and reciprocity can all lead to either direct or indirect accountings. This schema, and others like it, have proved very effective at classifying known examples of animal and human cooperation: all of the classical examples such as cooperative hunting, joint predator surveillance, cooperative breeding, and symbiosis can be assigned to their own spot in such a tree (Connor 1995; Leimar and Connor 2003; Bergmüller et al. 2007; Connor 2007; Bshary and Bergmüller 2008; Connor 2010). This partitioning has greatly clarified economic and game theoretic modeling of cooperation as different parts of the classification tree require different assumptions to generate ESSs and exhibit quite different evolutionary dynamics. One comment on terminology: some authors have lobbied for the restriction of the word “mutualism” to interspecific relationships and “cooperation” to intraspecific ones (Bronstein 2001; Bergmüller et al. 2007). In this module, we use the two terms as synonyms because such use is already widespread in the literature and because convergent inter- and intraspecific associations can occur at the same point in classification trees such as the one in Figure 1. However, readers should be alert to publications where the authors specifically restrict the term mutualism to interspecific associations.

### An overview of the basic problem

Before we examine specific cases, it is useful to stand back and identify the challenges that the evolution of direct benefit cooperation faces from the perspective of evolutionary game theory. There are two basic hurdles that usually hinder the evolution of cooperation. The first arises because many situations in which mutual cooperation could be beneficial are vulnerable to cheats and defectors. When modeled as a single-bout 2x2 symmetric contest, the result is
invariably a take game with the Cheat or Defector strategy as the pure ESS (Figure 2A). The first challenge is thus to identify conditions under which take games can be replaced by a give games (Figure 2B). Once achieved, the second problem arises. As we saw in Web Topic 10.5, give games in the absence of genetic compensations such as greenbeard or kin selection invariably have two possible ESSs: one in which everyone cooperates, and one in which everyone defects. A few games (e.g., the Snowdrift Game discussed below) can lead to a mixed ESS in which each party cooperates some of the time, and defects the rest of the time. However, it is extremely difficult to find any realistic model in which pure cooperation is the only ESS. The second challenge is thus to find conditions that might get a population to the all-cooperation ESS and hold it there despite drift or other factors outside the basic game. While some of the literature on cooperation mingles both hurdles, an equally diverse amount focuses on only one of the two steps. We shall try to keep straight which hurdle we are considering in the following discussion.

Figure 2. Take versus give games of cooperation. (A) The normal form matrix for this example shows that it is a take game: even though everyone cooperating would produce a higher per capita payoff (pareto optimum) than everyone defecting, the advantages to defecting when others are cooperating have an even higher payoff. The result is that cooperation degenerates until everyone is defecting as the best reply. Defect is then a pure ESS. (B) In this game, pure cooperation can be an ESS, but it is only one of two possible ones, the other being pure defect—which is found depends on starting conditions. If the initial condition is for everyone to be selfish and not cooperate, the system will stay there barring drift or some other outside factor.

**The prisoner’s dilemma**

Most modern treatments of the economics and/or evolution of cooperation begin with the **Prisoner’s Dilemma**. This model was first introduced to the evolutionary biology literature by Trivers (1971). In this game, two suspects are taken prisoner by the police and held separately. Each is urged to provide incriminating evidence about the other. The relevant game is typically modeled as a dyadic single-bout 2 x 2 symmetric contest. If both keep silent, they are likely to gain only a light sentence given the lack of detailed evidence (a per capita payoff of $R$). If one provides incriminating evidence on the other, that suspect get to go free ($T$) whereas the other gets sent to jail for a very long time ($S$). If both provide evidence on each other, they both go to jail but for less time than if they had not provided evidence ($P$). In this game $T > R > P > S$. The normal form matrix for the game is shown in Figure 3.
Given that $T > R$ and $P > S$, Defect is a pure ESS. See text for justification of relative payoff values.

As is clear in Figure 3, this is a classic take game in which pure Defect is the only ESS. Cooperation will never evolve even though the pareto optimum of joint cooperation yields a higher payoff than the ESS (e.g., $R > P$). One can imagine many other scenarios in both animals and humans that would also be characterized by this game. As we shall see below, the only way to convert the Prisoner’s Dilemma to a give game with cooperation as one of the ESSs is to alter the type of game that is considered.

**An intermediate: the Snowdrift game**

As we saw in [Web Topic 10.5](#), a single-bout 2 x 2 symmetric contest has four possible outcomes: each of the two strategies can be the only pure ESS, both can be pure ESSs, or there may be a stable mixture of the two strategies as the ESS. The Prisoner’s Dilemma game has a single pure ESS: defect. All give games have two pure ESSs, which one is seen depends on initial conditions or if populations are finite, on relative bases of attraction. The Snowdrift Game describes a cooperation context in which the ESS is a mixed strategy (Hauert and Doebeli 2004; Nowak 2006a). The scenario begins with two commuters trapped in deep snow on the highway. If both get out and shovel snow, they can get home sooner and split the effort. However, if one lets the other do all the work, he still gets home but evades all the work. If neither shovel snow, they remain stuck. Assuming a common currency for costs and benefits, let $R$ be the net payoff if both drivers dig and share the work equally, $T$ be the payoff to the driver who lets the other do all the work, $S$ be the payoff to the driver who does all the digging alone, and $P$ be the payoff to both drivers if neither shovels. This results in the same normal form matrix as the Prisoner’s Dilemma except that this time $T > R > S > P$. This is shown in Figure 4:

![Figure 4. Payoff matrix for Snowdrift game.](image)

The best response is always the opposite of your opponent’s response. The only equilibrium is a mixed ESS in which players cooperate a fraction $(S-P)/(S+T-P-R)$ of the time and otherwise defect.
This model gets us part of the way between the take game of the Prisoner’s Dilemma and a cooperative give game, but only part of the way. In addition, it only characterizes a subset of the types of cooperation seen in nature: specifically, those in which there is a common task whose completion benefits both parties. It does not help us with situations in which the two parties do not face a common task.

### Replacing the take game with a give game

As long as there are opportunities to cheat or defect, it appears impossible to get a give game involving cooperation from a single-bout contest like the Prisoner’s Dilemma. In some contexts, we might be able to find stable mixtures of the two strategies (Snowdrift game). However, we can achieve a situation in which cooperation is a pure ESS only if we change the type of game. One solution is to allow participants to interact repeatedly: if early bouts can provide information that allows players to adjust strategies, the resulting sequential game might have an ESS policy favoring cooperation that is unavailable to single-bout contexts. A second way to change the game is to replace dyadic contests with multiplayer scrambles: as we saw in Web Topic 10.5, adding more players can open up ESS possibilities not available to dyadic games. For both approaches, we begin with the assumptions that the relevant populations are infinite in size and strategies are “well-mixed” throughout the population.

- **Dyadic sequential models of cooperation.** Even when treated as a single bout interaction, reciprocity is intrinsically a sequential game. We shall focus here on repeated interactions between the same pair of players. This is called **direct reciprocity.** Direct reciprocity was initially modeled as an iterated Prisoner’s Dilemma game in which a player could elect to cooperate or defect in each successive bout depending on what had happened in earlier bouts. Axelrod and Hamilton (1981) pointed out that any such game based on a finite number of bouts was bound to degenerate into defection by both parties: the last player in an expected finite sequence will always do best by defecting; knowing this, the other player will defect in the penultimate bout, and so on until both parties defect in every bout as in the single-shot Prisoner’s Dilemma. However, if the total number of bouts is unknown, this unraveling risk can be avoided.

Unfortunately, this complicates the relevant models as there are an infinite number of possible choreographies that a player could adopt given an infinite sequence of bouts. Axelrod thus hosted a “tournament” and invited anyone to submit their favorite choreography or rule-of-thumb policy. These were then competed against each other in computer simulations. One policy that won frequently was Tit-for-Tat (TFT): a TFT player always started with cooperate, and then just did what their opponent did on the latter’s last move. If the probability that two players will interact again, \( w \), is constant, the expected payoff of using TFT against other simple policies over an infinite number of successive bouts can be calculated and these expected payoffs compared to similar calculations for alternative strategies in a normal form game matrix. For example, if the probability that two interactants will meet again is \( w \), then the probability that they meet twice more is \( w^2 \), three times more \( w^3 \), etc. Suppose a TFT player is playing against another TFT player. During each bout, both players cooperate. The expected payoff for either player over an infinite number of bouts using the same payoffs for the Prisoner’s Dilemma (\( T \), \( R \), \( P \), and \( S \)) is:

\[
PO(TFT,TFT) = R + wR + w^2R + w^3R + w^4R+ ..... = R(1 + w^2 + w^3 + w^4 +...) \\
\]

Since the infinite series \( 1 + w^2 + w^3 + w^4 +... \) is convergent, it can be replaced with its sum which is \( 1/(1–w) \). Thus the expected payoff for a TFT player repeatedly playing against another TFT player is

\[
PO(TFT,TFT) = R/(1–w) \\
\]

A similar logic can be used to compute the average payoff to a player using a policy of always defecting (AllD) against another who also always defects:
PO(AllD, AllD) = P/\((1-w)\)

The average payoffs for a TFT and AllD player playing against each other are a little trickier algebraically but the same infinite series can be extracted. The resulting normal form matrix for an entire infinite series of bouts is shown in Figure 5:

![Figure 5. Iterated Prisoner’s Dilemma game using Tit-for-Tat (TFT) and Always Defect (AllD) as policies for sequential game with infinite number of bouts.](image)

Since by assumption in a Prisoner’s Dilemma game, \(P > S\), it is easy to show that \(P/(1-w) > S + wP/(1-w)\), so there is always a dot in the lower right corner of the matrix. The question is then whether TFT or AllD is the best response to an opponent playing TFT. TFT is the best response only if \(R/(1-w) > T + wP/(1-w)\). This will be true when

\[w > (T-R)/(T-P)\]

Thus we can turn the Prisoner’s Dilemma from a take game into a give game by allowing an infinite number of bouts, playing TFT against AllD, and assuming that the probability of two animals interacting again (\(w\)) is sufficient high, and/or that the difference in payoffs between cooperating and being exploited (\(T-R\)) is sufficiently small.

While this type of result at least allows pure cooperating to evolve, there are serious caveats. The first is that, at best, a give game has two possible pure ESSs—which is found depends on initial conditions. TFT thus cannot be the only pure ESS possible. Secondly, TFT and AllD are only two of an infinite number of choreographies and policies. TFT may find itself playing against any of these other strategies in dyadic contests: would it generate a give game in all cases? Third, there is no reason why only two strategies should be present: might dynamics change when three or more strategies are available (like the Rock–Paper–Scissors game)? Fourth, errors in assessing the opponent’s last move can destabilize a TFT population. Last, the classical TFT models lock each pair of individuals into long term interactions in which complex accountings must be maintained so that both parties achieve net positive payoffs. A number of authors have argued that these caveats should and do make classically defined direct reciprocity a rare process in nature (Hammerstein 2003; West et al. 2007b; Clutton-Brock 2009). Given the apparent importance of reciprocity in human interactions, a very large number of alternatives to TFT for iterated Prisoner’s Dilemma games have been put forward. Some of these seem too complicated for most animal contexts; however, others (such as Win–stay/Lose–shift (also called PAVLOV) which focuses on the stability of a player’s own payoffs as the sequence proceeds) are general enough that they might be widely applicable. For overviews of these modified models, see Dugatkin (1998) and Nowak (2006a).

- **Multi-player scramble games of cooperation.** There are two main ways in which adding more players to a single interaction can change cooperation economics:
  - **Partner Choice and Biological Markets:** The first alteration is to keep interactions dyadic, but allow a player to choose its partner for any given interaction (partner choice), and decide for how many occasions it will continue to interact with any given player (partner fidelity) (Bull and Rice 1991). While this eliminates the
locked-in cumulative accounting of classically defined direct reciprocity, it imposes a new burden by requiring that players identify some suitable criteria on which to base these decisions.

One obvious criterion is the value to the chooser of the commodity or service that a potential partner has to offer relative to the value to the partner of what the chooser might provide in exchange. If the chooser can provide a rare or highly valued commodity/service, and there are many potential partners, the chooser can be highly selective and pick a partner that will give it the best deal in return. Exchanges based on such selective trades are called biological markets (Noë and Hammerstein 1994, 1995).

Note that the two parties need not exchange the same commodity or service: one species might offer grooming services in exchange for a sharing of food or help in mate defense (de Waal 1997; Barrett et al. 1999; Gumert 2007; Fruteau et al. 2009; Fruteau et al. 2011); males of a bird species might offer newly constructed nest sites in exchange for female matings (Metz et al. 2007); and unrelated floaters may be allowed to join a breeding group in exchange for helping to rear offspring (Gaston 1978; Kokko et al. 2002). Stable cooperative trading only requires that the net payoffs to players of similar status be of roughly equal utility (e.g., a fair trade) (Hoeksema and Schwartz 2003; Andre and Baumard 2011). The relative valuation of the commodities may be negotiated through an extended series of signal exchanges (Patricelli et al. 2011), or by the imposition of costs on the contributing player that would only be borne if the associated commodity were truly of high value (Archetti 2011). Players might also rely on personality traits exhibited by alternative potential partners (Bergmüller et al. 2010; McNamara and Leimar 2010). Players of unequal status can also engage in stable commodity/service exchanges as long as the most exploitable party retains the ability to terminate the exchange and find an alternative partner (Johnstone and Bshary 2008; Wang and Shi 2010). The freedom to seek alternatives can interact with coercion by dominants in complicated ways, resulting in a diverse mix of accountings when asymmetries are present (Buston et al. 2007).

An alternative, and not necessarily exclusive, criterion that players might use to choose cooperation partners is the reputation of each candidate as a reliable and effective partner. Players can rank relative reputations by eavesdropping on prior exchanges involving a particular candidate (Earley 2010), or in humans, through gossip (Sommerfeld et al. 2007; Sommerfeld et al. 2008). Players may even generate a market by competing to acquire the best reputations (Zahavi 1995; Roberts 1998). Costly investments in cooperation by an individual now may thus be repaid later not by the recipient of the original investment (as in direct reciprocity), but instead by third parties who preferentially seek out affiliations with high reputation individuals. A reliance on reputation to determine partner choice and fidelity is thus one way to generate indirect reciprocity. In a variety of experiments with humans, provision of information about the past generosity of fellow players leads to higher levels of sustained cooperation (Nowak and Sigmund 2005).

• Scrambles and Public Goods Games: The second way in which multiple players might change cooperation economics is if the relevant game becomes a scramble instead of a contest. All of the games discussed in prior section of this module are contests: they exhibit first-order frequency dependence in that the average payoff of adopting a given strategy across many interactions depends on the relative abundances of the alternative strategies; however, the payoffs for a particular combination of strategies in a single interaction are fixed. In a scramble, both the long-term average payoffs and those for a given pair of strategies during a single interaction depend on the relative abundances of alternative strategies. Situations in which members of a group contribute investments to create and maintain some common good, and all members benefit equally from that common good, are called public goods games. The payoffs both within and across bouts depend on how much each participant contributed to the public good; this game is thus a scramble. Models of public goods games vary in whether the N-players interact as dyads or instead as multiple players during a given bout. However, all public goods models share the assumption of multiple bouts in which players can change strategies as the game progresses.
While communal efforts might lead to highly synergistic benefits relative to what individuals might obtain on their own, a public goods game is subject to the same risks of defection that we saw with the Prisoner's Dilemma. Consider a shared pasture, called a “commons,” in which all local farmers are allowed to graze their cattle. If everyone adheres to the same small number of cattle, they will all enjoy a pareto optimum. However, if some farmers give in to the temptation to graze more than their share of cattle on the commons, others will suffer and may retaliate by increasing their cattle numbers to restore their shares of the total. Eventually everyone will be worse off as the commons becomes severely overgrazed and nobody’s cattle do well. This is called the tragedy of the commons (Hardin 1968), and is the N-player equivalent of the dyadic single bout Prisoner’s Dilemma. It applies to many joint animal activities including cooperative hunting, joint territorial defense, cooperative breeding, and shared refuge building (Killingback et al. 2006; Bshary and Bshary 2010).

While some cases of public goods behaviors in animals might be explained as by-product mutualism or pseudo-reciprocity (Bergmüller et al. 2007; Connor 2007), many others remain at risk of a tragedy of the commons collapse (Rankin et al. 2007). What can make these types of cooperative efforts evolutionarily stable? Three additional policies have been identified that can help promote the evolution of cooperation in public goods games. Note that these need not be exclusive and some models of public good games explicitly include two or even all three practices concurrently:

- **Punishment**: In a wide variety of experiments on human public goods games, players invariably begin cooperating, but over many successive bouts, defection gradually increases and finally replaces most cooperation. If partway through the game, while cooperation is still reasonably common, players are allowed to add punishment of defectors (or in milder terms, policing), punishment is adopted increasingly as the game progresses and the game usually stabilizes at a relatively high rate of cooperation (Fehr and Gachter 2002; West et al. 2002; Boyd et al. 2003; Frank 2003; Gardner and West 2004; Gurerk et al. 2006; Sigmund 2007). Punishment is widely used throughout human societies to stabilize cooperation (Herrmann et al. 2008). One caveat that arises in relevant models is that punishment is usually costly to the punisher (Dreber et al. 2008). This sets up a potential for second-order defection in which individuals who are otherwise cooperators fail to contribute to the punishing of defectors. Second-order defection can be curbed by second-order punishers, but again, there can be defections at this level as well (Jensen 2010; Sigmund 2010). Punishment is more likely to be stable if punishers are sometimes “corrupt”: that is, they are able to extort some additional benefits given their power or they themselves quietly become defectors and reduce contributions to the public good (Ubeda and Duenez-Guzman 2011). Where overt punishment can be replaced by less costly but credible threats, this can also stabilize punishment (Cant and Johnstone 2009; Cant 2011). In the absence of corruption or existing high levels of cooperation, some other concurrent policy may be required to stabilize punishment. As we discuss below, reputation can often fulfill that role (dos Santos et al. 2011). While punishment can help stabilize an existing population of cooperators, it cannot help a rare cooperative strategy invade a population of defectors (Lehmann et al. 2007; Rand et al. 2009; Forsyth and Hauert 2011).

- **Rewards**: Rewarding cooperators is the opposite of punishing defectors. Unlike punishment, reward can help rare cooperator mutants get established in a population of defectors, but reward is not effective at stabilizing cooperation when cooperators begin to dominate in the population (Hauert 2010; Forsyth and Hauert 2011). This is probably because a punishment policy is most costly when most of the population are defectors and is needed less often when cooperators are common; the reverse is true for a rewarding policy which becomes most expensive when cooperators are common (Sigmund 2010). Thus rewards are more likely to be effective when cooperation is rare, and punishment when it is common. As with punishment, combining rewards with reputation can lead to stable cooperation, although this combination can also generate cyclic and chaotic dynamics (Hauert 2010).

- **Reputation (Indirect Reciprocity)**: The provision of reputation information often stabilizes high
levels of cooperation in human public goods experiments (Milinski et al. 2002; Nowak and Sigmund 2005; Sigmund 2010; Sylwester and Roberts 2010). One particularly interesting set of studies alternated dyadic exchanges where participants could select their levels of generosity, and public goods games where all players could elect how much to contribute to the public good. When information about the players’ relative generosities in the first stage games were made public, reputations played a significant role in stabilizing high levels of cooperation in the subsequent public goods games (Milinski et al. 2002; Semmann et al. 2004, 2005). In a sense, selective choice of partners based on reputation is a form of mild punishment: instead of overt harm inflicted on defectors, the latter are excluded from participating in future exchanges. Selective partnering using reputations is thus less costly than having to inflict physical punishment (Rockenbach and Milinski 2006; Ohtsuki et al. 2009). Despite this, humans often recruit both selective partner choice based on reputation and overt punishment in experimental public goods games. One explanation is that being a punisher might improve a player’s reputation, and this would then compensate for the costs of inflicting punishment.

A number of related models provide additional perspectives. Using three alternative discrete strategies, instead of two, results in more complex dynamics: one model includes cooperators, defectors, and punishers as separate strategies (Szolnoki et al. 2011), and another lets the third strategy be the chance to abstain from any interactions (Semmann et al. 2003). Both models can produce stable cooperation, but also Rock–Paper–Scissors-like cycles depending on parameters and initial conditions. Making the level of cooperation into a continuous variable turns the discrete games above into a continuous one: stable equilibria tend to occur at intermediate levels of cooperation (Zhang et al. 2010). Adding error in assessment of reputations or value of commodities (perhaps due to reliance on rules of thumb), can also change the evolutionary dynamics (Nowak and Sigmund 2005; Brosnan et al. 2010; de Waal and Suchak 2010; Sigmund 2010).

Favoring the cooperation ESS in a give game

The models reviewed above typically assume that focal populations are effectively infinite in size and well-mixed, so that different strategies are encountered in proportion to their global abundances. The payoff matrices for these games usually do not change if these assumptions are violated, and thus the basic outcomes (e.g., a give game) remain. What can change if one or both assumptions is violated is which of the alternative ESSs in the resulting give game is most likely to occur. Finite populations are subject to drift, and the smaller the population, the greater the role of drift relative to selection. The dynamics become stochastic, and this can significantly weaken the usual predictions.

Populations can also have heterogeneous distributions of alternative strategies; such populations are often called “structured” and modeled as networks (see Chapter 15). A well-mixed population is a network in which every individual is potentially linked to every other individual. Heterogeneously structured populations have a much more sparse network than that for a well-mixed population: only some individuals in structured populations are linked to any focal individual. There are two general classes of heterogeneous networks: a) those in which there is at least one path (however convoluted) linking any two individuals (a giant component); or b) not all individuals are linked by a path and thus the population is subdivided into multiple components.

All of the models below assume finite populations; however, they vary as to whether they focus on small population sizes where drift is significant, or sufficiently large ones where drift is not a major factor. They also vary depending on whether their structure is well-mixed, fully but sparsely linked, or subdivided into separate components. We cannot review all combinations here, but will provide a sampling of the more widely cited ones.

- **Small well-mixed populations**: This combination is nicely reviewed by Nowak (2006a). Outcomes depend in part on the strength of selection associated with the relevant game: if the payoffs of the focal game constitute a small part of the player’s total fitness, selection is considered weak; if they constitute a large part, selection is
Consider a generalized single-bout 2 x 2 discrete contest with the normal form matrix shown in Figure 6:

If the population size is very small and stable, strategy A will most likely take over the population if \( b > c \). If the population is stable but large, and selection due to the game is weak, A will most likely take over the population if \( a + 2b > c + 2d \). Suppose A is the TFT strategy and B is the AllD strategy when computed across an infinite number of successive interactions. We were earlier able to identify appropriate values of the payoffs and a probability of repeated interactions that caused this to be a give game (e.g., \( a > c \) and \( d > b \)). As discussed in Web Topic 10.5, any give game will have a mixture consisting of a fraction \( f \) of A players and a fraction \((1-f)\) of B players at which playing either strategy A or B gives the same payoff. We called this equilibrium frequency of A \( f_e \); it depended only on the relative payoff values in the matrix. Combining the computation for \( f_e \) with the condition in finite stochastic populations that makes TFT more likely to take over the finite population, \( a + 2b > c + 2d \), and letting \( f \) be the fraction of cooperators in the population, it can be shown that TFT is more likely to take over the population than AllD if \( f_e < 1/3 \). This is known as the “one-third law” and reflects the relative basins of attraction of the two possible outcomes in finite and well-mixed populations (Nowak et al. 2004). Unlike the infinite population case, rare TFT mutants can invade a finite population of AllD strategists and drive the population all the way to a population entirely of TFT players. Limiting population size can thus make cooperation the more likely ESS without having to invoke reputation, punishment, or other additional policies.

- **Large finite populations with sparse linkage:** Here we assume that each individual has interaction connections with only a subset of the entire population; however, the network consists of one large component and not of relatively independent groups. The effects of this type of spatial structure are usually studied under the rubric of evolutionary graph theory. The simplest cases involve viscous population in which both offspring dispersal and availability of cooperative partners tends to be limited to the immediate neighborhood of a focal animal. The relative roles of fecundity and survival in promoting or hindering the evolution of cooperation in viscous populations are summarized in Chapter 15; readers can seek additional details in Nowak and May 1992; Nowak 2006a; Nowak et al. 2010. The basic gist is that given the right demographic schedules, a pure cooperation ESS is much more likely to arise in viscous populations than in well-mixed ones. Interestingly, viscosity is not the only way that structure favoring the cooperation ESS might be achieved: if animals can preferentially select and modify links in their interaction networks, even when chosen partners are not immediate neighbors, cooperation can again become the more favored ESS (Ohtsuki et al. 2006; Pacheco et al. 2006; Fletcher and Doebeli 2009; Cao et al. 2011; Zhang et al. 2011).

Several authors have re-examined some of the classical games described above but with all players linked in a lattice. Given sufficiently favorable parameters, pure cooperation can spread until it fills the entire lattice, irrespective of whether the payoff matrix is for the Prisoner’s Dilemma or the Snowdrift game. However, under less favorable conditions, a population playing the Prisoner’s Dilemma game evolves into many clusters of pure cooperators that gradually increase in size, whereas Snowdrift economics cause cooperators and defectors to be such frequent neighbors that cooperation is not globally favored (Hauert and Doebeli 2004; Fu et al. 2010).
another set of simulations, a public goods game with punishment was allowed to evolve on a lattice. Local feedbacks created a runaway process in which increased cooperation led to increased and more severe punishment, and vice versa. The result is an extremely strong combination of cooperation and punishment (Nakamaru and Dieckmann 2009).

- **Finite populations divided into many groups**: Dividing a population into relatively isolated groups can significantly change evolutionary outcomes. In one simple model by Killingback et al. (2006), the population is divided into many groups. Within each group, reproduction is a function of a public goods game whose productivity depends on the fractions of some maximal possible investment contributed by group members and the rate at which these investment are synergistically amplified. In each generation, a fixed fraction of offspring from each group is allowed to disperse and settle randomly among the existing groups. This generates variation in group sizes over time, and in the dispersion of individuals with different preferred levels of investment. There is no competition between groups, no punishment or rewards, no reputation effects, and no kin effects. At least for low but non-zero rates of offspring dispersal and synergistic amplification of two to three times investments, enough small groups whose members all have high rates of investment arise each generation to produce more offspring than are produced by the low investors in the population. The fraction of high investors thus increases over evolutionary time resulting in stable participation in the public goods game.

Several other models, again focusing on interactions within small groups, find that realistic parameters can lead to stable *generalized reciprocity*, in which a group member invests in any other group member if the investor’s most recent interaction resulted in it receiving an investment (Pfeiffer et al. 2005). This process can be enhanced if there is some physiological state that can track cumulative effects of recent interactions (Barta et al. 2011), partner choice is assortative rather than random (Rankin and Taborsky 2009), or social behaviors associated with exchanges are learned by copying group mates (Rutte and Pfeiffer 2009). A model by van Veelen et al. (2010) looks at the possible coevolution of group size and levels of cooperation and finds that only certain combinations of group size and cooperation are stable; this suggests that coevolution of group size and cooperation may be key to stable cooperative behaviors.

A final accounting in which division of a population into groups favors within-group cooperation is called *group selection*. As originally defined, the relevant models focused on the possibility that competition between groups can have fitness consequences that swamp out fitness effects due to competition within groups (Wilson 1975, 1990). However, high rates of offspring dispersal between groups can undermine this advantage (for similar reasons, the Killingback et al. model above only works for low rates of inter-group dispersal). The original discrete distinction between individual and group selection eventually evolved into quantitative accountings that partition selection into additive between-group and within-group effects (Wade 1985; Wilson 1997; Wilson and Dugatkin 1997; Kokko et al. 2001; Traulsen and Nowak 2006; O’Gorman et al. 2008; Wade et al. 2010). These are now called *multi-level selection* models. Other authors feel that this approach is simply an alternative accounting to the inclusive fitness method used widely for kin selection analyses (Gardner and Grafen 2009; Leigh 2010). Both accountings usually begin with the Price Equation, and as we discuss briefly in Chapter 9, the two approaches are often two ways to divide up the same pie. To be fair, most of the models we have reviewed in prior sections of this module do not even include competition between groups as a possible factor, whereas multi-level models include all possible factors. The question thus remains how often competition between groups contributes significantly to the overall fitness accountings once the within-group effects (e.g., those discussed in this module and those related to kin selection) have been accounted for. Further work will hopefully resolve this quantitative issue.

**Matching field data to theory**

Many examples of by-product mutualism and pseudo-reciprocity have been cited in the literature; we discuss some of these in a review of game theoretic models for environmental signaling in Web Topic 14.1. On the other hand, both direct and indirect reciprocity, policing, and market models have proved to be less easily assigned to specific biological examples. In part, this is because none of the alternative accountings listed above, this time including kin...
selection and greenbeard biases, need to be exclusive determinants of a given example of cooperative behavior. On the contrary, mixes of economics are likely the rule rather than the exception. For that reason, and because different authors cannot often agree on which accounting predominates for a given phenomenon, we do not provide a list of examples assigned to each accounting here. Instead, the reader is directed to several published reviews that attempt to make assignments including mixtures of accountings to examples of animal cooperation. General reviews include Dugatkin (1998, 2002), Sachs et al. (2004), Silk (2007), West et al. (2007a), Bshary and Bergmüller (2008), Clutton-Brock (2009), and Connor (2010). Melis et al. (2010) compare the mechanisms and accountings that justify cooperation in animals with those in humans.

Further reading

A number of authors have proposed taxonomies or classification trees for the economics of cooperation among unrelated animals. Stark (2010) combines all the classical 2 x 2 models into a common framework, with a particular focus on cases where partial cooperation may evolve in both single-shot and iterated contexts. More details on the tree shown in Figure 1 can be found in Connor (1995), Leimar and Connor (2003), Bergmüller et al. (2007), Connor (2007), Bshary and Bergmüller (2008), Connor (2010), and Bshary and Bronstein (2011). Nowak (2006b) adds spatial pattern effects to the list (but leaves out some of the other accountings). His introduction to spatial and finite game models (Nowak 2006a) is an excellent starting point. Nowak and Sigmund (2005) provide a thoughtful review of indirect reciprocity and its relationship to alternative accountings. Reviews of cooperation economics that compare direct benefit economics with kin selection are provided by Queller (1985), Sachs et al. (2004), Lehmann and Keller (2006), Queller and Strassmann (2006), West et al. (2007a), and Clutton-Brock (2009). Bowles and Hammerstein (2003) provide interesting contrasts in the application of market theory to human versus animal social contexts.

Given that cooperation is often at most a two-ESS give game, it should not be surprising to find that some mutualisms in the past have now degenerated into pure defect states. Sachs and Simms (2006) use molecular techniques to identify examples in which mutualism is now absent, but was likely present in the past. Sanfey et al. (2003) discuss which parts of human brains are involved in selected cooperation games, and Soares et al. (2010) provide similar perspectives on the neuroendocrine bases of cooperative behaviors.

Several recent books try to tie much of this together. The first chapter of Karl Sigmund’s *The Calculus of Selfishness* (2010) provides a good overview; the rest of the book relies very heavily on complex mathematics. Nowak and Highfield’s *SuperCooperators* (2011) is also readable, but somewhat polemical and not always even-handed.

Literature cited


Web Topic 13.2 Examples of Social Integration Signals

Introduction

Social integration signals include a wide range of visual, acoustic, chemical, and tactile signals. Some of the signals are complex and variable in structure so that they can encode individual identity, an important requirement for group-living species with repeated interactions among group members. These signals also serve specific functions, such as synchronizing parental care behaviors in biparental species, facilitating parent and offspring communication, and coordinating group movements. In this Web Topics unit we provide rich media examples of identity signals, pair integration signals, parent–offspring signals, and group integration signals, following the outline in the main text.

Identity signals

Three examples of visual identity signals were illustrated in Figure 13.5. There are of course many more examples of patterns and colors that have evolved to be especially variable between individuals to encode individual distinctiveness. The interested reader should see the article on this topic by Tibbetts and Dale (2007). Individual chemical signals, or signature odors, can only be illustrated with gas chromatograph tracings showing the different components and proportions of these multi-chemical blends (see Clapperton and 1988; Buesching et al. 2002; Smith 2006). Similarly, electric fish also show very slight but meaningful individual differences in electric organ discharge waveforms (McGregor and Westby 1992). Here we focus on acoustic signature signals for a few social species in which individual identity is important. We show the spectrograms of calls from several different individuals. Click on the species name to see an audio–visual movie of the call series.

- **Jungle crow** (*Corvus macrorhynchos*) “ka” contact calls; individual identity is encoded in fundamental frequency, dominant frequency, FM shape, and duration (Kondo et al. 2010). (Audio files Courtesy of Noriko Kondo.)

![Spectrogram of Jungle Crow Contact Calls](image)

- **Dhole** (*Cuon alpinus*) biphonic calls, containing a simultaneously given high-frequency *yap* and low-frequency *squeak*. Each component can also be given separately. The two-component call is given during peaceful group interactions and periods of group movement (Volodina et al. 2006). (Audio files courtesy of Elena Volodina.)

![Spectrogram of Dhole Contact Calls](image)
- **Bottlenose dolphin** (*Tursiops truncatus*) signature whistles; identity is encoded with a unique frequency modulation pattern that is learned or invented early in life and remains constant thereafter (see Sayigh et al. 2007; Sayigh and Janik 2010). (Audio files courtesy of Leala Sayigh.)

**Pair integration signals**

Here are some examples of pair integration signals in a variety of monogamous species, ranging from duets and other mutual displays to greeting ceremonies, courtship feeding, allopreening, and other tactile signals. Most of these examples are avian species. The majority of birds have a monogamous mating system, associated with pair bonds that last at least the duration of a breeding cycle and with biparental care of offspring. Pair bonds are maintained for
multiple years in sea birds, in which effective communication and coordination of parental activities are crucial for successful reproduction and improve as the pair gains breeding experience.

Mutual displays

- Greeting between mates in the closely related crested penguin (*Eudyptes pachyrhynchus*). This species is found in dense coastal forests along the New Zealand coast, in contrast to the rocky coastal and island habitat characteristic of its congeners: [http://www.arkive.org/fiordland-crested-penguin/eudyptes-pachyrhynchus/video-09c.html](http://www.arkive.org/fiordland-crested-penguin/eudyptes-pachyrhynchus/video-09c.html)
- Courtship, nesting, and chick feeding in Adélie penguins (*Pygoscelis adeliae*). This species typically lays a two-egg clutch which is incubated for about 34 days by both parents. A stint of incubation lasts an average of 12 days while the other parent is away foraging: [http://www.arkive.org/adelie-penguin/pygoscelis-adeliae/video-09b.html](http://www.arkive.org/adelie-penguin/pygoscelis-adeliae/video-09b.html)
- Pair bonding in the Northern royal albatross (*Diomedea sanfordi*). This very long-lived seabird (up to 40 years) doesn’t begin to reproduce until 8 years of age, and then breeds every other year, producing a single-egg clutch that requires 80 days to incubate and 240 days to fledge: [http://www.arkive.org/northern-royal-albatross/diomedea-sanfordi/video-09c.html](http://www.arkive.org/northern-royal-albatross/diomedea-sanfordi/video-09c.html)
- Courtship display in ostriches (*Struthio camelus*). The male forms a bond with one female who performs the diurnal incubation. He also attracts additional females who lay eggs in the pair’s nest but do not contribute toward incubation. The male incubates at night. This video clip shows the wing waving used by both sexes during courtship and the female mating solicitation display: [http://www.arkive.org/ostrich/struthio-camelus/video-09a.html](http://www.arkive.org/ostrich/struthio-camelus/video-09a.html)
- Duetting in Neotropical wrens of the “*Thryothorus*” taxon; this website shows a variety of spectrograms with accompanying audio files from the article by Mann et al. (2009). Species vary in how precisely male and female components are timed. Some species are cooperative breeders with complex choruses in which the males all sing one part in unison while the females sing a second part in unison: [http://media.brill.nl/behaviour/146/1/](http://media.brill.nl/behaviour/146/1/)
- Duetting by a pair of purple-crowned fairy-wrens (*Malurus coronatus*). Most of the *Malurus* fairy-wrens are cooperative breeders with subordinate helpers of variable relatedness to the breeding pair. Duetting occurs primarily or solely between the breeding pair. Extra-pair paternity is high in most species, but is low in the purple-crowned fairy-wren. This has been attributed to strong fidelity by the female and constant mate guarding by the male, and not to duetting, male courtship feeding, or frequent copulation (Hall and Peters 2008, 2009): [http://www.cell.com/current-biology/supplemental/S0960-9822%2807%2901263-8](http://www.cell.com/current-biology/supplemental/S0960-9822%2807%2901263-8)
- The precision of duets by magpie lark (*Grallina cyanoleuca*) pairs increases with age and experience; more precise duets are more threatening territory defense signals. This website contains two audio files from the article by Hall and Magrath (2007), the first one represents a well-coordinated pair, the second one a poorly coordinated pair (download the supplementary document to see spectrograms of these sounds): [http://www.cell.com/current-biology/supplemental/S0960-9822%2807%2901263-8](http://www.cell.com/current-biology/supplemental/S0960-9822%2807%2901263-8)
Courtship feeding: a common behavior in many species during the period of pair bond formation and courtship, in which the female may give a begging display much like the juvenile begging display and the male then feeds her. This behavior is not only a male mate attraction and copulation solicitation signal, but it also provides additional nutrition to the egg-laying female.

- Courtship feeding in crested kingfishers (*Megaceryle lugubris*): [video](http://www.youtube.com/watch?v=WkHX9p7vWnA)
- Courtship feeding in the blue-eared kingfisher (*Alcedo meninting*): [video](http://www.youtube.com/watch?v=jcknqisq_cs)
- Male black robin (*P Petroica traversi*) courtship feeding his mate: [video](http://www.youtube.com/watch?v=bZCOS7gYDQw)
- Courtship feeding in the woodpigeon (*Columba palumbus*) as a prelude to copulation: [video](http://www.youtube.com/watch?v=DPsfly1ej_o)

Mammalian examples

- A pair of hoolock gibbons (*Hoolock hoolock*) duetting together; the male has the dark fur color, the female has the buffy color. The monogamous pair jointly defends a stable territory and males often carry the offspring: [video](http://www.arkive.org/western-hoolock-gibbon/hoolock-hoolock/video-13.html)
- Titi monkeys (*Callicebus* spp) mate monogamously for life and live in nuclear family groups, with the pair’s accumulating offspring. The male performs more parental care than the female, carrying the young and bringing it to the female only for suckling. They often rest side by side with tails entwined, a tactile group bonding signal: [video1](http://www.arkive.org/western-hoolock-gibbon/hoolock-hoolock/video-13.html), [video2](http://vimeo.com/14735750), [video3](http://www.youtube.com/watch?v=U5-MP8nZlsU)
  They also duet or chorus as a pair or family group: [video](http://macaulaylibrary.org/audio/flashPlayer.do?id=88995)
- Male California mouse (*Peromyscus californicus*) grooming his mate; this is one of the few rodent species that forms monogamous pair bonds:
Examples of offspring begging signals: Here we show the offspring begging signals in a variety of species, primarily avian. In some species, adults bring a single item that can be given only to one offspring in a brood at a time; often the largest or closest chick is fed. Competition among chicks can be strong in such species. In other species, the adults bring a large amount of food, either multiple items or a semi-digested bolus that can be given to several or all of the young during a single provisioning trip. Note how in these species the parent attempts to feed each offspring.

- **Single-item feeders**
  - **Starling chicks** (*Sturnus vulgaris*) begging and being fed; note that only one chick is fed at a time): [http://www.arkive.org/european-starling/sturnus-vulgaris/video-09b.html](http://www.arkive.org/european-starling/sturnus-vulgaris/video-09b.html)
  - **Blue tits** (*Parus caeruleus*) feeding their large brood of begging chicks: [http://www.arkive.org/blue-tit/parus-caeruleus/video-09c.html](http://www.arkive.org/blue-tit/parus-caeruleus/video-09c.html)
  - **European robin chicks** (*Erithacus rubecula*); being fed (fecal sac removal keeps the nest clean): [http://www.arkive.org/robin/erithacus-rubecula/video-09c.html](http://www.arkive.org/robin/erithacus-rubecula/video-09c.html)
  - A second clip of bee-eater chicks showing intense sibling aggression when chicks vary significantly in size. In cavity and burrow nests, parents always enter from the same side and larger chicks can position themselves on this side and gain a significant feeding advantage: [http://www.arkive.org/white-fronted-bee-eater/merops-bullockoides/video-17b.html](http://www.arkive.org/white-fronted-bee-eater/merops-bullockoides/video-17b.html)

- **Multiple-item and regurgitation feeders**
  - **Bullfinch** (*Pyrrhula pyrrhula*) feeding chicks; this species is a seed regurgitation feeder, note that all chicks receive some food: [http://www.arkive.org/bullfinch/pyrrhula-pyrrhula/video-09.html](http://www.arkive.org/bullfinch/pyrrhula-pyrrhula/video-09.html)
  - **Skylark** (*Alauda arvensis*) chicks begging and being fed. Multiple items are brought and several chicks are often fed in one trip, note the care taken by the parent to spread the food around: [http://www.arkive.org/skylark/alauda-arvensis/video-09.html](http://www.arkive.org/skylark/alauda-arvensis/video-09.html)
  - **Begging by green-rumped parrotlets** (*Forpus passerinus*). Most parrots produce medium-sized clutches and begin incubation before all eggs have been laid, resulting in highly asynchronous hatching and large variation in nestling size. Parents make a special effort to inspect and feed all chicks, especially the smaller ones (Krebs 1999). This video clip was taken inside an artificial PVC tube nest occupied by a wild pair of birds: [http://www.vimeo.com/9117349](http://www.vimeo.com/9117349)
  - **Scarlet macaw** (*Ara macao*) feeding chicks. Note the absence of conspicuous begging in this case, possibly associated with the small clutch size, egalitarian provisioning, and low level of competition between the chicks: [http://www.arkive.org/scarlet-macaw/ara-macao/video-09a.html](http://www.arkive.org/scarlet-macaw/ara-macao/video-09a.html)
  - **Begging in Humboldt penguins** (*Spheniscus humboldti*). In many seabirds, parents forage for a long period at sea and store multiple food items in their crops; the chicks reach into the parent’s open mouth to grab the regurgitated semi-digested items: [http://www.arkive.org/humboldt-penguin/spheniscus-humboldti/video-09b.html](http://www.arkive.org/humboldt-penguin/spheniscus-humboldti/video-09b.html)
King penguin chicks (*Aptenodytes patagonicus*) in a creche beg to any passing adult. Adults call to locate their own single chick; note the two voice components of the adult call:

Gouldian finch chicks (*Erythrura gouldiae*) have a striking visual display, with blue phosphorescent beads on the sides of the beak, to guide the provisioning adult in the dark covered nest:
http://www.arkive.org/gouldian-finch/erythrura-gouldiae/video-09b.html

Coyote (*Canis lantrans*) female arriving at the den, she greets the pups, they beg for regurgitated food by nuzzling her mouth, and the pups fight over the food:
http://www.arkive.org/coyote/canis-latrans/video-09c.html

Burying beetles (*Nicrophorus* spp) adult and larvae:
http://www.youtube.com/watch?v=yRarHKpOf_8&NR=1 and
http://www.youtube.com/watch?v=httCgCeOpAQ

Parental directive signals

- **Red grouse pair** (*Lagopus lagopus*) guarding and leading precocial chicks:
  http://www.arkive.org/red-grouse/lagopus-lagopus/video-sc09b.html
- **Emperor goose** (*Chen canagica*) calling and leading chicks:
  http://www.arkive.org/emperor-goose/chen-canagica/video-09b.html
- **Parent American coots** (*Fulica americana*) leading their young chicks from the nest out to the pond. Note the bright red head coloring on the chicks, which has evolved as a consequence of parental preferences for feeding more ornamented chicks, and also assists parents in recognizing their own chicks in cases of conspecific brood parasitism (Lyon et al. 1994; Krebs and Putland 2004; Shizuka and Lyon 2010):
  http://www.vimeo.com/994897
- **Female impala** (*Aepyceros melampus*) calling out her young faun from the creche to suckle; the faun has learned to recognize its mother’s call:
  http://www.arkive.org/impala/aepyceros-melampus/video-09c.html

Group integration signals

Here are some examples of group integration signals, divided into categories of appeasement signals and group coordination signals.

Appeasement signals

- **Toque macaques** (*Macaca sinica*) allogrooming:
- **Black macaques** (*Macaca nigra*) allogrooming:
- **Red colobus** (*Piliocolobus badius*) grooming:
  http://www.arkive.org/red-colobus/piliocolobus-badius/video-05.html
- **White throated capuchins** (*Cebus capucinus*) grooming, including rubbing leaves on fur:
  http://www.arkive.org/white-throated-capuchin/cebus-capucinus/video-05a.html
- **Mandrills** (*Mandrillus sphinx*) allogrooming; note that in most of these incidents of allogrooming dominant male receives grooming by a subordinate male or female:
  http://www.arkive.org/mandrill/mandrillus-sphinx/video-05.html
- **Impala** (*Aepyceros melampus*) self grooming and allogrooming:
  http://www.arkive.org/impala/aepyceros-melampus/video-05.html

Badgers
Meles meles allogrooming:
http://www.arkive.org/badger/meles-meles/video-05b.html

Plains zebra (Equus quagga) allogrooming:

Aggressive and submissive behavior in wolves (Canis lupus), plus howling, a territorial and group integration signal:
http://www.youtube.com/watch?v=T8AFMq-j3w8&feature=related

Teeth chattering in the Barbary macaque (Macaca sylvanus) serves several different functions. It is a submissive signal given by subordinates toward more dominant individuals, it is used to de-escalate aggressive encounters, it signals affiliation and likelihood of positive social interaction or tolerance, and it is given by dominants as a reassurance gesture of non-aggression (Wiper and Semple 2007):

Social interactions in hamadryas baboons (Papio hamadryas) are focused around the dominant male, who aggressively chases off intruders and controls the females. Other troop members behave submissively around him, averting their gaze or presenting their bottoms. There is virtually no courtship by the male; estrous females present their bottoms to be inspected by the male and copulation ensues:

Group coordination signals

- Group aggregation in the Emperor penguin (Aptenodytes forsteri). All nearby individuals cluster and huddle in a tight pack to withstand adverse weather conditions. The first section of the video clip shows courtship and nesting behavior and an incubation exchange, and the ending section shows chick begging and provisioning. The two-voice contact call can be heard throughout the clip:

- Meerkat (Suricata suricatta) playful and defensive behaviors, including sentry duty:

- Group decision-making in macaque troops. This video shows how one individual gives a signal to move, and then looks to see how many other individuals agree (“vote”) to follow; a minimum number of group members must agree before the whole troop moves. The audio narrative is in French:
  http://www.youtube.com/watch?v=lCDTE9cVfJ4

- European starling swarms (Sturnus vulgaris). Excellent footage and explanation of swarm formation in this common species; coordination is based on cues, not signals, as explained in Chapter 15:
  http://www.youtube.com/watch?v=XH-groCeKbE

- Snow goose (Chen caerulescens) flock take-off; some of the preflight signals can be seen here, such as directional head orientation and walking:
  http://www.vimeo.com/16582104

- Dominance and leadership in horse herds. This insightful horse trainer explains the subtle group dynamics within a herd of mares:
  http://www.youtube.com/watch?v=tqI546leL6Y&NR=1

- Another interesting clip on horse social signals, showing herd movement coordination, aggressive, and submissive signals in the wild horses of the Camargue in southern France:
  http://www.vimeo.com/20403005

- Group movement and coordination signals in spinner dolphin groups (Stenella longirostris). The first ten minutes provides excellent footage and explanation of various acoustic signals such as tail, head, back, and side slaps, spinning leaps, and echolocation signals:
  http://directoryofkauai.net/blog/hawaiian-wildlife-and-spinner-dolphins-3

- Honeybees. As discussed in Chapter 13, honeybees use several signals to coordinate the fissioning of a colony
and the departure of the queen and part of the colony, first to an outside swarm, and then to a new home (Rangel and Seeley 2008; Rittschof and Seeley 2008; Rangel et al. 2010).

- The **waggle dance** is given by a successful forager upon returning to the hive and signals the angle and direction to the food source to other foragers. This well-narrated clip explains how the dance components encode this information:  
  http://www.youtube.com/watch?v=4NtegAOQpSs

- The **shaking signal** recruits more receiver bees to take the nectar from the incoming foragers. In this clip, the central bee moves around inside the hive and directs the shaking signal repeatedly toward different surrounding bees:  
  http://www.youtube.com/watch?v=Nf89PR_R8so

- The **piping signal** is given both by virgin queens to call out other queens for a fight to determine dominance, and by scout bees to activate the workers about an hour before departure to a new home. In this video clip, the queen has been constrained in a cage and several worker bees are giving the high-pitched piping signal to stimulate her to move with them (courtesy of Tom Seeley).

- The **buzz-run display** is produced in the one to two minutes before swarm departure and signals that lift-off is imminent. In this video, the queen of a colony that is about to fission has again been confined in a cage, and several buzz-runners are attempting to stimulate her to join the cluster of potential swarmers. The buzz-runner approaches lethargic bees in a zig-zag run, makes contact with them, and spreads her wings and buzzes, while pushing them to join the swarm (courtesy of Tom Seeley): <<Link to buzzrunning video in Bee folder>>

**Literature cited**


Introduction

As discussed in Chapter 13, females often use different signals and cues to ascertain whether a candidate male mate is a conspecific versus whether this male is of higher or lower quality than other candidates. However, in the end, a female’s evaluations of these two questions get melded into a single decision: she accepts him as a mate or she rejects him. This is but one example of decision-making in which multiple criteria must be evaluated before a choice can be made. For example, consider a forager that encounters a new food find: it will likely scan the abundance of food items, the average item quality, the degree of predator risk in that site, and the likelihood of finding better alternative sites before deciding to eat there. Two individuals facing a potential escalation during a conflict may try to assess the other’s relative body size, the condition of its armaments, the results of prior contests with the same individual, the proximity of coalition members, etc. When the cues or signals being evaluated reflect the same specific property in the sender, they will be correlated and thus redundant. Here, we are interested in cues, signals, and signal components that are used in a single decision but are largely uncorrelated and not redundant; they provide different information. This means that they might provide discordant information and that generates a cognitive challenge: what is the optimal way for a receiver to make a decision when one source of information favors one choice whereas another source favors the opposite choice? Because the issue was raised in Chapter 13 for females that are choosing mates, and this is the context that has largely been examined theoretically, we shall focus below on models of mate choice with conflicting inputs. However, it should be noted that this is a more general cognitive issue that may arise in a variety of animal communication contexts.

Cognitive options

We are interested here in multiple and uncorrelated sources of information about candidate males that a female might assess before deeming the male acceptable or not. These sources will likely include some mix of different male cues, signals, and components in multivariate signals. While several recent models focus on the different components in multivariate signals (e.g., dominant frequency and pulse rate versus call rate in anurans as indicators of species identity versus male quality), most of the models could handle any combination of the different possible sources. Before examining any of these models in detail, it is useful to list the various ways that a female might combine the uncorrelated information from multiple sources before making a decision (Jennions and Petrie 1997; Fawcett and Johnstone 2003; Castellano and Cermelli 2006; Phelps et al. 2006; Castellano 2010):

- **Single-source protocols**: Here, the female relies on information from only one of the sources and invokes only one criterion to make a decision (e.g., species identity); she is indifferent to information from the other source (e.g., once accepted as members of the correct species, males are chosen by chance). Most models assume females invoke a threshold for values of the favored source below which a male is rejected and above which he is accepted.

- **Amalgamated-source protocols**: In this case, observed values of each source’s data are transformed into some common currency and the transformed values from all sources are combined into a single composite score. There are several ways the transformed measures might be combined:

  - **Additively**: The simple algebraic sum of the transformed variables could be used, perhaps weighting each before addition according to relative impacts on fitness. Note that such a sum could be large even if one of the variables had a low value.

  - **Multiplicatively**: The score for a particular candidate male would here equal the product of the transformed values of the source variables. The product would be minimal if any variable were small in value, and maximal when all were large. Note that if stimuli are perceived by females on a logarithmic scale (e.g., Weber’s Law, Web Topic 8.6), the same relative scoring could be obtained by adding the logarithms of
the perceived stimuli.

- **Bayesian updating**: The conditional probabilities that a male was suitable given knowledge of a code and the female’s assessment of its values from each source could be entered separately into Bayes’ equation sequentially (in any order) to generate an updated probability that the male was suitable and then compared to some threshold value (Luttbeg 1996, 2002; Cheng et al. 2007). Unlike additive models, the magnitude of the change in a score after Bayesian updating depends on the value of the score before combining the new information.

- **Separate multiple-source protocols**: Here, the female would assess each source separately. Resolution of contrary outcomes would be resolved by assigning a priority order to the sources: only if a candidate male passed the highest priority criterion would he then be evaluated by the next criterion. There are two likely versions of this approach:
  - **Either**: The female accepts the male if the values of either source trait meet their respective criterion. This usually means assessing one trait first, and if the male passes the criterion, he is selected. If he fails that first criterion, the female then examines the second source and only accepts him if he passes the second criterion.
  - **Both**: In this protocol, the female examines the first trait and then goes on to assess the second trait only if the male passes the first criterion. A male must pass both criteria to be accepted. This is often called a **hierarchical** protocol. Note that this approach might generate outcomes that superficially look very like an additive model in which one criterion was much more highly weighted than the other before combination.

### Some relevant models

Early models of female mate choice tended to assume that: a) choice outcomes were binary (given male trait A then all females would always make the decision B), and b) females made no errors in assessing male attributes. While these related assumptions simplified the models, they are likely to be unrealistic. More recent work assumes that females make errors and thus that the outcome of any female decision will be stochastic. Outcomes are then best characterized along a continuous scale instead of a discrete one. Below, we examine several models that focus on female choice with these latter assumptions:

- **Fawcett and Johnstone (2003)**
  This model focuses on the tradeoff between the costs of female evaluation of male traits and the degree of error likely when assessing any given trait. It is an optimality model that seeks to identify the optimal protocol given a particular cost–error combination:
    - **Basic Assumptions**:
      - Males occur in two classes: suitable and not suitable. They display two traits that can take either of two states: if they are suitable, both traits exhibit the favorable state; if the male is unsuitable, both traits exhibit an unfavorable state.
      - Females are assumed to make errors in evaluating the state of each trait; the probability that a female errs by perceiving a favorable state in one of the traits of an unsuitable male (a false alarm) is assumed to equal that if she errs by perceiving an unfavorable state in a suitable male (miss).
      - Females pay a cost to search and find a male, and when they find one, then pay an additional but smaller cost for each trait that they take time to assess. If they find and accept a suitable male, they get a fitness benefit that must then be reduced by the costs spent in search and assessment. Females continue sampling males until they find one that is deemed suitable.
      - Alternative protocols considered in order of increasing “choosiness”:
        - **Anyone**: make no assessments and mate first male encountered
        - **Either**: assess favored trait first; accept if male appears to meet criterion; if not, examine second
and accept if male appears to meet criterion

—Single Trait: assess favored trait and accept male only if it appears to meet criterion

—Hierarchical: assess favored trait first and if male meets criterion, then assess second trait. Accept only if it meets both criteria in that order.

- Outcomes and interpretations:
  - When assessment costs and accuracies of assessment are the same for two traits, optimal protocol is for maximal choosiness (Hierarchical) when costs of finding new males is low and accuracies are high; as either assessment error or cost of finding new males increases, optimum shifts to each successive level of choosiness eventually arriving at the anyone protocol. Anyone is also favored if suitable males become much rarer than unsuitable ones.
  - When costs of finding males are low to moderate, and assessment costs are equal but one trait is much more accurate than the other, the optimal protocols are Hierarchical for intermediate accuracy combinations and single trait for more extreme ones. In each case, the favored trait is the one with the higher accuracy. As males become more rare, single trait becomes the dominant protocol over a wide range of accuracy combinations. At intermediate costs of finding males, either protocols replace Hierarchical protocols for intermediate accuracy combinations. Interestingly, when males are rare, Hierarchical protocols favor assessing the most accurate trait first whereas either protocols favor assessing the least accurate trait first; when males are common, the exact opposite orders are found for the two protocols. At very high costs of finding males, anyone is again the optimal protocol.
  - When costs of assessing traits are unequal, it is generally optimal for a female to begin by assessing the cheaper trait first.

- Summary: Whether it is better for a female to assess none, one, or two uncorrelated male traits, and if two, in which order, both vary with the costs of finding additional males, costs of assessing each trait, the accuracies with which each trait can be assessed, and the relative abundances of suitable and unsuitable males.

- Phelps et al. (2006)
  These authors ignore the costs experienced by choosing females and analyzed in the Fawcett and Johnstone (2003) paper. Instead, they investigate whether an amalgamated protocol can explain the frequently seen emphases on different male traits for recognition and discrimination studies of female choice. Many authors have cited such data as support for hierarchical protocols. Phelps et al. instead show that amalgamated protocols can produce very similar outcomes to the data cited in support of hierarchical protocols. Based on their model, they then examine data specifically designed to discriminate between the two protocols:

  - Basic Assumptions:
    - Females are assumed to combine information from uncorrelated male traits into a single preference score. How this is done is not examined.
    - Females use these assigned scores to both evaluate whether a given male has a score above a fixed threshold (e.g., species recognition) and compare two males who are both above threshold to identify the one with the higher score for mating (e.g., mate discrimination). A female encountering two males can either pick one for mating (as in forced choice experiments) or reject both (as can occur in free choice experiments).
    - Females make errors in assessment and the probability of assigning a given preference score value to a particular male is assumed to be distributed as a normal (bell-shaped) distribution with the mode equal to errorless judgments.
    - As in a signal detection model, females impose a threshold cutoff score below which no male is
The probability that a female will accept a given male is the total area under the normal curve for that male above the threshold cutoff. As the mode of the curve assigned to a male is increased, both the area under the curve and the probability he will be accepted for mating will increase. The function that describes how this probability increases with male mode will be nonlinear, as it depends on the shape of the function describing female error (in this case, a normal distribution). This aspect of the model is seen as the recognition process.

- The joint probability that a female will assign a given score to one male and another given score to a second male follows directly from the corresponding normal curves assigned to each male. The result can be graphed with the assigned value to one male on one axis, the assigned value to the other male on an orthogonal axis, and the joint probability of this combination on a third orthogonal axis. The volume under the resulting surface can be partitioned into those combinations of assigned male scores that lead to acceptance of one male by the female, those that lead her to accept the other male, and those that cause her to reject both. These volumes correspond to the overall probabilities that each of those outcomes will occur given the modal values for each male and the patterns of female error. These probabilities, which describe the discrimination process, are also nonlinear functions of male mode values. As noted by the authors, the discrimination function for a given male includes the special case for his recognition function when the modal value for a second male is very low relative to that of the first. That is, both recognition and discrimination tasks can be accommodated within a single continuous discrimination function.

- **Outcomes and interpretations:**
  - The authors argue that an apparent better fit of experimental data to hierarchical than to amalgamated protocols can be an artifact of performing forced choices that do not allow females to reject both males.
  - A second reason why hierarchical protocols have seemed to fit data better than amalgamated protocols is that the power of the tests applied by females varies depending on how close the evaluated male’s scores are to the recognition threshold. Recognition tests (comparing a single male’s score to a threshold value) are statistically more powerful than discrimination tests when tested male scores are close to the female’s recognition threshold. Similarly, discrimination tests (comparing two male’s scores to see which is larger) are statistically more powerful when males are far above a female’s threshold. This difference in statistical power can give the impression that the recognition and discrimination processes are hierarchical and separate when in fact, they need not be: an amalgamated protocol would give a similar result.
  - The authors then identify several tests that are designed specifically to generate different outcomes for amalgamated and hierarchical protocols. These both show strong evidence in favor of amalgamated protocols. These tests and their outcomes are reviewed in more detail in Chapter 13.

- **Summary:** A simple model incorporating female error and assuming an amalgamated female protocol is both capable of explaining apparent emphases from animal experiments on some traits for recognition and some for discrimination, and provides a better fit to data obtained in experiments specifically designed to generate different outcomes for different protocols.

- **Castellano and Cermelli (2006)**
  In this model, the authors assume that female choice always begins with a recognition step. If a discrimination step is involved, it is assumed to occur after the recognition process. In this sense, they assume a hierarchical ordering of steps. They then consider several different protocols for how females might implement a discrimination step.
  - **Basic Assumptions:**
    - Male advertisement signals are allowed to vary along multiple uncorrelated dimensions.
    - Females collapse the multiple-dimensioned male signals into a one-dimensional scale of recognition scores. The score for each male is then compared to a threshold value and males with less than
threshold values are rejected as unsuitable. Because females can err in assigning these scores, the outcome in this step is treated as a probability of accepting a male.

- If two males are deemed acceptable, the female then applies a second discrimination step to choose between them. This step might use the same score created for the recognition step, or it might require the assignment of a different set of scores based on the same signals.

- The model focuses on computing the probability that a focal male will be chosen as a mate when it and a competitor are both encountered by a female. There are two possibilities: a) the focal male is found acceptable during the recognition step but the other male is not; or b) both males are found acceptable and the female must then apply a separate discrimination function to select between them. The probability that a will occur is the product of the probability that the focal male is accepted during the recognition step, and the probability that the other male is rejected. The probability that b will occur is the product of the probabilities that each male will be found acceptable during recognition. The net probability for b is then this product times the probability that the focal male will be chosen during a discrimination comparison. The net probability that the focal male will be chosen is then the probability from a plus that for b (since this is an “OR” situation).

- Selection on the recognition function is assumed to be stabilizing (e.g., there is an optimal male trait value and males that are increasingly different from this optimal form are increasingly likely to be rejected. Selection on the discrimination function, if one is applied, is assumed to be directional: the more extreme a focal male’s signal in one direction, the more likely he is to be selected.

- The authors then examine an example in which each male’s signal has two uncorrelated components: one is more useful for recognition, and the other is more useful for discrimination. They examine two ways females might use these two components: a) multiplicatively amalgamating transformed values of each component’s values into a single composite score; or b) keeping the transformed values of each component separate and using them in different steps.

- They then consider three protocols:

  — One step amalgamated protocol: In this case, females use the amalgamated score for the recognition step, but do not bother with a discrimination step: if two males are both acceptable, one of them is chosen by chance. There is thus no discrimination function needed.

  — Two step amalgamated protocol: Females use the same amalgamated score for both a recognition step and a discrimination step. This is similar to the Phelps et al. model above.

  — Two step separated protocol: Females use the score extracted from the first component for the recognition step, and that extracted from the second component for a discrimination step. This is a classical hierarchical model in which only males that pass the first recognition filter are then considered for a discrimination contrast.

• Outcomes and interpretations:

By varying parameter values in simulations, the authors found that:

- Holding the directional selection component fixed, both two-step protocols provided a female with much more accurate recognition performance than did the one-step protocol.

- Holding the recognition selection fixed, the two-step separated protocol provided the most accurate discrimination between males, followed closely by the two-step amalgamated protocol; the one-step protocol, not surprisingly, provides only weak discrimination between males.

- In both two-step protocols, the degree of choosiness in the recognition step could be used to fine-tune the accuracy of the discrimination step; however, fine-tuning of the amalgamated two-step can be done in a way that does not decrease recognition accuracy whereas this is not true for the separated two-step protocol.
**Summary:** This model confirms the Phelps et al. conclusion that the same amalgamated scores of multivariate male signals with uncorrelated components can be used for both recognition and discrimination with high accuracy. Hierarchical models are only marginally more accurate, and may impose costs in other performance measures. As with both prior models, having access to multiple uncorrelated signal components can improve the accuracy of female choice.

- **Castellano and Cermelli Random Walk Models (2009, 2010)**
  In these papers, the authors compare the outcomes of female choice using additive amalgamation versus multiplicative amalgamation protocols. Note also that none of the prior models allow for refinement of female assessments due to cumulative sampling, even in the Fawcett and Johnstone model where females sample males successively until one meets the suitability criterion. As we discuss in Chapter 8, cumulative sampling is a common strategy that can improve the accuracy of decision makers. Here, the authors examine the two amalgamation protocols in a random-walk model in which females are able to sample repeatedly before making a choice.

  - **Basic Assumptions:**
    - Multivariate male advertisement signals contain two uncorrelated components: C1 is most useful for recognition, whereas C2 tends to be more useful for discrimination.
    - Females assign a single-dimensioned attractiveness score to each male by transforming each signal component’s value into a common “attractiveness” currency and then amalgamating the perceived attractiveness values for the two components. They can amalgamate transformed components additively or multiplicatively.
    - In a recognition task, successive sampling of the same male is assumed to generate a running sum of its attractiveness score values. The recognition step ends when a female accepts a male as suitable because his running sum exceeded some threshold, or rejects the male because some set maximum number of samplings is exceeded.
    - In a discrimination task, the differences between the perceived attractiveness scores of two males given successive samplings are added cumulatively (using the same individual’s score as the subtracted term). Sampling stops when the absolute value of this sum exceeds some threshold value: if that final sum is negative, the subtracted individual is chosen; if the sum is positive, the other male is chosen. The probability that any given male will be selected is a sigmoid function of the threshold value and the average difference between its score and that of the other male.

  - **Outcomes and interpretations:**
    - Because the recognition signal component is under stabilizing selection, two males that are both dubbed suitable will have similar values for C1. This means that C1 will contribute little to the difference between two males’ scores used in a discrimination step if amalgamation is additive. Additive amalgamation causes discrimination to depend mostly on the differences in the males’ values of their C2 component.
    - When multiplicative amalgamation is the basis of score computation, the running sum of differences between male signals depends on the values of both C1 and C2. Depending upon the error level with which females assess C2, the strength of the discrimination preference for one male over the other can increase continuously as the average values of C1 for the two males increases, or even peak at some intermediate value of C1. The C1 component can thus amplify the differences between males in the C2 component during female discriminations.
    - This predicts that if amalgamation is additive, the strength of female discrimination between males should be dominated by C2 values and be relatively independent of C1; if amalgamation is multiplicative, changing C1 values should change the strength of discrimination for any given difference in C2 values between the males. The author then presents experimental data on an Italian frog that supports the multiplicative amalgamation protocol.
Summary: This publication extends both the theory and the data in support of amalgamation over separated and hierarchical protocols in female choice of mates using uncorrelated multivariate signal components, and provides evidence that multiplicative amalgamation is the more likely process.

Perspectives

The examined models have provided a good start in understanding the optimal strategies for females choosing males, and more generally, deciders incorporating multiple sources of information into a single decision. Comparisons between these models and data support female choice protocols that amalgamate information from different sources into a single score before undertaking recognition and discrimination tasks. However, there are still many questions that have yet to be examined. While relative costs of finding mates and assessing them formed the focus of the Fawcett and Johnstone (2003) models, costs were not considered by any of the other models considered here. Combining costs with alternative protocols might be of value; an initial step in this direction can be found in a recent model by Castellano and Cermelli (2011). Although both Phelps et al. (2006) and Castellano and Cermelli (2010) mention possible Bayesian updating, this protocol for combining information across multiple sources and over time has yet to be formulated in a way that can be compared to the other protocols described here (but see Luttbeg 2002). A common focus in the models reviewed here is on multiple and uncorrelated sources of information; redundant multivariate signals were excluded from consideration. There are good reasons to believe that optimal multivariate signals have components that are partially but incompletely correlated (Ay et al. 2007). Examples that seem to fit this circumstance include the color intensity and size of genital swellings in receptive female baboons (Higham et al. 2008), and the individual compounds in complex pheromonal mixtures of lizards that indicate different but not always uncorrelated aspects of overall health (Lopez et al. 2006). A more general model than those considered here might vary the degree to which multiple information sources provide correlated information.

Further reading and examples

Each of the Phelps et al. and Castellano and Cermelli papers compares their models to their own favorite anuran species. Clearly, one would like additional contrasts on these species, and broader contrasts across other species. Among other experiments on anurans, Baugh et al. (2008) show that as many as seven continuously-varying components in male calls are combined by females into their threshold recognition tasks. Schul and Bush (2002) find that two closely related tree frog species both perform recognition tasks based on pulse duration and inter-pulse interval in male calls: in one species, these are amalgamated additively, and in the other they are amalgamated multiplicatively. Burke and Murphy (2007) show that females in another species of the same genus of tree frogs amalgamate male call rate and call duration data multiplicatively before accepting or rejecting mates.

Several studies have identified tradeoffs in the relative weighting of recognition criteria and discrimination criteria during mate choice. For example, crickets appear to emphasize recognition criteria over discrimination traits (Scheuber et al. 2004). Spadefoot toads also stress recognition tasks over discrimination tasks where populations historically overlap with related species, but may not if heterospecific overlap is uncommon (Pfennig 2000). Novel or rare hybrid Drosophila may be preferred over conspecifics if they carry a sufficiently favored discrimination trait; here the discrimination step can overrule the recognition one (Boake et al. 1997). In guppies and sticklebacks, the size and intensity of male colored areas and the rates of display can all affect female choices but with different weightings depending on contexts (Kodric-Brown and Nicoletto 2001; Kunzler and Bakker 2001; Karino and Urano 2008; Karino et al. 2010). Other examples are reviewed in Chapter 13.

Some additional theoretical treatments provide further insights. Candolin (2003) provides a broad review of why multiple criteria might be favored in female choice. The issues of costs and how these might affect optimal female decision protocols have been examined by several authors. Chittka et al. (2009) focus on the general problem of speed versus accuracy in decision-making, Bateson and Healy (2005) conclude that assessment costs and time pressures are likely to favor heuristic decisions by females and multiple criteria may facilitate these heuristics. Wiegmann and Angeloni (2007) examine the role of female search and assessment costs in setting suitability thresholds when not all fitness-
determining traits in males are assessable by females. While they focus on a single male trait, they point out that the outcomes will be similar if multiple traits are combined during female assessment into a single amalgamated score.

**Literature cited**


Karino, K. and Y. Urano. 2008. The relative importance of orange spot coloration and total length of males in female


The bigger game

The history of evolutionary game modeling of parental care strategies is unusually convoluted. Initially, researchers tended to focus on three separate games as if they were independent. The first involved conflict between the parents as to which one, if either, should do any parental care (Dawkins and Carlisle 1976; Maynard Smith 1977; Grafen and Sibly 1978). This initially was called the “desertion game” where the discrete alternative strategies for either parent were to fully commit to or do no parental care; later models were continuous games in which each parent could contribute a level of parental investment from a wide and continuous range. The second game recognized that high parental investment in any one brood would likely decrease a parent’s remaining lifetime resources for later broods. Since offspring in any given brood favored high parental investments, whereas parents were likely to be selected to hold back, the result was a parent–offspring conflict game (Trivers 1974; Macnair and Parker 1978; Parker and Macnair 1978, 1979). This was treated as a sequential game for the parents and ESSs were optimal policies by which a parent should distribute its resources over successive broods. Finally, offspring within the same brood are likely to compete with each other for whatever food and parental care was available (Macnair and Parker 1979; Harper 1986; Parker et al. 1989). The degree of kin relatedness among nestlings may set a limit on how selfish any one offspring should be (Hamilton 1964).

It soon became clear that these games were not independent, and strategic decisions by any one player in one game could significantly affect the options and payoffs of the same or other players in one of the other games. The earlier analyses are best thought of as sub-games within a much larger game. As reviewed by Parker et al. (2002b), the simplest global game is a two-role asymmetric scramble with more than one player in each role (e.g., an N-person game). The two asymmetric roles are parent and offspring. The game is an N-person scramble (instead of a dyadic contest) because the payoffs within a single interaction to a player in one role depend both on the relative abundances of alternative strategies in its own role as well as the relative abundances of alternative strategies in the other role. The sequential aspects of the parent–offspring conflict game are usually simplified by finding a single optimal policy that applies to all successive broods over a lifetime. The overall game can be diagrammed as in Figure 1:

---

**Figure 1.** Network of interactions defining larger parent–offspring conflict game.
Blue boxes are the players, double-headed arrows show the directions of conflict among them, and beige ovals show the traditional sub-game levels of conflict. (After Parker et al. 2002.)

When there are asymmetric sub-types within either or both of the parent and offspring roles, the game is even more complicated. For example, asynchronous hatching can result in nestlings of different body sizes and thus different competitive abilities when begging. If one sex is more likely to defect and leave parenting to the other, this generates an additional asymmetry.

There are so many conflicts in this game that the combined “battleground” is large (Godfray 1995a). These conflicts arise because the allocation of parental care that would maximize fitness for one party is invariably an allocation that is suboptimal for other parties. As with many evolutionary game models, the resulting ESSs are usually nobody’s optimal outcome; instead, ESS equilibria occur where each type of player’s current strategy is the best response to the best responses of all of the other types. At an ESS, each type of player will have its own best response: for offspring, that is a level of begging directed toward each type of parent, and for the parents, it is the amount of parental care (including feeding) provided to each type of offspring. An ESS will thus specify the best response for each of the types in the game.

There may be more than one ESS in a game this complex. In addition, these ESSs may be swayed toward or away from any given player type’s optimum as contexts change. If the optimum is an intermediate value, an ESS that pushes a party’s best response to either a lower or higher value will decrease its fitness below the maximum possible and thus be suboptimal. The ways in which the ESSs shift have been examined theoretically for the following contextual contrasts:

- One parent (usually male) deserts leaving the other to undertake all parental care versus both parents share in the raising of offspring
- Offspring are all closely related (same father and mother), only partially related (different fathers due to extra-pair copulations or different mothers due to egg dumping), or not at all related (nestling shares nest with offspring of nest parasite).
- Parent(s) are likely to produce many versus only a few broods per lifetime
- Parent(s) determine how much food will be provided to a given brood (supply determines demand) versus offspring begging determines amount of food a given brood receives (demand determines supply)
- Parent(s) control how much each offspring is fed versus offspring interactions determine the amount each offspring is fed
- Offspring compete with initial strong asymmetries (dominance effects) versus offspring initially have similar competitive abilities (resulting in so-called scramble competition)
- Offspring signal need honestly to parents versus do not signal

We shall take up some of these specific cases below.

**Some nuts and bolts of begging models**

There are several commonly invoked methods used to find ESSs in begging models. Nearly all analyses use a marginal value type of analysis to identify the optimal investment that should be spent in any one episode (whether foraging on a patch, defending an ephemeral resource from competitors, or investing in a particular brood). The optimal investment is the one that maximizes the realizable rate of benefits relative to costs (Parker 1974; Charnov 1976). On a plot of benefits versus costs, this can be drawn as a straight line with positive slope (Figure 2). Lines with increasing slopes but the same intercept represent lines with higher marginal benefits (i.e., a small increase in cost results in a higher increase in benefits). The problem is to find which line has the highest slope but is also realizable given the phenomenological relationship between benefits and costs. The latter is usually drawn on such a plot as an increasing
but asymptotic (or rapidly decelerating) function. The function may or may not have an initial minimal investment that is required before any benefits accrue: this is shown on the graph of the function as a straight line along the horizontal axis. The point on this phenomenological relationship that is just touched by one of the possible straight lines yields the optimal rate of benefit to cost that is feasible. If the animal uses this criterion for all successive episodes, it will maximize its net fitness.

Figure 2. Marginal value analysis of optimal strategy. (A) Beginning at some initial benefit level with no investment, benefits are assumed to increase linearly with costs. Blue dashed lines show three possible benefit/cost rates: higher marginal rates of benefit per unit cost produce lines with steeper slopes. (B) Phenomenological relationship between benefits and costs for some real situation. Initial function is flat because it requires some minimal initial investment before any benefits accrue. If the situation is foraging, the flat portion might represent the investment required to find a new food patch; if the situation is breeding, the investment might be building a nest; if the investment is parental provisioning, a minimal amount of food must be given to an offspring before it has any chance of surviving. (C) To find the optimal investment, given the phenomenological function, different possible rate lines are drawn on top of the phenomenological line. The high rate line never touches the phenomenological curve, so is not realizable. The low rate is realizable, but not the highest realizable rate. The middle line just touches the phenomenological curve and is thus the highest realizable rate. The corresponding investment (where vertical dotted line intercepts horizontal axis) is the optimal investment.

In begging models, it is assumed that fitness benefits (the vertical axis in the relevant marginal value plots) increase with increasing parental investment but with diminishing returns for both parents and offspring. Hence these models have decelerating phenomenological curves. The shape and height of the phenomenological curves can vary depending upon the various contexts listed earlier. In addition, both the slope and the intercept point for the possible benefit/cost lines can vary with contexts. For example, the slope of the lines will be less if only one parent instead of two cares for the young, but the intercept will be higher in single parent cases if this also creates a greater range of possible sibling relatedness values (Parker et al. 2002b). The task for a particular context is thus to identify the correct phenomenological curves and the correct rate lines.

These efforts only define the optimal amount of parental investment for each type of party in the bigger game. Where the optima for different types are different, there is conflict of interest over how large an investment parents should undertake. As noted earlier, the fitness payoff to any given role (parent or offspring) depends on what strategy it adopts, what strategies its fellow role members adopt, and what strategies are adopted by members of the other role. At an ESS, all members of any given role should adopt the same strategy. Thus at the ESS, the payoff for a member of a given role depends only on what strategy all of its members adopt, and what strategy all members of the opposite role adopt. At an ESS, selection does not favor either role changing its current strategy. Thus the first derivative of the continuous function describing lifetime fitness as a function of strategy value for a role will equal zero if all members of that role adopt the ESS strategy value. Since the first derivatives of the lifetime fitness functions for parents and offspring both contain terms representing their own and the other’s strategy values, and both equal zero at the ESS, there are two equations with two unknowns, so one can solve for the ESS strategy for the parents and the
corresponding strategy for the offspring. In many cases, this ESS pair of strategies will not sit on either role’s optimum as defined by the marginal value analysis. A graph of two optima and battleground region for the traditional parent–offspring conflict is illustrated in Figure 13.18 of the main text.

Other assumptions shared by most models include:

- Asymmetric game with continuous strategy sets: usually levels of begging for the offspring role and sensitivity to begging behaviors for the parent role.
- Parental investment in the current brood erodes parental resources available for future broods. The problem for the parent is undertaking the optimal number of breeding episodes, and then distributing their finite total available investment among them in an optimal way (see Chapter 9).
- Offspring fitness depends on the parental investment they receive directly, and that provided to siblings (both in the same brood and in their parents’ future broods) discounted by their relatedness to those siblings. Increasing one’s own share of the parental investment increases its own fitness but at the expense of the indirect fitness it might have received through its kin. The optimal balance depends on the shape of the functions relating offspring fitness to parental investments.
- To prevent begging from escalating to excessive levels, many models include an increasing cost for higher levels of begging. The cost may be an individually experienced cost such as higher energetic expenditure, or a cost shared by the whole brood, such as increased risk of attracting predators.

Below we describe several different parent–offspring conflict models, including the classical scramble competition and honest signaling models, tug-of-war models, quantitative genetics models, parental negotiation models, and a sibling negotiation model (see also Mock and Parker [1997] and Godfray and Johnstone [2000] for reviews of some of these models).

**Parent–offspring scramble competition model (Parker and Macnair 1979)**

**Context:**

Offspring compete for access to food brought by parent by giving conspicuous displays.

**Synopsis:**

- Siblings do not differ in competitive ability, but instead engage in scramble competition for parental attention. The offspring’s level of begging or jostling varies over a continuous range. The higher the level, the more food the individual receives relative to siblings.
- One parent provisions the offspring. The sensitivity of the parent to the offspring’s behavior varies over a continuous range from completely insensitive to highly sensitive. For any level of sensitivity greater than zero, the parent uses a matching rule, giving the offspring that behave more vigorously a larger amount of food relative to offspring behaving less vigorously. In other words, the parent passively feeds the offspring presenting the greatest stimulus.
- Begging is costly for the offspring, either energetically or by way of attracting predators, and therefore reduces offspring survival.
- The model assumes that offspring respond to higher levels of obtained resources by reducing the intensity of their begging or jostling behavior.
- Variants of the model consider monogamous or polygamous mating systems so that offspring are full or half sibs, respectively; one or two attending parents; and intra- or interbrood conflict. In the case of intrabrood conflict, the parent’s total investment for the whole breeding attempt is fixed, and an increase in feeding to one offspring in a brood reduces feeding to other offspring in the same brood. In the case of interbrood conflict, an increase in total parental investment for the current brood decreases parental survival or investment in future broods.
Conclusions:

- The optimal food intake rate for the offspring is greater than the optimal provisioning rate of the parent. This difference establishes the battleground for this subgame. The difference is greater when offspring are half sibs compared to full sibs, when there are two parents compared to one parent, and for intrabrood compared to interbrood conflict.

- At the resolution ESS, the parent provides more food than is optimal for it, but less food than the optimum preferred by the offspring, i.e., the solution is a compromise lying somewhere between the two optima.

- Because of the competition between offspring in the same brood, each individual offspring is selected to demand a greater share of the food brought to the brood than its sibs are selected to yield. Extravagant begging displays or sib–sib aggressive behaviors are expected to evolve to manipulate parents into bring more food than they would prefer.

- The compromise ESS predicted by this model depends on there being a moderate cost to the offspring of begging or jostling. If begging is extremely costly, such that the fitness of an offspring that solicits without receiving a reward drops off steeply, begging vigor must remain at a minimally low level, and the ESS provisioning level can shift to either the parental optimum or the offspring optimum. The parent-wins ESS is more likely to occur when begging is energetically costly, when the parent pays little cost for being insensitive, and with intrabrood conflict. The offspring-wins ESS is more likely to occur with high shared begging costs (i.e., risk of predation) and with interbrood conflict. In this context, Zahavi (Zahavi 1977b; Zahavi and Zahavi 1997) proposed the blackmail hypothesis to explain the function of loud begging (i.e., the parent is obliged to feed the offspring to keep their begging at a low level). However, it is competition, not blackmail, that pushes offspring to beg more vigorously.

Related models:

- The three earlier articles in the Parker-Macnair series (Macnair and Parker 1978; Parker and Macnair 1978; Macnair and Parker 1979) establish the different parental and offspring optima for different mating systems and types of brood conflict (intra or inter), but do not compute the resolution ESS as in Parker and Macnair (1979). The fifth article in the series (Parker 1985) considers the resolution outcome when one versus two parents perform the provisioning, and finds that conflict is greater (offspring are more demanding) when two parents collaborate.

- Harper (1986) developed a parent–offspring coevolutionary model in which the offspring can vary its rate of begging and rate of change in begging intensity with changes in parental investment level, and the parent can vary its provisioning investment with changes in begging intensity. These were assumptions in the Parker-Macnair model, but are allowed to evolve in this model. At the ESS, offspring reduce their begging when parental investment is high, and parents increase their provisioning rate when offspring begging rate is high. This means that any fluctuations in parental investment are compensated for on subsequent feeding trips. Total begging intensity of the brood should increase with brood size, but in larger broods begging intensity per offspring is lower.

- Godfray and Parker (1992) also expanded the Parker and Macnair model to examine the role of brood size. Larger brood size results in increased sibling competition. Parents may be able to increase their fitness by reducing brood size, in order to reduce this cost of competition.

Asymmetric sibling competition (Parker et al. 1989)

Context:

Offspring differ in competitive abilities.

Background:
The Parker-Macnair models were expanded to examine what happens when offspring differ in competitive abilities. This model was stimulated by the frequent occurrence in birds of asynchronous hatching, caused by the parent starting incubation before all of the eggs have been laid. The result is variation in nestling age and size. Older nestlings often gain a significant advantage over their nestmates, and in many cases younger nestlings fail to receive sufficient nutrition and die, especially if food availability is low. The question addressed is: how selfish should stronger sibs be?

**Synopsis:**

- Parental investment in a given brood is fixed, and only one breeding event is considered, so this model is based on intrabrood conflict only. Two time scales are considered: the whole period of parental care for the brood, and the short-term period (one day or a single feeding event). Parental sensitivity does not evolve in this model, so it is strictly a sib–sib conflict subgame.
- Offspring differ in competitive ability, and allocation of food depends on the asymmetry in offspring behaviors. Two types of asymmetries are considered in separate models.
- In the hierarchical model, siblings form an age or size-based dominance hierarchy at no cost to the dominants. Dominants physically outcompete subordinates for the food. Two- and three-sibling broods were modeled separately. The alpha (A) sib has first access to food resources brought by the parent, and the beta (B) sib gets what is left over. With three siblings, the gamma (C) sib gets what is left over after A and B have taken what they want.
- In the scramble competition model, the two sibs differ in the benefits and costs of display. The begging rate for the siblings is allowed to evolve independently, and each sib chooses its optimal level of display. The parent feeds the sib with the more stimulating target.
- The ESS solution occurs at the point where the marginal gains of the two (or three) siblings are equal.

**Conclusions:**

- **Hierarchical model results:** When food resources are plentiful, the A sib gains little by denying the B sib access to food, so A and B receive nearly equal shares and their fitnesses are similar and close to the maximum attainable. As food resources become more scarce, more food is allocated to the A sib. In the short-term version, A may even allow B to obtain the larger share, if B’s marginal fitness gain is higher.
- **Scramble competition model results:** If there is no cost to begging, begging levels escalate indefinitely and there is no ESS. An increasing individual energetic cost for higher begging levels is therefore required. At the ESS, the A sib expends less energy on begging than the B sib, but A’s display is more effective. As the difference in competitive ability increases, B’s begging costs increase. If food resources are somewhat limited, B’s begging costs may peak at an intermediate level of asymmetry in competitive ability; at this point, B is begging more vigorously than A and receiving a larger share of food, but this benefit is balanced by its greater begging costs. At low food levels, sib A takes the greater share of the food because the B sib cannot afford to display more vigorously.

**Related models:**

- Rodríguez-Gironés et al. (2001) expanded on this model by allowing sibs to employ different forms of competitive strategies. Two offspring strategies are considered—signaling (begging) behavior and non-signaling (aggressive) behavior—which can vary in their weighting. The model also includes a co-evolving parental response and interbrood effects. Increases in begging intensity can cause the parent to increase its rate of provisioning, whereas increases in non-signaling behaviors do not increase provisioning rate but do affect allocation of food between the siblings. In the model with only signaling behavior, more competitive offspring beg less but obtain a greater share of the food. In the model with only non-signaling behavior, parents deliver a fixed amount of food and more competitive offspring obtain a greater share. When both types of offspring behaviors are allowed and combined additively, the result depends on the degree of asymmetry: for low asymmetry, both sibs invest in signaling, and for high asymmetry, the less competitive sib invests in signaling and the more
competitive sib invests in non-signaling behavior. When behaviors are combined multiplicatively, competitive sibs invest more in signaling. Under most scenarios, more competitive sibs receive a greater share of food. The authors conclude that signaling is costly and wasteful (e.g., if a weak sib begs vigorously to induce the parent to bring more food, the stronger sib will just take a greater share of it), so signaling is expected to play a minor role in determining the allocation of food.

Honest signaling model (Godfray 1991, 1995b)

Context:
Offspring transmit information about their need for food to the parent using begging vigor as a signal.

Background:
Begging as a signal of offspring need was a common idea previous to Trivers (1974), but was abandoned in the first series of parent offspring models by Parker and Macnair (described above) that emphasized sibling competition. These models suffered from the fact that they assumed fixed behaviors by parents and offspring—parents respond to increased offspring solicitation by providing more resources, and offspring respond to increased resources by decreasing their solicitation level. A better model would allow these responses to evolve (e.g., as in Harper 1986). The development of honest signaling models by Zahavi and Grafen (Zahavi 1975; Zahavi 1977a; Grafen 1990a, 1990b) revived the idea that begging behaviors could be considered honest signals of need, as long as they were costly. The honest signaling framework allows the investigator to seek the conditions under which begging behaviors can become honest indicators of offspring need.

Synopsis:
- Offspring vary in their condition, a state variable that determines their need for food. An offspring in poor condition gains a higher marginal benefit when it receives food from the parent than an offspring in good condition.
- The (single) parent is selected to give more resources to young in poorer condition. The parent cannot assess condition directly via cues, but must obtain this information from signals given by the offspring.
- Offspring fitness depends on condition, level of solicitation, and amount of food resources obtained from the parent. As in other models, offspring fitness increases monotonically but at a decelerating rate with increasing resources received. The condition (or need) component is included as a constant that determines the rate at which the fitness benefit curve (Figure 2) approaches the asymptote; greater need means that the fitness curve rises faster and is more concave. Solicitation is costly, so fitness decreases linearly with increasing solicitation level. Offspring also obtain indirect fitness benefits from any future offspring the parent raises, discounted by relatedness.
- Parental fitness depends on the number of total offspring raised. An increase in provisioning to the current brood increases the current offspring’s chances of surviving, but decreases the parent’s future reproductive rate.
- The amount of resource obtained by the offspring is some function of the level of display; this function is initially unknown and allowed to evolve.
- In Grafen (1991), there is a single offspring raised per year over a succession of years, so only interbrood conflict is considered.
- In Grafen (1995b), two offspring simultaneously compete for a fixed amount of food brought by the parent, so only intrabrood conflict occurs. The parent adjusts its distribution of food to the offspring until the marginal gains from allocating more food to each offspring are equal. Each offspring is also sensitive to the nestmate’s begging level.

Conclusions:
There is an ESS pair of strategies in which the offspring signals at a rate that is determined by its condition and increases as its condition declines, and the parent allocates more resources to the offspring using the rate of display as an indicator of offspring condition.

Offspring are constrained from exaggerating their need by the cost of begging and the loss in indirect fitness from future sibs.

Honest signaling of need is locally stable as long as begging is costly and the value of obtaining more resources increases with diminishing returns. An offspring in maximum condition does not display at all, as it receives no benefit from doing so.

The ESS resides at the parental optimum; as long as the parent receives honest information of offspring need from the signals, the parent controls food delivery at its optimal level.

The extent of costly begging depends on the variance in the condition component of fitness. If condition varies very little within and between offspring, it is more difficult to achieve the honest signaling ESS.

In the case with two young per brood, if the cost of solicitation differs between offspring with the same need, the offspring for whom solicitation is cheaper will misrepresent their needs and obtain more resources, which moves the ESS off the parental optimum. Furthermore, if one is deprived of food and begs more, the other one increases its begging rate slightly.

Related models:

- Rodríguez-Gironés et al. (1996, 1998) re-examined Godfray’s models and found that a second equilibrium exists in which young do not signal and the parent provisions according to a statistically probable distribution of offspring need states; fitness for both parties is higher at this ESS than at the signaling ESS because signaling costs are avoided. They suggest that begging may have initially evolved through direct sibling fighting in multiple-young broods. If fighting is condition-dependent, parents could use this information even though it was not adapted for communication, and subsequent co-evolution could then lead to the signaling ESS. Payne and Rodríguez-Gironés (1998) suggested another way that signaling might invade the non-signaling state and reach the threshold need to converge on the signaling ESS—a pre-existing receiver bias in the parent which causes them to respond to low-intensity begging behaviors.

- The Sir Philip Sidney game was used to analyze signals of need, first by Maynard Smith (1991) (discrete strategy version) and then by Johnstone and Grafen (1992) (continuous strategy version). Because this game involves a related donor and receiver of aid, it fits the parent–offspring context very well. As with the Godfray models, these models also require that signals must be costly and honest in a way that is related to true need, in order to maintain the signaling ESS. Costs could be lower if relatedness was higher. But a non-signaling ESS was again found to be a likely alternative.

- Bergstrom and Lachmann (Bergstrom and Lachmann 1997, 1998; Lachmann and Bergstrom 1998) objected to the requirement of costly signaling to achieve a signaling ESS. They proposed instead the pooling equilibria model, in which signalers over large ranges of conditions give the same class of signal (described in main text Chapter 10 section on conventional signals).

Models combining competition and need

Background:

Several models combined variation in offspring need and competitive interactions among sibs to explore the interaction of these two factors.

- Motivated by the problem of the strong non-signaling equilibrium in Godfray’s handicap signaling model, Rodríguez-Gironés (1999) expanded Godfray’s model to include competition among sibs. Three main modifications were introduced. First, unlike Godfray’s model, where total parental investment was fixed, in this model the parent can increase its provisioning rate when begging increases and vice versa. Second, offspring
competition, and not begging intensity, determines the allocation of food among the siblings. Third, the begging intensity of an offspring is not affected by the need of its sib. With these modifications, the non-signaling equilibrium disappears and the signaling equilibrium becomes stable, but signaling affects provisioning rate, not allocation.

- Johnstone (1999) expanded the Sir Philip Sidney game to include any number of competitors. Signaling could affect both the likelihood of resources being delivered and the offspring to whom they were given. Each potential recipient knows its own need but not that of its competitors, except that it is drawn from a probability distribution (there is no exchange of information among the competitors). Each recipient independently chooses a signaling level from a continuous distribution, which determines the cost it will expend. All recipients are related to the donor by the same degree. The single donor must decide whether to relinquish the resource, and if so, to which potential recipient based on their signals. If the signalers are honest, the donor should give the resource to the recipient with the most intense signal, who presumably has the greatest need. The outcome is that signaling is more likely to be profitable when \( r \) is low, signaling is more profitable for very needy recipients, and the inclusion of competition decreases the overall intensity of signaling.

- Parker et al. (2002a) examined the interaction between need and competitive ability using an expanded version of the sibling competition model of Parker et al. (1989). As in the 1989 article, total parental input is fixed (parental responses don’t evolve). In addition to their inherent strength asymmetry, an additional asymmetric need or condition variable was added to the fitness equation for each sib. As in Godfray’s model, this component was included as a constant that determines the rate at which the fitness benefit curve approaches the asymptote, keeping the actual value of the asymptote equal for the sibs. Either sib could have the greater need. Costs per unit of begging are equal for the sibs, but the more competitive sib has a more effective begging display in terms of obtaining food. As in the Parker et al. (1989) article, the stronger A sib begs less but obtains more food, especially if it is the needier sib. If the weaker B sib is hungrier, it begs much more intensely and may sometimes obtain the greater share. The authors find that need and competitive ability have similar effects on begging intensity and may be difficult to distinguish in the field (Royle et al. 2002).

**Tug of war models**

*Background:*

Tug of war models were developed as an alternative to the transactional reproductive skew models to understand the division of reproductive benefits among cooperatively breeding individuals (Reeve and Shen 2006; Shen and Reeve 2010). In the general bordered tug of war model, two individuals compete for reproductive shares in a joint breeding effort. Their level of conflict is limited by their options for breeding outside the group. If there are good outside options, individuals will not tolerate a very high skew within the group, otherwise they will leave, so the range of conflict if the group is to remain stable is narrower than when outside options are poor. The limits of cooperation also depend on the relatedness of the two individuals. Once these “borders” for the region of stable cooperation have been established, the two individuals engage in a costly tug-of-war to compete for access to the group benefits. The two individuals are assumed to differ in dominance or competitive efficiency. The dominant can gain more benefits at the expense of the subordinate, but the costly efforts expended by both parties in this war use up a portion of the total group benefit. The ESS solution establishes the final share of reproduction for each party.

No formal tug-of-war model has been developed yet for parent–offspring conflict, but Yamamura and Higashi (1992) developed a model very much in the spirit of the tug-of-war approach. The borders of conflict are determined as before by the parental and offspring optima. This model assumes that both parent and offspring engage in costly behaviors in an attempt to achieve their optimal strategy. Any deviation by one party to move closer to their optimum results in a counter move by the other. For example, if the offspring develops a strategy to obtain more food, the parent counters by minimizing its response to the strategy. The ESS occurs where the marginal benefits of the behavioral deviations by the two parties exactly balance. The final parental investment level will lie somewhere between the two optima; it will be closer to the parental optimum if parental control is less costly, and closer to the offspring optimum if offspring can easily manipulate the parent.
Quantitative genetic models (Kölliker et al. 2005, 2010)

Context:
Genetic models evaluate the coevolutionary trajectories of parental and offspring traits and the likelihood of convergence.

Background:
When traits in one individual are influenced by traits in another individual, indirect genetic effects can play a role in their evolution. The phenotypic models discussed above do not take into account the genetic architecture of parent and offspring traits and the mutual correlations between them. For example, the fitness of offspring is affected by the combination of their own genes and the specific nesting environment generated by the parent’s genes; the fitness of parents is similarly affected by the nature of their responses to the stimuli presented by the offspring traits. Certain combinations of offspring genes and parentally generated environments can result in high offspring fitness, causing particular genes expressed in parents and offspring to become genetically linked. The first attempt to apply quantitative genetic models to parent–offspring conflict found that selection on offspring traits alone can lead to linkage disequilibrium and indirect selection on maternal traits, so that the offspring trait becomes coadapted with the parental performance for that trait (Wolf and Brodie 1998). However, this model did not capture the reciprocal nature of parent-offspring interactions.

Synopsis:
- In the models by Kölliker et al. (2005, 2010), selection acts on individuals during both the offspring and the parental life stages. Solicitation behavior is expressed and subjected to selection during the offspring stage, and provisioning behavior (i.e., sensitivity to solicitation behavior) is expressed and subject to selection for those individuals that survive to breed. These traits are envisioned as evolving reaction norms.
- Selection at these two life stages is antagonistic, as offspring benefit from obtaining more resources while parents lose in terms of future fitness if they provide more resources. Lifetime fitness is modeled as survival in the offspring stage, a function of the amount of parental investment received by offspring as shown in Figure 2, and fecundity in the adult stage, a function of the lifetime number of offspring successfully raised. Offspring do not differ in competitive ability or need, there is no cost of solicitation behavior, and all offspring in a brood receive the same amount of resources.
- In the 2010 article, a third variable, baseline parental provisioning, is included as an evolving trait, along with offspring solicitation level and parental sensitivity to solicitation. A deterministic model was developed, and then stochastic simulations with mutational variation examined the changes in parental and offspring behaviors over time.

Conclusions:
- Baseline parental investment stabilizes at an intermediate level.
- Offspring solicitation and parental sensitivity become genetically correlated through co-adaptation and mutual indirect effects, dictated by the proximate mechanisms that regulate offspring demand and parental supply.
- The simulations revealed a zone of equivalent co-adaptation outcomes between solicitation level and parental sensitivity to solicitation, with these two behaviors changing rapidly and in concert over time. The dynamic character of antagonistic parent–offspring conflict may explain the diversity of provisioning strategies in different species.

Related models:
- Early genetic approaches to PO conflict employed simple two-locus models. Such models can examine
trajectories and find evolutionary end points, but require assumptions about genetic mechanisms that are usually not known. In the model by Stamps et al. (1978), one locus determined the offspring’s level of signaling—one allele codes for a nonselfish offspring that signals for food in direct proportion to what it needs, and the other allele codes for a selfish offspring that signals for more food than it needs. The other locus determines the parent’s provisioning response to the begging signal—one allele codes for a rule of feeding in direct proportion to the signal, and the other allele codes for providing less than requested. Not surprisingly, offspring are more likely to be nonselfish (honest) when the costs of signaling and selfish behaviors are high.

- In another two-locus model by Feldman and Eshel (1982), the offspring locus determines the degree of sibling competitive interaction, and the parental locus determines the degree to which the parent interferes with offspring competition. As in Stamps et al. (1978), this model leads to a range of outcomes that depend on the specific costs and benefits and net selection on the genotypes.

- Eshel and Feldman (1991) expanded on the above model with a two-locus model that examined the conditions for the spread of handicapped (honest-signaling) offspring, compared to offspring without the handicap. They allowed for linkage between the parental and offspring loci. Handicapped offspring could spread in the population if linkage was low, but not when linkage was tight. Costly signaling in this model is favored when the offspring survival function is very concave (rapid rise to the asymptote) and parents have low fertility (small brood size).

**Sibling negotiation (Johnstone and Roulin 2003)**

**Context:**

Siblings negotiate over which one should obtain the next provisioned item.

**Background:**

The models outlined above considered begging as either a signal directed at parents to inform them of the offspring’s hunger level, or a form of scramble competition to gain access to the provisioning parent. Another function of begging could be to communicate with other siblings, and this interaction could even occur in the absence of the parent. For example, offspring could communicate their need to other sibs, so that less hungry sibs reduce their begging when the parent arrives with food. Or they could indicate that their need is high and that they will compete more intensely to obtain the next meal. This model thus deals exclusively with the sib–sib subgame.

**Synopsis:**

- The model assumes there are two sibs, who compete to acquire the food brought by the parent. The sibs communicate before the arrival of the parent by begging to each other, and adjust their begging and competitive behavior when the parent arrives based on the outcome of the negotiation. The model thus implies that the parent does not have complete control of food allocation.

- The offspring differ in both need and competitive ability (size). Begging is costly, but the cost of begging at a given level does not differ for the two sibs. However, the more competitive offspring can more effectively monopolize access to the parent with a given level of display than a less competitive offspring. A factor in the model specifies how easy it is to monopolize the parent, which is high in hole-nesting species where a stronger sib can push its way to the entrance and block food access by other nestmates, and low in open-nesting species where the parent can arrive from any direction.

- The hungrier sib obtains a larger marginal benefit from receiving the food than the less hungry sib. The need, or value of the food, for the two offspring varies independently according to a probability distribution.

- The model allows the level of investment in begging to the sib to evolve. Once informed of each other’s needs, they then use this information to decide how vigorously to beg once the parent arrives. The analysis is a two-step process: first, to determine the level of begging in the second stage given that the sibs have provided full information about each other’s needs, and then to determine the conditions under which the begging in the first
stage can honestly reflect true need.

Conclusions:

- When the food cannot be easily monopolized and offspring do not differ in competitive ability, offspring are predicted to inform the sib of their true need during the negotiation stage, and the less needy offspring is predicted to reduce its begging level during the second stage while the needier sibling continues to beg very vigorously. Half sibs are predicted to invest less in negotiation than full sibs.

- When offspring differ in competitive ability, the stronger sib is predicted to reduce its begging effort less in relation to a weaker and hungrier sib, but in the reverse situation when the stronger sib is hungrier, the weaker sib is predicted to reduce its begging effort more.

- Negotiation breaks down when it is relatively easy to monopolize the food (the stronger sib will compete for it without bothering to negotiate). Negotiation also breaks down when the food item is divisible; when it is indivisible, the siblings are more likely to negotiate and assess each other’s needs.

Parental negotiation (Johnstone and Hinde 2006)

Context:

When two parents share parental care, how should they respond to each other’s efforts?

Background:

As discussed in the main text, male and female parents are expected to disagree about their level of investment in their joint brood, as each would benefit if the other took on more of the costly effort (Trivers 1972; Maynard Smith 1977; Chase 1980; Houston and Davies 1985; Houston et al. 2005). On an evolutionary time scale, whether neither, both, or one parent should care depends on the costs and benefits of staying and leaving for the two sexes. Houston and Davies (1985) showed that biparental care is stable when a reduction in one parent’s effort selects for partial compensation (an increase) by the partner, and vice versa if one parent increases its effort. This compensatory response requires that the partners have some mechanism for communicating and coordinating their efforts. McNamara et al. (1999, 2003) suggest that conflicts of interest between two opponents should be modeled on a behavioral time scale as a series of interactions in which the parties negotiate to reach a resolution. What evolves is not a level of investment, but a stable negotiation rule, which specifies the nature of the reaction norms of each player towards the behavior of the other. Johnstone and Hinde (2006) extend this idea further in a specific model of parental coordination that also includes signals of need from the offspring. This model is therefore a male–female parent subgame with input from offspring.

Synopsis:

- Two parents provide food to their joint brood. Increasing the total provisioning by the combined parents increases total brood productivity, at the cost of future parental fitness. The benefit depends on the need of the brood, which can vary according to a probability distribution function. Parental state (condition) also varies according to independent probability distribution functions for male and female parents.

- Each parent bases its provisioning decisions on its own state (which it knows) and its estimate of brood need (which has some error). Parents may also adjust their provisioning level in response to the provisioning behavior of the partner.

- Payoff to each parent depends on the final outcome of negotiation. The analysis seeks stable pairs of linear response rules, each of which is optimal given the rule adopted by the other.

Conclusions:

- When a parent has only partial information about brood need, greater effort by the mate serves as a signal that
reduces the focal parent's error in the estimate of true brood need. This leads to a matching provisioning response by the focal parent. On the other hand, a net increase in provisioning by the mate results in a reduction in the marginal value of investment, which leads to compensatory response by the focal parent (reduced provisioning). The net outcome depends on the strength of these two opposing forces.

- When the variation in brood need is high relative to parental state, the parents are more likely to match each other’s provisioning rates, and less likely to show a compensatory response.
- Male and female parents might respond differently to each other if there is an informational asymmetry. The parent that is better informed about brood need is predicted to work harder, respond more strongly to changes in brood need, be less sensitive to the cost of provisioning, and compensate more strongly in response to changes in the partner’s provisioning effort. The poorly informed parent is predicted to match the changes in its partner’s effort.

**Related models:**

- McNamara et al. (2003) examined several forms of parental conflict over level of care with temporal patterns of decision-making. They find that when one parent decides about its level of care before the other and costs of care are asymmetric, the offspring may be better off with one parent than with two parents. When the level of care is reached by negotiation, there are again cases in which the offspring do better with one parent, even when costs of care are symmetric. These results differ from earlier models suggesting that two parents are always better than one (Houston and Davies 1985), but they are consistent with other models that found that offspring are likely to beg more competitively when there are two parents (Parker 1985; Mock and Parker 1997).

**Literature cited**


Web Topic 14.1
Models for Environmental Signaling

Introduction

Nearly all of the interactions between individuals discussed in Chapter 14 can be modeled with evolutionary game theory. It is instructive both for understanding the specific interactions and for getting a broader feel for game theoretical modeling to deconstruct the basic logic and outcomes of some of these models. We do not cover all the models cited in Chapter 14: those applicable to the evolution of crypsis, aposematism, and mimicry have been nicely reviewed elsewhere (Ruxton et al. 2004; Mappes et al. 2005; Skelhorn et al. 2010). However, the examples below provide a broad sampling of ESS approaches. They are presented in the order encountered in Chapter 14, and we classify the types of format used according to the schema outlined in Web Topic 10.5.

Predator notification (Bergstrom and Lachmann 2001)

The Question: When does it pay for a prey animal to notify a predator that it has been detected, and when does it pay for the predator to attend to such notification signals?

Background: If predators were to attend to such signals, what would keep a prey from giving the signal whether it has spotted a predator or not? In that case, predators would do better to ignore the signal and then there would be no selection for the prey to give it. There is clearly a conflict of interest between the two parties and some honesty guarantee will be required for communication to be favored. In this model the proposed cost that insures honesty is the risk that giving the signal will reveal the presence of the prey to a predator that had not yet detected it.

Game Format:

The game is obviously asymmetric (prey versus predator).

- The model is based on discrete alternative strategies (prey: signal or don’t; predator: chase prey or don’t).
- It is a single-shot model analyzed in extensive form with successive branch points: nature decides whether a predator is near a prey or not, the prey then decides whether to signal or not, and the predator (if present) then decides whether to chase or not.
- Payoffs at the terminus of each tree branch are fixed (a contest).

Special Assumptions: Only a single prey is considered; this excludes alarm signaling to conspecifics and relative vulnerabilities of multiple prey as factors. The model also excludes signals that indicate prey unsuitability for reasons other than predator detection (e.g., aposematism or relative escape abilities).

Synopsis:

- Prey are exposed to stimuli that may or may not be generated by a nearby predator. They then use Bayesian updating to combine prior probabilities of predator presence with the conditional probabilities of perceiving these stimuli to generate an updated probability that a predator is present.
- Prey can also estimate the risk of capture given the current context and likely predator. The product of this risk and the updated probability that a predator is present provides an overall measure of predation risk. Different stimuli and different occasions result in different overall risks.
- Emitting a predator detection signal costs a prey by increasing its risk of detection by other predators. Prey should only emit the signal if these signal costs are less than the appropriately scaled overall risk of capture. For a fixed average cost, this defines a threshold stimulus level below which no signals should be emitted and above which they should.
The predator's payoff depends on the probability of capturing a prey minus appropriately scaled costs of a chase.

The analysis sought equilibrium (ESS) strategies in which prey only signal if stimuli exceed the threshold, and predators avoid chasing prey that so signal. ESSs only exist if prey that estimate high overall risk (implying a predator is almost surely present) can use this information to reduce their risk of capture. That is, detecting a predator early in its hunt helps a prey to escape a subsequent chase. A related requirement is that prey that are exposed to stimuli below the threshold do not signal. If met, these conditions generate an ESS in which emission of a predator detection signal is relatively “honest” and thus valuable to both parties.

Two specific contexts were considered: in one case, predators always chased prey that did not signal. In the other, predators did not pursue prey until at least some signaled so that they could focus on the ones that did not.

**Conclusions:**

The analysis for both contexts showed that predator detection signals are only likely to be honest if there is a cost to senders (in this case, the risk of attracting other predators). In addition, the benefit to prey of predator detection signals only accrue if a prey is relatively sure that a predator is present, and if this knowledge improves the prey’s chances of escaping if chased. This means that animals that cannot reduce their risks even if they detect a predator early (e.g., sessile or poorly mobile species, animals far from refuges, etc.) should not bother to give predator detection signals. The model also predicts that the threshold for giving predator detection signals should decrease when predators are more common, and the minimal cost that guarantees sender honesty can be smaller if the costs of pursuit to predators are higher.

**Related Models:**

- Getty (2002) expanded on this model by noting that the prey face a signal detection problem in which each seeks to minimize costly false alarms; at the same time, the predator faces an optimal diet problem. Getty computed the relevant “red line” that the prey should adopt along the predator stimulus axis given a signal detection analysis. He also considers what happens when predators have alternative prey to pursue.

- This model only considered a single prey animal. When multiple prey are present, a variety of consequences follow. Those relating to relative vulnerabilities are outlined in the models presented below. However, the simplest change is that instead of giving predator notification signals independently, a group of prey does so in concert. This usually takes the form of group predator inspection behavior. A relevant model for this tactic can be found in Dugatkin and Godin (1992).

- A number of other models consider predator notification signals for other reasons. Readers may want to compare the above model to those in which prey signal to predators honest indications of their physical condition, stamina, palatability, or agility (Nur and Hasson 1984; Hasson 1991; Vega-Redondo and Hasson 1993; Maynard Smith and Harper 2003; Ruxton et al. 2004; Searcy and Nowicki 2005).

**Vigilance behaviors (Sirot and Touzalin 2009)**

**The Question:** What is the ESS level of vigilance in a foraging group if vigilant individuals are more likely to survive a predator attack than foraging individuals?

**Background:** Prey that are vigilant when a predator appears or attacks can often escape sooner and more effectively than prey that are foraging. This is because vigilant prey spot a predator first, know where it is, and thus know which direction would move them away from it. Foragers may only flee after vigilant animals have fled or given alarms, and then have no idea where the predator is. Given these differences, the risk to foragers increases as more and more members of their group become vigilant; the risk to vigilant animals decreases as more and more group members forage. The best strategy for escaping predation is always to be vigilant; however, the animal will then die of starvation. Is there an average level of vigilance which if adopted by all in the group optimizes the tradeoff between predation and
starvation risks?

**Game Format:**

- The strategy set is the fraction of time an individual spends being vigilant instead of foraging. It is thus continuous.
- The game is a scramble because the payoff of adopting a given strategy depends on the mix of strategies adopted by a defined circle of neighbors.
- Although relative location in a group may affect optimal strategy, all individuals in the same neighborhood begin the game with otherwise similar properties. The game is thus symmetric.

**Special Assumptions:** The game is spatially explicit; payoffs depend on abundances of different strategies adopted by local neighbors of a focal animal, not by the population as a whole. The game is played on cellular automaton grids such that each animal (at a node) has a fixed number of surrounding neighbors with which it shares predator detection ranges, alarms, and attack vulnerabilities. No coordination of vigilance based on mutualism, reciprocity, kin selection, or other cooperative economics is included in this game. Each animal just does what is best for itself (although noting what neighbors are doing is a key part of that process).

**Synopsis:**

- Prey animals need to acquire new energy through foraging and avoid being killed by predators. Vigilance is assumed to be incompatible with foraging and vice versa. There is thus a tradeoff in how to allocate actions.
- The payoff currency in this model is the amount of energy reserves attained at the end of the game. This depends on how fast energy can be obtained when foraging and lost when not foraging, and the fraction of time that was spent foraging instead of being vigilant. The potential final reserves are then discounted by the probability of having survived all predator attacks.
- Mortality risks due to a predator’s attack are varied along two axes. The first axis is the ratio between vigilant and foraging animal vulnerabilities: at one extreme, vigilant animals are never killed if there is at least one foraging animal in an attacked group; at the other extreme, vigilant and foraging group members have identical risks. The second axis is related to who, if anyone, first spots an approaching predator: risk is assumed to be least for an animal spotting and responding to an approaching predator, somewhat higher for animals who do not first spot the predator but are secondarily alerted by another member’s predator detection and response, and highest if nobody in the group spots the approaching predator.
- Two models are considered. The first (static) model looked for the ESS average fraction of time that each individual in the group should be vigilant given its position in the group, its initial energy reserves, and the current probability of predator attacks. The second (dynamic) model acknowledges that different animals may elect to be vigilant at different times, thus making the number of vigilant neighbors a stochastic variable. The second model allows a focal animal to track these unpredictable variations and adjust its own vigilance accordingly.
- The static model sought an overall ESS in which each group member adopted an average fraction of time spent vigilant that was optimal given their position in the group and the fractions adopted by other members in the group. At this mix of fractions, nobody benefited by adopting a different average fraction. These equilibria could not be identified analytically, but were identified for each set of initial conditions using asymptotic iterations. Of major interest was how ESS mixtures of fractions changed depending upon initial reserves at the start of the game, group size, and different relative vulnerabilities for foraging and vigilant group members.
- The second model focused on a fixed number of relevant neighbors (12) and a fixed value for initial reserves, and played the cellular automaton game on a torus to eliminate world boundary effects. The optimal (ESS) fraction of time vigilant was computed using the equations for the first model and plotted against successively greater fractions of the group being vigilant. Equilibria occurred whenever the ESS fraction for an individual equaled the fraction of the group that was currently vigilant (using a method called rational reaction sets [Simaan...
Equilibria were stable if a slightly higher fraction of vigilant neighbors benefitted from a lower optimal (ESS) vigilance, and a slightly lower fraction benefitted from a higher optimal vigilance. Any drift away from a stable equilibrium would thus trigger individual adjustments back to the equilibrium point.

- Even if all group members adopted the same optimal fraction of time vigilant, they might or might not synchronize their vigilance with that of neighbors. Simulations of the dynamic model were used to measure levels of spatial and temporal autocorrelation in vigilance for different conditions.

Conclusions:

- The static model predicted greater vigilance fractions for animals located on the outer edge of a group (since they had fewer neighbors to share the risk), for all neighbors given greater initial energy reserves (since they did not need to feed as much), when foragers were at higher risk than vigilant animals (since nobody wanted to be the straggler if attacked), and in contexts where predator attacks were more common.

- The dynamic model made different predictions depending upon: a) the relative vulnerability of foragers and vigilant animals, and b) the differences in risk when at least one neighbor spotted the predator versus when nobody did:
  - **Vigilant and foraging animals suffer similar capture probabilities, but detection of an approaching predator significantly reduces everyone’s risk.** Here, an individual’s optimal vigilance decreased to an asymptote as the number of vigilant neighbors increased. An intermediate level of vigilance was the stable optimum. Instantaneous snap shots of such populations showed little spatial or temporal synchrony in being vigilant. An individual animal’s timing of activity was little affected by that of its neighbors.
  - **Foraging animals suffer higher risks of being killed than vigilant animals, and detection of an approaching predator significantly reduces risk.** High levels of vigilance were optimal when few neighbors were vigilant since this increased chances of the few vigilant animals spotting the predator and alerting all. High levels of vigilance were also favored when most group members were vigilant since being one of the few foraging animals greatly increased risk. When intermediate numbers of neighbors were vigilant, there were enough eyes and ears at work to lower one’s own vigilance, and enough other foragers around to reduce chances of being the only suitable prey if attacked. Stable equilibria occurred at a somewhat less than intermediate fraction of time being vigilant, and another when everyone was vigilant. Instantaneous snapshots of the population showed dense patches of synchronously vigilant animals (since nobody wanted to be the sole vulnerable forager) but no similar contagion of foraging.
  - **Foraging animals suffer higher risks than vigilant animals of being killed, but early detection of a predator has little effect on anyone’s risk.** The only relevant factor here was the greater risk of being killed while foraging when most of the neighbors were being vigilant. A stable ESS occurred at a low average vigilance fraction. These conditions generated contagious patches of synchronous vigilance as in the prior case, but also contagious patches of foraging.

- See Figure 14.13 in the text for graphic comparisons of the three conditions.

Related Models:

- This model built on the results of a number of prior publications on vigilance in groups of animals including Pulliam et al. 1982; Lima 1987; McNamara and Houston 1992; Lima 1995b, a; Lima and Zollner 1996; Bednekoff and Lima 1998b, a; Bahr and Bekoff 1999; Hilton et al. 1999; Lima and Bednekoff 1999; Beauchamp 2007; Pays et al. 2007a; Pays et al. 2007b. Readers may want to consult these earlier treatments after having worked through this one.

- There are several parallel models examining the degree to which animals should coordinate vigilance by observing which neighbors are vigilant (Rodriguez-Girones and Vasquez 2002; Fernandez-Juricic et al. 2004; Jackson and Ruxton 2006).

- Sirot and Pays (2011) recently published a paper using similar logic to identify the optimal amount of time a solitary forager should devote to scanning for predators.
Sentinel behaviors (Bednekoff 1997)

The Question: Why and when should an animal in a foraging group act as a sentinel?

Background: It is not initially obvious why any animal in a group would give up foraging and take on sentinel duty to the benefit of other group members. One answer is that, as suggested by the previous model, being a sentinel is often safer than being a forager. Also following the logic of the previous model, being a sentinel all the time would then be the safest anti-predator strategy, but the consequence would be death by starvation. Is there an optimal mix of foraging and sentinel duty and how might this mix vary depending upon the way in which sentinels and foragers do or do not alert each other when a predator is spotted?

Game Format:

- The strategy set is discrete: individuals can either act as a sentinel or forage.
- The game is a dynamic one: it consists of many successive bouts, during each of which each player decides to act as a forager or a sentinel for that bout. The goal is to identify the optimal choreography or “policy” to follow over a long series of bouts.
- All group members start out with similar resources so the game is symmetric.
- The payoffs are discrete: either an animal survived the series of bouts or it did not. It could die during any bout as a result of either predation or starvation.
- Because the outcome of any bout depends not only on a focal player’s actions in that bout but also on the fractions of the foraging group that have adopted forager and sentinel roles, the game is a scramble.

Special Assumptions: Reciprocity and kin considerations are excluded from this model. By-product mutualism (in which doing what is best for oneself incidentally helps others) does come into play.

Synopsis:

- The model assumes that sentinels are less likely to be killed during a predator attack than are foragers. The ratio of relative vulnerabilities is a critical variable in the model.
- ESS policies for a given set of conditions were derived using both forward and backward induction methods (see Web Topic 10.5 and Mangel and Clark (1988) for background).
- ESS policies were independently derived for different patterns of “information sharing.” Three cases were considered: a) no sharing: detection of a predator by either a sentinel or forager never alerts other group members, either because the detector fails to give an alarm, or because its flight is not detectable by others; b) sentinel sharing: detection of a predator by sentinels but not foragers alerts other group members; and c) full sharing: detection of a predator by either sentinels or foragers alerts other group members. Being alerted reduces predation risk.
- The median and distribution of fractions of the group acting as sentinels were computed for each ESS policy derived.
- The initial model assumed a group size of five animals, four-fold lower vulnerability to attacks for sentinels than for foragers, and 90% probability that foragers would fail to detect approaching predators on their own. These values were later varied in a sensitivity analysis to determine the robustness of the initial results.

Conclusions:

- No information sharing: Here, predation risk for either sentinels or foragers increased as more of the individuals in a group became sentinels. This is because the risk to remaining foragers increases as they become more rare (same outcome as vigilance model of Sirot and Touzalin [2009] above), and once foragers are sufficiently rare, being a sentinel becomes just as risky and it no longer pays to be a sentinel. As a consequence, the threshold in
energy reserve that was required before a forager switched to being a sentinel decreased as the number of sentinels in the group increased. The distribution in number of sentinels over time was highly variable with a mode at 0 sentinels. In short, sentinel behavior is unlikely to pay, even if sentinels have lower vulnerabilities than foragers, when there is no information sharing.

- **Sentinel sharing and full sharing**: These two cases had very similar outcomes. Predation risk when no sentinels were present decreased dramatically to a very low value when one sentinel was present; adding additional sentinels further reduced risk for either foragers or sentinels but at a much decelerated rate. As a result, the threshold energy reserves required before an animal switched from foraging to being a sentinel rose dramatically once one other group member became a sentinel and increased only slightly when higher fractions acted as sentinels. Observed distributions of numbers of sentinels in multiple simulations showed a significantly narrow peak at one sentinel per group. Sentinel behavior can thus be favored as long as information is shared in some way.

- **Alternation**: When information is shared, a threshold ratio between sentinel and forager vulnerabilities can be determined at which a well-fed animal could do equivalently well as a sentinel or forager. This sets the scene for alternation in which different individuals take successive turns in the sentinel role (a common observation in natural groups). Given the other parameters examined in these models, this threshold was always met if the conditions favoring being a sentinel at all were met.

- **Robustness of results**: The optimal fraction of the group that should act as sentinels was independent of group size. Reduced mean food intake when foraging or greater unpredictability of food intake reduced the ESS fraction of the group serving as sentinels. Variation in other model parameters such as predator attack rates, failure of foragers to detect approaching predators, and initial energy reserves had only minor effects on predictions.

- See Figure 14.14 in the text for graphic representation of some of these results.

**Related Models:**

- This author developed the same model further, including the special, but common, case of pairs of animals in Bednekoff (2001).
- The literature citing or building on this paper overlaps extensively with those papers cited for the prior game model (Sirot and Tourazin 2009).
- A large number of field studies have since examined the assumptions and predictions of this model. Citations can be found in the main text.
- Some authors have extended the information sharing aspect of sentinel behavior by noting that many sentinels emit repeated calls while on duty. The evidence that this enhances forager efficiency versus has some alternative function remains conflicting (Bednekoff et al. 2008; Hollen et al. 2008; Ellis 2009).

**False alarm call rates (Beauchamp and Ruxton 2007)**

**The Question**: Why and how often should members of a foraging group attend to false alarms?

**Background**: It is well known that foraging groups of some species experience frequent false predator alarms. Some false alarms are simply due to mistaken stimulus interpretation, but others may be intentional manipulations by some members of a group. Given the relative risks of predator attack and starvation, what is the optimal level of false alarms that can be tolerated by a foraging group, and are there other modulators of responses to false alarms (in particular, how many group members have taken flight) that alarm responders should consider?

**Game Format:**

- The strategy set was discrete: individuals could either forage or be vigilant.
- Individuals did not begin the game identically, but because their parameters were drawn from the same random
distributions, the game was stochastically symmetric.

- Payoffs were 0 if an individual starved or was killed by a predator, and equal to its accumulated energy reserves if a survivor. Because both factors could vary depending on what others in the group did, this game was a scramble.

- Like the prior example, the game was dynamic: within a generation, each individual's lifetime was divided into many successive bouts within each of which it could choose to be vigilant or forage.

**Special Assumptions:** As with prior examples, kin selection and reciprocity were not allowed in the models.

**Synopsis:**

- As in prior examples, individuals were assigned to groups. Different group sizes were examined but held constant for a given analysis.

- Vigilant animals never erred in detecting an approaching predator or by fleeing or alarm signaling to innocuous stimuli (i.e., they never gave false alarms). Foragers however, could err in both ways. They could thus generate false alarms.

- Each individual was haploid and carried three “genes” on a single chromosome: one specified the probability that the individual would be vigilant at any given time, one specified the probability that the individual would flee if it heard one group member’s alarm or saw one group member fleeing, and the third gene specified the probability that the individual would flee when it saw two or more group members fleeing. A model run began by assigning random values to each individual’s genes.

- Unlike the prior example, evolution was then allowed to proceed through many successive generations. At the end of each generation, its members were ranked according to individual payoffs. The next generation was created by: a) replicating the top half of the ranks in the ending generation, and b) replacing the lower half of the prior generation by random choices of individuals from the top half of the rankings. Mutations were then added to each gene at a low rate.

- After a sufficient number of generations, each of the three gene loci tended to converge on a stable distribution with a distinct mode and median. The combination of the three asymptotic medians for a given set of initial conditions defined an ESS. This type of multi-generational analysis is called genetic algorithm modeling (Vrugt and Robinson 2007).

- ESS mixes were identified for a variety of different ambient conditions. Sensitivity analyses were then run to evaluate the robustness of these outcomes.

**Conclusions:**

- The median value of the vigilance gene decreased as group size was increased, and also when the rate at which foragers misclassified stimuli was increased.

- The median value of the gene controlling the probability of fleeing after two group members fled was always higher than that for the probability of fleeing after only one member’s flight. The ESS thus favored reliance on more than one alarmed individual before fleeing.

- The median value for both genes controlling the probability of flight decreased as group size increased, but the reduction was much smaller for the gene controlling flight after multiple alarms. Animals tend to rely on multiple flights or alarms at about the same rate regardless of group size or rates of forager misclassification of stimuli.

- The average fraction of false alarms at the ESS was 20% for a group of six but as high as 55% for a group of 20 individuals. Larger groups thus tolerated surprisingly high rates of false alarms.

**Related Models:**

- Lima (1994) and Proctor et al. (2001) consider similar questions, but include a number of constraints that are relaxed in this model.
Flower marking by bees (Stout and Goulson 2002)

Question: Why and when should foraging bees avoid nectar collection at a flower marked by an earlier forager?

**Background:** Bees visiting flowers remove most of the available nectar per visit. The flower then begins replenishing the nectar at a species-specific rate. The bee will obtain the largest subsequent nectar load if it postpones a return visit until the depleted flower has sufficiently replenished its nectar. Many bee species mark flowers with complex pheromone footprints whose components volatilize at different rates (Goulson et al. 2000). A bee can then monitor the changing composition of a footprint to determine how long a flower has had to replenish. Since different flower species have different nectar replenishment rates, optimal return times will differ for different species of flowers. Bumblebees have been shown to adjust return rates according to flower species. One cost of the time mark system is that other bees, even other species, detect and monitor the marks. A “cheater” bee can drain a flower before the marking bee would normally return. The cheater only gains a partial load, but they at least get this nectar before any competitor does. The resulting arms race could, in principle, undermine the entire marking strategy by forcing competitors to make earlier and earlier visits to marked flowers. For a given flower replenishment rate and density of competitors, is there an ESS that maximizes individual rates of nectar acquisition but preserves the flower marking strategy?

**Game format:**

- The strategy set is the estimated time since a flower was last visited above which a bee will collect the nectar and below which it will reject the flower for now. The set is thus continuous.
- The payoff to be maximized is the rate at which nectar is acquired per unit time.
- All bees have equal access to the same set of flowers and the game is thus symmetric.
- Because the nectar return for a bee adopting a given return time depends on the return times adopted by other bees foraging in the same neighborhood, the game is a scramble.

**Special Assumptions:** The authors further assumed that:

- The time required to remove most of a flower’s nectar (the handling time) was independent of the amount of nectar being removed.
- The bees always removed all the available nectar per visit.
- The energetic costs of flight and handling were similar enough to use time spent as a common currency.
- The interacting bees were unrelated excluding kin cooperation from playing a role.

**Synopsis:**

- The average rate of nectar acquisition was computed based on the search and handling times per flower, the relative fraction of flowers that were currently acceptable (above threshold nectar replenishment), and the average nectar load obtainable from those flowers.
- It was assumed that there were enough bees and a finite number of flowers such that all flowers in the neighborhood had been visited recently. New flower production was thus assumed to be negligible over short time periods. The fraction of flowers that were acceptable thus depended on the rate of replenishment and return visit times used by the average bee.
- If all bees cooperated by observing the same return time, a pareto optimal time could be computed that maximized individual forager intakes for a given flower species. No bee could do better without reducing the intake of others.
• Assuming most bees had adopted the pareto return time, the rate of nectar accrual was then computed for a single “mutant” bee that visited flowers randomly without regard to return time marks. Using observed values for the relevant parameters, the rate of nectar accrual given random visits was usually higher than that for bees respecting the pareto rate of return. This meant that there was ample incentive for some bees to “cheat” by visiting flowers earlier than that proscribed by the pareto strategy.

• An equilibrium occurs when bees using chemical marks to select or reject flowers adopt a return time that yields a nectar accrual rate equal to that enjoyed by bees that visit flowers randomly. If the bees relying on marking were to use a longer time, cheaters would be favored and increase in numbers; if cheaters were to use a shorter period, bees that relied on marks would do better. The equilibrium is thus stable and an ESS.

• The study then compared the actual return times of wild bumblebees to the pareto optimal and ESS predicted values.

Conclusions:

• As expected, average return times to different flower species by wild bumblebees were inversely related to the rate of nectar replenishment.

• For several combinations of bumble species and flower species, particularly flowers with rapid replenishment times, the pareto return time was considerably longer than the ESS time. Where this was the case, the average return time exhibited by the wild bumblebees was much closer to the ESS value than to the pareto value.

• Where the pareto and ESS return times were very similar, the wild bumblebees exhibited average return times very close to the shared value of both predictions.

Related Models:

No related models since this publication were found. However, there remains much debate over whether the “footprint marks” of foraging bees are signals actively deposited on flowers to allow replenishment time monitoring or instead are inadvertent cues. Relevant citations are given in the text.

Literature cited


Introduction

Approximately 760 species of bats echolocate as do 74 species of toothed (Odontocete) cetaceans (Nowak 1999). As we note in Chapter 14 of the text, both groups have exploited this sensory system to master an enormous variety of ecological niches (Thomas et al. 2002; Kunz and Fenton 2003). The degree of coordination that many species display to locate and capture prey is remarkable. Even for those videos that do not translate the ultrasonic echolocation sounds used by these animals down to frequencies that humans can hear, it is still valuable to see how various echolocating taxa use their unique sensory systems in natural and laboratory contexts. Below, we provide links to some illustrative videos across a wide variety of habitats and diets.

Echolocating bats:

- **Tadarida brasiliensis**: Brazilian free-tailed bats are shown leaving one of their large colonies at sunset. These bats echolocate by emitting pulses through their mouths as shown in several close-ups. They then forage at very high altitudes away from any clutter echoes. Surprising numbers of insects are found high above the ground. Pulses tend to be low frequency CF with FM when close to prey. ([http://www.arkive.org/brazilian-free-tailed-bat/tadarida-brasiliensis/video-00.html](http://www.arkive.org/brazilian-free-tailed-bat/tadarida-brasiliensis/video-00.html))

- **Pipistrellus pipistrellus**: These tiny insectivorous bats largely feed in zones of intermediate clutter echoes, often in the vicinity but not right up against foliage or walls. They emit their pulses through their mouths. This clip includes their echolocation pulses translated down to frequencies that we can hear. Pulses are FM. ([http://www.arkive.org/pipistrelle-bats/pipistrellus-pipistrellus-and-pipistrellus-pygmaeus/video-00.html](http://www.arkive.org/pipistrelle-bats/pipistrellus-pipistrellus-and-pipistrellus-pygmaeus/video-00.html))

- **Plecotus auritus**: These insectivorous FM bats are gleaners: they hunt for insects close to the foliage and must contend with high levels of clutter echoes. This clip shows them hunting moths. ([http://www.arkive.org/brown-long-eared-bat/plecotus-auritus/video-08a.html](http://www.arkive.org/brown-long-eared-bat/plecotus-auritus/video-08a.html))

- **Myotis daubentonii**: Daubenton’s bats largely forage over water where they capture insects near the surface or small fish creating ripples just at the surface. They emit their CF-FM pulses through their mouths. This excellent BBC video shows Daubenton’s bats hunting in various circumstances, and concludes with a short clip of the prior species, Plecotus auritus, feeding. The clip converts the bats’ pulses into the audible range for humans. ([http://www.youtube.com/watch?v=p08Y0oRAX3q](http://www.youtube.com/watch?v=p08Y0oRAX3q))

- **Noctilio leporinus**: These large neotropical fishing bats use a combination of CF-FM pulses to detect ripples generated by fish just under the water’s surface. This National Geographic clip shows the bats foraging over rivers and pools. Echolocation pulses were not included (in fact, you may prefer to mute this video entirely when watching it...). ([http://video.nationalgeographic.com/video/player/animals/mammals-animals/bats/bat_fishing.html](http://video.nationalgeographic.com/video/player/animals/mammals-animals/bats/bat_fishing.html))

- **Desmodus rotundus**: The common vampire bat of the neotropics routinely drinks blood obtained from small incisions made in large mammals. The first clip shows vampires trying to feed on a domestic pig ([http://www.arkive.org/common-vampire-bat/desmodus-rotundus/video-08b.html](http://www.arkive.org/common-vampire-bat/desmodus-rotundus/video-08b.html)). A second similar clip ([http://www.arkive.org/common-vampire-bat/desmodus-rotundus/video-08c.html](http://www.arkive.org/common-vampire-bat/desmodus-rotundus/video-08c.html)) shows how the vampires swell up (like ticks!) while drinking. This makes flight difficult, but they have special kidneys that immediately begin removing the water from the ingested blood so that after urination, flight is more feasible. Making the cut without disturbing the host is clearly a tricky business. The third clip shows the bats in their day-roost colonies ([http://www.arkive.org/common-vampire-bat/desmodus-rotundus/video-00.html](http://www.arkive.org/common-vampire-bat/desmodus-rotundus/video-00.html)). Echolocation pulses are slowed down so they are audible to us. Vampires have to eat every other day to survive; if they fail to find blood they return to the roost and beg from successful “buddies” who regurgitate enough to keep their reciprocal partners alive. At least two regurgitation sharings are shown in this clip. Pulses are very high frequency FM.
Mystacina tuberculata: This New Zealand FM bat does most of its foraging on the ground where it runs around as easily as a rodent or a shrew. The following two clips show it foraging (http://www.arkive.org/lesser-short-tailed-bat/mystacina-tuberculata/video-08.html) and both foraging and flying. (http://www.arkive.org/lesser-short-tailed-bat/mystacina-tuberculata/video-00.html)

Rhinolophus ferrumequinum: Greater horseshoe bats are CF bats that use Doppler shifts to detect moving prey. The first clip shows a small group in their roost (http://www.arkive.org/greater-horseshoe-bat/rhinolophus-ferrumequinum/video-03.html), and the second shows a single bat on a hunt (http://www.arkive.org/greater-horseshoe-bat/rhinolophus-ferrumequinum/video-08.html). Both clips include translation of pulse frequencies into our audible range.

Hipposideros diadema euotis: A short clip showing a small group of these large roundleaf bats preparing to take flight. Like the prior horseshoe bats, these bats emit their largely CF pulses through their nostrils and a rounded noseleaf. They typically hang from a perch and make short sallies to “hawk” nearby flying prey (mostly beetles, but the bats occasionally eat birds). (http://www.arkive.org/diadem-roundleaf-bat/hipposideros-diadema/video-00.html)

Leptonycteris nivalis: This new world leaf-nosed bat feeds nearly exclusively on flower nectar for energy and certain types of pollen grains as nitrogen sources. It emits its FM pulses through its nose-leaf. Here it is seen approaching and obtaining nectar from a flower Nectar feeding bats are major pollinators for many tropical plant species. (http://www.arkive.org/mexican-long-nosed-bat/leptonycteris-nivalis/video-08.html)

Various species: The following video with German narration has some excellent slow motion footage of hunting bats with their pulse frequencies translated to our range. You do not need to understand German to appreciate most of this clip. (http://www.youtube.com/watch?v=31WMJMFqUG4&feature=related)

Neotropical bat footage: This short video is narrated entirely in Spanish and ends with a focus on Bolivian bats. Whatever your language skills, it includes footage from around the world with sequences not seen in other sources in our list. (http://www.youtube.com/watch?v=4fKgHhOhP3Y&feature=youtube&noredirect=1)

Echolocation and insect capture: Here are several examples from the laboratory of Dr. Cynthia Moss of an insectivoruous bat capturing an insect showing the trajectories of bat and prey and playing the emitted pulses transposed down to the human audible range:

- **Example 1**: Bat takes direct track to intercept prey. (http://www.bsos.umd.edu/psyc/batlab/insect_capture_trials/1999.11.01.1.07.a.avi)
- **Example 2**: Bat has to make second loop to catch prey. (http://www.bsos.umd.edu/psyc/batlab/insect_capture_trials/1999.10.29.1.12.a.avi)
- **Example 3**: Bat misses target on first pass but quickly circles back and grabs it on second. (http://www.bsos.umd.edu/psyc/batlab/insect_capture_trials/2000.06.07.1.10.a.avi)
- **Example 4**: And here is a last minute detection. (http://www.bsos.umd.edu/psyc/batlab/insect_capture_trials/2000.06.07.1.11.a.avi)

Echolocating cetaceans

- **Basic dolphin sonar intro**: This short clip introduces the basic mechanisms of cetacean echolocation and provides some interesting footage of species detecting hidden prey buried in sea and river bottoms. (http://www.youtube.com/watch?v=ZoNDW0zSRNo&feature=related)
- **Mechanisms of sonar pulse generation in odontocete cetaceans**: While the songs of humpback whales and other mysticete cetaceans appear to be generated by a specialized larynx (Reidenberg and Laitman 2007, 2008), echolocating odontocetes such as dolphins, killer whales, and sperm whales produce their echolocation pulses and various social sounds in the pair of nasal passages that link the blowhole to the pharynx and larynx (Cranford 2000; Cranford and Amundin 2003). In all odontocetes except sperm whales, each of the paired nasal passages hosts “phonic lips” that can occlude that passage but can be forced open when air pressure on one
side of the lips is greater than on the other (see Figure 2.36 in text; Cranford et al. 1996). Perhaps to maximize
the intensity and range of their echolocation clicks, sperm whales only have phonic lips in one highly elaborated
nasal passage. When vocalizing under water, contractions of pharyngeal muscles increase the pressure of
stored air below the phonic lips. This forces the lips to open and vibrate briefly. Because the blowhole is closed,
the air that is forced through the lips collects in an inflatable sac. After producing sounds, muscles around the
inflatable air sac can then force the air back towards the pharynx (without making sounds) where it is available
for the next set of sonar clicks or social calls. Like some birds, cetaceans other than sperm whales have two
relatively independent sound sources. Cranford and colleagues (2011) recently used endoscopy and concurrent
audio recording to verify that bottlenose dolphins (Tursiops truncatus) can produce echolocation clicks using
either set of phonic lips; they can do so either separately or concurrently. Ted Cranford generously provided a
video clip that shows the endoscopic images of the two sets of phonic lips (the right side tinted red and the left
side green) along with oscillographic images of the waveforms of concurrent clicks and social whistles. The
dolphins were trained to produce sounds on command; the voices of the trainer and supervising veterinarian can
also be heard on the sound track.

**Dolphin Dual Sources**

- **Dolphin** (Tursiops truncatus) target discrimination: These experiments with captive bottlenose dolphins
demonstrate how an echolocating cetacean can identify the shape (and/or composition) of targets that it cannot
see. ([http://www.youtube.com/watch?v=51G83jaeNC4](http://www.youtube.com/watch?v=51G83jaeNC4))

- **Lagenorhynchus obliquidens**: First clip shows underwater videos of Pacific white-sided dolphins
([http://www.arkive.org/pacific-white-sided-dolphin/lagenorhynchus-obliquidens/video-00.html](http://www.arkive.org/pacific-white-sided-dolphin/lagenorhynchus-obliquidens/video-00.html)). Initial sequence
includes typical echolocation pulses and final sequences shows dolphins diving through fish schools. Second
clip includes more examples of echolocation buzzes. ([http://www.arkive.org/pacific-white-sided-
dolphin/lagenorhynchus-obliquidens/video-06a.html](http://www.arkive.org/pacific-white-sided-dolphin/lagenorhynchus-obliquidens/video-06a.html))

- **Lagenorhynchus obscurus**: Dusky dolphin group herding and compressing a school of fish before feeding on
them ([http://www.arkive.org/dusky-dolphin/lagenorhynchus-obscurus/video-08b.html](http://www.arkive.org/dusky-dolphin/lagenorhynchus-obscurus/video-08b.html)). A second clip shows
similar behavior ([http://www.arkive.org/dusky-dolphin/lagenorhynchus-obscurus/video-08c.html](http://www.arkive.org/dusky-dolphin/lagenorhynchus-obscurus/video-08c.html)). In the final clip,
the dolphins use their echolocation to examine a southern right whale (which does not echolocate)
audible in all three clips.

- **Sotalia fluviatilis**: This clip shows Tucuxi dolphins foraging for fish. Echolocation sounds were not recorded. This
species lives in estuaries and rivers in northeastern South America and Caribbean sides of Central America, and
must often have to hunt in murky waters with low visibility. ([http://www.arkive.org/tucuxi-dolphin/sotalia-fluviatilis/video-08.html#text=Range](http://www.arkive.org/tucuxi-dolphin/sotalia-fluviatilis/video-08.html#text=Range)).

- **Inia geoffrensis**: The Amazon river dolphin (or boto), like the Tucuxi dolphin, must often forage in murky waters
([http://www.arkive.org/boto/inia-geoffrensis/video-01b.html](http://www.arkive.org/boto/inia-geoffrensis/video-01b.html)). Also like most river dolphins, it has a distinctive head
with a long narrow rostrum and narrowly rounded head melon ([http://www.arkive.org/boto/inia-geoffrensis/video-
00.html](http://www.arkive.org/boto/inia-geoffrensis/video-00.html)).

- **Stenella frontalis**: First clip has great footage and echolocation buzzes of Atlantic spotted dolphins including
several mother-calf interactions ([http://www.arkive.org/atlantic-spotted-dolphin/stenella-frontalis/video-00.html](http://www.arkive.org/atlantic-spotted-dolphin/stenella-frontalis/video-00.html)).
Second shorter clip shows this species hunting flying fish at night when visibility is likely nil.

- **Delphinapterus leucas**: The beluga whale uses its echolocation sonar both to find food and to monitor it position
relative to arctic ice above it. In the following clip, it is hard to hear echolocation clicks, although a short buzz
can be heard in the final sequence ([http://video.nationalgeographic.com/video/player/animals/mammals-
animals/whales/whale_beluga.html](http://video.nationalgeographic.com/video/player/animals/mammals-
animals/whales/whale_beluga.html)). A related species, the narwhal (Monodon monoceros) also forages under
the arctic ice and presumably has similar echolocation pulses. While the following clip only records social
sounds, the animals are so unusual that it is worth watching anyway.
Orca orcinus: Killer whales use their sonar to localize non-cetacean prey, but will also turn off their echolocation if hunting other cetaceans that might detect them before they can start their chase. The following Ocean Adventures video shows killer whales hunting sting rays, seals, sea lions, and whales (http://www.pbs.org/koed/oceanadventures/video/orcas). Two National Geographic clips show killer whales hunting sea lions off the coasts of California (http://www.youtube.com/watch?v=k5M3gs76fzA) and Argentina (http://www.youtube.com/watch?v=DWsN63PRCW8). No echolocation pulses are heard in these videos, but the methods of hunting are clearly demonstrated.

Physeter macrocephalus: The sperm whale is the largest of the odontocete cetaceans. They echolocate like other members of the group. The following clip shows a calf being tended by adults; when the adults aim their heads at the photographer, one can hear their echolocation pulses (http://www.arkive.org/sperm-whale/physeter-macrocephalus/video-09c.html). A second clip includes audible echolocation clicks toward the end (http://www.arkive.org/sperm-whale/physeter-macrocephalus/video-06.html), and a third shows closeups of the whales with high rate buzzes (http://www.arkive.org/sperm-whale/physeter-macrocephalus/video-13.html). Similar footage with narration can be found at the National Geographic site. (http://video.nationalgeographic.com/video/player/animals/mammals-animals/whales/whale_sperm.html)

Literature cited


Web Topic 15.1
Software for Network Measures

Introduction

A number of different fields have developed network analysis tools independently of each other. Often, the same measure has been developed for different users and given different names. The recent perception that different fields were often interested in the same network properties has encouraged general reviews that clarify equivalences and similarities (Albert and Barabási 2002; Barabási and Bonabeau 2003; Dorogovtsev and Mendes 2003; Newman 2003; Boccaletti et al. 2006; Croft et al. 2008; Dorogovtsev et al. 2008; Krause et al. 2009). Even with a correction for synonyms, network measures remain diverse. A network can have structure at any of multiple hierarchical levels: one can extract separate measures for each level independently, or apply measures that examine the degree to which the structures at different levels are correlated. In this regard, network measurement tools face the same challenges as tools used to characterize spatial pattern (Bivand et al. 2008).

Below, we provide general background on network measurement software and URLs to a variety of network analysis packages:

Software environments

There are several standard environments in which one can run network analysis software. Before listing options, it is important to decide which environment may be best for you:

- **Platform specific software**: Programs written to run natively on computers are often platform specific (e.g., Linux, Mac, Windows). In general, there are more network software programs written for Windows, but the increasing availability of platform-independent software (see below) is changing that bias rapidly.

- **Java-based software**: Programs written in the Java programming language can usually run on any platform as long as Java interpreting tools are loaded on the host computer. Most modern computers now come with Java interfacing when purchased. If yours does not, check with relevant websites to see what you need to download the suitable Java software and install it.

- **Matlab routines**: Matlab is a programming environment originally designed for matrix mathematics. Matlab comes in separate versions for Linux, Macs, and Windows computers and is fairly expensive (although many universities and colleges have site licenses and reduced rates for faculty and students). Matlab provides thousands of very useful built-in functions that one can then combine, using a simple Matlab language, into very complex sequences of procedures. These can be saved and distributed widely. Once you have Matlab installed on your computer, there are thousands of Matlab procedures that you can download (usually free). Several of the network packages listed below run in Matlab. You can find out more about Matlab at their website (http://www.mathworks.com/products/matlab/).

- **R**: R is a programming environment similar to Matlab except that it was originally aimed at statistics. Like Matlab, it comes with many useful functions that can be combined using a simple R language into very complex procedures. Also like Matlab, there are thousands of useful procedures that others have written and are happy to share. Unlike Matlab, R is free. It runs on all platforms. Check out the R site for details (http://www.r-project.org/).

- **GRAPHML**: XML defines standards for the format and syntax in which data can be shared over the internet (http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/). Such standards are needed if multiple programs are to be able to exchange data. GRAPHML is a subset of the overall XML schema that specifies how network data should be formatted for exchange and use in XML-based analysis programs. To understand these standards, one should be roughly familiar with the general XML principles and syntax. GXL (http://en.wikipedia.org/wiki/GXL) is a similar XML standard for graphical data used by some network graphics programs.
Overview sites

- Wikipedia ([http://en.wikipedia.org/wiki/Social_network_analysis_software](http://en.wikipedia.org/wiki/Social_network_analysis_software)): The Wikipedia article on “Social network analysis software” begins with a brief overview of the kinds of network software available, and then provides a table with a list of different software packages, for what purposes they can be used, the available platforms and environments, costs, and relevant details. Links in the table can take you directly to the source site.

- International Network for Social Network Analysis ([http://www.insna.org/software/index.html](http://www.insna.org/software/index.html)): This site also provides a list of network analysis software along with detailed annotations on each package, and like the Wikipedia site, is largely aimed at social network studies.

Data setup

The data needed to construct a network usually include a set of nodes (usually individuals), a set of links between nodes (which may be unweighted or weighted, signed or unsigned, fixed or dynamically valued, etc.), and a list of attributes for each node (e.g., sex, age, morph, etc.). Many network analyses begin with the construction of an association matrix. This is a two-dimensional matrix that codes the relationships between each node (usually an individual) and each other node. Codes can be discrete (usually 1’s and 0’s) or continuous (weighted links). There are many pitfalls and complexities in setting up association matrices properly. Alternatively, one can provide a program with node and edge lists: if the network is sparse, node and edge lists can be much more economical than setting up the full association matrix. Instead, one generates separate lists of nodes, links, and attributes. Such lists are usually provided as simple ASCII files but one has to be careful to follow any syntax or format restrictions for the particular program that will use the lists.

Some relevant resources for setting up network data:

- **SOCPROG** ([http://myweb.dal.ca/hwhitehe/social.htm](http://myweb.dal.ca/hwhitehe/social.htm)): This extensive series of association matrix tools (and some network analysis measures) was developed by Hal Whitehead at Dalhousie University. The package is free, but note that it consists entirely of Matlab procedures. You thus would need to have Matlab (particularly the Statistics Toolbox) installed on your computer to use this package. Because the cell values in an association matrix are not independent of each other, one cannot use standard statistics to test hypotheses (e.g., whether the distribution of values in the matrix differs significantly from random assignments). Whitehead thus provides a number of tools for performing such tests (e.g., Mantel tests) that are valid despite the dependence of values. He also describes methods that extract latent orthogonal variables from similarity/distance matrices which can be used in standard parametric statistics (e.g., multidimensional scaling, principal coordinates analysis).

- **NETWORK** ([http://www.jstatsoft.org/v24/i02/](http://www.jstatsoft.org/v24/i02/)): This site provides access to a package written for use in R that describes the format and syntax in which network data can be summarized into lists before feeding it to various network analysis programs in R.

- **GRAPHML** ([http://graphml.graphdrawing.org/primer/graphml-primer.html](http://graphml.graphdrawing.org/primer/graphml-primer.html)): This site provides a very illustrative primer on how to format network data in GRAPHML. A number of platform-independent programs can import GRAPHML data files.

- **PAJEK** ([http://pajek.imfm.si/doku.php?id=pajek](http://pajek.imfm.si/doku.php?id=pajek)): PAJEK is both an analysis package (see below) and has its own format for importing and exporting network data. Some other analysis programs will import PAJEK data (e.g., GELPHI).

Network analysis software

Most modern software packages provide various tools for importing data (association matrices or data lists), applying filters, various modes of visualizing the input networks, a variety of network connectivity, centrality, and attribute assortativity measures, and tools for identifying communities, modules, and cliques. Where statistics are feasible,
suitable statistical test tools may also be included. Most programs allow one to export images of the drawn network and various measure values. Note that network analysis continues to be a very dynamic field: most of the packages listed below are continually being upgraded and extended. Here is a sampling of some of the current software packages:

- **UCINET** ([http://www.analytictech.com/ucinet6/ucinet.htm](http://www.analytictech.com/ucinet6/ucinet.htm)): This program is probably the easiest to master and most of the tools that most people will need. The program is currently available for the Windows platform only (although it works just as well in Windows emulator environments on Macintosh and Linux machines). It can be downloaded for free for 60 days and reduced rates on licenses are available for faculty and students. Data can be entered as an association matrix or as a list of nodes and links. It has been widely used in a variety of disciplines including behavioral ecology (Croft et al. 2008).

- **MULTINET** ([http://www.sfu.ca/personal/archives/richards/Multinet/Pages/multinet.htm](http://www.sfu.ca/personal/archives/richards/Multinet/Pages/multinet.htm)): This package has not been updated in some time, but has some nice functions, particularly eigen-analysis tools. It only runs on Windows platforms.

- **PAJEK** ([http://pajek.imfm.si/doku.php?id=pajek](http://pajek.imfm.si/doku.php?id=pajek)): This is another free Windows program that has been widely used in biological contexts. Books are available explaining how to extract and interpret PAJEK measures. Although it has been available for some time, it was recently updated.

- **STATNET** ([http://statnetproject.org/](http://statnetproject.org/)): STATNET is a free suite of procedures written for R. You must have R (or an equivalent derivative such as SPLUS) installed on your computer to use this software. However, because R can be installed on any platform, this means the units are platform independent. A nice feature is that the possible tasks one might consider are largely broken down into separate modules. These are easily interfaced, but one need download only those units that are relevant to your particular interests. There are good references available for most modules.

- **TNET** ([http://toreopsahl.com/tnet/](http://toreopsahl.com/tnet/)): These free units are also written to run in R and thus can be used on any platform that has R installed. They focus particularly on weighted networks and networks where there are two concurrent sets of attributes present.

- **JUNG** ([http://jung.sourceforge.net/index.html](http://jung.sourceforge.net/index.html)): This free package is written in JAVA and so should run on any platform with suitable JAVA interpretation tools installed. Its introductory page says that it implements “a number of algorithms from graph theory, data mining, and social network analysis, such as routines for clustering, decomposition, optimization, random graph generation, statistical analysis, and calculation of network distances, flows, and importance measures (centrality, PageRank, HITS, etc.).” JUNG also provides a number of network visualization tools. It will handle both weighted and unweighted networks.

- **GEPHI** ([http://gephi.org/](http://gephi.org/)): GEPHI is a free JAVA package that will run on any platform. It accepts a wide variety of input data formats and provides a wide range of tools for coloring, rotating, expanding and annotating images of networks. It also provides a standard set of measurement tools.

- **NETWORKBENCH** ([http://nwb.cns.iu.edu/](http://nwb.cns.iu.edu/)): NETWORKBENCH is a free JAVA-based package that should run on any platform. However, check the FAQ for extra items that may need to be installed to run these modules on a Mac. This program accepts a wide variety of import data files include GRAPHML, PAJEK, and various list files. The package is now quite extensive and it may take one a bit of time to master some of the more sophisticated tools.

- **VISONE** ([http://visone.info/html/about.html](http://visone.info/html/about.html)): This free package (for non-commercial users) is also written in JAVA and should run on any platform. Some documentation and relevant publications are available. As with GEPHI, the package emphasizes visual inspection of networks.

**Visualization tools**

Some available software largely focuses on visualizing networks. Some useful sources:

- **NETDRAW** ([http://www.analytictech.com/netdraw/netdraw.htm](http://www.analytictech.com/netdraw/netdraw.htm)): This is a free Windows program that draws network structures given a variety of input data file types. It complements the UCINET program that is produced
by the same source.

- **Y-ED** ([http://www.yworks.com/en/products_yed_about.html](http://www.yworks.com/en/products_yed_about.html)): This is part of an extensive JAVA library called **yFiles** ([http://www.yworks.com/en/products_yfiles_about.html](http://www.yworks.com/en/products_yfiles_about.html)). The free yED graphic editor has its own GUI for importing and manipulating data. The program accepts a number of import file types, though favoring GraphML, and can save the resulting images in various formats.

- **GRAPHVIZ** ([http://graphviz.org/About.php](http://graphviz.org/About.php)): These routines are largely designed for implementation in other programs and packages. Readers who want to import them into their own programs or procedures may want to consult this site.

### Extensions

Here are a few additional sites that may be of interest:

- **CFINDER** ([http://cfinder.org/](http://cfinder.org/)): This free JAVA program has versions for any platform. It is designed to identify communities and modules in networks. Input files are simple text file lists of links with an optional modifier indicating link weights.

- **DAMSONS** ([http://www-users.york.ac.uk/~df525/damsons.html](http://www-users.york.ac.uk/~df525/damsons.html)): Many researchers on networks want to create artificial examples that meet specific conditions (e.g., random networks, scale-free networks with a particular slope in the degree distribution plot, etc.). This Windows-only free software provides tools for generating a variety of types of networks for model testing, measurement design, etc.

- **NETWORK FRACTALS**: As discussed in Chapter 15, one can use a box-covering algorithm to compare the number of boxes needed to cover a network at various box sizes to determine the relative fractal versus small world nature of that network (Song et al. 2005, 2006; Song et al. 2007; Rozenfeld and Makse 2009; Rozenfeld et al. 2010). The algorithms used by these authors can be obtained at H.A. Makse’s website ([http://www-levich.engr.ccny.cuny.edu/~hmakse/soft_data.html](http://www-levich.engr.ccny.cuny.edu/~hmakse/soft_data.html)). At the top of this page under “Research on Complex Networks” the first two paragraphs provide links for downloading the relevant routines. Note that these must be run in the Python NetworkX environment ([http://networkx.lanl.gov/](http://networkx.lanl.gov/)). There may be other routines at NetworkX of interest. For another algorithm for determining fractal dimension of networks see Locci et al. ([http://portal.acm.org/citation.cfm?id=1852511](http://portal.acm.org/citation.cfm?id=1852511)).

### Literature cited


Web Topic 15.2
Additional Information on Evolutionary Graph Theory

Although the field is new, a number of websites provide additional background on evolutionary processes on graphs and help explain some of the models outlined in Chapter 15. Suggested links include:

Virtual Labs in Evolutionary Game Theory
This site contains a suite of evolutionary game models. In most cases, the authors present a well-mixed classical game theory or adaptive dynamics case and then examine how the outcomes might change if the game is played on a structured network (lattice) with local interactions only. (http://www.univie.ac.at/virtuallabs/)

Moran Process Demonstration
(http://demonstrations.wolfram.com/TheMoranProcess/)

Cooperation in Heterogeneous Populations
(http://demonstrations.wolfram.com/CooperationInHeterogeneousPopulations/)

Lecture on evolutionary graph theory methods
PDF of Powerpoint slides by Gregory Puleo (www.math.uiuc.edu/~puleo/mighty.pdf)

Other
A search on Google for “evolutionary graph theory” will turn up a large number of PDFs and journal article abstracts that go beyond the general background at the sites above. Many of these are cited in the bibliography of our text, but others will surely be posted after the publication of the text.

Web Topic 16.1
Social Microbes

Introduction

Microbes of two completely unrelated taxa have evolved a convergent system of aggregating, migrating, and forming fruiting bodies: the myxobacteria and the social amoebae. Their similar soil environment is undoubtedly responsible for this convergence. In any seasonal environment, the soil periodically dries up and becomes inhospitable for many organisms. To survive, organisms must either move, or reproduce and send out airborne or otherwise mobile propagules. Reproduction via fruiting bodies invariably requires some individuals to altruistically sacrifice their reproduction to build the fruiting structure. Thus aggregations should be restricted to kin, and mechanisms should evolve for identifying and selectively joining kin groups. The web sites listed here are from various laboratories studying these organisms, and show the remarkable movement patterns of these very small entities.

Myxobacteria

The best studied social bacterium, *Myxococcus xanthus*, engages in cooperative swarming, predation, and multicellular development. The cells are normally solitary, but aggregate when they run low on amino acids, their primary food source. The aggregations have been likened to wolf packs, as they collectively cover prey organisms and secrete toxic and lytic compounds to degrade and digest the prey.

Suggested sites

- **Velicer Laboratory, University of Indiana**
  [http://sites.bio.indiana.edu/~velicerlab/Main2/index.html](http://sites.bio.indiana.edu/~velicerlab/Main2/index.html)
- **Zusman Laboratory, University of California Berkeley**
  [http://mcb.berkeley.edu/labs/zusman/zusman%20lab%20web%20page/Myxococcus%20xanthus.html](http://mcb.berkeley.edu/labs/zusman/zusman%20lab%20web%20page/Myxococcus%20xanthus.html)
- **VideoSurf**
- **YouTube**
  [http://www.youtube.com/watch?v=0ALM7X1_LqA](http://www.youtube.com/watch?v=0ALM7X1_LqA)

Social amoeba

The social amoeba, or slime molds, *Dictyostelium* spp., are single-celled eukaryotes. They initiate their aggregation in response to poor environmental conditions when a core group of cells sends out waves of an aggregating chemical signal, cyclic AMP. The aggregation develops through several stages to become a slug, which has the ability to move to a better location. The slug can also form a fruiting body, which is raised above the soil surface and can attach itself to passing mobile animals.

Suggested sites

- **Firtel Laboratory, University of California San Diego**
- **DictyBase website**
- **DictyBase videos**
  [http://dictybase.org/Multimedia/development/development.html](http://dictybase.org/Multimedia/development/development.html)
Web Topic 16.2
Plant Movement

Introduction

Plants are in constant motion, but they live on a different time scale from that of animals. They move as they grow, search for light and nutrients, avoid predators, exploit neighbors, and reproduce. Time-lapse photography speeds up this movement so we can view it on our time scale. These websites have some remarkable video clips of plant motion.

Suggested sites

- **Plants-in-Motion, by Roger Hangarter**
  This is a well-organized website with many examples of plant movements, including germination, photomorphogenesis, tropisms, nastic movements such as trap closing by the Venus flytrap, circadian responses, general growth, and flower movement (be patient, it takes a minute to load).

- **Dnatube site**

- **Mimosa on Wikipedia**

- **Vine searching for support**
  [http://www.youtube.com/watch?v=sLHeFmLJoLg](http://www.youtube.com/watch?v=sLHeFmLJoLg)

Highly recommended reading:

Web Topic 16.3
Body Language

Introduction

Humans transmit a great deal of information to others through their body language—gestures, postures, and body movements. Sending and receiving of these signals is largely unconscious, although when pressed, people can verbalize their perceptions and evaluation of body language signals, and senders can sometimes voluntarily modulate their signals. It is very challenging to conduct objective studies of the perception and evaluation of body language in humans because perceivers are strongly influenced by other aspects of the sender's appearance, such as skin color, facial features, gender, height, and clothing. Over the years, innovative research techniques have been devised to overcome this problem. This Web Topics unit focuses on studies of human nonverbal communication, especially body movements, that have employed these techniques. We first briefly describe the techniques, and then highlight some of the most interesting recent studies that used these techniques to examine human gait patterns, dancing, dominance postures, and speech-associated gestures. The interested reader can obtain more detailed information on this topic in the recent review by Hugill et al. (2010).

Overview of techniques

The earliest technique used to obtain objective evaluations of body movement patterns is called the point-light (PL) technique. This simple but elegant and effective technique was developed by Johansson (1973, 1976). Approximately 10–13 small lights are placed on the major joints of the body (ankles, knees, hips, shoulders, elbows, wrists, and head), and ambient lighting is dimmed so that only the point lights are visible while subjects perform prescribed movements. Humans are remarkably able to extract information about gender, age, emotion, and activity, as long as the subjects are moving (Barclay et al. 1978; Bertenthal and Pinto 1994).

Once graphical computer software became widely available to researchers, more sophisticated techniques became possible. Video motion-capture technology was first developed for sports training applications to study detailed body movements (Josefsson et al. 1996). Subjects wear tight (or little) clothing, and up to 40 small infrared reflectors are attached to key joints and body regions. Eight to 12 infrared high-speed video cameras simultaneously record the performer. Software programs are then used to extract the trajectories of each reflector during the movement episode. Vicon™ (Oxford, UK) is a widely used commercial product that provides both the cameras and the analysis software. If only an analysis of movement patterns is required, these trajectories can be directly submitted to statistical analysis. However, for behavioral research on the signal value of movements, another software program is employed to transform the trajectories into a video clip of a moving three-dimensional human-like figure. The kinematics of the movement is based on a model of the constrained movements of the human skeleton with input from the joint trajectories. In some cases, the figure is a jointed stick-like structure or wooden manikin, whereas in more recent studies a smoothed human form is constructed, popularly called an avatar. The very great advantage of this technique is the ability to experimentally modify the movement patterns to test hypotheses about the critical components of specific body movements.

Some highly sophisticated video-capture and software systems are now able to extract the movement kinematics directly from the video recording without the need for the subject to wear any reflectors (Bente 1989; Bente et al. 2001; Bente et al. 2008). This is advantageous when studying conversational interactions, where wearing tight clothing and reflectors can be inhibiting for the actors. The avatar directly mimics the postures and movements of the video-recorded interactants. A final strategy is to link fine-scale gestures of speaking subjects to the spoken words, an application designed to make animated figures in computer games show more realistic movement patterns while speaking phrases of text (Levine et al. 2009; Levine et al. 2010).

Analysis of gait patterns
Point-light studies of walking humans have shown that perceivers can derive a great deal of socially relevant information from this highly simplified form of visual input. For example, observers can assess an individual’s age, based on the happier and more powerful gaits of younger individuals (Montepare and Zebrowitz-McArthur 1988). The gender of a walker is easily distinguished by types of lateral sway: males tend to swing their shoulders from side to side more than their hips, whereas women swing their hips more than their shoulders (Barclay et al. 1978; Mather and Murdoch 1994; Sumi 2000). Sexual orientation is also distinguishable from simple point-light displays of walkers (Ambady et al. 1999). Individual identity can often be recognized from point-light depictions based on differences in stride length, bounce, rhythm, speed, and swagger, and observers can usually identify their own point-light image (Stevenage et al. 1999; Loula et al. 2005). Various types of action, such as different modes of locomotion, instrumental actions, and social activities such as playing, chasing, fighting, courting, and following, can also be distinguished from simplified visual input (Johansson 1976; Dittrich 1993; Norman et al. 2004; Barrett et al. 2005). Even two-day old human infants can discriminate biological versus non-biological point-light animations (Simion et al. 2008). These studies lead to the conclusion that humans possess intrinsic visual and cognitive adaptations that enable them to discriminate human and animal motion and infer the identity and actions of other people from a distance using gait patterns.

It has long been known that emotional feelings are reflected in a person’s gait, along with facial expressions and other nonverbal body postures and gestures. Psychological studies with actor–walkers and evaluators are able to identify sadness, anger, happiness, and pride from gait information at better than chance levels. Features such as the amount of arm swing, stride length, heavy-footedness, vertical head position, and walking speed are the cues that were found to differentiate these emotions (Montepare et al. 1987; de Meijer 1989; Sogon and Masutani 1989; Schouwstra and Hoogstraten 1995; Wallbott 1998; Montepare et al. 1999; Coulson 2004; Michalak et al. 2009). Studies based on computer-animated avatars have been able to zero in on the precise body movements that characterize different emotions. Roether et al. (2009) employed the Vicon motion-capture technique to extract the key joint trajectories that characterize anger, happiness, sadness, and fear gaits. Observers were able to correctly classify these four gaits with 70 to 90% accuracy. Multivariate analyses found that relatively few postural and kinematic features of gaits were required to characterize these emotions. Head inclination distinguished sadness from the others, elbow flexion characterized both fear and anger, speed of movement was high for anger and happiness and low for fear and sadness, upper arm retraction and knee flexion were increased for fear, and hip and shoulder movement distinguished anger and happiness. Avatar stimuli with only these few diagnostic features were then generated and presented to observers, who were able to correctly identify the emotions nearly as well as with the stimuli obtained from natural walkers. Movies from this study can be viewed at the website given below. In a similar study, Karg et al. (2010) found that emotions can be very well differentiated both in terms of basic emotional categories and in terms of the three emotional axes—valence (pleasure), arousal (activation), and potency (power or degree of control), as illustrated in main text Figure 16.26. Kapur et al. (2005) also demonstrated excellent emotion category classification by both human observers and machine learners. These studies highlight how important the signaling of emotion is in humans; gait-related components are visible at a greater distance than facial components and provide early information on the intentions of an approaching person. Roether et al. (2009) supplemental material can be accessed here: http://www.journalofvision.org/content/9/6/15.full#media-1.

Although women do not give obvious signals of fertility, they may produce some subtle fertility cues that are detectable and highly attractive to men (Morris and Udry 1970; Doty et al. 1975; Roberts et al. 2004). Another relevant gait study asked whether women in the fertile versus non-fertile stage of their menstrual cycle differed in gait, and if so, whether men expressed any preferences (Provost et al. 2008). This study employed the Vicon motion-capture system to extract information on joint trajectories during normal walking. Women subjects using birth control pills performed one walking session, while subjects not taking pills performed two walking sessions, one during the luteal phase (non-fertile), and another during the follicular phase (optimal fertile period 14–16 days before the start of the next menstrual period). A discriminant function analysis on the combined trajectory variables found a significant difference overall between fertile and non-fertile women, and 71% of women were correctly classified. Although hypothetical extreme walkers showed visible differences, involving lateral distances between the knee and ankle joints and hip movement, no significant difference was found between fertile and non-fertile women for these three body regions. The differences were either very subtle, or different subjects differed in the body region that varied as a function of phase. In the male perception...
experiment, point-light displays of the non-birth-control women were rendered and presented as stimuli in random order. Men scored the attractiveness of the walkers on a six-point scale. Women in the non-fertile stage were rated higher than women in the fertile stage. Thus despite the results of other studies showing that some cues given by women while fertile are detected by men and are rated as more attractive, these cues are more likely to be detected only by an intimate partner; women may subconsciously make a broadly advertised cue such as walking more attractive when they are not fertile, and avoid attracting unwanted suitors when they are fertile.

Dance performance

The performance of vigorous courtship displays is an important mate choice criterion for many animals (see Behavioral Display section of main text Chapter 12 and review by Byers et al. [2010]). Movement displays have a greater potential to reveal underlying aspects of genetic and phenotypic quality of the sender because they require the integration of many morphological and physiological pathways and more developmental genes than is the case for most structural display ornaments. In humans, dancing is a set of intentional rhythmic body movements generally considered to play a role in sexual attraction for both sexes. Dancing likely evolved in parallel with bipedalism, which facilitates a greater range of movement possibilities than found in quadrupeds. The torso can twist and bend, the arms can throw and swing in many directions, the legs can assume a variety of positions, and the head can swivel and nod in many ways, permitting a wide range of dynamic creative movements (Sheets-Johnstone 2005). For men in particular, dancing ability could signal aspects of male quality such as physical strength and agility (reviewed by Hugill et al. [2010]). Several studies have employed computer-animated models to discover possible underlying determinants of dance performance while removing all other aspects of physical appearance.

Brown et al. (2005) were the first to quantitatively assess the possible signal value of dancing performance. This group asked whether limb symmetry (fluctuating asymmetry) affected the dance quality of Jamaican young men and women, as judged by evaluators from the same population. They used the Vicon video recording system with eight high-speed cameras, and the subjects were outfitted with 41 infrared reflectors attached to key joints. The trajectory data were transformed into a virtual 3D animation of a wooden manikin that closely paralleled the original movements without containing any other visual cues. In addition, multiple morphometric measurements were collected from right and left sides of the subjects, including wrists, ankles, elbows, third to fifth digits, ears, feet, and knees; a composite measure of relative asymmetry was computed, and the upper third and bottom third of individuals from this distribution were retained as final subjects for the dance stimuli (10 symmetric and 10 asymmetric stimuli for each sex). The subjects danced for 30 seconds to the same currently popular music tune while being recorded. Evaluators viewed all 40 audio–visual stimuli and ranked each one on a scale from bad to good dancer. Dances by symmetric individuals were rated significantly higher than dances by asymmetric individuals, but the effect was stronger for male dancers. Women evaluators rated dances by symmetrical men more positively than did male evaluators, and more symmetrical men valued symmetry in woman dancers more than did less symmetrical men. The authors concluded that if fluctuating asymmetry indicates developmental stability, dancing in this population could reveal important information about dancer genetic quality and appeared to be subject to sexual selection by female preferences. Video clips of symmetric and asymmetric stimuli, as well as a video demonstrating the methodology used, can be viewed at the website of Brown et al. (2005) supplemental material: http://www.nature.com/nature/journal/v438/n7071/suppinfo/nature04344.html

Another study of dancing in humans by Neave et al. (2011) focused on female evaluation of male dancers and attempted to discover the specific moves that females found attractive. This study employed similar recording technology to the one described above, with some improvements. The system consisted of 12 cameras that tracked 38 reflectors on the subjects. In addition, the reconstructed 3D animated stimuli were more human-like avatars, rather than skeletal stick figures, but contained no cues of body mass, height, or facial features. There were 19 male subjects, who danced to a core drumbeat to remove any impact of music likability, and 37 heterosexual female evaluators, who rated the dances on a seven-point scale. The biomechanical trajectory data from the reflectors were used to quantify different aspects of the movement, such as the amplitude, speed, duration, and variability of different limb movements. Several key movement measures were correlated with dance quality: variability and amplitude of neck and trunk movements, and speed of movements of the right knee. These results pointed to greater attractiveness of more vigorous dancing patterns, again supporting the conclusion that dancing reveals potentially useful information about
dancer quality. Video clips of the stimuli can be viewed at the website of Neave et al. (2001) supplemental material: http://rsbl.royalsocietypublishing.org/content/suppl/2010/09/07/rsbl.2010.0619.DC1.html

Analysis of dominance interactions

The analysis of dyadic interactions between a dominant and subordinate individual presents a special challenge because of the two-way effects of each participant on the other. We reviewed the definitions of power, and some of the body postures and conversational management differences between subordinates and dominants in Chapter 16. Research on the nonverbal signals given in such interactions is further confounded by cultural differences in the value people assign to the vertical dimension of social relations. Some cultures accept that inequalities in power and status are natural, so those with power tend to emphasize it and distinguish themselves from subordinates, while other cultures view power and status as man-made and artificial and therefore de-emphasize it and share power (Storti 1999). Despite different perceptions of the value of status, individuals from different cultures generally agree on the nonverbal signals that express power (Triandis and Gelfand 1998; Kowner and Wiseman 2003; Guerrero and Floyd 2006; Burgoon et al. 2010). To resolve these differences in perception and evaluation of power from nonverbal signals, Bente et al. (2010) conducted a study of dominance interactions in three cultures (Germans, Americans, and Arabs) using culture-free video clips converted into avatars. Subjects from the three countries enacted a supervisor–employee conflict resolution scenario while being video recorded. Specialized software extracted the body movements of both participants and recreated a video clip of a one minute section of the interaction between wooden manikin figures (Bente 1989; Bente et al. 2001; Bente et al. 2008). These clips were shown to receivers from the same and different countries. Status roles were only distinguishable above chance in the German sample. There was no evidence for an in-group advantage. Nevertheless, there were significant correlations in the ratings of dominance across observers, with a slightly stronger in-group effect. Microanalysis of movement behaviors demonstrated that certain behaviors predicted the dominance rating: vertical head posture and an open position of upper and lower limbs. In Germany only, less movement activity and stretching of the lower limbs was perceived as more dominant. Germany is known to be a high-power-distance society, and America is known to be a lower-power-distance society, so results for these countries met expectations. Arabs are a high-power-distance society, but they are also a collectivist society, so expressions of power could be repressed for this reason.

Other techniques have been developed to study conversational interactions with computer-generated faces. Wilms et al. (2010) used eye-tracking technology to generate a gaze-dependent avatar face whose behavior becomes responsive to being looked at by the participant. This allows the participant to engage in “online” interaction with the avatar in real time. In some applications of this technology, brain activity of the subject is simultaneously monitored with fMRI. Čereković and Pandžić (2011) developed what they call Embodied Conversation Agents, which graphically embody virtual characters that can engage in meaningful conversation with human participants. Multimodal data are extracted from the speech and nonverbal behaviors of subjects performing short two-second expressions. These clips form a large library, from which facial movements and emotions are synthesized. The goal of this complex project is to create believable and expressive virtual characters in order to enhance the communication abilities of machines. Similarly, Levine et al. (2009, 2010) extract correlations between spoken words and speech gestures in real speakers, and then use these associations to produce text-specific movements for animated figures in games. A video clip explaining the methods and objectives of this research can be found at the website of Levine et al. (2009) supplemental material (click on the Source Materials tab): http://portal.acm.org/citation.cfm?id=1618518

Further reading


**Literature cited**


