15

The value of information in signals and cues

MICHAEL LACHMANN

In this chapter I will discuss some uses of information measures in animal behaviour and genetics. Instead of delving into the question of when, whether and how a signalling system evolves, I will examine a simpler definition of information transfer and mutual information from statistical mechanics and the theory of communication. This circumvents, but does not solve, the question of whether or not signalling systems exist in biology. However, in my opinion, understanding the evolution of information use and transfer in biology is an important step in understanding the evolution of signals and cues.

In describing the many possible ways in which animals can interact, manipulate or maybe transmit information, the field of animal signalling borrowed many different concepts, which sadly still do not even begin to cover all the different possible types of biological 'signals', their origin and stability. Two of these concepts are 'signals' and 'cues'. Maynard Smith and Harper (2003) define a signal as "any act or structure that alters the behavior of other organisms, which evolved because of that effect, and which is effective because the receiver's response has also evolved". They also add: "It follows that the signal must carry information – about the external world – that is of interest to the receiver."

A cue, on the other hand, is defined as "a feature of the world, animate or inanimate, which can be used by an animal to guide future actions" (Maynard Smith & Harper, 2003, after Hasson, 1994).

Thus both signal and cue carry information about the world. When an organism evolves to respond to a cue its fitness can only increase, since a

Animal Communication Theory: Information and Influence, ed. Ulrich Stegmann. Published by Cambridge University Press.

strategy of ignoring the cue must have been less fit. The same is not true for a signalling system or when interacting with other organisms. Once signaller and receiver evolve to signal and respond, their interaction can change so that both might lose (Hirschleifer, 1971; Box 15.1).

Box 15.1 Is information always beneficial?

When an organism has a cue, i.e. a feature of the environment that it can respond to, it usually also has the option to ignore it. If a strategy of responding to the cue invades, it means that

$$\text{Fitness(responding to cue)} > \text{Fitness(ignoring cue)}$$

which of course means that there is a fitness gain from responding to the cue.

When we deal with an interaction between two or more actors, there is not always a gain. Imagine two female, F1 and F2, competing over the mating possibilities with two males. Assume that one of the males has higher quality than the other, though it is not necessarily known which, and that F1 had first choice. F2 can then decide whether to mate with the same male as F1, or with the other male. Let us assume that the fitness gain from mating alone with the high-quality male is 5, and with the low-quality male 1. Assume that two females mating with the same high-quality male get a fitness of 2 each, and both mating with the low-quality male get a fitness of 0 each.

If no female has information about male quality, the second female does best in choosing the other male. Half the time she would then mate with the high-quality male, and half of the time with the low-quality male, for an average fitness of $(5 + 1)/2 = 3$. Mating with the same male as F1 would give her a fitness of 2 or 0, both lower than 3.

What happens if F1 has information about male quality? In this case F1 would choose the high-quality male. If F2 now took the other male, she would always get the low-quality male, and an average fitness of 0. Therefore, she should prefer to choose the same male as F1, for a fitness of 2. In this case, F1 would also get a fitness of 2 – lower than her original average fitness of 3, without information.

Notice that for F1 the strategy of ignoring the information and flipping a coin in her choice of male is not stable: in that case F2 would choose the other male, and F1 would be tempted to switch to the strategy of choosing the high-quality male, and gain a fitness of 5 instead of 3. Thus in this case gaining additional information has reduced F1's average fitness at equilibrium.

In the following I will introduce the notion of mutual information, and explain its simple connection to the notion of fitness or relative growth rate. I will then present some of the uses of information measures in genetics. This chapter is not aimed as a review of the uses of information measures in biology (see instead, for example, Adami, 2004; Dall *et al.*, 2005; Dall Schmidt & Van Gils, 2010 and other articles in the same issue of *Oikos*; Sherwin, 2010). Instead, I want to explain why such measures have their value in theoretical biology and in particular in the theory of animal communication.

15.1 Entropy and mutual information

Uncertainty is closely related or identical to the concept of 'entropy' in various fields. In physics, two different but closely related definitions of entropy are used. The first, introduced by Clausius (1867), is based on macroscopic measures of the system, such as temperature, pressure and volume. Clausius defines the change in entropy as the integral of dQ/T, the heatflow divided by the temperature, and writes

> I propose to call the magnitude S the entropy of the body, from the Greek word $\tau\rho o\pi\acute{\eta}$, transformation. I have intentionally formed the word entropy so as to be as similar as possible to the word energy; for the two magnitudes to be denoted by these words are so nearly allied their physical meanings, that a certain similarity in designation appears to be desirable. (Clausius, 1867, p. 357).

The other, based in statistical mechanics, is a measure of the number of possible states a system could be in given its macroscopic properties. To describe a glass of water fully is impossible – we would need to know the position and velocity of more than 10^{24} atoms, and we cannot even measure the position or wavefunction of a single atom precisely. Instead we know some macroscopic properties of the water – volume, temperature, weight etc. We do not know a system's complete description, and yet, when someone tells us the temperature of the water, we gain information. Entropy in statistical mechanics is a measure for the 'number'[1] of possible states a system could be in given its macroscopic properties, or more precisely the log of the number of states. The importance of this measure in physics comes mainly from what is known as 'Liouville's theorem', which states that for isolated physical systems, two distinct states of the system cannot converge. So if we have a system that could be in one of

[1] For simplicity I talk in this chapter about number of states, instead of talking about the volume of an ensemble in phase space.

1000 possible states, and let it evolve for three hours, we will still be looking at 1000 different possible states. Thus, for example, if we start with all 1000 possible states associated with the current macroscopic properties of the system, then at any time in the future we must still be looking at 1000 different possible states. This means that the system cannot reach a condition where its macroscopic properties are associated with only 500 possible states. It must have macroscopic properties associated with *at least* 1000 possible states. This is one possible explanation why entropy, the number of possible states of the system given its macroscopic properties, cannot decrease – this is the second law of thermodynamics.

In the theory of communication, entropy is also a name for uncertainty, e.g. the number of possible states a signal could be associated with. Claude Shannon developed a measure for the uncertainty, and found it to be very similar to Boltzmann's derivation of entropy. Tribus and McIrvine (1971, p. 180) cite Shannon as saying:

> My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage."

Entropy measures are also seeing an increased use in statistics, mainly in Bayesian and likelihood analysis. Jaynes' (1957a, 1957b) maximum entropy principle suggests a way to choose a prior distribution for Bayesian analysis: one that maximises our uncertainty but is consistent with what we know.

The use of a measure related to entropy makes sense when we are interested in a count or a quantification of the number of states of a system, and especially when one is interested in the change in the number of states. Here we should distinguish two ways to calculate the difference in uncertainty.

The first is called 'mutual information'. Mutual information tells us how much knowing one thing helps us in knowing something else – for example, how much knowing the season tells us about the temperature. The second way of calculating a difference in uncertainty, which does not seem to have a name, but is used often, is to calculate the difference in entropy after a *particular* change in our knowledge has occurred. We might, for example, hear that

the forecast is for rain tomorrow, and as a result of hearing that particular message "rain tomorrow" our uncertainty changes. So in the first way of calculating the difference in uncertainty we ask how much our uncertainty changes on average when we hear the weather report, in the second way we ask how much it changed after we heard a specific weather report, namely that it will rain.

Of these two, mutual information is the one that is more intuitive to understand. As we would expect from gaining information, this measure is never below zero – being given information should not increase our uncertainty. Thus, knowing the month reduces our uncertainty about the temperature. The second measure described above *can* be negative – imagine living in a place where it is sunny most of the time, say Palo Alto, California, and hearing a weather report saying that tomorrow there is a 50% chance of rain. Without this information, there was a 99% chance of sunshine, 1% chance of rain – the average weather. But the particular report gave rain a 50% chance, so now we are much less certain whether to take an umbrella with us or not.

To state this again: on average, listening to a weather report whose precision we know cannot increase our uncertainty, averaging over all possible reports and their probability. But a certain report for a certain day can increase our uncertainty.

If we write $H(X)$ for the entropy, or uncertainty associated with not knowing X (for example not knowing if it will rain), $H(X|Y)$ for the uncertainty in X given that we know the state of Y (Y could, for example, be the weather report) and $H(X|Y = s)$ for the uncertainty of X when Y is in a certain state s, then the mutual information between X and Y, written as $I(X;Y)$, can be written as

$$I(X;Y) = \sum_{\text{all states } s \text{ of } Y} p(s)\,[H(X) - H(X\,|Y = s)] \tag{15.1}$$

where $p(s)$ stands for the probability that Y is in state s (for more explanation see Box 15.2). We see the connection between the mutual information and the difference in entropy conditioning on a single state – the first is an average of the second. Mutual information is the weighted average, over all states of Y, of the difference in entropy conditional on each single state. Mutual information has some very convenient properties. Thus $I(X;Y) = I(Y;X)$: how much the month tells us about the temperature is equal to how much the temperature tells us about the month. Mutual information is also independent of recoding the variables. So, if instead of measuring temperature in Celsius we measure it in Fahrenheit, or maybe we look at the log of the temperature, the

Box 15.2 Mutual information

Mutual information tells us how much our uncertainty about one thing is reduced when we are told about another. In the figure, we are looking at two variables X and Y. Not all values of X and Y are possible. The white square outlines the possible values of X and Y, and I assume that all rectangles are equally likely. Thus Y can take a value of 4 when X is 5, 6, 7 or 8, but not when X is 1, 2, 3 or 4.

The down-pointing arrow shows what happens if we are interested in X, and ignore Y, or have no information about Y. For example, X can be 1 when Y is 1 or 2. In this case we don't know Y, but the chance for $X = 1$ is equal to the sum of these two rectangles, or 1/8. So, when Y is not known X can take 8 different values, all equally likely, and therefore the missing information is $H(X) = \log(8) = 3$ bits. If we know Y, we know something about X. Thus, when we know that $Y = 4$, X can only be one of 5, 6, 7 or 8. For any value of Y, X can take on 4 values, so once we know Y, we have $H(X|Y) = \log(4) = 2$ bits of information missing. The difference in the missing information when we don't know Y and when we know Y is the difference between these two values, so $I(X;Y) = H(X) - H(X|Y) = 3 - 2 = 1$. Knowing Y gives us 1 bit of information about X; it halves the number of possible states of X from 8 to 4.

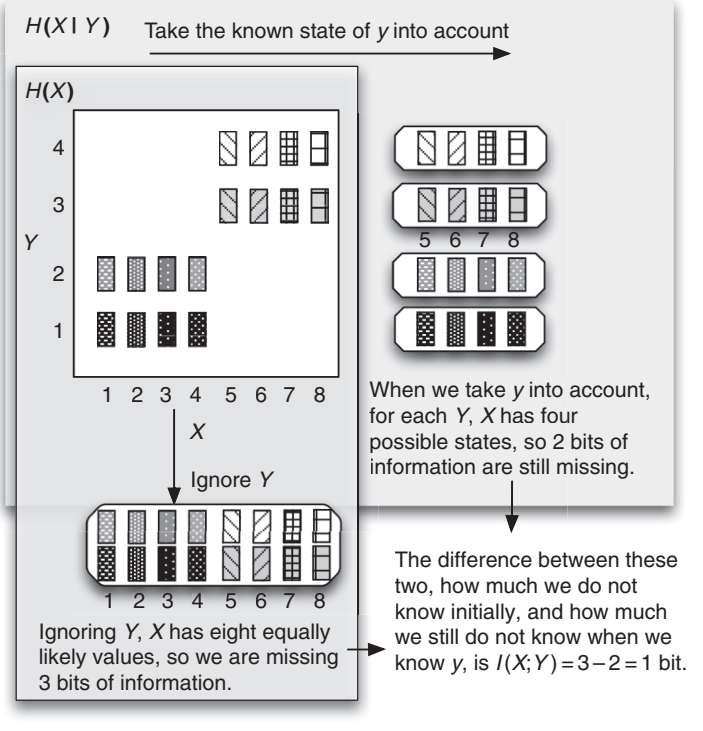

Figure 15.2

mutual information stays the same. The same is not true, for example, for the closely related concept of statistical correlation, which does change if we remap our variables. Box 15.3 explains why this is the case. It is also true that whenever there is a statistical correlation between two variables, they carry mutual information about one another. The converse is not true – it could be that two variables have mutual information and yet have a correlation of zero.

Box 15.3 Effect of rescaling a variable on mutual information and on correlation

Most variables we measure have no natural scale. One researcher might look at the weight of the organism in kilograms and another in pounds, or maybe in log scale, when comparing many different orders of magnitude. Such changes have an effect on correlation measures of variables, but not on the mutual information between them. Let us take X and Y from Box 15.1, and rescale Y so that the possible values are 1, 2, 8 or 9. Rescale X so that the possible values are 1, 2, 3, 4 or 15, 16, 17, 18. We can now complete the same calculation we did in Box 15.1: when ignoring Y, X still has 8 possible equally likely values, and for every value of Y there are 4 possible values of X. Nothing has changed from Box 15.1: we still gain $3 - 2 = 1$ bit of information about X when we know Y, so $I(X;Y) = 1$. It is easy to understand why rescaling Y and X will never have an effect – in calculating the mutual information we are interested in the probabilities of the different values, and not in where they sit.

Correlation changes when we rescale the values. Graphical representations of the correlation coefficient are not easy to understand (see "Thirteen ways to look at the correlation coefficient" by Rodgers & Nicewander, 1988.) One interpretation of the correlation draws a rough ellipse around the sample points, and compares the height of the ellipse in the Y direction to its width in the middle (see grey arrows below). The smaller the width of the ellipse in the middle relative to its height, the larger the correlation. When we rescale the variables we will transform this ellipse, and the correlation will change. In the example I gave above, the correlation between X and Y increases.

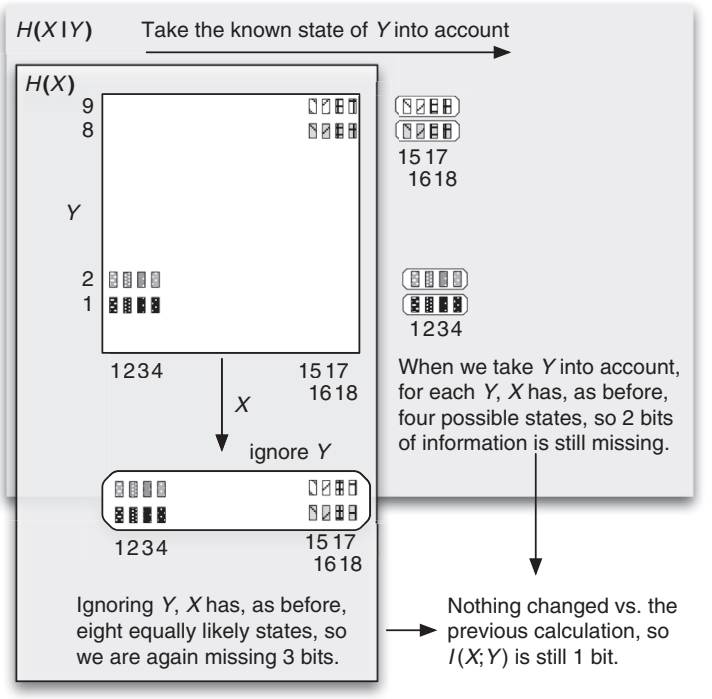

Figure 15.3A

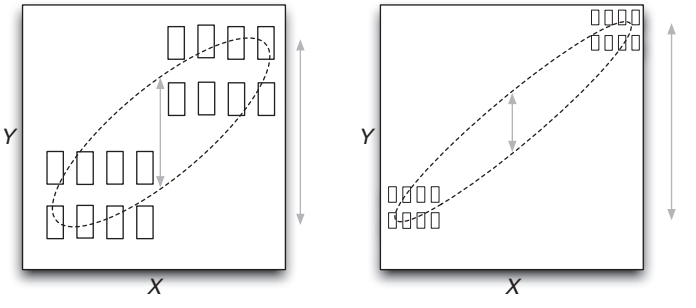

Figure 15.3B

15.2 The information value of cues and signals

In the following, I will examine the information content of a cue or signal given to an organism. Following arguments used in previous papers (Bergstrom & Lachmann, 2004; Donaldson-Matasci, Bergstrom & Lachmann, 2010), I will argue that even in cases where we are interested in the fitness of organisms, information theoretic measures will be related to fitness gain of lineages. I will look at two measures of the information content. The first is the mutual information between the cue and some aspect of the organism's environment. The second looks at the effect of the signal or cue on the organism's fitness. I begin with the latter.

Consider the difference between the optimal fitness without the cue, and the optimal fitness with the cue. Let us write $F(X)$ for the optimal fitness possible when the state of X is not known to the organism, and $F(X|c)$ for the optimal fitness when X is not known, but a cue c is received – i.e. the fitness conditional on X. We can then look at the "value of information" (Stephens, 1989).

$$\text{Fitness gain} = \sum_{\text{all cues } c} p(s)[F(X) - F(X|c)] \tag{15.2}$$

The similarity of this equation to the definition of mutual information in Equation 15.1 is apparent, except that here we measure fitness, and above bits. Is there any connection between these two?

In information theory we look at the frequency of events, for example how often the word 'me' is used in the English language versus the word 'encyclopaedia', or how often the bit 0 is stored in a file versus the bit 1. Using these frequencies we can then find an optimal coding – how long should the word that stands for 'me' be versus the word that stands for an encyclopaedia. What is not taken into account is the meaning of the signals. Thus, we calculate the information capacity of a channel, without taking into account whether it serves as a police dispatch, or a television channel for late-night infomercials. Let us look at a simple example: we could use information theory to reduce the number of key presses one needs to use on average to dial a number, making often-used numbers shorter. Imagine that our dial pad had just two keys with the digits 1 and 2 on them, and we want to be able to quickly dial three contacts, Aaron, Ben and Claudia, but we call Aaron twice as often as each of the other two. In this case it would be optimal to assign Aaron the number '1', and assign Ben and Claudia the numbers '21' and '22'. With this assignment, half the time we would have to dial just one digit (when dialling Aaron we would dial '1'), and half the time we would have to dial two digits (when reaching one of the other two we would dial '21' or '22'). Our average dial length would be 1.5 digits,

which is optimal in this case. But usually the assignment of telephone numbers also takes into account other considerations: the reason that emergency numbers are short and easy to remember is that they are important, and not that they are dialled more often (though they probably are). In the above example, if an expecting couple was assigning the numbers, and Ben was their contact in the maternity ward, the couple might use the shorter number '1' for Ben, even though they do not need to reach him very often – because when they do need to reach him it is urgent. We see that in many real-world examples, one needs to take into account not only frequency, but also meaning.

One can ask a similar question in biology. Why would a measure of the mutual information between a cue and the environment be at all relevant to understanding an organism, if it is not specified what this cue is about? Why would the information content of a cue carrying two bits of information about the location of an ant be important if we do not know whether it is perceived by a beetle or by a lion? In a world that cares about survival and reproduction, is the information content of a cue as measured in bits relevant?

The surprising result is that entropy-based measures are also relevant to such cases. In economics it turns out that the information in a signal about stocks is closely related to how much faster your money will double when using the information (Kelly, 1956; Cover & Thomas, 1991). What we and others have shown (Cohen, 1966; Bergstrom & Lachmann, 2004; Kussell & Leibler, 2005; Donaldson-Matasci *et al.*, 2010) is that a similar result also holds in biology: under certain conditions, the Shannon information of an environmental cue is exactly the fitness benefit one could gain from heeding it. Heeding a cue giving one bit of information will then allow a two-fold increase in fitness, without specifying what this cue is about, or the fitness consequences of the different strategies available to the organism! In the following I will try to explain this somewhat unintuitive result.

When dealing with the evolution of strategies of organisms, we are interested in which strategy will out-survive its competitors. The simple case examined is an organism living in a variable environment. We assume it reacts to a cue reducing its uncertainty about the environment, and we want to ask how relevant is the information content of the cue about the environment as measured by Shannon's information measure, i.e. the mutual information between the environment and the cue.

It turns out that for evolutionary questions one needs to specify not only whether an environment is variable, but also how the uncertainty is distributed between individuals in the population (Donaldson-Matasci, Lachmann & Bergstrom, 2008; Rivoire & Leibler, 2011). At the one extreme are environments that are shared by all individuals in the population, for example whether it was a

cold or warm winter. At the other extreme are variable environments that are different for every individual in the population, for example whether a predator is hiding close by. Between these extremes are environments whose variability is partly shared, or ones whose variability is only shared between some individuals – for example all those living in a certain location. For simplicity, I assume that the variability is shared between all individuals in the current generation, and that if an individual is adapted to the wrong environment, it dies. Here I also assume that each generation is exactly one year long.

A central concept in information theory is the idea of 'typical sequences' (Shannon & Weaver, 1948; McMillan, 1953; Cover & Thomas, 1991). Imagine an environment that has 'cold years' with probability 70% and 'warm years' with probability 30%: then for long enough sequences of years we will see with high probability a sequence that has around 70% cold years and 30% warm years. See Box 15.4 for more explanation on typical sequences.

A hypothetical organism with no information about the environment has little choice but to either have a fixed genetically defined phenotype, or have its phenotype randomly determined by a 'roll of dice' – the juvenile develops into an adult adapted to cold years with probability q, and adapted to warm years with probability $(1 - q)$. The genes will determine only the chance q, but not the phenotype of each offspring. Thus, each organism will have a fraction q of offspring adapted to cold and $1 - q$ adapted for warmth, and each of those again will have a fraction q of its offspring adapted to cold and $1 - q$ to warmth and so on. In Box 15.5 I explain why, in this case, it is optimal for the organism to have a genotype so that q is equal to p. The optimal strategy thus involves bet-hedging. Without information, a lineage has to divide its bets equally between all possible typical environments *and match gthe proportion of phenotypes to the proportion of corresponding environments*. When having the wrong phenotype does not result in

Box 15.4 Typical sequences

If we look at a sequence of independent events, each of which has a chance p of occurring, then for long enough sequences we will mostly see sequences that have a ratio of around p to $1 - p$ of the two event types. Thus, if we throw a coin with even chances for heads and tails, most of the time we will see a sequence that has around 50% heads and 50% tails.

In the figure we see the four possible outcomes of two coin tosses (with heads and tails represented by black and white), all equally likely. A ratio of 0.5 occurs most often, but not by a huge margin. With four tosses, we will

have 6/16 possibilities with 50% heads and 50% tails. Now 14/16 = 87% of the possibilities are between 25% and 75%.

With a large enough sequence, almost all sequences we observe will have that ratio; they will all be approximately equally likely. In N events, Np times an event with chance p happens, and $N(1-p)$ times one with chance $(1-p)$, so the chance for each sequence is $p^{Np}(1-p)^{N(1-p)}$, which can be written as $2^{N[p \log p + (1-p)\log(1-p)]}$. In general there will be 2^{NH} typical sequences all equally likely, with H the entropy of each event.

Typical sequences have a central importance in understanding information theory. Thus, if we wish to compress the outcomes of N coin tosses, we can just number the events that are likely to occur from 1 to 2^{NH}. Since other events almost never occur, it does not matter how we encode those. This means that to encode N tosses we will only need NH bits. In a similar way we can calculate how many different messages the English language can transmit. If we assume that words occur at certain frequencies, independent of each other, then we will mainly be observing typical sequences of words, and we know how many of those there are: 2^{NH}. Words, of course, do not occur independently of each other, so we can expand our treatment to pairs of words, sentences and paragraphs. (This example is very similar to those used in Shannon's original paper.)

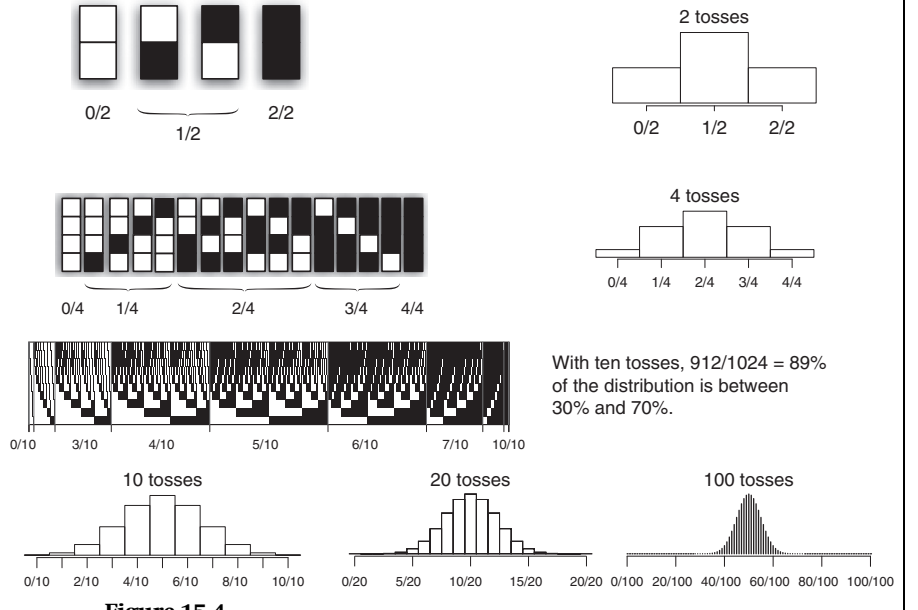

Figure 15.4

death, just in lower fitness, the optimal strategy can still be one with bet-hedging, but finding the optimal strategy is a bit more complicated. In this case there will be some frequencies of warm and cold years for which there will be no bet-hedging – instead the individuals will always develop the same phenotype. Other, intermediate, frequencies will still have bet-hedging as the optimal strategy.

Real organisms will, of course, always have some information about the relevant state of the environment. Below I discuss the optimal strategy in those cases, which will, for minuscule amounts of information, be very close to that without any information.

How will additional information about the environmental state help the lineage? As mentioned above, information is a measure of the (fold) reduction in uncertainty. (I will use the word 'fold' in this chapter for the amount multiplied or divided by. A four-fold reduction thus means that we divide the number of possible states by four, and I will talk about the 'fold reduction' to speak about an unspecified amount of reduction.) If there are 100 equally likely states, and we get one bit of information about them, we can see our updated uncertainty as 50 equally likely states; two bits of information will leave us with 25 states and so on. Thus, if a lineage has to divide its resources equally between 100 possible outcomes because it is relying on bet-hedging, and it gains one bit of information, it will need to divide its resources only between 50 possible outcomes, and thus have twice as many surviving lineages. Therefore, if along the tree of future possibilities a lineage gains every generation one bit of information it means that it can exactly double its fitness – the growth rate is multiplied by two. The fitness gain of x bits of information is in these cases exactly a 2^x-fold gain in fitness. Quite an amazing result: without knowing what the system in question is, lions or bacteria, without knowing the exact conditions or fitness effects and how many different environments the organism experiences, we can say that one bit of information given to the organism can gain it a two-fold increase in fitness.

How can one bit of information about an important or unimportant part of the individual's life lead to the same fitness benefit? The trick is that one bit of information equals one bit of fitness increase only for the case where the organism bet-hedges on the outcomes before and after the information is provided. If the optimal strategy before having information is to bet-hedge on the outcomes, and after information is received the optimal strategy is still to bet-hedge (with modified probabilities since additional information modified the probabilities of the different outcomes), then the value of information is equal to the mutual information. When the information is about a feature of the organism's environment that is of little consequence, the optimal strategy will not be to bet-hedge, and the value of the information in terms of fitness will be less than the information content. This pre-selects on the type of information

Box 15.5 Typical sequences in the environment and in a lineage

In our example, an organism survives only if its phenotype (fill pattern shown in the circle) matches the environment in the generation (fill pattern shown in the background rectangles). If the environment is white with probability p and striped with probability $(1 - p)$, a typical sequence of environments will have almost exactly that ratio of white to striped states. For a future lineage of an organism to survive (all lineages originating from circle on left), its phenotypes must match the environmental phenotypes exactly.

If in every generation a fraction q of the offspring are white and a fraction $(1 - q)$ striped, then a typical sequence of phenotypes along a lineage will have almost exactly that fraction of white versus striped phenotypes. For the phenotype to match the environment every generation, q has to be equal to p (see, for example, along the marked surviving lineage).

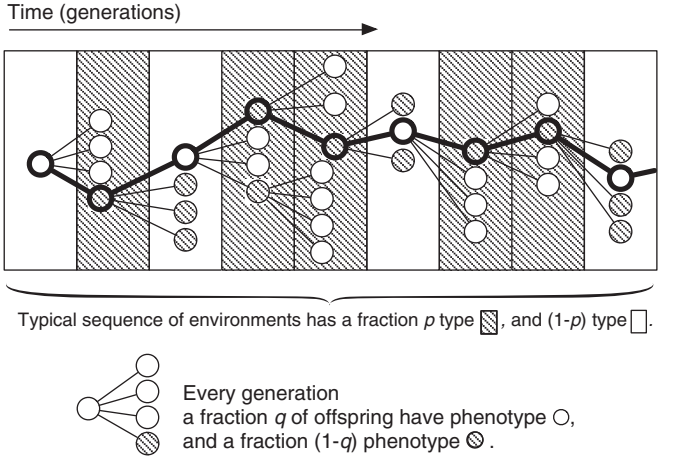

Figure 15.5

that our claim applies to. When we give a lion information about the location of an ant, then because the location of the ant is not important enough for the lion, it will not have bet-hedged on the location of the ant. It might, for example, have bet-hedged on the location of antelopes, so information about antelopes will fulfil our criterion. A beetle will not bet-hedge on the location of antelopes, but might do so on the location of ants. Thus bet-hedging is an equalising criterion. Only when we provide a certain amount of information about a system to

organisms that bet-hedge on the state of the system can we get exactly the same fitness effects as the information content. Since the organism will not bet-hedge on unimportant aspects of its environment, these cases are excluded from our result.

The connection between information and fitness comes from counting. Shannon's information measure is about a fold reduction of possible states. When we ask about compression of a file, we have to count how many possible equally likely states there are. When we talk about the difference in entropy between a cup of tea at 100 or 25 °C, we ask about the difference in the number of possible states between these two temperatures. Fitness can also be seen as a measure of the number of possible future lineages, and since we again look at a measure of equally likely possibilities, information and entropy can be applied.

It is interesting to note that Wagner (2007) used quite a different approach to reach very similar results. He looked at the fitness effects of the sensing of a limiting nutrient's abundance, using a model for the metabolic network. In the model, he again arrived at the result of a connection between information content and fitness.

15.3 Information and genomes

Many bacteria live in a constant arms-race of producing toxins, antitoxins and other countermeasures. Some of the countermeasures to antibiotic toxins, such as tetracycline produced by actinobacteria, involve arrested or slowed growth – obviously a phenotype with a big cost. In these cases, an environmental cue about the presence of a toxin could be 'lineage-saving'. Often, however, the action of the toxin could be so quick that the cells die before they can respond to it. Bacteria might also be using bet-hedging, producing a subpopulation of individuals that grow more slowly, to save their lineages from unexpected exposure to toxins. Kussel and Leibler (2005) looked at the efficiency of response to a cue versus a bet-hedging strategy, and found the conditions under which one or the other gives a faster average growth rate. They noticed the connection between information and fitness effect. We can do a quick analysis based on the tools described above. When the bacteria have no information about the environment, not even the phenotypic/genotypic state of their parents, their fitness will be reduced by at least H(Env) relative to ones that know the exact state of the environment. If there are cues available about the environment, their fitness can increase by up to I(cue; Env) – the mutual information between the cue and the environment (the 'relevant' environment, i.e. the environmental states over which bet-hedging is used). But the bacteria can use an alternate method – evolutionary switching between states. We can

then see the state of the genome with respect to arrested growth as a 'cue' to the state of the environment. The increase in fitness from using this strategy will be I(Genome; Env) or less – the mutual information between the genome and the environment. What creates a correlation between the genome and the environment is selection, and the fact that recent environmental states, those during which selection acted, contain information about the current environment.

Kimura (1961) tried to calculate the rate at which natural selection inserts information into the genome:

> If those individuals which are to be eliminated by natural selection in the process of progressive evolution were kept alive and allowed to reproduce at the same rate as the favored individuals, the population number would become, after t generations, e^{-Lt}. This means that natural selection allows an incident to occur with probability one, which, without selection, could occur only with a probability of e^{-Lt} . . . and therefore information gained per generation is L/log(2) bits.

Here L refers to the 'load', a measure saying how many lineages are lost to the population because not all individuals have the same fitness as the best in the population (Kimura used the notation L_e for this). The load L is defined so that if in a generation half the lineages are lost through their low fitness then $\exp(L)= 2$, so that $L = \log(2)$. Kimura then balanced the above measure with the rate at which deleterious mutations destroy information per generation, to estimate the total number of bits gained since the Cambrian explosion – around 10^8 bits. Recently, Adami (2004) re-examined such an approach to estimate the selection that acts on a sequence. Sites that are neutral are under no evolutionary constraints, and thus are free to take any possible sequence. Over enough evolutionary time we should see every one of them. Missing sequences hint at selection, for example if at a certain site we only see a 'G' across many species. As before, we can measure the fold reduction in the number of sequences (again, I use the word 'fold' to stand for the amount by which we divide the number of sequences). The exact reduction is sometimes hard to calculate, because the number of species is so much smaller than the number of possible sequences. As a proxy, we can look base by base, ignoring interactions between bases. As above, the additivity of mutual information would allow us to add the measure for each base to get the overall amount of conservation of the site. Adami also showed how one can find interacting sites in an RNA molecule by looking at their mutual information.

An interesting related use of entropy introduced to biology is for counting possible states of the genome fulfilling some condition. We can measure the specificity of a binding motif by counting the number of possible sequences that

bind to a certain molecule relative to the number of all possible sequences of the genome in a region of the same size. Thus the method introduced by Schneider and Stephens (1990) represents each base with a height proportional to the information content of the position, i.e. a height related to how constrained the motif in that position is. If the motif can have only one possible base, say an 'A', then we reduce the number of possibilities from 4 to 1, i.e. a four-fold reduction. This four-fold reduction is represented by its log in base 2, so we say that the position specifies two bits of information about the possible binding motif, and its height is proportional to 2. If a position is not constrained at all, and can have any of the four bases, there is no reduction. Then that position specifies zero bits of information, and the height of that position is zero. Because of the additive properties of information, one can add the heights of all bases in the motif to get its overall specificity – two motifs with similar total height put the same amount of constraint on the sequence for binding. (This method of representation, only looking at single positions, ignores correlations between different bases.) We can also use a similar representation to specify motifs of amino acids in proteins.

Notice that the measure used here is not mutual information, but instead the reduction in uncertainty only for the regions that bind to the motif. The expression for mutual information between the binding of a molecule and the sequence of the genome will involve at least two possible outcomes of the binding state, e.g. either the molecule binds or it does not bind in the region, and then we will have a second term involving the reduction in uncertainty when we know the molecule does not bind in a region. Or it could be that we had a certain chance to be told that the molecule binds in the region in case it does, so the two possible events would be that we are told and we are not told that the molecule binds in the region. The measure introduced by Schneider and Stephens only includes the difference in binding regions. However, since the effect on non-binding regions is so slight, the second term has an insignificant contribution to the mutual information.

15.4 Conclusion

I have tried to highlight some of the uses of quantitative measures of information and entropy in cues and communication between organisms. Many biologists, including myself, hope that at some point entropy measures could take as central a role in the theory of evolution as they do in physics and information theory. But for now these promises are unfulfilled – few would argue that a course in information theory should be in the standard biology curriculum. On the other hand, small local uses of entropy are appearing, such

as describing the specificity of a binding motif. In the narrow sense described in this chapter, it is easy to agree that there is mutual information between a certain cue in the environment or signal given by another organism and the relevant environmental state of an organism. One might also look at the information in the genome in the same way – mutual information between the genome and the environment. The sense in which the term 'information' is then used is similar to its use in statistical mechanics and the theory of communication. Beyond these static measures, if we try to analyse the behaviour of organisms providing cues to one another, the mathematical analysis becomes much harder. One needs to then take into account factors such as manipulation or the stability of the signalling system to mutations and strategy changes by the participants. Other important factors would be how the detection of the cue originates, or how the detection of a signal or the emission of a signal originates in evolution. All these, however, are mathematical or modelling questions, more easily addressed than the question of whether or not the organism really sent a 'signal' or was just manipulating the receiver.

References

Adami, C. (2004). Information theory in molecular biology. *Physics of Life Reviews*, **1**, 3–22.

Bergstrom, C. & Lachmann, M. (2004). Shannon information and biological fitness. *Proceedings of the Information Theory Workshop, 2004. IEEE*, pp. 50–54.

Clausius, R. (1867). *The Mechanical Theory of Heat: With its Applications to the Steam-engine and to the Physical Properties of Bodies*. London: John Van Voorst.

Cohen, D. (1966). Optimizing reproduction in a randomly varying environment, *Journal of Theoretical Biology,* **12**, 119–129.

Cover, T. & Thomas, J. (1991). *Elements of Information Theory*. New York: Wiley.

Dall, S. R., Giraldeau, L.-A., Olsson, O., McNamara, J. M. & Stephens, D. W. (2005). Information and its use by animals in evolutionary ecology. *Trends in Ecology and Evolution*, **20**, 187–193.

Dall, S. R. X., Schmidt, K. A. & Van Gils, J. A. (2010). Biological information in an ecological context. *Oikos*, **119**, 201–202.

Donaldson-Matasci, M., Bergstrom, C. T. & Lachmann, M. (2010). The fitness value of information. *Oikos*, **119**, 219–230.

Donaldson-Matasci, M., Lachmann, M. & Bergstrom, C. T. (2008). Phenotypic diversity as an adaptation to environmental uncertainty. *Evolutionary Ecology Research*, **10**, 493–515.

Hasson, O. (1994). Cheating signals. *Journal of Theoretical Biology* **167**, 223–238.

Hirschleifer, J. (1971). The private and social value of information and the reward to inventive activity. *American Economic Review*, **61**, 561–574.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review*, **106**, 620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics. II. *Physical Review*, **108**, 171–190.

Kelly Jr, J. (1956). A new interpretation of information rate. *Bell System Technical Journal*, **35**, 917–926.

Kimura, M. (1961). Natural selection as the process of accumulating genetic information in adaptive evolution. *Genetics Research*, **2**, 127–140.

Kussell, E. & Leibler, S. (2005). Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, **309**, 2075–2078.

Maynard Smith, J. & Harper, D. (2003). *Animal Signals*. New York: Oxford University Press.

McMillan, B. (1953). The basic theorems of information theory. *Annals of Mathematical Statistics*, **24**, 196.

Rivoire, O. & Leibler, S. (2011). The value of information for populations in varying environments. *Journal of Statistical Physics*, **142**, 1124–1166.

Rodgers, J. L. & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician*, **42**, 59–66.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, **18**, 6097–6100.

Shannon, C. (1948). The mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423.

Sherwin, W. B. (2010). Entropy and information approaches to genetic diversity and its expression: genomic geography. *Entropy*, **12**, 1765–1798.

Stephens, D. (1989). Variance and the value of information. *American Naturalist*, **134**, 128–140.

Tribus, M. & McIrvine, E. (1971). Energy and information. *Scientific American*, **225**, 179–188.

Wagner, A. (2007). From bit to it: how a complex metabolic network transforms information into living matter. *BMC Systems Biology*, **1**, 33.