Unit	Expression that has those units
bit	$\log_2 p$
nat	$\log_e p$
ban	$\log_{10} p$
deciban (db)	$10 \log_{10} p$

Table 18.8. Units of measurement of information content.

The *bit* is the unit that we use most in this book. Because the word 'bit' has other meanings, a backup name for this unit is the *shannon*. A *byte* is 8 bits. A megabyte is $2^{20} \simeq 10^6$ bytes. If one works in natural logarithms, information contents and weights of evidence are measured in *nats*. The most interesting units are the *ban* and the *deciban*.

The history of the ban

Let me tell you why a factor of ten in probability is called a ban. When Alan Turing and the other codebreakers at Bletchley Park were breaking each new day's Enigma code, their task was a huge inference problem: to infer, given the day's cyphertext, which three wheels were in the Enigma machines that day; what their starting positions were; what further letter substitutions were in use on the steckerboard; and, not least, what the original German messages were. These inferences were conducted using Bayesian methods (of course!), and the chosen units were decibans or half-decibans, the deciban being judged the smallest weight of evidence discernible to a human. The evidence in favour of particular hypotheses was tallied using sheets of paper that were specially printed in Banbury, a town about 30 miles from Bletchley. The inference task was known as Banburismus, and the units in which Banburismus was played were called bans, after that town.

▶ 18.4 A taste of Banburismus

The details of the code-breaking methods of Bletchley Park were kept secret for a long time, but some aspects of Banburismus can be pieced together. I hope the following description of a small part of Banburismus is not too inaccurate.[1]

How much information was needed? The number of possible settings of the Enigma machine was about 8×10^{12}. To deduce the state of the machine, 'it was therefore necessary to find about 129 decibans from somewhere', as Good puts it. Banburismus was aimed not at deducing the entire state of the machine, but only at figuring out which wheels were in use; the logic-based bombes, fed with guesses of the plaintext (cribs), were then used to crack what the settings of the wheels were.

The Enigma machine, once its wheels and plugs were put in place, implemented a continually-changing permutation cypher that wandered deterministically through a state space of 26^3 permutations. Because an enormous number of messages were sent each day, there was a good chance that whatever state one machine was in when sending one character of a message, there would be another machine *in the same state* while sending a particular character in another message. Because the evolution of the machine's state was deterministic, the two machines would remain in the same state as each other

[1] I've been most helped by descriptions given by Tony Sale (`http://www.codesandciphers.org.uk/lectures/`) and by Jack Good (1979), who worked with Turing at Bletchley.

for the rest of the transmission. The resulting correlations between the outputs of such pairs of machines provided a dribble of information-content from which Turing and his co-workers extracted their daily 129 decibans.

How to detect that two messages came from machines with a common state sequence

The hypotheses are the null hypothesis, $\mathcal{H}_0$, which states that the machines are in *different* states, and that the two plain messages are unrelated; and the 'match' hypothesis, $\mathcal{H}_1$, which says that the machines are in the *same* state, and that the two plain messages are unrelated. No attempt is being made here to infer what the state of either machine is. The data provided are the two cyphertexts $\mathbf{x}$ and $\mathbf{y}$; let's assume they both have length T and that the alphabet size is A (26 in Enigma). What is the probability of the data, given the two hypotheses?

First, the null hypothesis. This hypothesis asserts that the two cyphertexts are given by

$$\mathbf{x} = x_1x_2x_3\ldots = c_1(u_1)c_2(u_2)c_3(u_3)\ldots \tag{18.17}$$

and

$$\mathbf{y} = y_1y_2y_3\ldots = c'_1(v_1)c'_2(v_2)c'_3(v_3)\ldots, \tag{18.18}$$

where the codes c_t and c'_t are two unrelated time-varying permutations of the alphabet, and $u_1u_2u_3\ldots$ and $v_1v_2v_3\ldots$ are the plaintext messages. An exact computation of the probability of the data $(\mathbf{x}, \mathbf{y})$ would depend on a language model of the plain text, and a model of the Enigma machine's guts, but if we assume that each Enigma machine is an *ideal* random time-varying permutation, then the probability distribution of the two cyphertexts is uniform. All cyphertexts are equally likely.

$$P(\mathbf{x}, \mathbf{y} \,|\, \mathcal{H}_0) = \left(\frac{1}{A}\right)^{2T} \quad \text{for all } \mathbf{x}, \mathbf{y} \text{ of length } T. \tag{18.19}$$

What about $\mathcal{H}_1$? This hypothesis asserts that a *single* time-varying permutation c_t underlies both

$$\mathbf{x} = x_1x_2x_3\ldots = c_1(u_1)c_2(u_2)c_3(u_3)\ldots \tag{18.20}$$

and

$$\mathbf{y} = y_1y_2y_3\ldots = c_1(v_1)c_2(v_2)c_3(v_3)\ldots\ . \tag{18.21}$$

What is the probability of the data $(\mathbf{x}, \mathbf{y})$? We have to make some assumptions about the plaintext language. If it were the case that the plaintext language was completely random, then the probability of $u_1u_2u_3\ldots$ and $v_1v_2v_3\ldots$ would be uniform, and so would that of $\mathbf{x}$ and $\mathbf{y}$, so the probability $P(\mathbf{x}, \mathbf{y} \,|\, \mathcal{H}_1)$ would be equal to $P(\mathbf{x}, \mathbf{y} \,|\, \mathcal{H}_0)$, and the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ would be indistinguishable.

We make progress by assuming that the plaintext is not completely random. Both plaintexts are written in a language, and that language has redundancies. Assume for example that particular plaintext letters are used more often than others. So, even though the two plaintext messages are unrelated, they are slightly more likely to use the same letters as each other; if $\mathcal{H}_1$ is true, two synchronized letters from the two cyphertexts are slightly more likely to be identical. Similarly, if a language uses particular bigrams and trigrams frequently, then the two plaintext messages will occasionally contain the same bigrams and trigrams at the same time as each other, giving rise, if $\mathcal{H}_1$ is true,

```
      u  LITTLE-JACK-HORNER-SAT-IN-THE-CORNER-EATING-A-CHRISTMAS-PIE--HE-PUT-IN-H
      v  RIDE-A-COCK-HORSE-TO-BANBURY-CROSS-TO-SEE-A-FINE-LADY-UPON-A-WHITE-HORSE
matches:  .*....*..******.*.............*..........*...............*..........
```

Table 18.9. Two aligned pieces of English plaintext, **u** and **v**, with matches marked by ∗. Notice that there are twelve matches, including a run of six, whereas the expected number of matches in two completely random strings of length $T = 74$ would be about 3. The two corresponding cyphertexts from two machines in identical states would also have twelve matches.

to a little burst of 2 or 3 identical letters. Table 18.9 shows such a coincidence in two plaintext messages that are unrelated, except that they are both written in English.

The codebreakers hunted among pairs of messages for pairs that were suspiciously similar to each other, counting up the numbers of matching monograms, bigrams, trigrams, etc. This method was first used by the Polish codebreaker Rejewski.

Let's look at the simple case of a monogram language model and estimate how long a message is needed to be able to decide whether two machines are in the same state. I'll assume the source language is monogram-English, the language in which successive letters are drawn i.i.d. from the probability distribution $\{p_i\}$ of figure 2.1. The probability of $\mathbf{x}$ and $\mathbf{y}$ is nonuniform: consider two single characters, $x_t = c_t(u_t)$ and $y_t = c_t(v_t)$; the probability that they are identical is

$$\sum_{u_t, v_t} P(u_t) P(v_t) \, \mathbb{1}[u_t = v_t] \;=\; \sum_i p_i^2 \;\equiv\; m. \tag{18.22}$$

We give this quantity the name m, for 'match probability'; for both English and German, m is about 2/26 rather than 1/26 (the value that would hold for a completely random language). Assuming that c_t is an ideal random permutation, the probability of x_t and y_t is, by symmetry,

$$P(x_t, y_t \,|\, \mathcal{H}_1) \;=\; \begin{cases} \frac{m}{A} & \text{if } x_t = y_t \\ \frac{(1-m)}{A(A-1)} & \text{for } x_t \neq y_t. \end{cases} \tag{18.23}$$

Given a pair of cyphertexts $\mathbf{x}$ and $\mathbf{y}$ of length T that match in M places and do not match in N places, the log evidence in favour of $\mathcal{H}_1$ is then

$$\log \frac{P(\mathbf{x}, \mathbf{y} \,|\, \mathcal{H}_1)}{P(\mathbf{x}, \mathbf{y} \,|\, \mathcal{H}_0)} \;=\; M \log \frac{m/A}{1/A^2} + N \log \frac{\frac{(1-m)}{A(A-1)}}{1/A^2} \tag{18.24}$$

$$\;=\; M \log mA + N \log \frac{(1-m)A}{A-1}. \tag{18.25}$$

Every match contributes $\log mA$ in favour of $\mathcal{H}_1$; every non-match contributes $\log \frac{A-1}{(1-m)A}$ in favour of $\mathcal{H}_0$.

Match probability for monogram-English	m	0.076
Coincidental match probability	$1/A$	0.037
log-evidence for $\mathcal{H}_1$ per match	$10 \log_{10} mA$	3.1 db
log-evidence for $\mathcal{H}_1$ per non-match	$10 \log_{10} \frac{(1-m)A}{(A-1)}$	−0.18 db

If there were $M = 4$ matches and $N = 47$ non-matches in a pair of length $T = 51$, for example, the weight of evidence in favour of $\mathcal{H}_1$ would be +4 decibans, or a likelihood ratio of 2.5 to 1 in favour.

The *expected* weight of evidence from a line of text of length $T = 20$ characters is the expectation of (18.25), which depends on whether $\mathcal{H}_1$ or $\mathcal{H}_0$ is true. If $\mathcal{H}_1$ is true then matches are expected to turn up at rate m, and the expected weight of evidence is 1.4 decibans per 20 characters. If $\mathcal{H}_0$ is true

then spurious matches are expected to turn up at rate $1/A$, and the expected weight of evidence is -1.1 decibans per 20 characters. Typically, roughly 400 characters need to be inspected in order to have a weight of evidence greater than a hundred to one (20 decibans) in favour of one hypothesis or the other.

So, two English plaintexts have more matches than two random strings. Furthermore, because consecutive characters in English are not independent, the bigram and trigram statistics of English are nonuniform and the matches tend to occur in bursts of consecutive matches. [The same observations also apply to German.] Using better language models, the evidence contributed by runs of matches was more accurately computed. Such a scoring system was worked out by Turing and refined by Good. Positive results were passed on to automated and human-powered codebreakers. According to Good, the longest false-positive that arose in this work was a string of 8 consecutive matches between two machines that were actually in unrelated states.

Further reading

For further reading about Turing and Bletchley Park, see Hodges (1983) and Good (1979). For an in-depth read about cryptography, Schneier's (1996) book is highly recommended. It is readable, clear, and entertaining.

▶ 18.5 Exercises

▷ Exercise 18.3.[2] Another weakness in the design of the Enigma machine, which was intended to emulate a perfectly random time-varying permutation, is that it never mapped a letter to itself. When you press `Q`, what comes out is always a different letter from `Q`. How much information per character is leaked by this design flaw? How long a crib would be needed to be confident that the crib is correctly aligned with the cyphertext? And how long a crib would be needed to be able confidently to identify the correct key?

[A *crib* is a guess for what the plaintext was. Imagine that the Brits know that a very important German is travelling from Berlin to Aachen, and they intercept Enigma-encoded messages sent to Aachen. It is a good bet that one or more of the original plaintext messages contains the string `OBERSTURMBANNFUEHRERXGRAFXHEINRICHXVONXWEIZSAECKER`, the name of the important chap. A crib could be used in a brute-force approach to find the correct Enigma key (feed the received messages through all possible Engima machines and see if any of the putative decoded texts match the above plaintext). This question centres on the idea that the crib can also be used in a much less expensive manner: slide the plaintext crib along all the encoded messages until a perfect *mismatch* of the crib and the encoded message is found; if correct, this alignment then tells you a lot about the key.]

19

Why have Sex? Information Acquisition and Evolution

Evolution has been happening on earth for about the last 10^9 years. Undeniably, *information has been acquired* during this process. Thanks to the tireless work of the Blind Watchmaker, some cells now carry within them all the information required to be outstanding spiders; other cells carry all the information required to make excellent octopuses. Where did this information come from?

The entire blueprint of all organisms on the planet has emerged in a teaching process in which the teacher is natural selection: fitter individuals have more progeny, the fitness being defined by the local environment (including the other organisms). The teaching signal is only a few bits per individual: an individual simply has a smaller or larger number of grandchildren, depending on the individual's fitness. 'Fitness' is a broad term that could cover

- the ability of an antelope to run faster than other antelopes and hence avoid being eaten by a lion;
- the ability of a lion to be well-enough camouflaged and run fast enough to catch one antelope per day;
- the ability of a peacock to attract a peahen to mate with it;
- the ability of a peahen to rear many young simultaneously.

The fitness of an organism is largely determined by its DNA – both the coding regions, or genes, and the non-coding regions (which play an important role in regulating the transcription of genes). We'll think of fitness as a function of the DNA sequence and the environment.

How does the DNA determine fitness, and how does information get from natural selection into the genome? Well, if the gene that codes for one of an antelope's proteins is defective, that antelope might get eaten by a lion early in life and have only two grandchildren rather than forty. The information content of natural selection is fully contained in a specification of which offspring survived to have children – an information content of *at most one bit per offspring*. The teaching signal does not communicate to the ecosystem any description of the imperfections in the organism that caused it to have fewer children. The bits of the teaching signal are highly redundant, because, throughout a species, unfit individuals who are similar to each other will be failing to have offspring for similar reasons.

So, how many bits per generation are acquired by the species as a whole by natural selection? How many bits has natural selection succeeded in conveying to the human branch of the tree of life, since the divergence between

Australopithecines and apes 4 000 000 years ago? Assuming a generation time of 10 years for reproduction, there have been about 400 000 generations of human precursors since the divergence from apes. Assuming a population of 10^9 individuals, each receiving a couple of bits of information from natural selection, the total number of bits of information responsible for modifying the genomes of 4 million B.C. into today's human genome is about 8×10^{14} bits. However, as we noted, natural selection is not smart at collating the information that it dishes out to the population, and there is a great deal of redundancy in that information. If the population size were twice as great, would it evolve twice as fast? No, because natural selection will simply be correcting the same defects twice as often.

John Maynard Smith has suggested that the rate of information acquisition by a species is independent of the population size, and is of order 1 bit per generation. This figure would allow for only 400 000 bits of difference between apes and humans, a number that is much smaller than the total size of the human genome – 6×10^9 bits. [One human genome contains about 3×10^9 nucleotides.] It is certainly the case that the genomic overlap between apes and humans is huge, but is the difference that small?

In this chapter, we'll develop a crude model of the process of information acquisition through evolution, based on the assumption that a gene with two defects is typically likely to be more defective than a gene with one defect, and an organism with two defective genes is likely to be less fit than an organism with one defective gene. Undeniably, this is a crude model, since real biological systems are baroque constructions with complex interactions. Nevertheless, we persist with a simple model because it readily yields striking results.

What we find from this simple model is that

1. John Maynard Smith's figure of 1 bit per generation is correct for an *asexually-reproducing* population;

2. in contrast, *if the species reproduces sexually*, the rate of information acquisition can be as large as $\sqrt{G}$ bits per generation, where G is the size of the genome.

We'll also find interesting results concerning the maximum mutation rate that a species can withstand.

▶ 19.1 The model

We study a simple model of a reproducing population of N individuals with a genome of size G bits: variation is produced by mutation or by recombination (i.e., sex) and truncation selection selects the N fittest children at each generation to be the parents of the next. We find striking differences between populations that have recombination and populations that do not.

The genotype of each individual is a vector $\mathbf{x}$ of G bits, each having a good state $x_g = 1$ and a bad state $x_g = 0$. The fitness $F(\mathbf{x})$ of an individual is simply the sum of her bits:

$$F(\mathbf{x}) = \sum_{g=1}^{G} x_g. \tag{19.1}$$

The bits in the genome could be considered to correspond either to genes that have good alleles ($x_g = 1$) and bad alleles ($x_g = 0$), or to the nucleotides of a genome. We will concentrate on the latter interpretation. The essential property of fitness that we are assuming is that it is locally a roughly linear function of the genome, that is, that there are many possible changes one

could make to the genome, each of which has a small effect on fitness, and that these effects combine approximately linearly.

We define the normalized fitness $f(\mathbf{x}) \equiv F(\mathbf{x})/G$.

We consider evolution by natural selection under two models of variation.

Variation by mutation. The model assumes discrete generations. At each generation, t, every individual produces two children. The children's genotypes differ from the parent's by random mutations. Natural selection selects the fittest N progeny in the child population to reproduce, and a new generation starts.

[The selection of the fittest N individuals at each generation is known as truncation selection.]

The simplest model of mutations is that the child's bits $\{x_g\}$ are independent. Each bit has a small probability of being flipped, which, thinking of the bits as corresponding roughly to nucleotides, is taken to be a constant m, independent of x_g. [If alternatively we thought of the bits as corresponding to genes, then we would model the probability of the discovery of a good gene, $P(x_g = 0 \rightarrow x_g = 1)$, as being a smaller number than the probability of a deleterious mutation in a good gene, $P(x_g = 1 \rightarrow x_g = 0)$.]

Variation by recombination (or crossover, or sex). Our organisms are haploid, not diploid. They enjoy sex by recombination. The N individuals in the population are married into $M = N/2$ couples, at random, and each couple has C children – with $C = 4$ children being our standard assumption, so as to have the population double and halve every generation, as before. The C children's genotypes are independent given the parents'. Each child obtains its genotype $\mathbf{z}$ by random crossover of its parents' genotypes, $\mathbf{x}$ and $\mathbf{y}$. The simplest model of recombination has no linkage, so that:

$$z_g = \begin{cases} x_g & \text{with probability } 1/2 \\ y_g & \text{with probability } 1/2. \end{cases} \tag{19.2}$$

Once the MC progeny have been born, the parents pass away, the fittest N progeny are selected by natural selection, and a new generation starts.

We now study these two models of variation in detail.

▶ 19.2 Rate of increase of fitness

Theory of mutations

We assume that the genotype of an individual with normalized fitness $f = F/G$ is subjected to mutations that flip bits with probability m. We first show that if the average normalized fitness f of the population is greater than $1/2$, then the optimal mutation rate is small, and the rate of acquisition of information is at most of order one bit per generation.

Since it is easy to achieve a normalized fitness of $f = 1/2$ by simple mutation, we'll assume $f > 1/2$ and work in terms of the excess normalized fitness $\delta f \equiv f - 1/2$. If an individual with excess normalized fitness δf has a child and the mutation rate m is small, the probability distribution of the excess normalized fitness of the child has mean

$$\overline{\delta f}_{\text{child}} = (1 - 2m)\delta f \tag{19.3}$$

and variance

$$\frac{m(1-m)}{G} \simeq \frac{m}{G}. \tag{19.4}$$

If the population of parents has mean $\delta f(t)$ and variance $\sigma^2(t) \equiv \beta m/G$, then the child population, before selection, will have mean $(1-2m)\delta f(t)$ and variance $(1+\beta)m/G$. Natural selection chooses the upper half of this distribution, so the mean fitness and variance of fitness at the next generation are given by

$$\delta f(t+1) = (1-2m)\delta f(t) + \alpha\sqrt{(1+\beta)}\sqrt{\frac{m}{G}}, \tag{19.5}$$

$$\sigma^2(t+1) = \gamma(1+\beta)\frac{m}{G}, \tag{19.6}$$

where α is the mean deviation from the mean, measured in standard deviations, and γ is the factor by which the child distribution's variance is reduced by selection. The numbers α and γ are of order 1. For the case of a Gaussian distribution, $\alpha = \sqrt{2/\pi} \simeq 0.8$ and $\gamma = (1-2/\pi) \simeq 0.36$. If we assume that the variance is in dynamic equilibrium, i.e., $\sigma^2(t+1) \simeq \sigma^2(t)$, then

$$\gamma(1+\beta) = \beta, \text{ so } (1+\beta) = \frac{1}{1-\gamma}, \tag{19.7}$$

and the factor $\alpha\sqrt{(1+\beta)}$ in equation (19.5) is equal to 1, if we take the results for the Gaussian distribution, an approximation that becomes poorest when the discreteness of fitness becomes important, i.e., for small m. The rate of increase of normalized fitness is thus:

$$\frac{\mathrm{d}f}{\mathrm{d}t} \simeq -2m\,\delta f + \sqrt{\frac{m}{G}}, \tag{19.8}$$

which, assuming $G(\delta f)^2 \gg 1$, is maximized for

$$m_{\mathrm{opt}} = \frac{1}{16G(\delta f)^2}, \tag{19.9}$$

at which point,

$$\left(\frac{\mathrm{d}f}{\mathrm{d}t}\right)_{\mathrm{opt}} = \frac{1}{8G(\delta f)}. \tag{19.10}$$

So the rate of increase of fitness $F = fG$ is at most

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \frac{1}{8(\delta f)} \text{ per generation.} \tag{19.11}$$

For a population with low fitness ($\delta f < 0.125$), the rate of increase of fitness may exceed 1 unit per generation. Indeed, if $\delta f \lesssim 1/\sqrt{G}$, the rate of increase, if $m = 1/2$, is of order $\sqrt{G}$; this initial spurt can last only of order $\sqrt{G}$ generations. For $\delta f > 0.125$, the rate of increase of fitness is smaller than one per generation. As the fitness approaches G, the optimal mutation rate tends to $m = 1/(4G)$, so that an average of $1/4$ bits are flipped per genotype, and the rate of increase of fitness is also equal to $1/4$; information is gained at a rate of about 0.5 bits per generation. It takes about $2G$ generations for the genotypes of all individuals in the population to attain perfection.

For fixed m, the fitness is given by

$$\delta f(t) = \frac{1}{2\sqrt{mG}}(1 - c\,e^{-2mt}), \tag{19.12}$$

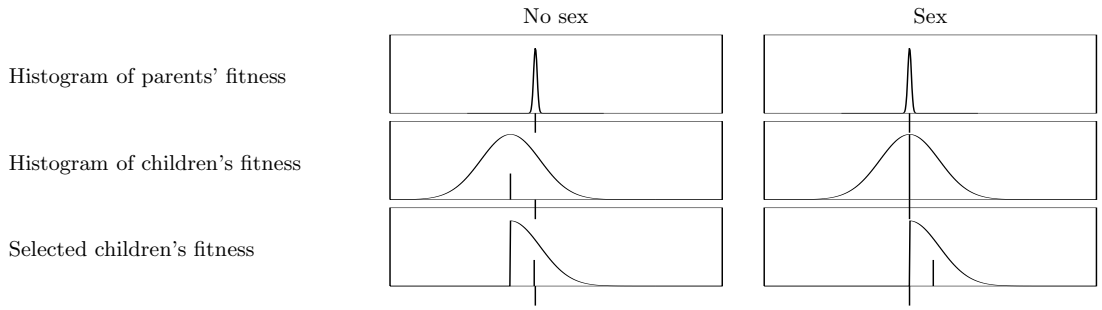

Figure 19.1. Why sex is better than sex-free reproduction. If mutations are used to create variation among children, then it is unavoidable that the average fitness of the children is lower than the parents' fitness; the greater the variation, the greater the average deficit. Selection bumps up the mean fitness again. In contrast, recombination produces variation without a decrease in average fitness. The typical amount of variation scales as $\sqrt{G}$, where G is the genome size, so after selection, the average fitness rises by $O(\sqrt{G})$.

subject to the constraint $\delta f(t) \leq 1/2$, where c is a constant of integration, equal to 1 if $f(0) = 1/2$. If the mean number of bits flipped per genotype, mG, exceeds 1, then the fitness F approaches an equilibrium value $F_{\rm eqm} = (1/2 + 1/(2\sqrt{mG}))G$.

This theory is somewhat inaccurate in that the true probability distribution of fitness is non-Gaussian, asymmetrical, and quantized to integer values. All the same, the predictions of the theory are not grossly at variance with the results of simulations described below.

Theory of sex

The analysis of the sexual population becomes tractable with two approximations: first, we assume that the gene-pool mixes sufficiently rapidly that correlations between genes can be neglected; second, we assume *homogeneity*, i.e., that the fraction f_g of bits g that are in the good state is the same, $f(t)$, for all g.

Given these assumptions, if two parents of fitness $F = fG$ mate, the probability distribution of their children's fitness has mean equal to the parents' fitness, F; the variation produced by sex does not reduce the average fitness. The standard deviation of the fitness of the children scales as $\sqrt{Gf(1-f)}$. Since, after selection, the increase in fitness is proportional to this standard deviation, *the fitness increase per generation scales as the square root of the size of the genome,* $\sqrt{G}$. As shown in box 19.2, the mean fitness $\bar{F} = fG$ evolves in accordance with the differential equation:

$$\frac{{\rm d}\bar{F}}{{\rm d}t} \simeq \eta\sqrt{f(t)(1-f(t))G}, \tag{19.13}$$

where $\eta \equiv \sqrt{2/(\pi+2)}$. The solution of this equation is

$$f(t) = \frac{1}{2}\left[1 + \sin\left(\frac{\eta}{\sqrt{G}}(t+c)\right)\right], \quad \text{for } t+c \in \left(-\tfrac{\pi}{2}\sqrt{G}/\eta, \tfrac{\pi}{2}\sqrt{G}/\eta\right), \tag{19.14}$$

where c is a constant of integration, $c = \sin^{-1}(2f(0)-1)$. So this idealized system reaches a state of eugenic perfection ($f = 1$) within a finite time: $(\pi/\eta)\sqrt{G}$ generations.

Simulations

Figure 19.3a shows the fitness of a sexual population of $N = 1000$ individuals with a genome size of $G = 1000$ starting from a random initial state with normalized fitness 0.5. It also shows the theoretical curve $f(t)G$ from equation (19.14), which fits remarkably well.

In contrast, figures 19.3(b) and (c) show the evolving fitness when variation is produced by mutation at rates $m = 0.25/G$ and $m = 6/G$ respectively. Note the difference in the horizontal scales from panel (a).

Box 19.2. Details of the theory of sex.

How does $f(t+1)$ depend on $f(t)$? Let's first assume the two parents of a child both have exactly $f(t)G$ good bits, and, by our homogeneity assumption, that those bits are independent random subsets of the G bits. The number of bits that are good in both parents is roughly $f(t)^2G$, and the number that are good in one parent only is roughly $2f(t)(1-f(t))G$, so the fitness of the child will be $f(t)^2G$ plus the sum of $2f(t)(1-f(t))G$ fair coin flips, which has a binomial distribution of mean $f(t)(1-f(t))G$ and variance $\frac{1}{2}f(t)(1-f(t))G$. The fitness of a child is thus roughly distributed as

$$F_{\text{child}} \sim \text{Normal}\left(\text{mean} = f(t)G, \text{variance} = \frac{1}{2}f(t)(1-f(t))G\right).$$

The important property of this distribution, contrasted with the distribution under mutation, is that the mean fitness is equal to the parents' fitness; the variation produced by sex does not reduce the average fitness.

If we include the parental population's variance, which we will write as $\sigma^2(t) = \beta(t)\frac{1}{2}f(t)(1-f(t))G$, the children's fitnesses are distributed as

$$F_{\text{child}} \sim \text{Normal}\left(\text{mean} = f(t)G, \text{variance} = \left(1+\frac{\beta}{2}\right)\frac{1}{2}f(t)(1-f(t))G\right).$$

Natural selection selects the children on the upper side of this distribution. The mean increase in fitness will be

$$\bar{F}(t+1) - \bar{F}(t) = [\alpha(1+\beta/2)^{1/2}/\sqrt{2}]\sqrt{f(t)(1-f(t))G},$$

and the variance of the surviving children will be

$$\sigma^2(t+1) = \gamma(1+\beta/2)\frac{1}{2}f(t)(1-f(t))G,$$

where $\alpha = \sqrt{2/\pi}$ and $\gamma = (1-2/\pi)$. If there is dynamic equilibrium $[\sigma^2(t+1) = \sigma^2(t)]$ then the factor in (19.2) is

$$\alpha(1+\beta/2)^{1/2}/\sqrt{2} = \sqrt{\frac{2}{(\pi+2)}} \simeq 0.62.$$

Defining this constant to be $\eta \equiv \sqrt{2/(\pi+2)}$, we conclude that, under sex and natural selection, the mean fitness of the population increases at a rate *proportional to the square root of the size of the genome*,

$$\frac{\mathrm{d}\bar{F}}{\mathrm{d}t} \simeq \eta\sqrt{f(t)(1-f(t))G} \text{ bits per generation.}$$

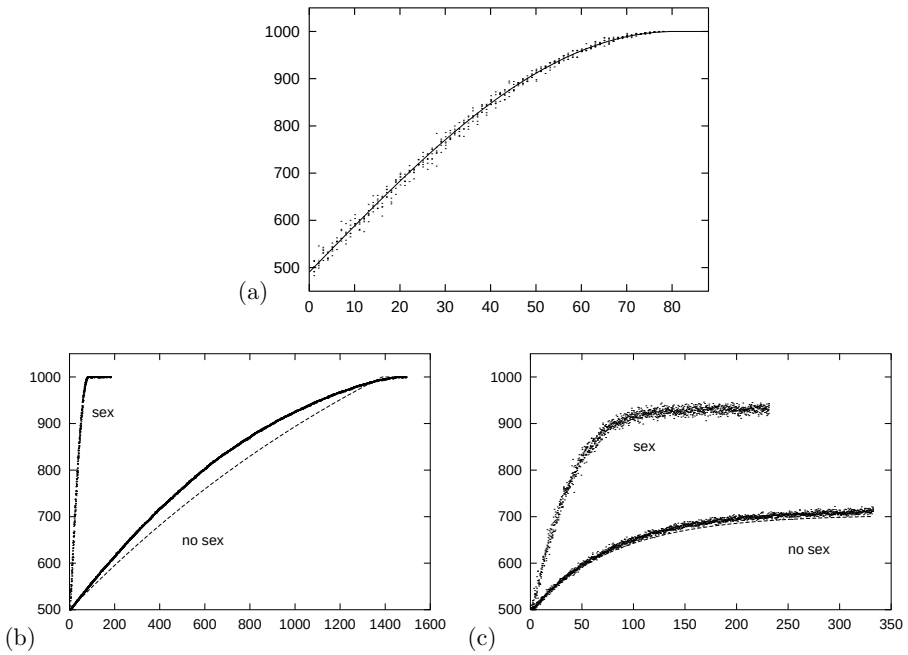

Figure 19.3. Fitness as a function of time. The genome size is $G = 1000$. The dots show the fitness of six randomly selected individuals from the birth population at each generation. The initial population of $N = 1000$ had randomly generated genomes with $f(0) = 0.5$ (exactly). (a) Variation produced by sex alone. Line shows theoretical curve (19.14) for infinite homogeneous population. (b,c) Variation produced by mutation, with and without sex, when the mutation rate is $mG = 0.25$ (b) or 6 (c) bits per genome. The dashed line shows the curve (19.12).

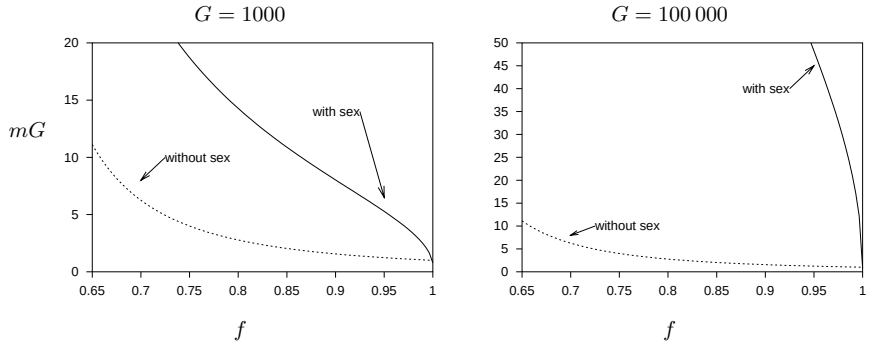

Figure 19.4. Maximal tolerable mutation rate, shown as number of errors per genome (mG), versus normalized fitness $f = F/G$. Left panel: genome size $G = 1000$; right: $G = 100\,000$. Independent of genome size, a parthenogenetic species (no sex) can tolerate only of order 1 error per genome per generation; a species that uses recombination (sex) can tolerate far greater mutation rates.

Exercise 19.1.[3, p.280] Dependence on population size. How do the results for a sexual population depend on the population size? We anticipate that there is a minimum population size above which the theory of sex is accurate. How is that minimum population size related to G?

Exercise 19.2.[3] Dependence on crossover mechanism. In the simple model of sex, each bit is taken at random from one of the two parents, that is, we allow crossovers to occur with probability 50% between any two adjacent nucleotides. How is the model affected (a) if the crossover probability is smaller? (b) if crossovers occur exclusively at *hot-spots* located every d bits along the genome?

▶ 19.3 The maximal tolerable mutation rate

What if we combine the two models of variation? What is the maximum mutation rate that can be tolerated by a species that has sex?

The rate of increase of fitness is given by

$$\frac{\mathrm{d}f}{\mathrm{d}t} \simeq -2m\,\delta f + \eta\sqrt{2}\sqrt{\frac{m + f(1-f)/2}{G}}, \tag{19.15}$$

which is positive if the mutation rate satisfies

$$m < \eta\sqrt{\frac{f(1-f)}{G}}. \tag{19.16}$$

Let us compare this rate with the result in the absence of sex, which, from equation (19.8), is that the maximum tolerable mutation rate is

$$m < \frac{1}{G}\frac{1}{(2\,\delta f)^2}. \tag{19.17}$$

The tolerable mutation rate with sex is of order $\sqrt{G}$ times greater than that without sex!

A parthenogenetic (non-sexual) species could try to wriggle out of this bound on its mutation rate by increasing its litter sizes. But if mutation flips on average mG bits, the probability that no bits are flipped in one genome is roughly e^{-mG}, so a mother needs to have roughly e^{mG} offspring in order to have a good chance of having one child with the same fitness as her. The litter size of a non-sexual species thus has to be exponential in mG (if mG is bigger than 1), if the species is to persist.

So the maximum tolerable mutation rate is pinned close to $1/G$, for a non-sexual species, whereas it is a larger number of order $1/\sqrt{G}$, for a species with recombination.

Turning these results around, we can predict the largest possible genome size for a given fixed mutation rate, m. For a parthenogenetic species, the largest genome size is of order $1/m$, and for a sexual species, $1/m^2$. Taking the figure $m = 10^{-8}$ as the mutation rate per nucleotide per generation (Eyre-Walker and Keightley, 1999), and allowing for a maximum brood size of 20 000 (that is, $mG \simeq 10$), we predict that all species with more than $G = 10^9$ coding nucleotides make at least occasional use of recombination. If the brood size is 12, then this number falls to $G = 2.5 \times 10^8$.

▶ 19.4 Fitness increase and information acquisition

For this simple model it is possible to relate increasing fitness to information acquisition.

If the bits are set at random, the fitness is roughly $F = G/2$. If evolution leads to a population in which all individuals have the maximum fitness $F = G$, then G bits of information have been acquired by the species, namely for each bit x_g, the species has figured out which of the two states is the better.

We define the information acquired at an intermediate fitness to be the amount of selection (measured in bits) required to select the perfect state from the gene pool. Let a fraction f_g of the population have $x_g = 1$. Because $\log_2(1/f)$ is the information required to find a black ball in an urn containing black and white balls in the ratio $f : 1-f$, we define the information acquired to be

$$I = \sum_g \log_2 \frac{f_g}{1/2} \text{bits}. \tag{19.18}$$

If all the fractions f_g are equal to F/G, then

$$I = G \log_2 \frac{2F}{G}, \tag{19.19}$$

which is well approximated by

$$\tilde{I} \equiv 2(F - G/2). \tag{19.20}$$

The rate of information acquisition is thus roughly two times the rate of increase of fitness in the population.

▶ 19.5 Discussion

These results quantify the well known argument for why species reproduce by sex with recombination, namely that recombination allows useful mutations to spread more rapidly through the species and allows deleterious mutations to be more rapidly cleared from the population (Maynard Smith, 1978; Felsenstein, 1985; Maynard Smith, 1988; Maynard Smith and Száthmary, 1995). A population that reproduces by recombination can acquire information from natural selection at a rate of order $\sqrt{G}$ times faster than a parthenogenetic population, and it can tolerate a mutation rate that is of order $\sqrt{G}$ times greater. For genomes of size $G \simeq 10^8$ coding nucleotides, this factor of $\sqrt{G}$ is substantial.

This enormous advantage conferred by sex has been noted before by Kondrashov (1988), but this meme, which Kondrashov calls 'the deterministic mutation hypothesis', does not seem to have diffused throughout the evolutionary research community, as there are still numerous papers in which the prevalence of sex is viewed as a mystery to be explained by elaborate mechanisms.

'The cost of males' – stability of a gene for sex or parthenogenesis

Why do people declare sex to be a mystery? The main motivation for being mystified is an idea called the 'cost of males'. Sexual reproduction is disadvantageous compared with asexual reproduction, it's argued, because of every two offspring produced by sex, one (on average) is a useless male, incapable of child-bearing, and only one is a productive female. In the same time, a parthenogenetic mother could give birth to *two* female clones. To put it another way, the big advantage of parthenogenesis, from the point of view of the individual, is that one is able to pass on 100% of one's genome to one's children, instead of only 50%. Thus if there were two versions of a species, one reproducing with and one without sex, the single mothers would be expected to outstrip their sexual cousins. The simple model presented thus far did not include either genders or the ability to convert from sexual reproduction to asexual, but we can easily modify the model.

We modify the model so that one of the G bits in the genome determines whether an individual prefers to reproduce parthenogenetically ($x = 1$) or sexually ($x = 0$). The results depend on the number of children had by a single parthenogenetic mother, $K_{\rm p}$ and the number of children born by a sexual couple, $K_{\rm s}$. Both ($K_{\rm p} = 2$, $K_{\rm s} = 4$) and ($K_{\rm p} = 4$, $K_{\rm s} = 4$) are reasonable models. The former ($K_{\rm p} = 2$, $K_{\rm s} = 4$) would seem most appropriate in the case of unicellular organisms, where the cytoplasm of both parents goes into the children. The latter ($K_{\rm p} = 4$, $K_{\rm s} = 4$) is appropriate if the children are solely nurtured by one of the parents, so single mothers have just as many offspring as a sexual pair. I concentrate on the latter model, since it gives the greatest advantage to the parthenogens, who are supposedly expected to outbreed the sexual community. Because parthenogens have four children per generation, the maximum tolerable mutation rate for them is twice the expression (19.17) derived before for $K_{\rm p} = 2$. If the fitness is large, the maximum tolerable rate is $mG \simeq 2$.

Initially the genomes are set randomly with $F = G/2$, with half of the population having the gene for parthenogenesis. Figure 19.5 shows the outcome. During the 'learning' phase of evolution, in which the fitness is increasing rapidly, pockets of parthenogens appear briefly, but then disappear within a couple of generations as their sexual cousins overtake them in fitness and

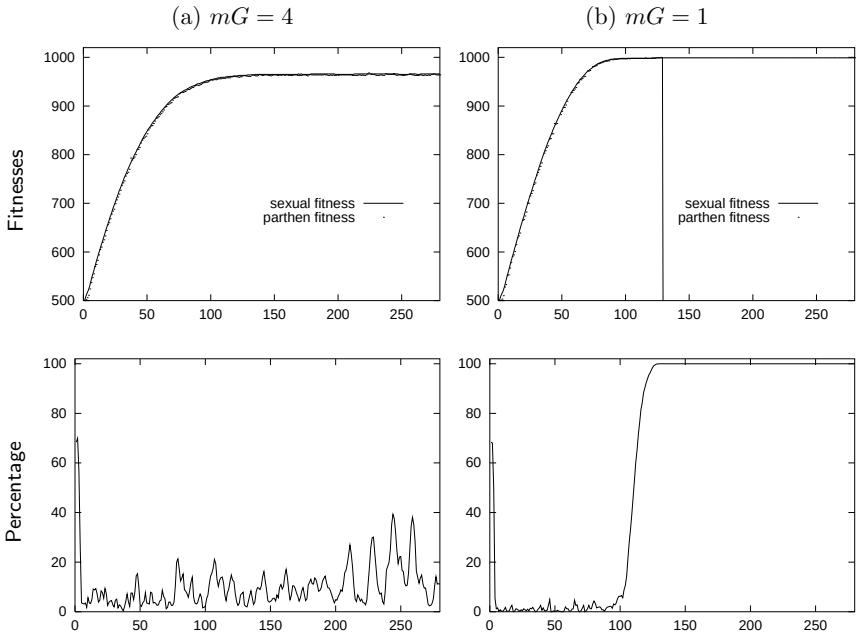

Figure 19.5. Results when there is a gene for parthenogenesis, and no interbreeding, *and single mothers produce as many children as sexual couples.* $G = 1000$, $N = 1000$. (a) $mG = 4$; (b) $mG = 1$. Vertical axes show the fitnesses of the two sub-populations, and the percentage of the population that is parthenogenetic.

leave them behind. Once the population reaches its top fitness, however, the parthenogens can take over, if the mutation rate is sufficiently low ($mG = 1$).

In the presence of a higher mutation rate ($mG = 4$), however, the parthenogens never take over. The breadth of the sexual population's fitness is of order $\sqrt{G}$, so a mutant parthenogenetic colony arising with slightly above-average fitness will last for about $\sqrt{G}/(mG) = 1/(m\sqrt{G})$ generations before its fitness falls below that of its sexual cousins. As long as the population size is sufficiently large for some sexual individuals to survive for this time, sex will not die out.

In a sufficiently unstable environment, where the fitness function is continually changing, the parthenogens will always lag behind the sexual community. These results are consistent with the argument of Haldane and Hamilton (2002) that sex is helpful in an arms race with parasites. The parasites define an effective fitness function which changes with time, and a sexual population will always ascend the current fitness function more rapidly.

Additive fitness function

Of course, our results depend on the fitness function that we assume, and on our model of selection. Is it reasonable to model fitness, to first order, as a *sum* of independent terms? Maynard Smith (1968) argues that it is: the more good genes you have, the higher you come in the pecking order, for example. The directional selection model has been used extensively in theoretical population genetic studies (Bulmer, 1985). We might expect real fitness functions to involve interactions, in which case crossover might reduce the average fitness. However, since recombination gives the biggest advantage to species whose fitness functions are additive, we might predict that *evolution will have favoured species that used a representation of the genome that corresponds to a fitness function that has only weak interactions.* And even if there are interactions, it seems plausible that the fitness would still involve a sum of such interacting terms, with the number of terms being some fraction of the genome size G.

Exercise 19.3.[3C] Investigate how fast sexual and asexual species evolve if they have a fitness function with interactions. For example, let the fitness be a sum of exclusive-ors of pairs of bits; compare the evolving fitnesses with those of the sexual and asexual species with a simple additive fitness function.

Furthermore, if the fitness function were a highly nonlinear function of the genotype, it could be made more smooth and locally linear by the Baldwin effect. The Baldwin effect (Baldwin, 1896; Hinton and Nowlan, 1987) has been widely studied as a mechanism whereby *learning* guides evolution, and it could also act at the level of transcription and translation. Consider the evolution of a peptide sequence for a new purpose. Assume the effectiveness of the peptide is a highly nonlinear function of the sequence, perhaps having a small island of good sequences surrounded by an ocean of equally bad sequences. In an organism whose transcription and translation machinery is flawless, the fitness will be an equally nonlinear function of the DNA sequence, and evolution will wander around the ocean making progress towards the island only by a random walk. In contrast, an organism having the same DNA sequence, but whose DNA-to-RNA transcription or RNA-to-protein translation is 'faulty', will occasionally, by mistranslation or mistranscription, accidentally produce a working enzyme; and it will do so with greater probability if its DNA sequence is close to a good sequence. One cell might produce 1000 proteins from the one mRNA sequence, of which 999 have no enzymatic effect, and one does. The one working catalyst will be enough for that cell to have an increased fitness relative to rivals whose DNA sequence is further from the island of good sequences. For this reason I conjecture that, at least early in evolution, and perhaps still now, the genetic code was not implemented perfectly but was implemented noisily, with some codons coding for a distribution of possible amino acids. This noisy code could even be switched on and off from cell to cell in an organism by having multiple aminoacyl-tRNA synthetases, some more reliable than others.

Whilst our model assumed that the bits of the genome do not interact, ignored the fact that the information is represented redundantly, assumed that there is a direct relationship between phenotypic fitness and the genotype, and assumed that the crossover probability in recombination is high, I believe these qualitative results would still hold if more complex models of fitness and crossover were used: the relative benefit of sex will still scale as $\sqrt{G}$. Only in small, in-bred populations are the benefits of sex expected to be diminished.

In summary: Why have sex? Because sex is good for your bits!

Further reading

How did a high-information-content self-replicating system ever emerge in the first place? In the general area of the origins of life and other tricky questions about evolution, I highly recommend Maynard Smith and Száthmary (1995), Maynard Smith and Száthmary (1999), Kondrashov (1988), Maynard Smith (1988), Ridley (2000), Dyson (1985), Cairns-Smith (1985), and Hopfield (1978).

▶ 19.6 Further exercises

Exercise 19.4.[3] How good must the error-correcting machinery in DNA replication be, given that mammals have not all died out long ago? Estimate the probability of nucleotide substitution, per cell division. [See Appendix C.4.]

Exercise 19.5.[4] Given that DNA replication is achieved by bumbling Brownian motion and ordinary thermodynamics in a biochemical porridge at a temperature of 35 C, it's astonishing that the error-rate of DNA replication is about 10^{-9} per replicated nucleotide. How can this reliability be achieved, given that the energetic difference between a correct base-pairing and an incorrect one is only one or two hydrogen bonds and the thermal energy kT is only about a factor of four smaller than the free energy associated with a hydrogen bond? If ordinary thermodynamics is what favours correct base-pairing, surely the frequency of incorrect base-pairing should be about

$$f = \exp(-\Delta E/kT), \tag{19.21}$$

where ΔE is the free energy difference, i.e., an error frequency of $f \simeq 10^{-4}$? How has DNA replication cheated thermodynamics?

The situation is equally perplexing in the case of protein synthesis, which translates an mRNA sequence into a polypeptide in accordance with the genetic code. Two specific chemical reactions are protected against errors: the binding of tRNA molecules to amino acids, and the production of the polypeptide in the ribosome, which, like DNA replication, involves base-pairing. Again, the fidelity is high (an error rate of about 10^{-4}), and this fidelity can't be caused by the energy of the 'correct' final state being especially low – the correct polypeptide sequence is not expected to be significantly lower in energy than any other sequence. How do cells perform error correction? (See Hopfield (1974), Hopfield (1980)).

Exercise 19.6.[2] While the genome acquires information through natural selection at a rate of a few bits per generation, your brain acquires information at a greater rate.

Estimate at what rate new information can be stored in long term memory by your brain. Think of learning the words of a new language, for example.

▶ 19.7 Solutions

Solution to exercise 19.1 (p.275). For small enough N, whilst the average fitness of the population increases, some unlucky bits become frozen into the bad state. (These bad genes are sometimes known as hitchhikers.) The homogeneity assumption breaks down. Eventually, all individuals have identical genotypes that are mainly 1-bits, but contain some 0-bits too. The smaller the population, the greater the number of frozen 0-bits is expected to be. How small can the population size N be if the theory of sex is accurate?

We find experimentally that the theory based on assuming homogeneity fits poorly only if the population size N is smaller than $\sim\sqrt{G}$. If N is significantly smaller than $\sqrt{G}$, information cannot possibly be acquired at a rate as big as $\sqrt{G}$, since the information content of the Blind Watchmaker's decisions cannot be any greater than $2N$ bits per generation, this being the number of bits required to specify which of the $2N$ children get to reproduce. Baum *et al.* (1995), analyzing a similar model, show that the population size N should be about $\sqrt{G}(\log G)^2$ to make hitchhikers unlikely to arise.