

Natural selection

Gil McVean

The time-scale of evolution

Think of the 3.5 billion years that makes the history of life on earth as stretching for a single year. On the first of January, there is nothing more than primitive bacteria, and this remains the case for the next eight months. Multicellular eukaryotic life appears in early September, vertebrates invade the land at the beginning of December and dinosaurs become extinct sometime on Christmas eve. The history of humans can be squeezed into the last day. Just before noon, the lineage leading to modern humans makes the split from that leading to chimpanzees. And just 15 minutes before midnight on the 31st December, modern humans appear. By way of comparison, the life span of a human is hardly more than half a second.

The time scale on which species appear, change and become extinct is beyond the limits of human experience, or even comprehension. So how can we learn about processes that operate over such immeasurable periods of time? There are two possibilities. First, we may collect evidence about long-term historical processes, either through the fossil record or the genetic differences that separate extant species. Second, we can take a snap-shot of the microscopic details of evolution and hope that the forces that have shaped the diversity of life on earth are at work, right now, under our noses.

Such reductionism is the rationale for population genetics. By studying contemporary patterns of genetic and phenotypic variation, we hope not just to infer details of the recent past, but also provide evidence for what forces have been important on much larger time scales. For example, humans and chimps diverged in the region of 250,000 generations ago. Population genetics gives us a window stretching back at most about $4N_e$ generations. The effective population size of humans is probably about 10,000. If so, current levels of diversity will tell us about events stretching back 20% of the time to divergence.

The effect of natural selection on allele frequencies

Evolution is usually defined as modification with descent. At the finest time scale possible, this is nothing more than a change in allele frequencies over time. Reversing the argument, if we can describe how natural selection changes allele frequencies, we can describe evolution.

Consider a population in which there are two alleles segregating, and that these alleles influence the number of offspring that an individual will produce – perhaps one allele is resistant to a pathogen, or one is a disease allele. The expected genotype frequencies in a randomly mating population are given by the Hardy-Weinberg equilibrium, and the number of offspring a given genotype is expected to produce is called its fitness or viability. The number of descendents left by a given allele is the sum of the expected number of offspring for each genotype weighted by the frequency of that genotype and the number of the alleles of interest in it. However, we are interested in allele frequency rather than absolute number (we don't really care if the population is growing, shrinking, or maintaining a constant size). So all that matters is the number of offspring left by an

allele, relative to all other alleles. So we can arbitrarily set the relative fitness of one genotype to 1. A conventional way of representing fitness and relative fitness is

Genotype	AA	Aa	aa
Frequency	$(1 - x)^2$	$2x(1 - x)$	x^2
Viability	w_{AA}	w_{Aa}	w_{aa}
Relative fitness	1	$1 - hs$	$1 - s$

Working with relative fitnesses, the expected frequency of a in the next generation is the marginal fitness of an allele divided by the average population fitness.

$$E[x'] = \frac{x(1-x)(1-hs) + x^2s}{\bar{w}}$$

where

$$\bar{w} = 1 - 2x(1-x)hs - x^2s$$

is the average fitness of the population (actually the average relative fitness). We can also express this relationship as the expected change in allele frequency due to selection. Written in terms of h and s it is not very illuminating. However, Wright found an alternative expression

$$E_s[\Delta x] = \frac{x(1-x)}{2\bar{w}} \frac{d\bar{w}}{dx}$$

The subscript s on the expectation means it is the expected change due to selection. This formula has two key points. First, the allele is expected to increase in frequency when an increase in frequency means an increase in the average fitness of the population. Second, the rate of change in allele frequency is greatest for intermediate frequencies. Selection acting on very rare or very common alleles will tend to have less impact on allele frequencies.

Fisher's fundamental theory

This last point is related to an observation initially made by Fisher. As a consequence of his work on quantitative genetics, Fisher showed that the rate of change in a character due to natural selection is directly proportional to the additive genetic variance for that character. More heritable variation means that a population will respond faster to selection for a different value. This is related to Wright's formulation, because additive genetic variance is greatest when alleles are at intermediate frequency. So the rate of change in allele frequency is greatest at the same point as when the rate of change in character is also greatest. The paradox is that such directional selection acts to increase the frequency of favourable alleles, and as a result heritable variation is lost and the rate of evolution slows down. Eventually, evolution can only proceed further by the input of new variation through mutation.

Natural selection can take many forms, each of which may leave different traces on patterns of genetic variability. The three most important from the point of view of understanding genetic variability are selection against recurring deleterious mutations, balancing selection and directional, adaptive evolution.

Deleterious mutations in populations

There is one type of selection that is evident in all populations – deleterious mutation. And deleterious mutations are ubiquitous. Studies of inbreeding in humans suggest the every individual carries several deleterious mutations. For example, in Sweden, about 4% of offspring born to unrelated parents have some form of disease or physical or mental defect, whereas among first-cousin marriages the proportion is 16%. A crude estimate of the number of loci leading to such defects suggests that every individual carries at least 5 recessive deleterious mutations.

In *Drosophila*, an experiment in which a number of wild strains were made homozygous for one of their chromosomes revealed that over 20% of all 2nd chromosomes carried a lethal recessive allele (Rayla and Crow, 1960). Given that the 2nd chromosome represents only 1/5th of the total DNA, on average, every *Drosophila* carries at least one lethal mutation. Many more 2nd chromosomes carried sub-lethal, deleterious alleles that considerably reduced viability.

The prevalence of disease, or other deleterious mutations depends on three factors; whether mutations are dominant or recessive, the rate of mutation to disease alleles, and the decrease in viability due to the mutation. The last effect can vary enormously. Cystic fibrosis is a crippling condition that essentially prevents reproduction. However other severe diseases, such as Duchenne’s muscular dystrophy have much less effect on viability because they occur late in life. Other conditions, such as thalasseмии, haemophilia or even colour-blindness have intermediate effects.

We can use population genetics to predict the relationship between these three factors and the prevalence of disease alleles. Consider autosomal dominant mutations such as those that cause Huntington’s disease. The fitness of the genotypes can be represented as

Genotype	AA	Aa	aa
Frequency	$(1 - x)^2$	$2x(1 - x)$	x^2
Relative fitness	1	$1 - s$	$1 - s$

So the average fitness of the population is

$$\bar{w} = 1 - 2sx(1 - x) - sx^2$$

The change in allele frequency per generation is

$$E_s[\Delta x] = \frac{x(1 - x)}{2\bar{w}}[-2s + 4xs]$$

Assuming that the average fitness is close to one, and that selection against the disease is so strong that almost all disease alleles are found in heterozygotes, this simplifies to

$$E_s[\Delta x] \approx -sx(1 - x)$$

The equilibrium frequency of disease alleles will occur when the rate of mutation to new alleles, which occurs at the rate

$$E_u[\Delta x] = (1 - x)u$$

is equal to the rate of loss of alleles due to selection. This gives

$$sx(1-x) = (1-x)u$$

$$\tilde{x} = \frac{u}{s}$$

The equilibrium frequency of disease alleles is the mutation rate divided by the strength of selection against the disease (the decrease in reproductive success). For example, if the rate of mutation to new disease alleles is about 1 per million copies, and the disease reduces reproductive success by 2%, the expected allele frequency is about 1 in 20,000. These figures are roughly compatible with what is known about the mutation rate and incidence of Huntington's disease in Western Europe.

A very different result occurs when we consider recessive autosomal diseases such as cystic fibrosis. Writing the genotype fitnesses as

Genotype	AA	Aa	aa
Frequency	$(1-x)^2$	$2x(1-x)$	x^2
Relative fitness	1	1	$1-s$

The average population fitness is

$$\bar{w} = 1 - sx^2$$

And the expected change in allele frequency due to selection is approximately

$$E_s[\Delta x] = -sx^2(1-x)$$

Solving for the equilibrium allele frequency

$$\tilde{x} = \sqrt{\frac{u}{s}}$$

For example, with the same parameters as for the dominant case, the expected allele frequency is about 1 in 140. The reason this is so much higher is that selection against recessive alleles can only occur in homozygotes, which are very rare when allele frequencies are low. So the disease allele can drift to moderate frequency. For example, the incidence of cystic fibrosis is about 1 in 4000 in the US, so even though the disease is extremely debilitating, in the region of 3% of individuals carry the disease allele.

Mutation load arguments

The issue of how much deleterious mutation affects the overall fitness of a population has received considerable attention. The concept known as genetic load can be defined as the reduction in the total reproduction rate due to the mutation. This is equal to the incidence of a disease multiplied by its effect on fitness. So, for an autosomal dominant disease this is

$$L = 2\tilde{x}(1-\tilde{x})s$$

$$\approx 2\frac{u}{s}s = 2u$$

In other words, the genetic load is independent of the strength of selection against the disease. For a recessive disease we have

$$L = \tilde{x}^2s$$

$$= u$$

Again, this is independent of the selection coefficient and is half that of a dominant disease. If genetic loads are multiplicative across loci, the total genetic load is approximately

$$L = 1 - \exp[-(U_{rec} + 2U_{dom})]$$

Where the capital U represents the total genomic mutation rate to recessive and dominant disorders. In other words, if the total number of deleterious mutations per haploid genome per generation is much more than one, the mutation load should be more than a population with a low reproductive rate could sustain. In humans it appears that the total genomic deleterious mutation rate is almost certainly greater than 1 (Eyre-Walker and Keightley, 1999; Nature 397:344). If the theory is correct, the population cannot maintain mutation-selection balance and will accumulate deleterious mutations. Does this mean that the population is inexorably doomed to extinction? The answer is probably no, perhaps most importantly because loads are unlikely to be additive – if many mutations can be eliminated in a few individuals (fitness is hardly affected if you have a few deleterious mutations, but is very reduced if you have lots), genetic load is much reduced. See Crow (PNAS 94:8380; 1997) for evolutionary arguments concerning the impact of deleterious mutations on human health.

Selection that promotes genetic variation

Selection against deleterious mutations will act to eliminate them from the population – selection removes genetic variability. However, there is another type of selection that results in the long-term maintenance of genetic variation. Polymorphisms that are maintained in populations are the result of balancing selection, which may be caused by either heterozygote advantage (having two different alleles is better than having two of the same kind) or frequency dependent selection (rare alleles are favoured simply because they are rare).

The classic example of heterozygote advantage is that of the beta haemoglobin S variant that causes sickle-cell anaemia when homozygous (usually fatal), but protects against one of the agents of malaria, *Plasmodium falciparum* when heterozygous. In parts of Africa where malaria is common, the S allele can reach frequencies of up to 30%. Estimates of the relative viability of the different alleles are

Genotype	AA	AS	SS
Frequency	$(1 - x)^2$	$2x(1 - x)$	x^2
Relative fitness	$1 - t$	1	$1 - s$
Estimated fitness	0.85	1	0

Using the same approach as before

$$\bar{w} = 1 - t(1 - x)^2 - sx^2$$

$$E_s[\Delta x] = 2t - 2x(t + s)$$

We will ignore new mutations to or from the S allele. Because the heterozygote is the most fit genotype, we expect the allele frequency to reach an equilibrium at an intermediate frequency. To find this frequency, we find the value of x for which the expected allele frequency change is zero

$$2t - 2x(t + s) = 0$$

$$\tilde{x} = \frac{t}{t + s}$$

Putting in the numerical results for sickle-cell anaemia, t is 0.15 and s is 1. So the expected equilibrium frequency is 0.13. This is a reasonably good match to the average frequency of the allele across countries with malaria. In any real population, we would expect the frequency to fluctuate, but to average out at this equilibrium.

The *AS* polymorphism was one of the first genetic variants to be described, and one of the first to be understood in evolutionary terms. As a consequence, for a long time it was thought that balancing selection would be the answer to the presence of many other genetic polymorphisms. However, since then, there have only been a handful of cases where there is incontrovertible evidence for heterozygote advantage, the notable example being the *HLA* loci in humans. In the case of *HLA*, there is a good understanding of why having two different alleles may be advantageous. However, for the majority of other genes, sufficiently little is understood of their function and interaction to predict the effects of different alleles. Furthermore, even if having different alleles is beneficial, the selective effects may be too slight to pick up in an experiment, or fluctuate over space and time. For the majority of loci, we need indirect methods for detecting balancing selection.

The effect of balancing selection on gene genealogies

Genetic polymorphisms that are maintained by balancing selection may have been in the population for many thousands of years. More importantly, they may have been maintained over a much longer time-scale than the rate of coalescence in a standard Fisher-Wright population. Consequently, for loci closely linked to the sites under balancing selection, the population has effectively been subdivided, such that coalescence can occur within a subpopulation, but not between. For the sites actually under balancing selection, you have to go back as far as the origin of the polymorphism to find the MRCA, however for sites that are not completely linked, recombination between genes introduces some migration between the subpopulations.

In the limit of strong selection, so that allele frequencies at the balanced locus have not fluctuated over time, the history of a sample of alleles taken from the population is entirely equivalent to the history of a sample taken from two demes, with some migration between them. It is therefore a simple manner to construct coalescent simulations in order to investigate the effect of balancing selection on linked neutral genealogies, and consequently diversity. The key points though are very intuitive. A sample of sequences that all possess the same balanced allele will tend to coalesce rapidly with each other, but will coalesce at a very slow rate with sequences that have the different balanced allele. As a consequence we expect genealogies to be characterised by having very long internal branches, high levels of polymorphism and positive Tajima D values. As you move away from the balanced locus, migration between the types due to recombination increases, and the genealogies are less distorted.

Testing for balanced selection

A natural way to test for balancing selection would therefore be to look for evidence of high diversity and positive Tajima D values. However, there is a potential problem. I have just said how the effects

of balancing selection are equivalent to a two-deme model of population structure. So significant departures from Fisher-Wright expectations at a single locus may be nothing to do with deviation from neutrality, and simply the result of deviation from the randomly-mating model.

The deeper problem is that we want to test not the Fisher-Wright, standard neutral model, but the neutral theory of molecular evolution. The neutral theory, as formulated and promoted by Kimura (1983), states that the vast majority of mutations segregating within species, or different between species, are of no selective importance. It says nothing about the type of population structure, or what mutation model is expected, just about the nature of mutations in samples.

Being more general, it is more difficult to test the neutral theory than the standard neutral model. But the key is that deviations from the standard neutral model can be tested on a per locus basis, whereas deviations from the neutral theory can only be tested by comparing different loci, or different classes of mutation (e.g. synonymous versus nonsynonymous). In the case of balancing selection, population structure is expected to affect all loci, but balancing selection should only affect a minority. By comparing patterns of genetic variability at multiple loci, and particularly by comparing patterns of divergence to patterns of polymorphism at multiple loci, we can devise powerful tests to identify the action of natural selection.

The Hudson, Kreitman and Aguadé test

The first attempt to use population genetic data to test the neutral theory was that of Hudson, Kreitman and Aguadé (1987). The HKA test compares patterns of polymorphism and divergence at two or more loci. The simplest form of the test uses polymorphism data from just one species.

Suppose we have samples of sequences from two unlinked regions, and outgroup sequences for both sets. Under the Fisher-Wright model, the expected number of segregating sites at locus A is

$$E[S_A] = \theta_A a_n$$

Where

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Allowing for different mutation rates at different loci, the expected number of segregating sites at locus B is

$$E[S_B] = \theta_B a_n$$

Under the neutral theory, the rate of substitution is equal to the rate of neutral mutation, so the expected divergence between species is

$$E[D_i] = T\theta_i$$

Where time is expressed in terms of $2N_e$ generations. We have four observations, and three parameters to estimate (θ_A , θ_B and T), which can be done by least squares. The HKA test then calculates the χ^2 goodness-of-fit statistic, and uses simulation to estimate the probability of observing a value as extreme as that seen in the data, under the assumptions of the model. The original demonstration of the test was on a collection of 81 samples of the *Adh* and 5' flanking region from *D. melanogaster*, using *D. sechellia* as an outgroup. For the 5' region, 210 differences and 9 polymorphisms were found. For the coding region, there were 18 differences and 8 polymorphic sites. The HKA test rejects the neutral hypothesis with $P = 0.016$. By comparing other regions, the authors were able to show that the

deviation is due to an excess of polymorphism at the *Adh* locus. This is compatible with the action of balancing selection in the gene. Biochemical analyses of the *Fast/Slow* polymorphism have shown that there is a two-fold difference in enzyme activity between alleles. Furthermore, geographical studies have found that there is a North-South decline in frequency of the *Fast* allele across America, and that the allele is also more common at higher altitudes. These data suggest that selection is acting on the gene in response to the diet of flies. Remarkably, there is some evidence that *Adh* also experiences balancing selection in the plants *Arabidopsis* and maize.

Trans-specific polymorphism

A form of balancing selection that is particularly striking is shown by two very different systems; self-incompatibility in plants and the MHC genes in mammals. Some plants have a genetic mechanism for avoiding inbreeding that is called self-incompatibility (SI). Gametes express recognition genes, such that if the pollen and ovule both express the same allele, fertilisation is prevented. As a consequence, individuals are always heterozygous for SI alleles, and there are many different alleles maintained in a population (at least 30 in *Brassica*). These loci are also maintained in populations for long periods of time, because they have an advantage when rare (they never meet with themselves). Consequently, some alleles have been maintained over speciation events, so that alleles sampled from one species are more similar to those from another species, than to any other alleles in the same population. A similar picture is found in HLA genes within the MHC cluster. Some alleles have been maintained for millions of years. There is a claim that some even predate the divergence of prosimian and anthropoid primates some 65 MYA (Klein 1986; *Natural History of the Major Histocompatibility Complex*, John Wiley). HLA loci are likely to benefit from both heterozygote advantage and rare-advantage.

Adaptation and directional selection

Neither the removal of deleterious mutations, nor balancing selection contribute to the long-term evolution of species. The most fascinating mutations are undoubtedly those that make species different from each other; mutations that have become fixed in a species because they confer a selective advantage. There are many examples of genetic microadaptations – the different oxygen binding properties of the different haemoglobins; the parallel adaptation of the digestive enzyme lysozyme in the guts of ruminants and leaf-eating monkeys; drug resistance in HIV viruses. However, two factors restrict a more general understanding of the genetic basis of adaptation. First, the genetic basis of the majority of traits that differ between species is poorly understood. Second, so little is known about the relationship between gene expression, protein structure and organismal fitness, that predicting the selective effects of observed genetic differences between species is an impossible task.

Population genetics, and more generally, molecular evolution, provides indirect methods for identifying the action of directional selection from patterns of DNA sequence polymorphism and divergence. Directional, or Darwinian selection leaves characteristic footprints in patterns of genetic variation that can be tested for without any functional understanding of the genes being analysed

We can describe the way directional selection influences allele frequency for beneficial mutations. Writing genotypes as

Genotype	AA	Aa	aa
Frequency	$(1 - x)^2$	$2x(1 - x)$	x^2
Relative fitness	1	$1 + hs$	$1 + 2s$

Where h is the dominance factor; when $h = 0.5$, heterozygotes are intermediate in fitness between the two homozygous genotypes – termed genic selection. The question of whether advantageous mutations are dominant, genic or recessive has received considerable attention, but little concrete evidence either way. Given that mutations of small deleterious effect tend to be genic (an empirical observation), we will conveniently assume that this is true for mutations of small beneficial effect. The average population fitness is

$$\bar{w} = 1 + 2x(1 - x)s + 2sx^2$$

And the expected change in allele frequency due to selection is approximately

$$E_s[\Delta x] = sx(1 - x)$$

Beneficial mutations do not reach an equilibrium frequency in populations, but are either lost or fixed. In very large populations, above a critical frequency, the rate at which an allele spreads to fixation is well approximated by the deterministic equation. However, when alleles are rare, genetic drift and chance still play a crucial role in the fixation process.

The fixation probability of an advantageous mutation in a finite population

Consider an advantageous mutation that arises as the result of a single mutation. Although this allele is expected to increase in frequency due to selection, just by chance, it is possible that the allele does not reproduce and it is lost from the population. Haldane (1927) was the first to attempt an analytical treatment of how likely a beneficial mutation is to persist in a population and spread to fixation. We will follow the theory for haploids, but a very similar argument holds for diploids. He assumed that the number of offspring an individual leaves behind is Poisson distributed with an expected value of 1 for the wild type genotype (A) and $1+s$ for the new genotype (a). The Poisson distribution effectively assumes that all individuals reproduce independently – this is not the Fisher-Wright model, but it is a good approximation over the time period we will consider. The probability of producing k offspring for individuals carrying the beneficial allele is

$$\Pr\{k\} = \frac{e^{-(1+s)}(1+s)^k}{k!}$$

What happens after two generations? If the beneficial allele left behind two copies in the first generation, these might each produce two copies in the next generation, or they might produce none. The distribution of the number of offspring after two generations can be obtained by summing over all possible events. This is called a chain distribution – as each element depends on preceding elements. There is no neat solution for the complete distribution, but Haldane noticed that there was a way of describing the probability that there were no copies of the allele left.

Let us call the probability of ultimate extinction of the allele E . In the first generation after it appears in the population it may leave 0 offspring (in which extinction has been achieved), or it may produce 1 or more. If the offspring of this allele reproduce independently, each offspring has probability E of ultimately going extinct. So we can write the probability of ultimate extinction as

$$E = P_0 + P_1E + P_2E^2 + P_3E^3 + \dots + P_iE^i$$

Where the P_i s are the probabilities of leaving i offspring in the first generation. Using the Poisson distribution

$$E = e^{-(1+s)} + (1+s)e^{-(1+s)}E + \frac{(1+s)^2E^2}{2!}e^{-(1+s)} + \dots + \frac{(1+s)^iE^i}{i!}e^{-(1+s)}$$

$$= e^{-(1+s)} \left\{ 1 + (1+s)E + \frac{(1+s)^2E^2}{2!} + \dots + \frac{(1+s)^iE^i}{i!} \right\}$$

For those familiar with Taylor series expansions, the sum in the brackets will be recognisable as a way of expressing the exponential formula. This allows us to simplify

$$E = e^{-(1+s)(1-E)}$$

What does this tell us about fixation? If a mutation becomes fixed in a population it will not become extinct. If a mutation does not get fixed, it will ultimately go extinct. So the probability of fixation, U , is just $1-E$.

$$1-U = e^{-(1+s)U}$$

Now let us expand the exponential as a Taylor series again

$$1-U = 1 - (1+s)U + \frac{(1+s)^2U^2}{2!} + \dots + (-1)^i \frac{(1+s)^iU^i}{i!}$$

If the selection coefficient is small, and the probability of fixation is much less than 1, we can focus on the first two terms in the expansion.

$$1-U \approx 1-U + sU + \frac{U^2}{2}(1+2s+s^2)$$

$$U \approx \frac{2s}{1+2s+s^2}$$

If $s \ll 1$, we are left with the approximation

$$U \approx 2s$$

The key point about Haldane's result is that even in extremely large populations, drift plays a crucial role in the fixation of beneficial mutations. This is because rare alleles, no matter how beneficial, may be lost from a population simply by chance. A mutation that confers a 1% advantage when heterozygous is about 50 times more likely to be lost, than to spread to fixation.

Diffusion theory treatment of the fixation process

Haldane's argument, although elegant and powerful, should not work. Critically, the assumption that individuals reproduce independently from each other is blatantly not true. It works because the critical point for selected alleles is when they are very rare, at which point they behave more or less independently of each other. Above a critical number of copies (about $1/2s$), the probability of loss is negligible. In addition, Haldane's argument doesn't allow us to work out anything else about the fixation process – how long should it take? What time do alleles spend in different parts of the allele frequency spectrum en route to fixation?

A much more powerful description of the fixation process was developed by Kimura and Ohta in the 1960s using diffusion theory. In the second lecture, I introduced the forward Kolmogorov

equation to describe the allele frequency distribution in the Fisher-Wright model. Introducing selection into the forward equation presents no technical difficulty. The distribution of allele frequencies under reversible mutation and selection can be worked out along the following lines.

The mean change in allele frequency each generation is sum of the effects of selection and mutation. Consider the case of genic selection, where the fitness of heterozygotes is $1+s$ and the fitness of homozygotes is $1+2s$. Mutation occurs to the beneficial allele at a rate u and beneficial alleles mutate back to the wild-type at a rate v . The mean change in allele frequency per generation is

$$M_{\delta x} = sx(1-x) + (1-x)u - xv$$

The only factor leading to variance in allele frequency change is assumed to be that due to binomial sampling in a finite population. If the deterministic effects on allele frequency are small, the variance is given by

$$V_{\delta x} = \frac{x(1-x)}{2N_e}$$

Using Kimura's adaptation of Wright's formula for the stationary distribution of allele frequencies under diffusion theory

$$\phi(x) = \frac{C}{V_{\delta x}} \exp\left[2\int \frac{M_{\delta x}}{V_{\delta x}} dx\right]$$

We find

$$\phi(x) = Ce^{4N_e sx} x^{4N_e u - 1} (1-x)^{4N_e v - 1}$$

Where C is a normalising constant. What should we make of this formula? The most important feature is that the formula only includes terms multiplied by the effective population size. So the effects of selection, like mutation, depend on the strength of selection relative to genetic drift, rather than the absolute value.

The backward Kolmogorov equation

While the forward process tells us about allele frequency distributions, it is cumbersome to use to describe the entire fixation process. Kimura introduced a method for describing the fixation probability and features of the fixation process based on the backwards Kolmogorov equation.

Let us call the probability of fixation, given some initial starting frequency x , $u(x)$. As for the forward equation result, consider what happens in the next generation. We can write the probability of fixation as the integral over all possible changes in allele frequency in the first generation, times the probability of fixation given that change.

$$u(x) = \int_{\delta x} u(x + \delta x) g(\delta x, x, \delta t)$$

where the function g means the probability of the change in allele frequency δx occurring in the time δt , when the allele is at frequency x . As before, we can make a Taylor series approximation to $u(x + \delta x)$, keeping only the first two terms (effectively assuming weak selection and large population size).

$$u(x + \delta x) \approx u(x) + \delta x \frac{du(x)}{dx} + \frac{\delta x^2}{2} \frac{d^2 u(x)}{dx^2}$$

Substituting this in the previous integral

$$u(x) \approx \int_{\delta x} u(x)g(\delta x, x, \delta t) + \int \frac{du(x)}{\delta x}(\delta x)g(\delta x, x, \delta t) + \int \frac{1}{2} \frac{d^2u(x)}{dx^2}(\delta x)^2 g(\delta x, x, \delta t)$$

The first term on the RHS is just $u(x)$. In addition, terms involving $u(x)$ do not depend on δx and can be taken out of the integral. We can write

$$M_{\delta x} = \int_{\delta x} (\delta x)g(\delta x, x, \delta t)$$

As the mean change in allele frequency per generation, and

$$V_{\delta x} = \int_{\delta x} (\delta x)^2 g(\delta x, x, \delta t)$$

As the variance in allele frequency change (assuming the mean change is very small). Cancelling the terms in $u(x)$, we get

$$0 = M_{\delta x} \frac{du(x)}{dx} + \frac{1}{2} V_{\delta x} \frac{d^2u(x)}{dx^2}$$

This is known as the backwards Kolmogorov equation. Solving the equation is moderately straightforward. First make the substitution

$$f(x) = \frac{du(x)}{dx}$$

We can write the differential equation as

$$sx(1-x)f(x) = \frac{-x(1-x)}{4N_e} f'(x)$$

Multiply both sides of the equation by $4N_e$, and write $4N_e s$ as S . Divide both sides by $x(1-x)$ to get

$$Sf(x) = -f'(x)$$

This has an exponential solution

$$f(x) = Ae^{-Sx} + B$$

Where A and B are constants. However, we will also make the switch back to working with $u(x)$, so that

$$u(x) = Ce^{-Sx} + D$$

Where C and D are constants. To obtain the constants we need to inject some cases where we know the solution. In particular, we know $u(0) = 0$ and $u(1) = 1$. So

$$C + D = 0$$

$$Ce^{-S} + D = 1$$

From which we find

$$C = \frac{1}{e^{-S} - 1}$$

Putting things together

$$u(x) = \frac{e^{-Sx} - 1}{e^{-S} - 1}$$

When $x = 1/2N$, and s is small, we can use the approximation

$$u(1/2N) \approx \frac{2sN_e / N}{1 - e^{-4N_e s}}$$

There are three key points about Kimura's result. First, it makes the link between neutral and selected mutations. For completely neutral mutations, the fixation probability is $1/2N$ in a diploid population. With genic selection, the fixation probability is about $2sN_e/N$ for strongly selected mutations. Second, weakly deleterious mutations have a non-negligible fixation probability. Third, the critical factor is not the absolute selection coefficient, but the product of the effective population size and the selection coefficient. For example, in a population of one million individuals, a mutation that has a 0.1% effect on fitness has about a 0.2% probability of fixation if beneficial, and an almost zero fixation probability if deleterious. The same mutation in a population of 100 individuals has a fixation probability of about 0.5%, irrespective of whether it is beneficial or deleterious, because drift is the dominant force in determining its fate. The point where selection and drift swap as being the dominant force is about $N_e s = 1$.

Detecting directional selection from population genetic data

There are several ways to detect the action of directional selection from DNA sequence data. The most widely used to restrict analyses to interspecific comparisons. Repeated bursts of adaptive evolution in a gene will lead to a large number of differences in genes between related species. Purifying selection, in contrast, will prevent mutations becoming fixed between species. By comparing the rate of amino-acid substitutions to the rate of supposedly neutral mutations (typically synonymous mutations at third positions in codons), genes, or regions of genes with unusually high rate of divergence can be identified. There are three problems with such analyses. First, in many species there is evidence that synonymous mutations are not neutral, but experience weak selective constraint due to codon-tRNA interactions during translation. This is not too restricting, as the selective constraints on synonymous mutations are likely to be weaker than those acting on amino acid changes. Second, exclusive focus on nonsynonymous – synonymous differences within coding regions ignores the possibility that adaptive can proceed through changes in regulatory regions. Thirdly, and perhaps most important, a high rate of nonsynonymous change can either be caused by adaptive evolution, or by low levels of selective constraint.

The traditional solution to this problem has been to try to identify genes, or regions of genes, in which the rate of amino acid change is greater than the rate of neutral (synonymous) evolution. Repeated bouts of adaptation may lead to the accumulation of functional differences at a rate faster than that expected from neutral mutations. The antigen recognition site of HLA proteins is the model for such an approach. Hughes and Nei (1988) found ratios of nonsynonymous to synonymous substitution far in excess of one in both humans and mice. Low selective constraint should never show this pattern.

The great limitation of this solution is that for the vast majority of genes, our model of adaptation is not one of recurrent substitution. More reasonable models would expect changes at one or a few sites in a gene to be important – it is only under antagonistic coevolution, or disruptive selection (extremes and novel variants are successful) that sustained rapid divergence is expected. More sensitive phylogenetic methods that use multiple species data are likely to be better at detecting sites showing unusual patterns of evolution. But without biological understanding of the functional

consequences of mutations, it will always be difficult to discriminate between adaptive evolution and low constraint.

An alternative to phylogenetic methods is to use population genetic data. The simplest possible test for the action of different types of selection was devised by McDonald and Kreitman (1991); known as the MK test. As with the HKA test, the starting point is the idea that under neutrality, the rate of sequence divergence and levels of polymorphism are simple functions of the neutral mutation rate. However, rather than compare different loci, the MK test compares different types of mutation at the same locus; typically synonymous versus nonsynonymous mutations, although other types (e.g. indels, different classes of nonsynonymous mutation, etc) can also be used. The use of interspersed mutations makes the test more robust than the HKA test, because the two types of mutation will have interspersed histories. Consequently, the test is insensitive to deviations from the standard neutral model. All that is being compared is one type of mutation against another, typically by using standard contingency table analysis.

The effects of different types of selection can be caricatured by the extremes of purifying selection against deleterious mutations and recurrent adaptive evolution. Deleterious mutations can persist in populations for several generations due to drift, but are very unlikely to become fixed. Consequently, they can contribute to polymorphism, but not divergence. In contrast, adaptive mutations become fixed in populations sufficiently fast that they contribute little to polymorphism, but appear as fixed differences between species. By comparing the number of fixed to polymorphic differences for synonymous and nonsynonymous mutations deviations from the neutral theory can be detected.

The difference between nuclear and mitochondrial loci

The MK test has had much success in detecting deviations from the neutral theory in a wide range of organisms. Variants have also been used to detect the action of selection on synonymous codon usage (Akashi 1995; *Genetics* 139:1067) and insertion-deletion polymorphisms (Akashi 1994; *Genetics* 144:1297). One striking result is the consistent difference in test results between nuclear loci and mitochondrial loci. Typically, mtDNA genes show an excess of replacement polymorphism compared to divergence, indicating purifying selection. In contrast, nuclear loci often show an excess of replacement divergence, indicating directional selection. The reason for the discrepancy is not clear, but perhaps the most likely explanation is that the mtDNA genes have a highly conserved function, such that almost no new mutations are beneficial. There are, however, other important differences between mtDNA and nuclear loci – most notably the lack of recombination in mtDNA.

The hitch-hiking effect of beneficial mutations

The MK test assumes that beneficial mutations rapidly sweep to fixation so are not observed as polymorphisms, while neutral mutations are. However, if the beneficial mutation has only recently become fixed in the population, the independence between the selected and neutral mutations may be violated. Specifically, the selective sweep of a beneficial mutation through a population will affect linked neutral loci (Maynard Smith and Haigh 1974).

Perhaps the easiest way to think about selective sweeps is from a genealogical perspective. Suppose that a few generations previous to our sample a beneficial mutation had become fixed in the

population, so that looking further back in time it will have rapidly increased from a low frequency. This distorts the genealogy of linked sites, because the ancestor of all sites in the sample must have come from a chromosome that carried the beneficial mutation. Looking back in time, there are fewer and fewer of such chromosomes, so the rate of coalescence must increase. In fact, the situation is exactly like that of population growth – we can think of the coalescent process occurring in population that at some point in the past was no more than a single individual. The recent history of the sample is little affected, because the beneficial allele represented the majority of the population. But further back in time coalescence has to occur rapidly. This leads to a great shortening of the genealogy of linked sites, and those mutations that are segregating are expected to be external branch mutations represented as singletons in the sample. In short, the signature of recent adaptive evolution in the genome is regions in which there is very low diversity.

The relationship between recombination and diversity

The picture of hitchhiking described is not quite complete. In regions of low recombination, sites physically distant to the locus at which selection is occurring may be affected. However, in regions of high recombination, associations between the selected allele and linked neutral mutations are rapidly broken down, and hitchhiking will be restricted to very small regions. The finding of a correlation between the local recombination rate and levels of polymorphism (Begun and Aquadro 1992) in *Drosophila* was therefore hailed as evidence that adaptive evolution is a common phenomenon in the fruit-fly genome. Similar findings have since been reported in other organisms including humans and tomato.

There is, however, an alternative explanation for the observation. Charlesworth et al (1993) pointed out that the elimination of deleterious mutations could also lead to the reduction of linked neutral polymorphism – a process known as background selection. Consider a population in which deleterious mutations are maintained at mutation-selection balance. Novel, neutral mutations occurring on chromosomes that carry deleterious mutations are destined to be lost from the population unless they can recombine onto a mutation free chromosome. Therefore in regions of low recombination the fraction of new mutations that can contribute to polymorphism is reduced. It is unlikely that the point-mutation deleterious mutation rate is high enough to create the observed relationship, but if selection against transposable elements is considerable, the data fit expectations well.

How can the hypotheses be distinguished? The obvious answer is to compare the effect of hitchhiking and background selection on the frequency distribution of linked neutral variants. The problem is that when polymorphism is low, the sampling variance in statistics relating to the frequency distribution is high, and there is little power to distinguish models, except with a lot of data. Now that sequencing is relatively fast and cheap, we can hope to get an idea of the extent of adaptive evolution

The nearly neutral theory of molecular evolution

The idea that deleterious mutations may be an important force in determining rates of molecular evolution and patterns of genetic variation has been extensively developed by Tomoko Ohta (1995). She advocates a theory called the nearly neutral theory of molecular evolution, the central feature of which is that the majority of substitutions in proteins are not neutral, but very slightly deleterious.

The theory was proposed in response to a criticism of the neutral theory. One of the key observations that originally prompted Kimura was the relatively constant rate of amino acid substitution in very different species; the molecular clock. However, if mutation occurs on a per generation rate, rather than a per year rate (as is likely if mutations occur during DNA replication in dividing germ-cells), organisms with a shorter generation time, should have higher per year rates of evolution. Indeed, if you compare the rate of synonymous mutation to that of nonsynonymous mutation in mammals, there is a considerable generation time effect on synonymous mutations, but not for nonsynonymous ones (Ohta, 1995). Such a pattern cannot be explained by a completely neutral model for amino acid evolution. Ohta's solution was to suggest that mutations in proteins are either strongly deleterious (and never get fixed) or weakly deleterious, in which case the fixation probability depends on the effective population size. Species with larger populations have more efficient selection against weakly deleterious mutations, and will have lower rates of substitution. In order to recover the molecular clock, it is necessary that species with shorter generation times have larger population sizes. To an extent the assumptions of Ohta's model must be true. No mutation can be truly neutral, and elephants have longer generation times and smaller population sizes than mice. But whether the parameters are so finely balanced to create a molecular clock is a debated question.

Evidence from population genetic surveys in several species suggests that amino acid changes segregating in populations tend to weakly deleterious (Cargill et al 1999, *Nature Genetics* 22:231; Akashi 1999, *Genetics* 151:229). However, at least in *Drosophila*, there is evidence for more fixed differences than a weakly deleterious model would predict (Akashi 1999). Estimating the contributions of adaptive evolution and deleterious mutation to patterns of molecular evolution and diversity is the key challenge now facing population genetics.