Some Mathematical Models from Population Genetics

Alison Etheridge

March 13, 2009

Contents

Historical Background

The main purpose of theoretical population genetics is to understand the complex patterns of genetic variation that we observe in the world around us. Its origins can be traced to the pioneering work of Fisher, Haldane and Wright. Their contributions were fundamental in establishing the *Modern Evolutionary Synthesis*, in which Darwin's theory of evolution by natural selection was finally reconciled with Mendelian genetics. Darwin's theory of evolution (Darwin 1859) can be simply stated: 'Heritable traits that increase reproductive success will become more common in a population'. Thus, in order for natural selection to act, there must be variation within a population and offspring must be similar to their parents. So to fully understand evolution we need a mechanism whereby variation is created and inherited. This is provided by Mendelian genetics (Mendel 1859). Again the idea can be simply stated. Traits are determined by *genes*. Each gene occurs in finitely many different types that we call *alleles* and different alleles may produce different traits. Offspring are similar to their parents because they inherit genes from their parents. The difficulty is that Darwin had argued that evolution of complex, well-adapted organisms depends on selection acting on a large number of slight variants in a trait and much of Mendel's work deliberately focussed on discontinuous changes in traits determined by a single gene.

The resolution lay in the foundations of theoretical population genetics. In 1918, Fisher showed how correlations between relatives that had been measured by biometricians could be explained by multiple Mendelian factors together with random, non-genetic, influences. In the process he developed the statistical theory of *analysis of variance*. He went on to show that Mendelian genetics was consistent with the idea of evolution by natural selection. Thus, if traits depend on multiple genes, each making a small contribution, the apparently discontinuous nature of Mendelian inheritance is reconciled with continuous variation and gradual evolution. Starting in 1924, Haldane published a series of papers that provide a detailed theoretical analysis of how differences in survival or reproduction due to one or two Mendelian genes would affect a population. He used examples like the evolution of the peppered moth to show that natural selection could act extremely fast. [1] In a series of papers starting in 1921, Wright quantified the way in which the random process of reproduction in a finite population would lead to changes in allele frequency and examined how this *random genetic drift* interacted with selection, mutation and migration. He introduced the notion of an adaptive landscape in which natural selection would drive a population towards a local maximum but genetic drift could push the population away from such a peak paving the way for natural selection to push it towards a different peak. Through this mechanism the population explores the evolutionary landscape. Mathematical modelling played a crucial rôle in the work of all three men.

Of course many details remained (and indeed remain) unclear. For example, what is the relative importance of mutation, selection, random drift and population subdivision for standing genetic

[1] The peppered moth was originally predominantly light-coloured, providing camouflage on the lichen coloured trees on which it rests. As a result of pollution the lichens died out and the trees became blackened by soot, making the light coloured moths vulnerable to predation, and dark-coloured moths flourished. With improving environmental conditions, light coloured moths have once again become common.

variation? Fisher emphasized gradual changes in a single large population due to selection acting on small variants, Haldane placed more importance on strong selection acting on single genes and Wright argued that adaptation would be most effective in a population that was subdivided into many small subpopulations in his *shifting balance* theory.

At the time of the evolutionary synthesis, genetic variability could not be observed directly. Early work was restricted to genes that happened to be detectable in an observable way, but over the subsequent fifty years things changed dramatically. In 1953, Watson and Crick showed that DNA forms a double helix in which one sequence of bases pairs with the complementary sequence. [2] Over the next decade it was established how DNA codes for proteins through the genetic code. Beginning in the mid-1960s scientists began to study evolution at the scale of DNA, RNA and proteins. Studies revealed an unexpectedly high level of variation within species. It was also possible to compare the evolution of the same protein across different species. By comparing species that diverged from one another at a known time, it was found that any given protein had evolved at a steady rate, even if it was evolving in very different organisms. In other words, there is a *molecular clock*. In 1968, Kimura famously (and controversially) argued that there is too much genetic variation in species for more than a small fraction to be subject to natural selection. He also argued that the molecular clock was best explained by the steady accumulation of mutations that have no effect on fitness. Others, by contrast, were emphasizing the rôle of selection in explaining even very small changes in phenotype. The theory remains controversial. It is simply not known what proportion of differences between individuals or species are maintained by selection, but it provides a valuable 'null' model against which data can be tested.

If we are to assess the relative importance of mutation, selection, drift, spatial structure and so on, then the first step is to distill our understanding of how these processes operate into a workable mathematical model whose predictions can be compared to data. Even now we don't know exactly how genes combine to influence a whole organism or what maintains the variation in those genes. But with advances in molecular biology, a wealth of data is available and mathematicians have a key rôle to play. Over the last three decades, in parallel with advances in DNA sequencing technology, new mathematical models have been introduced focussing on the genealogy (that is the ancestral history) of a random sample of genes from a population. These 'coalescent' models have a rich and beautiful mathematical structure and in addition to providing the necessary tools for the interpretation of genetic data they have become a popular playground for mathematicians.

In these lectures we will introduce and study some models (both old and new) that have their origins in theoretical population genetics. We will try to minimise the use of biological jargon, but we end this section with a note on terminology. The 'atom' of genetics is the single base pair or *nucleotide*. It is often refered to as a *site*. The term *locus* is used to refer in a general way to a location in the genome. It may refer to anything from a few hundred bases to a long stretch of DNA containing several genes. Whereas in classical Mendelian genetics a gene was a single well-defined unit, it is now loosely

[2] DNA stands for deoxyribonucleic acid. The four bases are A (adenine), T (thymine), C (cytosine) and G (guanine) and A bonds to T and C to G.

defined as a stretch of DNA that includes sequences that code for a protein (or a functional RNA molecule) and regulatory sequences. The genes are organised on chromosomes and, for mathematical convenience, we shall consider chromosomes to be linear. An excellent introduction to the underlying biology is Barton et al. (2007).

1 Mutation and Random genetic drift

1.1 The Wright-Fisher model and the Kingman coalescent

Evolution is a random process. Random events enter in many ways, from errors in copying genetic material to small and large scale environmental changes, but the most basic source of randomness that we must understand is due to reproduction in a finite population leading to *random genetic drift*.

The simplest model of random genetic drift was developed independently by Sewall Wright and R.A. Fisher and is known as the Wright-Fisher model. We consider a population in which every individual is equally likely to mate with every other and in which all individuals experience the same conditions. Such a population is called *panmictic*. We also suppose that the population is *neutral* (everyone has an equal chance of reproductive success). Most species are either *haploid* meaning that they have a single copy of each chromosome (for example, most bacteria), or *diploid* meaning that they have two copies of each chromosome (for example, humans). We suppose that the population is haploid, so that each individual has exactly one parent. Although in a diploid population individuals have two parents, each *gene* can be traced to a single parental gene in the previous generation and so it is customary in this setting to model the genes in a diploid population of size N as a haploid population of size $2N$. As we shall see later, this device fails once we are interested in tracing several genes at the same time.

Definition 1.1 (The neutral Wright-Fisher model) *The neutral Wright-Fisher model for a panmictic, haploid population of constant size N is described as follows. The population of N genes evolves in discrete generations. Generation $(t+1)$ is formed from generation t by choosing N genes at random with replacement. i.e. each gene in generation $(t+1)$ chooses its parent independently at random from those present in generation t.*

From this definition it is an elementary matter to work out the genealogical trees that relate individuals in a sample from the population. Suppose first that we take a sample of size two. The probability that these two individuals share a common parent in the previous generation is $\frac{1}{N}$. If they do not, then the probability that their parents had a common parent is $\frac{1}{N}$, and so on. In other words, the time to the most recent common ancestor (MRCA) of the two individuals in the sample has a geometric distribution with success probability $\frac{1}{N}$. (The probably that their most recent common ancestor was T generations in the past is pq^{T-1} where $p = \frac{1}{N}$ and $q = 1-p$.) In particular, the expected number of generations back to their MRCA is N. Now typically we are interested in large populations, where our rather crude models have some hope of having something meaningful to say. Then it makes sense to measure time in units of size N and in those units the time to the MRCA of a sample of size two is approximately exponentially distributed with parameter one. More generally, consider a sample of size $k \geq 2$. The probability of three (or more) individuals from the sample sharing a common parent is $\mathcal{O}\left(\frac{1}{N^2}\right)$ and similarly the chance that two separate pairs of individuals are 'siblings' is $\mathcal{O}\left(\frac{1}{N^2}\right)$. This means that the time we must wait before we see such an event is $\mathcal{O}(N^2)$ generations. But before this happens (with probability tending to one as $N \to \infty$) all our ancestral lineages will have coalesced through pairwise coalescence events (each of which occurs within $\mathcal{O}(N)$ generations). Thus the time (in units of size N) before the present at which we first see a 'merger' of lineages ancestral to our sample is approximately exponentially distributed with rate $\binom{k}{2}$ and, when that merger takes place, it affects exactly two lineages chosen uniformly at random from the $\binom{k}{2}$ pairs available. After that we just trace the remaining $\binom{k-1}{2}$ pairs of lineages and the same picture holds.

Remark 1.2 *Notice that since we are dealing with a haploid population, each individual has only one parent and the genealogical trees get* smaller *as we go backwards in time, in contrast to our usual understanding of family trees (for a diploid population) which grow as we trace backwards in time. We'll return to this point in §1.6.*

We shall loosely refer to the system of coalescing lineages that we have just described as Kingman's coalescent, but let us give a more formal definition. If we label individuals in our sample $\{1, 2, \ldots, k\}$, then our process of coalescing lineages defines a continuous time Markov process, $\{\pi_t\}_{t\geq 0}$, on the equivalence relations on $[k] = \{1, 2, \ldots, k\}$. Each equivalence class of π_t corresponds to an ancestor alive at time t before the present. It consists of the labels of all individuals in our sample descended from that ancestor.

Definition 1.3 (Kingman coalescent) *A k-coalescent is a continuous time Markov chain on $\mathcal{E}_k$, the space of equivalence relations on $[k]$, with transition rates $q_{\xi,\eta}$ $(\xi, \eta \in \mathcal{E}_k)$ given by*

$$q_{\xi,\eta} = \begin{cases} 1 & \text{if } \eta \text{ is obtained by coalescing two of the equivalence classes of } \xi, \\ 0 & \text{otherwise.} \end{cases}$$

The Kingman coalescent *on $\mathbb{N}$ is a process of equivalence relations on $\mathbb{N}$ with the property that, for each k, its restriction to $[k]$ is a k-coalescent. By convention, we take the initial condition to be the trivial partition into singletons.*

Remark 1.4 (Consistency) *If we take a $(k+l)$-coalescent and restrict it to $[k]$, then we obtain a k-coalescent. In particular, if we take a sample of size $k+l$ from the population and restrict the genealogical trees relating the full sample to a randomly chosen subsample of size k, then we arrive at the same trees as if we had just taken a smaller sample in the first place. This sampling consistency is an essential part of the interpretation of the model.*

Existence of the k-coalescent is clear (it is a finite state space Markov chain with bounded rates). The consistency allowed Kingman (1982) to pass to a projective limit.

Remark 1.5 (Terminology) *In what follows we will sometimes say that the geneaology of a sample (or population) of size k is determined by the Kingman coalescent. By this we mean that it is given by a k-coalescent.*

Now let's examine what happens when we let $N \to \infty$ in our Wright-Fisher model. Suppose that the gene in question has two alleles which we label a and A say. We try to characterise the process, $\{p_t\}_{t\geq 0}$, which records the proportion of a-alleles in the population at each time $t \geq 0$. Notice that in the prelimiting model, $\{p_t\}_{t\geq 0}$ is a discrete time Markov chain on a finite state space with traps at 0 and 1.

Definition 1.6 (Fixation) *If the proportion of one of the alleles in the population is one, then we say that the allele has* fixed. *The probability that a becomes fixed is its* fixation probability.

To characterise the distribution of $\{p_t\}_{t\geq 0}$, we consider how $\mathbb{E}[u(p_t)]$ changes with time for sufficiently nice functions $u : [0,1] \to \mathbb{R}$. In the rescaling that we took to obtain the Kingman coalescent, the model evolves at time intervals of length $1/N$. Evidently if a proportion p of the population is of type a in the current generation, then the expected *number* of a genes in the next generation is Np and the variance of that number is Npq (where $q = 1 - p$). Thus the mean allele frequency remains the same and the variance is pq/N. Moreover $\mathbb{E}[(p_{1/N} - p)^k \,|\, p_0 = p] = \mathcal{O}(1/N^2)$ for all $k \geq 3$. Now the evolution of the process is homogeneous in time, so it is enough to consider what happens close to time zero. Using Taylor's Theorem, we obtain

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}\left[u(p_t)\,|\,p_0 = p\right]\Big|_{t=0} &\approx N\left\{\mathbb{E}\left[u(p_{1/N})\,\big|\,p_0 = p\right] - u(p)\right\} \\
&= N\left\{u'(p)\mathbb{E}[(p_{1/N} - p)\,|\,p_0 = p] + \frac{1}{2}u''(p)\mathbb{E}[(p_{1/N} - p)^2\,|\,p_0 = p] + \mathcal{O}\left(\frac{1}{N^2}\right)\right\} \\
&= \frac{1}{2}p(1-p)u''(p) + \mathcal{O}\left(\frac{1}{N}\right).
\end{aligned}$$

Thus, in the limit as $N \to \infty$, if the process of allele frequencies converges to a well-defined stochastic process, then we expect that

$$\frac{d}{dt}\mathbb{E}\left[u(p_t)\,|\,p_0 = p\right]\Big|_{t=0} = \frac{1}{2}p(1-p)u''(p). \tag{1}$$

That is, we expect that in the limit, the distribution of the allele frequencies is governed by the solution to the Wright-Fisher stochastic differential equation:

$$dp_t = \sqrt{p_t(1-p_t)}dB_t, \tag{2}$$

where $\{B_t\}_{t\geq 0}$ is a standard Brownian motion.

What we have *shown* is that, at least for large populations, if we measure time in units of N generations, then the distribution of allele frequencies should be approximately governed by the partial differential equation (1) and the genealogy of a sample from the population should be well-approximated by the Kingman coalescent. Notice that it is the random genetic drift, that is the random change in allele frequencies caused by the random variation in individual reproduction, that causes coalescence of ancestral lineages as we trace backwards in time.

So how does this do as a model? Of course it is too simplistic to apply to most naturally occurring populations, but we can compare it to experimental data. Buri (1956) reports an experiment on populations of *Drosophila melanogaster*. About one hundred populations were propagated, each from eight males and eight females. The experiment measures the frequency of an allele of a gene that slightly alters the eye colour (without affecting fitness or reproductive success of the carrier). We'll denote it by a. He reports the change in the *variance* in allele frequency across the different populations with time. All populations are started with exactly half a and half A (which in this context just means 'not a') alleles. The variance starts at zero (all populations have the same frequencies) and then grows because of the random genetic drift until it reaches a maximum when each population consists either entirely of a-alleles or entirely of A alleles.

Writing v_t for the variance at time t in our rescaled time units, $v_t = \mathbb{E}[p_t^2] - \mathbb{E}[p_t]^2$ and using equation (1) we have that $\frac{d}{dt}\mathbb{E}[p_t] = 0$, $\frac{d}{dt}\mathbb{E}[p_t^2] = \mathbb{E}[p_t(1-p_t)]$ and $\frac{d}{dt}\mathbb{E}[p_t(1-p_t)] = -\mathbb{E}[p_t(1-p_t)]$. Combining these gives that $v_t \approx p_0(1-p_0)(1-\exp(-t))$. Writing V_t for the variance after t generations (in other words changing back to 'real' time units) this becomes

$$V_t \approx p_0(1-p_0)(1-\exp(-t/2N)).$$

The $2N$ is because Drosophila are diploid and in this case $N = 16$. The theoretical prediction for the rate of increase in the variance turns out to be not very accurate, but it becomes good when instead of substituting the actual population size, one substitutes a smaller *effective* population size (Buri reports a best fit of $N_e = 11.5$).

A variety of factors affect the rate of genetic drift and these are often summarised by using an effective population size. But why did we need to do that here? At first sight our populations appear to satisfy the assumptions of the Wright-Fisher model: they are panmictic and constant size, generation times are discrete and the allele under consideration does not affect fitness. In fact it is the Wright-Fisher reproduction mechanism itself that is at fault. It forces the variance of the offspring of a single individual to be one, but this does not reflect the true offspring distribution in the population. To see how offspring variance feeds into the effective population size we must consider a slightly more general model.

1.2 The Cannings model

First a definition.

Definition 1.7 (Exchangeable random vector) *A random vector $(\nu_1, \ldots, \nu_N)$ is said to be* exchangeable *if*

$$(\nu_1, \ldots, \nu_N) \stackrel{d}{=} (\nu_{\pi(1)}, \ldots, \nu_{\pi(N)})$$

for any permutation $\pi = (\pi(1), \ldots, \pi(N))$ *of* $\{1, \ldots, N\}$.

Definition 1.8 (Neutral Cannings Model) *Consider a panmictic, haploid population of constant size* N. *Labelling the individuals in generation* t *by* $\{1, \ldots, N\}$, *in a neutral* Cannings model, *generation* $t+1$ *is determined by an exchangeable random vector* $(\nu_1(t), \ldots, \nu_N(t))$ *with* $\sum_{k=1}^{N} \nu_k(t) = N$. *Here,* $\nu_k(t)$ *denotes the number of children of the kth individual and the vectors* $\{(\nu_1(t), \ldots, \nu_N(t))\}_{t \in \mathbb{N}}$ *are assumed to be independent and identically distributed.*

Notice that, mathematically, neutrality is captured by exchangeability.

The Wright-Fisher model is the special case of the Cannings model in which $(\nu_1(t), \ldots, \nu_N(t))$ has the multinomial distribution with N trials and equal weights.

Let's examine the genealogy of a sample from a large population evolving according to a more general Cannings model. Let c_N denote the probability that two individuals chosen at random from some generation have a common parent in the previous generation. Then (dropping the argument t)

$$c_N = \frac{\mathbb{E}[\nu_1(\nu_1 - 1)]}{N-1}.$$

To see this, condition on the vector $(\nu_1, \nu_2, \ldots, \nu_N)$ that determines the division of offspring into families. The chance that two offspring (sampled at random and *without* replacement) both fall among the ν_1 individuals that make up the first family is just $\nu_1(\nu_1 - 1)/N(N-1)$. Now average over the distribution of the vector $(\nu_1, \nu_2, \ldots, \nu_N)$. This gives the probability that both offspring are in the first family. Using exchangeability, the probability that they both belong to the same family (but any one of the N available) is just N times this probability, that is $\mathbb{E}[\nu_1(\nu_1 - 1)/(N-1)]$ as required. (For the Wright-Fisher model, $c_N = 1/N$.) The time until the MRCA of a random sample of size two from the population will be geometric with success probability c_N. This will determine the right time scaling to get convergence to a nontrivial limit as $N \to \infty$. We are going to assume that $c_N \to 0$ as $N \to \infty$. Now consider a sample of size three. The chance that they *all* have a common parent is

$$\frac{\mathbb{E}[\nu_1(\nu_1 - 1)(\nu_1 - 2)]}{(N-1)(N-2)}.$$

Thus, if we measure time in units of $1/c_N$, provided that

$$c_N \to 0 \quad \text{and} \quad \frac{\mathbb{E}[\nu_1(\nu_1 - 1)(\nu_1 - 2)]}{N^2 c_N} \to 0 \qquad \text{as } N \to \infty, \tag{3}$$

in the limit as $N \to \infty$ we will only ever see pairwise mergers and we recover Kingman's coalescent.

Lemma 1.9 *If we sample* k *individuals from a population evolving according to the neutral Cannings model of Definition 1.8 and if the conditions (3) are satisfied, then for large* N, *when measured in time units of* $1/c_N$ *generations, the genealogy of the sample is approximately a* k*-coalescent.*

Similar calculations to those we did for the Wright-Fisher model show that again measuring time in units of $1/c_N$ generations and under assumption (3), the distribution of allele frequencies for a sufficiently large population evolving according to the Cannings model will be governed (approximately) by the partial differential equation (1). The only difference from the Wright-Fisher setting is that now when we wish to compare to data we must remember that c_N is (approximately) $Var(\nu_1)/N$, where Var denotes variance. In our previous language, the *effective population size* is $N/Var(\nu_1)$. In particular, the greater the variance in offspring number, the smaller the effective population size and the faster the rate of random drift.

Remark 1.10 (Robustness of Kingman's coalescent) *In passing to an infinite population limit, we aim to find an approximation that reflects the key features of our population (in this case that it is neutral, panmictic and of constant size), but which is insensitive to the fine details of the prelimiting model. As we can already see, the Kingman coalescent approximates a wide variety of local structures and it is this robustness that makes it such a powerful tool. Forwards in time we have taken a* diffusion approximation*, approximating the Wright-Fisher model by a Wright-Fisher diffusion. The importance of diffusion approximations in population genetics can be traced to the seminal work of Feller (1951).*

1.3 Adding mutations

A mutation is formally defined as a 'heritable change in the genetic material (DNA or RNA) of an organism'. Mutations occur in many forms, but for simplicity we concentrate on *point mutations* which occur when there is a change from one base pair to another at a single position in the DNA sequence. Because of the redundancy in the genetic code some point mutations do not lead to a change in the sequence of amino acids. These are called *synonymous mutations.* Mutations are the ultimate source of all genetic variation; without them there would be no evolution. Although mutation rates are relatively slow, the mixing of mutations from different lineages that results from genetic recombination (see §2.5) rapidly leads to an enormous number of combinations on which natural selection can act. Mutation rates vary according to the type of mutation, the location on the genome and the organism involved, with the highest rates being in viruses.

Typically in our models we assume a constant probability μ per individual per generation of a mutation at a given base or more generally at a given locus. If we follow a particular ancestral lineage in our population, then we must wait a geometrically distributed number of generations (with mean $1/\mu$) until we see a mutation. Assuming that $2N_e\mu$, that is the mutation rate multiplied by the effective population size, is of order one, this will, in rescaled time be approximately exponential. Moreover, under this condition, the probability that we see both a coalescence and a mutation in our sample in a single generation is of order $1/N_e^2$. So just as in our derivation of the Kingman coalescent, we see that if there are currently k lineages ancestral to our sample, the time (in rescaled units) we must trace back until we see *some* event is (approximately) the minimum of k independent exponential random variables each with parameter $2N_e\mu$ and an independent exponential random variable with parameter $\binom{k}{2}$. Another way to say this is we can add mutations to Kingman's coalescent by simply superposing