Stochastic Models of Evolution in Genetics, Ecology and Linguistics

R. A. Blythe

SUPA, School of Physics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ

A. J. McKane

School of Physics and Astronomy, University of Manchester, Manchester M13 9PL

Abstract. We give a overview of stochastic models of evolution that have found applications in genetics, ecology and linguistics for an audience of nonspecialists, especially statistical physicists. In particular, we focus mostly on neutral models in which no intrinsic advantage is ascribed to a particular type of the variable unit, for example a gene, appearing in the theory. In many cases these models are exactly solvable and furthermore go some way to describing observed features of genetic, ecological and linguistic systems.

fact is a special case of the general model of population subdivision outlined here with all subpopulation sizes $N_i = 1$.

6. Selection

So far we have considered only neutral models of evolution, that is, those for which there is no preference for a particular allele. Despite being apparently a reasonable model for some aspects of genetic, ecological or linguistic behaviour (as we have previously discussed) geneticists in particular have been interested in the fate of alleles that are selected for or against.

The relationship between the genetic make-up of an individual and its survival is of course very complicated. However, one can explore the effects of selection by simply introducing parameters that determine how many offspring an individual carrying a particular allele (or combination of alleles when diploid organisms are being considered) has on average. In this section we offer a small taste of some evolutionary models that encompass selection.

Let us return to the model that has a randomly-mating haploid population of N individuals with two alleles, denoted A and B. In the neutral model, the parent of each individual in an offspring generation is chosen uniformly from all possible parents. When selection is active, it is supposed that an individual with allele A (B) is chosen to be a parent with a weight w_A (w_B). That is, if there are n' A alleles in the parent generation, each offspring has a probability

$$\frac{n'w_A}{n'w_A + (N-n')w_B}$$

of carrying allele A (at least, if no mutations occur). Since each individual is assigned a parent independently, we have that the probability for there to be n A alleles in the next generation is

$$p_{n\,n'} = \binom{N}{n} \left(\frac{n'w_A}{n'w_A + (N-n')w_B} \right)^n \left(\frac{(N-n')w_B}{n'w_A + (N-n')w_B} \right)^{N-n} \tag{127}$$

$$= \frac{1}{(\bar{w}')^N} \binom{N}{n} \left(\left[\frac{n'}{N}\right] w_A \right)^n \left(\left[1 - \frac{n'}{N}\right] w_B \right)^{N-n} \tag{128}$$

where in the second line we have simplified the notation by introducing the mean *fitness* of a population with n' A alleles

$$\bar{w}' = \frac{1}{N} \left(n'w_A + [N-n']w_B \right) \ . \tag{129}$$

Note that when the two alleles have equal fitness, i.e., $w_A = w_B$, (128) reduces to (2) for a neutral population.

It is usual to take a pre-existing 'wild-type' allele (we'll take this to be B) to have fitness $w_B = 1$, and the 'mutant' (A) to have fitness $w_A = 1 + s$. Then, the mean change in the number of mutants in one generation, given that there are $n(t)$ mutants in generation t, is

$$\frac{\langle n(t+1)\rangle - n(t)}{N} = s\frac{n(t)}{N}\left(1 - \frac{n(t)}{N}\right) + O(s^2)\ . \tag{130}$$

Denoting the frequency of mutant A alleles as x, the Fokker-Planck equation can be shown (e.g., via a Kramers-Moyal expansion or a large-N expansion) to be

$$\frac{\partial}{\partial t}P(x,t) = -s\frac{\partial}{\partial x}x(1-x)P(x,t) + \frac{1}{2N}\frac{\partial^2}{\partial x^2}x(1-x)P(x,t) \tag{131}$$

when the relative fitness of the mutant s is small. In this instance, small means $s \ll 1$, and not relative to $1/N$. Hence, if at some fixed s one has the (effective) population size $N \gg 1/s$, the effects of drift can typically be neglected. Then the mean allele frequency satisfies the deterministic equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = sx(1-x) \tag{132}$$

and one finds a logistic growth in the number of mutants

$$x(t) = \frac{x(0)}{x(0) + [1 - x(0)]\mathrm{e}^{-st}}\ . \tag{133}$$

Typically it is assumed that the stochastic effects of drift are important only when one of the allele frequencies is small, and so the logistic growth is taken to be representative of the change in frequency of a beneficial mutation in a randomly mating population once its frequency has reached some threshold (see e.g., [82]). This period of rapid growth is often referred to as a *selective sweep*. Of course, a mutant allele is present only in small numbers when it first appears, and here one can work out the probability that a mutant allele will be lost due to genetic drift. To do this one needs to solve the stationary *backward* Fokker-Planck equation corresponding to (131)

$$0 = x(1-x)\left[s\frac{\mathrm{d}}{\mathrm{d}x}Q(x) + \frac{1}{2N}\frac{\mathrm{d}^2}{\mathrm{d}x^2}Q(x)\right] \tag{134}$$

subject to the boundary conditions $Q(0) = 0$ and $Q(1) = 1$, since $Q(x)$ is the probability that a mutant allele becomes fixed given that its initial frequency is x. Integrating once gives $Q'(x) \propto \mathrm{e}^{-2Nsx}$ and again gives $Q(x) = A\mathrm{e}^{-2Nsx} + B$ where A and B are to be fixed by the boundary conditions. This yields [83]

$$Q(x) = \frac{1 - \mathrm{e}^{-2Nsx}}{1 - \mathrm{e}^{-2Ns}}\ . \tag{135}$$

In particular, if there is initially only a single mutant, $x = 1/N$, in the limit of an infinite population one has

$$\lim_{N\to\infty} Q(1/N) = \begin{cases} 2s & 0 < s \ll 1 \\ 0 & s \le 0 \end{cases} \tag{136}$$

where we recall that to obtain (131) it was assumed that s is small. From the backward Fokker-Planck equation (134) one can also find the mean number of generations τ until fixation of a selectively advantageous allele. It is given approximately as

$$\tau = \frac{2\ln N}{s} \tag{137}$$

when the combination Ns is large (e.g., s some fixed value and $N \to \infty$) [24].

One may ask how selection affects the backward-time formulation of the population dynamics that is couched in terms of genealogies. It turns out that the complexity of the calculations increases considerably, because as one goes back in time one needs to keep track of the number of alleles of each type present in the population. One can sometimes find situations in which there is an equilibrium in these frequencies, for example, when a selectively disadvantageous allele ($s < 0$) is maintained in a population due to recurrent mutations generating it [84]. When there is a selective sweep, it can be shown that the relevant genealogies are those that have *multiple* lineages coalescing simultaneously [85]—compare with the case of neutral evolution when the probability of a triple merger is suppressed by a factor of $1/N$ compared to that of a pairwise coalescence. This is consistent with the fact that for fixation to be achieved on a timescale sublinear in N (137): recall, that in a state of fixation *all* N individuals share a common ancestor. One way to formulate the genealogical process with selection is through the "Λ-coalescent" [86]; meanwhile, certain aspects have recently been elucidated by considering the properties of noisy travelling waves [87].

Given the discussion of Section 5, one may also be interested in determining how effects of selection and population subdivision combine. In a number of ways the situation is rather similar to the neutral case, at least if the selective advantage s (or disadvantage, if negative) is not spatially dependent. For example, when migration is strong one expects fixation probabilities and times to be given (135) and (137) but with N being an effective population size of the order of the total population size [88]. In the slow migration limit, the mean time for a lineage to hop between subpopulations $\sim N$ is much longer than the duration of a selective sweep $\sim \ln N$ and so typically each subpopulation is taken to be fixed in either the wild-type or mutant state. One thus calculates fixation probabilities and times by having wild-type subpopulations invaded by their mutant neighbours (and vice versa) on the migration timescale, and a flip from the wild-type to mutant state occurring with the probability given by $Q(1/N)$, and in the other direction with the same expression but with s replaced by $-s$. When conservative migration is in force, it can be shown that the probability a mutant allele fixes is independent of the location of the initial mutation [53, 88]: a similar situation occurred in the neutral case (although the fixation probability is different). To see deviations from this behaviour, one needs either to introduce additional processes (such as extinction and recolonisation of subpopulations [89]) or relax the assumption of conservative migration [90]. In the latter case one finds that the network structure connecting the subpopulations strongly influences whether selection or drift is the dominant process. For example, star-like structures amplify the effects of selection and can be constructed such that an advantageous mutation appearing almost anywhere is guaranteed to fix. On the other hand, it is also possible to contrive networks in which fixation occurs whenever the mutation occurs within a particular subpopulation, and never if it appears elsewhere. This latter type of behaviour does not, in fact, depend on the mutation having a selective advantage: even neutral mutations appearing in "well-connected" parts of a system can fix with high probability when migration is a non-conservative process [21].

Shifting focus from networks to continuous space, one can write down an equation for the advance of an advantageous mutation. Recall that in the absence of drift, one has the deterministic equation (132) for the mutant gene frequency x in a subpopulation. With isotropic migration in one dimension with a coordinate u one would augment this equation with a diffusion term

$$\frac{\partial}{\partial t} x(u,t) = D \frac{\partial^2}{\partial u^2} x(u,t) + s x(u)[1 - x(u)] \, . \tag{138}$$

This is known as the Fisher or KPP (Kolmogorov-Petrovksy-Piskunov) equation [91, 92] and admits travelling wave solutions of the form $x(u,t) = f(u - vt)$ where v is the wave velocity. If one anticipates that the leading edge of the wave has an exponential decay $f(\xi) = \mathrm{e}^{-\lambda\xi}$, one finds a range of velocities $v = D\lambda + s/\lambda \geq 2\sqrt{s/D}$ are possible. It turns out that in such an equation, if the interface between the stable and unstable phases (here, regions of high and low mutant frequencies) is sufficiently sharp, the front velocity selected is the smallest allowed [93]. If genetic drift is reintroduced, e.g., by adding a white noise to (138) with zero mean and variance $x(1-x)/N$ (cf. (131), but note the usual problems that arise from introducing multiplicative noise in such an *ad-hoc* way), one expects strong fluctuations in the leading edge of the travelling wave as a consequence of the possibility that a newly introduced mutant may go extinct. This causes a very small shift in the velocity of the front and a more diffuse profile than in the deterministic case [94].

One can also consider models with spatially varying fitnesses. A simple example would be if a mutant had fitness advantage $+s$ at position $u > 0$ and disadvantage $-s$ at $u < 0$ (s here is taken to be a positive number) [95]. At $u \to \infty$ one anticipates that the mutant allele would be fixed ($x = 1$), whilst at $u \to -\infty$ only the wild type would be found ($x = 0$). Using (138) the steady state can be found by solving the nonlinear equation

$$D \frac{\mathrm{d}^2}{\mathrm{d}u^2} x(u) = \pm s x(u)[1 - x(u)] \tag{139}$$

where the positive sign is taken for $u < 0$ and the negative sign for $u > 0$. The symmetry of the problem implies that $x(u) = 1 - x(-u)$, and in particular that $x(0) = \frac{1}{2}$. However, one finds the step in the fitness landscape at $u = 0$ induces a discontinuity in the gradient of $x(u)$ at that point. To see this, one must solve this differential equation which is achieved by multiplying both sides by $\frac{\mathrm{d}x}{\mathrm{d}u}$ and integrating twice. This procedure leads to [95]

$$x(u) = \begin{cases} \frac{3}{2}\left[1 - \tanh^2\left(\frac{1}{2}\sqrt{\frac{s}{D}} u - \kappa\right)\right] & u < 0 \\ \frac{3}{2}\tanh^2\left(\frac{1}{2}\sqrt{\frac{s}{D}} u + \kappa\right) - \frac{1}{2} & u > 0 \end{cases} \tag{140}$$

where

$$\kappa = \frac{1}{2} \ln\left(\frac{\sqrt{3} + \sqrt{2}}{\sqrt{3} - \sqrt{2}}\right) = 1.146216\ldots \, . \tag{141}$$

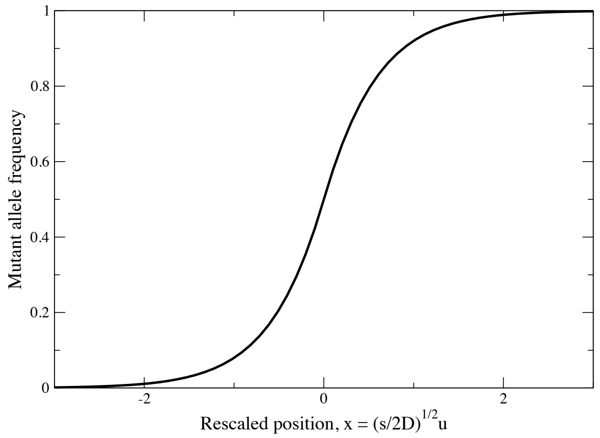

Figure 7. Variation of mutant allele frequency with position in a steady state where selection is balanced by migration. The mutant allele has a selective advantage s at positions $u > 0$, a disadvantage $-s$ at positions $u < 0$. Migration is characterised by a diffusion constant D. Although hard to discern on the plot, there is a discontinuity in the gradient of the allele frequencies at $u = 0$.

This function is plotted in Fig. 7. The spatial variation of allele frequencies due to a balance between migration and a changing fitness landscape has been called a *cline* [96].

Finally one can consider variation in fitness not in real space, but in the space of genotypes (i.e., possible allele combinations). Let us return to a randomly-mating diploid population, in which a fraction x_i of all genes are of allele A_i, $i = 1, 2, \ldots, n$. It is fairly straightforward to show [23] that, in the absence of drift, the change in allele frequency per generation is

$$\Delta x_i = \frac{x_i(1-x_i)}{2} \frac{\partial}{\partial x_i} \ln \bar{w} \tag{142}$$

where the mean fitness of the population is

$$\bar{w} = \sum_{i,j} x_i x_j w_{ij} \tag{143}$$

and w_{ij} is the fitness of an individual carrying the pair of alleles (A_i, A_j). One can therefore view the function $f(x_1, x_2, \ldots, x_n) = -\ln \bar{w}$ as a kind of free energy defined over the space of allele frequencies which the evolutionary dynamics seeks to minimise. If we now extend the discussion to multiple gene *loci* (i.e., genes coding for different traits) with interactions between them, it is possible for an extremely rugged fitness landscape in the space of all possible allele frequencies to emerge [97]. One can then find a transition between two distinct regimes induced by a change in mutation rate: if the total number of mutants appearing per generation across the whole population (which is governed by the population size N multiplied by the mutation rate u) is small, fixation of a selectively advantageous allele is likely to occur before the onset of the next mutation (see e.g, [98, 99]). Thus, all individuals in the population are then likely to have the same genotype, and one which tends towards a local fitness maximum. On the other hand, if the number of mutants appearing per generation is large, and one is likely to see many different genotypes in the population, each corresponding to a different fitness maximum. In the latter case the individuals are said to form a *quasi-species* [100, 98]

7. Conclusion

This article has been a review of the ideas and formalism used to model stochastic processes in fields that statistical physicists are not typically acquainted with, specifically population genetics, ecology and linguistics. As a consequence, some parts of the discussion will seem familiar, other parts will not. We have tried, and we hope that we have succeeded, to explain the background ideas and motivation, since this will be the greatest obstacle to understanding among a readership of statistical physicists. On the other hand the degree of mathematical sophistication that has been assumed is greater than would be typical outside physics or mathematical biology. This makes review quite different to others in the same area, and while we expect the readership to be mainly statistical physicists, we hope that some of those working particularly in ecology and linguistics will find our approach to their subject interesting and stimulating.

In our discussions of the mathematical models, we have mostly used the language of population genetics, but through of the mappings discussed in Section 2.5 the results obtained are more widely relevant. As the evolutionary paradigm becomes even more widely applied, there may be other areas in which analogies can be drawn. It is interesting how neutral processes turn out to have greater importance in all three areas we discussed; at the very least neutral theories can be thought of a null models, against which data and other models can be compared. Most textbooks in population genetics begin their discussion of genetic drift with the Wright-Fisher model, although for physicists the use of non-overlapping generations and a "time" measured in number of generations will not appear so natural. The Moran model, which has exactly the same limit when the number of genes become large, is far more familiar, resembling a birth/death processes where a death is immediately followed by a birth. In addition, the continuous time limit may easily be taken, leading to a master equation of a kind well-known in statistical physics.

We spent some time explaining the relationship between the discrete-time approach, based on transition probabilities, and the continuous-time master equation approach, based on processes occurring independently at a given rate. The use of master equations is apparently rare in the population genetics literature, which is one of the reasons why we have gone into this approach in such detail. Furthermore, this latter formalism usually turns out to be more efficient. For example, in his book Hubbell [7] formulated his model by analogy with the corresponding discrete-time genetic models, and had to perform simulations to generate the stationary probability distribution. However, if the model is formulated as a master equation, the stationary probability distribution can be obtained analytically [101]. However, the relationship between the discrete- and continuous-time formulations is not necessarily straightforward. For example, in