1 Sequence

1.1 Probability & Information

We are used to dealing with information presented as a sequence of letters. For example, each word in the English language is composed of $m = 26$ letters; the text itself includes also spaces and punctuation marks. Similarly in biology the blueprint for any organism is the string of bases along DNA, e.g. $AGTTCCAG\cdots$, where at each position there is a choice of $m = 4$ possible characters. A portion of this information is then transcribed into proteins, made of sequences of $m = 20$ amino acids. Clearly any of these sequences is far from random and there are constraints and correlations at many scales that conspire to make them meaningful. Nonetheless, as a means to unravel such constraints, it may be helpful to start with simple models which assume that sequences are randomly generated according to simple rules. Comparisons of such models with the actual sequences may then provide insights that help unravel their meaning.

As a simple example, let us consider a sequence of N characters, each chosen independently with probabilities $\{p_\alpha\}$, with $\alpha = 1, 2, \cdots, m$. (This choice is sometimes referred to as IID, for *identical, independently distributed* random variables.) Since the probabilities must be normalized, we require

$$\sum_{\alpha=1}^{m} p_\alpha = 1. \tag{1.1}$$

The probability of finding a sequence $S = \{\alpha_1, \cdots, \alpha_N\}$ is then given by the product of probabilities for its elements, as

$$p(S|\{p_\alpha\}) = \prod_{\ell=1}^{N} p_{\alpha_\ell}. \tag{1.2}$$

How many other sequences S' have this exact probability? Clearly as long as the number of occurrences $\{N_\alpha\}$ of each character is the same, the probability will be identical, i.e. the order of the elements does not matter in calculating the probability for this simple model. The number $\mathcal{N}$ of possible permutations of the elements in S is

$$\mathcal{N} = \frac{N!}{\prod_{\alpha=1}^{m} N_\alpha!}. \tag{1.3}$$

This is known as the multinomial coefficient as it occurs in the expression

$$(p_1 + p_2 + \ldots + p_m)^N = \sum_{\{N_\alpha\}}{}' p_1^{N_1} p_2^{N_2} \cdots p_m^{N_m} \times \frac{N!}{\prod_{\alpha=1}^{m} N_\alpha!}, \tag{1.4}$$

where the sum is restricted so that $\sum_{\alpha=1}^{m} N_\alpha = N$. Note that because of normalization, both sides of the above equation are equal 1. The terms within the sum on the right-hand side are known the *multinomial probabilities*

$$p(N_1, N_2, \cdots, N_m) = p_1^{N_1} p_2^{N_2} \cdots p_m^{N_m} \times \frac{N!}{\prod_{\alpha=1}^{m} N_\alpha!}. \tag{1.5}$$

With the assumption of independence, the probability of a sequence is determined entirely by the set $\{N_\alpha\}$ according to Eq. (1.5). It is easy to check that the most likely set (the mode $\{N_\alpha^*\}$) coincides with the average (mean $\{\langle N_\alpha \rangle\}$), and given by

$$N_\alpha^* = \langle N_\alpha \rangle = p_\alpha N. \tag{1.6}$$

Indeed, in the limit of large N, the overwhelming number of sequences generated will have the above composition. The number of sequences with character counts $N_\alpha = p_\alpha N$ is given by Eq. (1.3). Crudely speaking, this number $\mathcal{N}$ helps quantify the "information" contained within a sequence of length N, as it indicates how many different sequences have the same composition of characters (and hence the same *a priori* probability). We expect a good measure of information content to scale roughly linearly with the message length. (In the absence of context clues or syntax rules, a message twice as long should carry about twice as much information.) As a convenient measure, and taking clues from Statistical Mechanics, we take the logarithm of Eq. (1.3), which gives

$$\begin{aligned}
\log \mathcal{N} &= \log N! - \sum_\alpha \log N_\alpha! \\
&\approx N \log N - N - \sum_\alpha (N_\alpha \log N_\alpha - N_\alpha) \\
&= -N \cdot \sum_\alpha \left(\frac{N_\alpha}{N} \right) \log \left(\frac{N_\alpha}{N} \right).
\end{aligned}$$

(Stirling's approximation for $N!$ is used for all $N_\alpha \gg 1$.) The above formula is closely related to the *entropy of mixing* in thermodynamics, and quite generally for any set of probabilities $\{p_\alpha\}$, we can define a *mixing entropy*

$$\mathcal{S}\left[\{p_\alpha\}\right] = -\sum_\alpha p_\alpha \log p_\alpha. \tag{1.7}$$

Entropy is typically envisioned as a measure of disorder, and the information content $\mathcal{I}\left[\{p_\alpha\}\right]$ (picking up a specific element amongst a jumble of possibilities) is related to $-\mathcal{S}\left[\{p_\alpha\}\right]$.

Let us illustrate the relations among entropy and information in the context of DNA. To transmit a sequence, $ACTG\cdots$, along a binary channel we need to encode $2N$ bits, as there are $(2^2)^N$ possibilities. However, suppose that from prior analysis of DNA of a particular organism, we know that a typical sequence of length N has a likely composition $\langle N_A \rangle \neq \langle N_G \rangle \neq \cdots$. Given *a priori* knowledge of the probabilities $p_\alpha = N_\alpha/N$, the number of such likely sequences is

$$\mathcal{N} = \frac{N!}{\prod_{\alpha=1}^m N_\alpha!} \ll (2^2)^N,$$

or, upon taking the logarithm,

$$\log_2 \mathcal{N} = -N \sum_\alpha p_\alpha \log_2 p_\alpha < 2N.$$

We gain a definite amount of knowledge by having advance insight about $\{p_\alpha\}$. Instead of having to specify 2 bits per "letter" of DNA, we can get by with a smaller number. The information gained (in bits) per letter is given by

$$\mathcal{I}(\{p_\alpha\}) = 2 - \sum_\alpha p_\alpha \log_2 \left(\frac{1}{p_\alpha}\right). \tag{1.8}$$

If $p_\alpha = 1/4$, then Eq. (1.8) reduces to 0, which is consistent with the expected no gain in information. On the other hand, if $p_A = p_T = 0$ and $p_C = p_G = \frac{1}{2}$, then

$$\mathcal{I} = 2 - \sum_{G,C} \frac{1}{2} \log_2 2 = 1 \text{ bit per base.}$$

1.2 Evolving Probabilities

As organisms reproduce the underlying genetic information is passed on to subsequent generation. The copying of the genetic content is not perfect, and leads to a diverse and evolving population of organisms after many generations. The changes are stochastic, and are thus appropriately described by evolving probability distributions. After motivating such evolving probabilities in the contexts of DNA and populations, we introduce the mathematical tools for treating them.

1.2.1 Mutations

Consider the flow of information from DNA, transcribed to messenger RNA, and eventually translated to an amino acid chain. Suppose we begin with the DNA fragment

$$\text{ATT CGC ATG} \quad ,$$

which when unwound and transcribed to mRNA, appears as the complementary messenger chain

$$\text{UAA GCG UAC} \quad .$$

The protein building machinery (ribosome) translates this to a *peptide* chain consisting of a leucine, an alanine, and a tyrosine molecule, symbolically,

$$\text{Leu Ala Tyr} \quad .$$

Suppose, however, that a replication mistake causes the DNA strand's last "letter" to change. Instead of ATG, the last codon now reads ATC, which is a "stop signal"

$$\text{Leu Ala STOP.}$$

Such a mutation, let's say in the middle of a protein chain, will stop the translation process. The mutation is *deleterious* and the off-spring will not survive. However, as a result of the

redundancy in the genetic code, there are also mutations that are *synonymous,* in that they do not change the amino acid which eventually results. Because these synonymous mutations do not affect the biological viability of the organism, we can find genes whose exact DNA varies from individual to individual. This has opened up the field of DNA "fingerprinting:" blood can be matched to a particular individual by comparing such *single nucleotide polymorphisms* (SNPs). Non-synonymous mutations are not necessarily deleterious and may also lead to viable off-spring.

1.2.2 Master Equation

Let us consider the evolution of probabilities in the context of the simplified model introduced earlier of N independently distributed sites. We model mutations by assuming that at subsequent time-steps (generations) each site may change its state (independent of the other sites), say from α to β with a *transition probability* $\pi_{\beta\alpha}$. The $q \times q$ such elements form the *transition probability matrix* $\overleftrightarrow{\pi}$. (Without the assumption that the sites evolve independently, we would have constructed a much larger $(q^N \times q^N)$ matrix $\overleftrightarrow{\Pi}$. With the assumption of independence, this larger matrix is a direct product of transition matrices for individual sites, i.e. $\overleftrightarrow{\Pi} = \overleftrightarrow{\pi}_1 \otimes \overleftrightarrow{\pi}_2 \otimes \cdots \otimes \overleftrightarrow{\pi}_N$.) Using the transition probability matrix, we can track the evolution of the probabilities as

$$p_\alpha(\tau+1) = \sum_{\beta=1}^{m} \pi_{\alpha\beta} p_\beta(\tau), \quad \text{or in matrix form} \quad \vec{p}(\tau+1) = \overleftrightarrow{\pi}\vec{p}(\tau) = \overleftrightarrow{\pi}^\tau \vec{p}(1), \tag{1.9}$$

where the last identity is obtained by recursion, assuming that the transition probability matrix remains the same.

Probabilities must be normalized to unity, and thus the transition probabilities are constrained by

$$\sum_\alpha \pi_{\alpha\beta} = 1, \quad \text{or} \quad \pi_{\beta\beta} = 1 - \sum_{\alpha\neq\beta} \pi_{\alpha\beta}. \tag{1.10}$$

The last expression formalizes the statement that ithe probability to stay in the same state is the complement of the probabilities to make a change. Using this result, we can rewrite Eq. (1.9) as

$$p_\alpha(\tau+1) = p_\alpha(\tau) + \sum_{\beta\neq\alpha} \left[\pi_{\alpha\beta} p_\beta(\tau) - \pi_{\beta\alpha} p_\alpha(\tau)\right]. \tag{1.11}$$

In many circumstances of interest the probabilities change slowly and continuously over time, in which case we introduce a the time interval Δt between subsequent generations, and write

$$\frac{p_\alpha(\tau+1) - p_\alpha(\tau)}{\Delta t} = \sum_{\beta\neq\alpha} \left[\frac{\pi_{\alpha\beta}}{\Delta t} p_\beta(\tau) - \frac{\pi_{\beta\alpha}}{\Delta t} p_\alpha(\tau)\right]. \tag{1.12}$$

In the limit of small Δt, $[p_\alpha(\tau+1) - p_\alpha(\tau)]/\Delta t \approx dp_\alpha/dt$, while

$$\frac{\pi_{\alpha\beta}}{\Delta t} = R_{\alpha\beta} + \mathcal{O}(\Delta t) \quad \text{for } \alpha \neq \beta, \tag{1.13}$$

are the off-diagonal elements of the matrix $\overleftrightarrow{R}$ of *transition probability rates*. The diagonal elements of the matrix describe the depletion rate of a particular state, and by conservation of probability must satisfy, as in Eq. (1.10),

$$\sum_{\alpha} R_{\alpha\beta} = 0, \quad \text{or} \quad R_{\beta\beta} = -\sum_{\alpha\neq\beta} R_{\alpha\beta}. \tag{1.14}$$

We thus arrive at

$$\frac{dp_\alpha(t)}{dt} = \sum_{\beta\neq\alpha} \left(R_{\alpha\beta}p_\beta(t) - R_{\beta\alpha}p_\alpha(t)\right) \quad , \tag{1.15}$$

which is known as the *Master equation*.

1.2.3 Steady state

Because of the conservation of probability in Eqs. (1.10) and (1.14), the transition probability matrix $\overleftrightarrow{\pi}$, and by extension the rate matrix $\overleftrightarrow{R}$ have a left-eingenvector $\overleftarrow{v^*} = (1, 1, \cdots, 1)$ with eigenvalues of unity and zero respectively, i.e.

$$\overleftarrow{v^*}\overleftrightarrow{\pi} = \overleftarrow{v^*} \quad , \quad \text{and} \quad \overleftarrow{v^*}\overleftrightarrow{R} = 0. \tag{1.16}$$

For each eigenvalue there is both a left eigenvector and a right eigenvector. The matrices $\overleftrightarrow{\pi}$ and $\overleftrightarrow{R}$ thus must also have a right-eigenvector $\overrightarrow{p^*}$ such that

$$\overleftrightarrow{\pi}\overrightarrow{p^*} = \overrightarrow{p^*} \quad , \quad \text{and} \quad \overleftrightarrow{R}\overrightarrow{p^*} = 0. \tag{1.17}$$

The elements of the vector $\overrightarrow{p^*}$ represent the *steady state probabilities* for the process. These probabilities no longer change with time. From Eq. (1.11) we observe that the steady state probabilities satisfy the so-called condition of *detailed balance*,

$$\pi_{\alpha\beta}p^*_\beta = \pi_{\beta\alpha}p^*_\alpha. \tag{1.18}$$

The remaining eigenvalues of any transition matrix have magnitude less than unity; they determine how an initial vector of probabilities approaches steady state.

As a simple example, let us consider a *binary* sequence (i.e. $m = 2$) with independent states A_1 or A_2 at each site.[1] Let us assume that the state A_1 can "mutate" to A_2 at a rate μ_2, while state A_2 may change to A_1 with a rate μ_1. The probabilities $p_1(t)$ and $p_2(t)$ now evolve in time as

$$\frac{d}{dt}\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} -\mu_2 & \mu_1 \\ \mu_2 & -\mu_1 \end{pmatrix}\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}. \tag{1.19}$$

[1] Clearly with the assumption of independence we are really treating independent sites, and the insistence on a sequence may appear frivolous. The advantage of this perspective, however, will become apparent in the next section.

The above 2×2 transition rate matrix has the following two eigenvectors

$$\begin{pmatrix} -\mu_2 & \mu_1 \\ \mu_2 & -\mu_1 \end{pmatrix} \begin{pmatrix} \frac{\mu_1}{\mu_1+\mu_2} \\ \frac{\mu_2}{\mu_1+\mu_2} \end{pmatrix} = 0, \quad \text{and} \quad \begin{pmatrix} -\mu_2 & \mu_1 \\ \mu_2 & -\mu_1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = -(\mu_1+\mu_2) \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{1.20}$$

As anticipated, there is an eigenvector $\overrightarrow{p^*}$ with eigenvalue of zero; the elements of this vector are normalized to add to unity, as required for probabilities. We have not normalized the second eigenvector, whose eigenvalue $-(\mu_1+\mu_2)$ determines the rate of approach to steady state.

To make this explicit, let us start with a sequence that is purely A_1, i.e. with $p_1 = 1$ and $p_2 = 0$ at $t = 0$. The formal solution to the linear differential equation (1.19) is

$$\begin{pmatrix} p_1(t) \\ p_2(t) \end{pmatrix} = \exp\left[t \begin{pmatrix} -\mu_2 & \mu_1 \\ \mu_2 & -\mu_1 \end{pmatrix} \right] \begin{pmatrix} p_1(0) \\ p_2(0) \end{pmatrix}. \tag{1.21}$$

Decomposing the initial state as a sum over the eigenvectors, and noting the action of the rate matrix on each eigenvector from Eq. (1.20), we find

$$\begin{aligned} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} &= \exp\left[t \begin{pmatrix} -\mu_2 & \mu_1 \\ \mu_2 & -\mu_1 \end{pmatrix} \right] \left[\begin{pmatrix} \frac{\mu_1}{\mu_1+\mu_2} \\ \frac{\mu_2}{\mu_1+\mu_2} \end{pmatrix} + \frac{\mu_2}{\mu_1+\mu_2} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right] \\ &= \begin{pmatrix} \frac{\mu_1}{\mu_1+\mu_2} + e^{-(\mu_1+\mu_2)t} \frac{\mu_2}{\mu_1+\mu_2} \\ \frac{\mu_2}{\mu_1+\mu_2} - e^{-(\mu_1+\mu_2)t} \frac{\mu_2}{\mu_1+\mu_2} \end{pmatrix}. \end{aligned} \tag{1.22}$$

At long times the probabilities to find state A_1 or A_2 are in the ratios μ_1 to μ_2 as dictated by the steady state eigenvector. The rate at which the probabilities converge to this steady steady is determined by the eigenvalue $-(\mu_1+\mu_2)$.

1.2.4 Enzymatic reaction

The appeal of the formalism introduced above is that the same concepts and mathematical formulas apply to a host of different situations. For example consider the reactions

$$\mathrm{A} + \mathrm{E} \rightleftharpoons_{b'}^{a'} \mathrm{B} + \mathrm{E} \quad , \tag{1.23}$$

where the enzyme E facilitates the conversion of A to B at a rate a', and the backward reaction at rate b'. In a well mixed system, the numbers N_A and $N_B = N - N_A$ of the two species evolve according to the "mean-field" equation

$$\frac{dN_A}{dt} = -a' N_E N_A + b' N_E N_B = -a N_A + b(N - N_A)\,, \tag{1.24}$$

where $a = N_E a'$ and $b = N_E b'$. In this approximation, the fluctuations are ignored and the mean numbers of constituents evolve to the steady state with $N_A^*/N_B^* = b/a$.

However, in a system where the number of particles is small, for example for a variety of proteins within a cell, the mean number may not be representative, and the entire distribution is relevant. The probability to find a state with $N_A = n$ and $N_B = N - N_A$,

then evolves precisely according to Eq. (1.27) introduced above in the context of mutating populations. From the equivalence of this equation to the independently evolving binary states, we know that the final steady steady state solution also describes a chain of binary elements independently distributed with probabilities $p_A^* = b/(a+b)$ and $p_B^* = a/(a+b)$. Hence, the steady state solution to the complicated looking set of equations (1.27) is simply

$$p^*(n) = \left(\begin{array}{c} N \\ n \end{array} \right) \frac{b^n a^{N-n}}{(a+b)^N} \,. \tag{1.25}$$

In fact, this analogy enables following the full evolution of the probability to this state, starting let's say with an initial state that is all A (see Assignment #1).

1.3 Population Genetics

The study of heredity began long before the molecular structure of DNA was understood. Several thousand years of experience breeding animals and plants led, eventually, to the idea that hereditary characteristics are passed along from parents to offspring in units, which are termed *genes.* While *genotype* refers to the inherited genetic blueprint of an individual, *phenotype* refers to observable traits (such as height or eye color) distinguishing members of a population.

For theoretical studies, the term *locus* refers to a basic genetic element that is variable in a population, for example a single site along DNA, or an amino acid for a protein. Different states of the locus are called *alleles.* A particular phenotype may be the outcome of interplay amongst several alleles. However, for simplicity we shall typically deal with the most elementary example of one locus with two possible states, say A_1 and A_2.

A further complication in relating genotype to phenotype arises since humans, among other *diploid* organisms, carry two copies of each gene. (*Haploid* organisms, such as bacteria typically have only one copy.) Thus even in the above simplest case, there are three possible genotypes: A_1A_1, A_1A_2 and A_2A_2. (A_1A_1 and A_2A_2 are *homozygotes* while A_1A_2 is *heterozygote.*) A particular allele can be *dominant* or *recessive*; the presence of dominant allele outweighing the recessive one. For example, suppose that A_1 codes for brown eyes, while the variant A_2 leads to blue eyes. Brown eyes turn out to be dominant in humans, so a person with an A_1A_2 mix of alleles has brown irises, just like one whose alleles read A_1A_1. Only an A_2A_2 individual develops blue irises.

It is common in genetics to assume a fixed *population size* N, and inquire about the evolution of its genetic makeup with time. For example, within a haploid population we may chart the changes of the allele fraction $x_1 \equiv N_1/N$ with the number of generations. Such quantities evolve from an interplay of *mutation*, *reproduction*, and *selection*, as discussed in the following sections.

1.3.1 Mutation

In a previous example, we considered the case of a binary sequence of length N evolving by potential mutations on each site. From this perspective the model represents a collection of

N independent binary loci. In fact, with a simple reinterpretation, the same mathematical model can represent a single allele in a population of fixed size as follows. Let us assume that A_1 and A_2 denote two forms of a particular allele. In each generation any individual is replaced by an offspring that mostly retains its progenitor's allele, but may mutate to the other form at some rate. In this model the total population size is fixed to N, while the sub-populations N_1 and N_2 may vary. A particular state of the population is thus described by $N_1 = n$ and $N_2 = N - n$, and since $n = 0, 1, \cdots, N$ there are $N + 1$ possible states. At a particular time, the system may be in any one of these states with probability $p(n, t)$, and we would like to follow the evolution of these probabilities.

After an individual replication event (A_1 to A_1 at rate $-\mu_2$, A_1 to A_2 at rate μ_2, A_2 to A_1 at rate μ_1, or A_2 to A_2 at rate $-\mu_1$), the number N either stays the same, or changes by unity. Thus the transition rate matrix only has non-zero terms along or adjoining to the diagonal. For example

$$R_{n,n+1} = \mu_2(n + 1), \quad \text{and} \quad R_{n,n-1} = \mu_1(N - n + 1), \tag{1.26}$$

where the former indicates that a population of $n + 1$ A_1s can decrease by one if any one of them mutates to A_2, while the population a population with $n-1$ A_1s increases by one if any of A_2s mutates to A_1. The diagonal terms are obtained from the normalization condition in Eq. (1.14) resulting in the Master equation

$$\frac{dp(n,t)}{dt} = \mu_2(n + 1)p(n + 1) + \mu_1(N - n + 1)p(n - 1) - \mu_2 np(n) - \mu_1(N - n)p(n)\,, \tag{1.27}$$

for $0 < n < N$, and with boundary terms

$$\frac{dp(0,t)}{dt} = \mu_2 p(1) - \mu_1 N p(0), \quad \text{and} \quad \frac{dp(N,t)}{dt} = \mu_1 p(N - 1) - \mu_2 N p(N)\,. \tag{1.28}$$

1.3.2 Reproduction

The dynamics of a population depends upon births of new individuals, with possibly novel mutations. To maintain a constant population size this must be accompanied by death of members of previous generations. Even without mutations ($\mu_1 = \mu_2 = 0$ in the previous example), reproduction by birth/death introduces stochasticity in the dynamics (say of the proportion x_1 of allele A_1). To emphasize the role of reproduction, in this section we shall ignore the role of mutations, assuming a preexisting diversity of alleles in the population.

Hardy-Weinberg equilibrium: Within diploid organisms, sex and *mating* present additional complications, which we shall ignore by adapting a gene-centered perspective. To see why this may be justified in at least some limit, consider a very large population ($N \to \infty$) where diploid organisms mate randomly with no preference for phenotypic or geographic considerations. The initial population is characterized by the proportions x_{11}, x_{12}, and x_{22} of the genotypes A_1A_1, A_1A_2 and A_2A_2, with $x_{11} + x_{12} + x_{22} = 1$. The composition of the next generation is obtained by considering all possible matings and their outcomes. For example, a pairing of two homozygotes A_1A_1 individuals occurs with probability x_{11}^2, and

leads to A_1A_1 offspring. However, a mating of A_1A_1 with A_1A_2, with probability $x_{11}x_{12}$ may lead to either an A_1A_1 offspring, or an A_1A_2 offspring. Assuming no selective advantage for either such offspring, each happens with probability of 1/2. Similarly, the pairing of two heterozygotes A_1A_2 may result in A_1A_1, A_1A_2 and A_2A_2 with probabilities of 1/4, 1/2, and 1/4, respectively. Including all 9 (3×3) pairing, we arrive at

$$\begin{aligned} x'_{11} &= x_{11}^2 + 2 \cdot \frac{x_{11}x_{12}}{2} + \frac{x_{12}^2}{4}\,, \\ x'_{12} &= 2x_{11}x_{22} + 2 \cdot \frac{x_{11}x_{12}}{2} + 2 \cdot \frac{x_{22}x_{12}}{2} + \frac{x_{12}^2}{2}\,, \\ x'_{22} &= x_{22}^2 + 2 \cdot \frac{x_{22}x_{12}}{2} + \frac{x_{12}^2}{4}\,. \end{aligned} \tag{1.29}$$

(Note that pairings of distinct genotypes involve an additional factor of two, from the degeneracy in their order of selection.) It is easy to check that the above results are completely equivalent to $x'_1 = x_1$ and $x'_2 = x_2$, where $x_1 = x_{11} + x_{12}/2$ and $x_2 = x_{22} + x_{12}/2 = 1 - x_1$ are the the proportions of alleles A_1 and A_2 in the diploid population. (For example, the first equation above can be recast as $x'_{11} = x_1^{'2} = x_1^2$.) Thus, within one generation the alleles are mixed by random reproduction such that the proportion of the three possible genotypes merely reflects the proportion of the allele in the entire population. This so-called *Hardy-Weinberg equilibrium* justifies the gene-centered perspective as a theoretical limit. In fact, within a population of finite size N the frequency x_1 will change stochastically due to random reproduction events as discussed next.

Fisher-Wright (binomial) process: Consider a population with two forms of an allele, say A_1 and A_2 corresponding to blue or brown eye colors. The probability for a spontaneous mutation to occur that changes the allele for eye color is extremely small, and effectively $\mu_1 = \mu_2 = 0$ in Eq. (1.27). Yet the proportions of the two alleles in the population does change from generation to generation. One reason is that some individuals do not reproduce and leave no descendants, while others reproduce many times and have multiple descendants. This is itself a stochastic process and the major source of rapid changes in allele proportions. In principle this effect also leads to variations in population size. In practice, and to simplify computations, it is typically assumed that the size of the population is fixed.

Continuing with the gene-centered perspective, consider the following, so called *Fisher-Wright* process starting from the $2N$ alleles in a diploid population of size N. In the model of *binomial selection*, the process or reproduction from one generation to the next is assumed to be as follows: One allele is random selected, an exact copy is made for the next generation, while the original allele is returned to the original pool. This process is repeated $2N$ times to produce the next generation. Let us assume that in the initial population of $2N$ alleles, $N_1 = n = 2Nx_1$ are A_1, and the remaining $2N - n$ are A_2. The population at the next generation may have m individuals with allele A_1, with (transition) probability

$$\Pi_{mn} = \left(\frac{n}{2N}\right)^m \left(1 - \frac{n}{2N}\right)^{2N-m} \left(\begin{array}{c} 2N \\ m \end{array} \right). \tag{1.30}$$

The process leading to such probability is like reaching into a bag with n balls of blue color and $2N - m$ balls of brown color, recording the color of the selected ball and throwing it back to the bag. After repeating such selection N times, the probability that the blue color is recorded m times is given by the above binomial distribution. (The probability of getting a blue ball in each trial is simply $n/2N$, and $1 - n/2N$ for brown.) On average, the number of alleles does not change, since $\langle m \rangle = n$ from the binomial distribution (i.e. $\langle x_1' \rangle = x_1$ consistent with Hardy-Weinberg equilibrium). However, there is now a range of possible values of m; clearly the stochasticity arises since some balls can be picked up multiple times (multiple descendants), while some balls are never picked (no offspring). The mathematical consequences of Eq. (1.30) will be explored later on.

1.3.3 Selection

We assumed so far that the two alleles are completely equivalent, corresponding to *neutral* evolution. It is likely that one allele is better in the sense of conferring a selective advantage to the individual carrying it. The selective advantage of a genotype is parameterized through an associated *fitness* that quantifies its number of likely progeny (relative to other genotypes). In our diploid binary allele example, we may associate fitness values of f_{11}, f_{12} and f_{22} to the three genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. Indicating the proportion of allele A_1 in the population by $x \equiv x_1 = n/2N$, the average fitness is given by

$$\overline{f}(x) = x^2 f_{11} + 2x(1-x) f_{12} + (1-x)^2 f_{22} \,. \tag{1.31}$$

The expected numbers of off-spring for the three genotypes are thus $f_{11}/\overline{f}$, $f_{12}/\overline{f}$ and $f_{22}/\overline{f}$, respectively.

After one generation, the frequency x on average changes to

$$\langle x' \rangle = \frac{f_{11}}{\overline{f}} x^2 + \frac{1}{2} \frac{f_{12}}{\overline{f}} \cdot 2x(1-x) \,. \tag{1.32}$$

The expected change in the proportion of the allele is thus given by

$$\begin{aligned}
\Delta x \equiv \langle x' \rangle - x &= \frac{1}{\overline{f}} \left[f_{11} x^2 + f_{12} x(1-x) - \overline{f} x \right] \\
&= \frac{1}{\overline{f}} \left[f_{11} x^2 + f_{12} x(1-x) - f_{11} x^3 - 2 f_{12} x^2 (1-x) - f_{22} x (1-x)^2 \right] \\
&= \frac{1}{\overline{f}} \left[f_{11} x^2 (1-x) + f_{12} x(1-x)(1-2x) - f_{22} x (1-x)^2 \right] \\
&= \frac{x(1-x)}{\overline{f}} \left[\frac{1}{2} \frac{d\overline{f}(x)}{dx} \right] \\
&= \frac{x(1-x)}{2} \frac{d \ln \overline{f}}{dx} \,.
\end{aligned} \tag{1.33}$$

The above result, known as *Wright's equation* implies that allele frequencies always change so as to maximize the average fitness function $\overline{f}(x)$. A corresponding result holds for a multi-loci situation with a corresponding *fitness landscape* $\overline{f}(x_1, x_2, \cdots, x_n)$.

For ease of computations, in the following sections we shall write the selective advantage for allele A_1 ias

$$\Delta x = \frac{x(1-x)}{2} s \,, \tag{1.34}$$

typically ignoring any x dependence of s.

1.4 Continuum Limit

1.4.1 Forward Kolmogorov equation

Let us now consider a more general case where the states are still ordered along a line, such as in the previous examples with population size $n = 0, 1, 2 \cdots, N$. The general form of the Master equation is

$$\frac{dp_n}{dt} = + \sum_{m \neq n} R_{nm} p_m - \sum_{m \neq n} R_{mn} p_n \,. \tag{1.35}$$

In many relevant circumstances the number of states is large, and the probability varies smoothly from one site to the next. In such cases it is reasonable to replace the discrete index n with a continuous variable x, the probabilities $p_n(t)$ with a probability density $p(x,t)$, and the rates R_{mn} with a rate function $R(x', x)$. The rate function R depends on two variables x and x', denoting respectively the start and end positions for a transition along the line. We have the option of redefining the two arguments of this function, and it is useful to reparameterize it as $R(x' - x, x)$ indicating the rate at which, starting from the position x, a transition is made to a position $\Delta x = x' - x$ away. As in the case of mutations, there is usually a preference for changes that are *local*, i.e. the rates decay rapidly when the separation $x' - x$ becomes large.

These transformations and relabelings,

$$n \to x, \;\; p_n(t) \to p(x,t), \;\; R_{mn} \to R(x' - x, x) \,, \tag{1.36}$$

enable us to transform Eq. (1.35) to the continuous integral equation

$$\frac{\partial}{\partial t} p(x,t) = + \int^* dx' R(x - x', x') p(x', t) - \int^* dx' R(x' - x, x) p(x, t) \,. \tag{1.37}$$

Some care is necessary in replacing the sums with integrals, as the summations in in Eq. (1.35) exclude the term with $m = n$, To treat this restriction in the continuum limit, we focus on an interval y around any point x, and consider the change in probability due to incoming flux from $x - y$ and the outgoing flux to $x + y$, thus arriving at

$$\frac{\partial}{\partial t} p(x,t) = \int dy \left[R(y, x - y) p(x - y) - R(y, x) p(x) \right] \,. \tag{1.38}$$

Note that the contribution for $y = 0$ is now clearly zero. The flux difference for small y is now estimating by a Taylor expansion of the first term in the square bracket, *but only*

with respect to the location of the incoming flux, treating the argument pertaining to the separation of the two points as fixed, i.e.

$$R(y, x-y)p(x-y) = R(y,x)p(x) - y\frac{\partial}{\partial x}\left(R(y,x)p(x)\right) + \frac{y^2}{2}\frac{\partial^2}{\partial x^2}\left(R(y,x)p(x)\right) + \cdots . \quad (1.39)$$

While formally correct, the above expansion is useful only in cases where typical values of y are small (i.e. only almost *local* transitions occur). Keeping terms up to the second order, Eq. (1.38) can be rewritten as

$$\frac{\partial}{\partial t}p(x,t) = -\int dy\, y\frac{\partial}{\partial x}(R(y,x)p(x)) + \frac{1}{2}\int dy\, y^2\frac{\partial^2}{\partial x^2}(R(y,x)p(x)). \quad (1.40)$$

The integrals over y can be taken inside the derivatives with respect to x,

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}\left[p(x)\left(\int dy\, yR(y,x)\right)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[p(x)\left(\int dy\, y^2R(y,x)\right)\right], \quad (1.41)$$

after which we obtain

$$\boxed{\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[v(x)\; p(x,t)\right] + \frac{\partial^2}{\partial x^2}\left[D(x)p(x,t)\right].} \quad (1.42)$$

We have introduced

$$v(x) \equiv \int dy\, yR(y,x) = \frac{\langle\Delta(x)\rangle}{\Delta t}, \quad (1.43)$$

and

$$D(x) \equiv \frac{1}{2}\int dy\, y^2R(y,x) = \frac{1}{2}\frac{\langle\Delta(x)^2\rangle}{\Delta t}. \quad (1.44)$$

Equation (1.42) is a prototypical description of *drift* and *diffusion* which appears in many contexts. The *drift* term $v(x)$ expresses the rate (velocity) with which transitions change (on average) the position from x. Given the probabilistic nature of the process, there are variations in the rate of change of position captured by the position dependent *diffusion* coefficient $D(x)$. The drift–diffusion equation is known as the *forward Kolmogorov* equation in the context of populations. As a description of random walks it appeared earlier in physics literature as the *Fokker–Planck* equation.

1.4.2 Population dynamics

<u>*Mutation:*</u> In the context of population dynamics, the relevant variable is the allele frequency $x = n/2N$, such that in the continuum limit x is limited to the interval $[0,1]$. The rates in Eq. (1.26) change n by ± 1, and hence

$$v(x) = \frac{\langle\Delta n\rangle}{2N} = \frac{R_{n+1,n}\times(+1) + R_{n-1,n}\times(-1)}{2N} = \frac{1}{2N}\left[\mu_1(N-n) - \mu_2 n\right] = \mu_1(1-x) - \mu_2 x\,, \quad (1.45)$$

while

$$D(x) = \frac{\langle \Delta n^2 \rangle}{2(2N)^2} = \frac{R_{n+1,n} + R_{n-1,n}}{8N^2} = \frac{1}{8N^2}\left[\mu_1(N-n) + \mu_2 n\right] = \frac{\mu_1(1-x) + \mu_2 x}{8N}\,. \quad (1.46)$$

Reproduction: The process of binomial reproduction *in the absence of mutation and selection*, was introduced before and leads to Eq. (1.30) for the probability R_{mn} to obtain the random variable m, given an initial value of n. It is easy to deduce from standard properties of the binomial distribution that

$$\langle m \rangle = 2N \times \frac{n}{2N} = n\,, \quad \text{i.e} \quad \langle (m-n) \rangle = 0\,, \quad (1.47)$$

while

$$\left\langle m^2 \right\rangle_c = \left\langle (m-n)^2 \right\rangle = 2N \times \frac{n}{2N}\left(1 - \frac{n}{2N}\right)\,. \quad (1.48)$$

We can construct a continuum evolution equation by setting $x = n/N \in [0,1]$, and replacing $p(n, t+1) - p(n,t) \approx dp(x)/dt$, where t is measured in number of generations. Clearly, from Eq. (1.47), there is no drift

$$v(x) = \langle (m-n) \rangle = 0\,, \quad (1.49)$$

while the diffusion coefficient is given by

$$D_{\text{diploid}}(x) = \frac{1}{4N} x(1-x)\,. \quad (1.50)$$

(For Haploids we merely need to replace $2N$ with N in the above equations.)

Chemical analog & Selection: Through the reactions in Eq. (1.23), we introduced a chemical mixture that mimicks a mutating population. Consider a system where a reaction between molecules A and B can lead to two outcomes:[2]

$$\mathrm{A} + \mathrm{B} \rightharpoonup_c \mathrm{A} + \mathrm{A} \quad \text{or} \quad \mathrm{A} + \mathrm{B} \rightharpoonup^d \mathrm{B} + \mathrm{B}\,, \quad (1.51)$$

at rates c and d. In a "mean-field" approximation the number of A molecules changes as

$$\frac{dN_A}{dt} = (c-d) N_A N_B = (c-d) N_A (N - N_A)\,. \quad (1.52)$$

Equation (1.52) predicts steady states $N_A^* = 0$ for $c < d$, $N_A^* = N$ for $c > d$, while any composition is permitted for the symmetric case of $c = d$. As we shall demonstrate, fluctuations modify the latter conclusion.

As before, let us denote $N_A = n$, $N_B = N - N_A$, and follow the change in composition after a single reaction. The number of A species may change by ± 1 with rates

$$R_{n,n+1} = d(n+1)(N-n-1), \quad \text{and} \quad R_{n,n-1} = c(n-1)(N-n+1)\,, \quad (1.53)$$

[2]These reactions mimic an important element of the mating process which stochastically modifies the proportion of alleles in a fixed–size population: The offspring from mating a *heterozygote* (a diploid organism with different alleles A_1 and A_2) with a *homozygote* (say with two copies of allele A_1) may be either heterozygote (A_1A_2) or homozygote (A_1A_1).

where the product is over the number of possible pairs of A-B particles that can participate in the reaction. The diagonal terms are again obtained from the normalization condition in Eq. (1.14) resulting in the Master equation

$$\frac{dp(n,t)}{dt} = d(n+1)(N-n-1)p(n+1)+c(n-1)(N-n+1)p(n-1)-dn(N-n)p(n)-cn(N-n)p(n)\,, \tag{1.54}$$

for $0 < n < N$, and with boundary terms

$$\frac{dp(0,t)}{dt} = d(N-1)p(1), \quad \text{and} \quad \frac{dp(N,t)}{dt} = c(N-1)p(N-1)\,. \tag{1.55}$$

When the number N is large, it is reasonable to take the continuum limit and construct a Kolmogorov equation for the fraction $x = n/N \in [0,1]$. The rates in Eq. (1.53) change n by ± 1, and hence

$$\begin{aligned} v(x) &= \frac{\langle \Delta n \rangle}{N} = \frac{R_{n+1,n} - R_{n-1,n}}{N} = \frac{1}{N}\left[cn(N-n) - dn(N-n)\right] \\ &= N(c-d)x(1-x)\,, \end{aligned} \tag{1.56}$$

while

$$\begin{aligned} D(x) &= \frac{\langle \Delta n^2 \rangle}{2N^2} = \frac{R_{n+1,n} + R_{n-1,n}}{2N^2} = \frac{1}{2N^2}\left[cn(N-n) + dn(N-n)\right] \\ &= \frac{c+d}{2}x(1-x)\,. \end{aligned} \tag{1.57}$$

Comparison with Eqs.(1.49) and Eq. (1.50) indicates that the above reaction has the same behavior as binomial selection provided that $c = d = 1/(4N)$. Indeed the superficial difference in factor of N between the two cases is because in the latter we followed the reactions one at a time (at rate $c = d$), while in the former we computed the transition probabilities after a whole generation (N steps of reproduction and removal). The selection process characterized by Eq.(1.30) treats the two alleles as completely equivalent. Including, selection as in Eq. (1.34) leads to a form similar to Eq. (1.51) with $c \neq d$, related to the selection parameter s by

$$c = \frac{1}{4N}(1+s) \quad \text{and} \quad d = \frac{1}{4N}(1-s)\,. \tag{1.58}$$

In the following, we shall employ the nomenclature of population genetics, such that

$$v(x) = \frac{s}{2}x(1-x)\,, \quad \text{and} \quad D(x) = \frac{1}{4N}x(1-x)\,. \tag{1.59}$$

1.4.3 Steady states

While it is usually hard to solve the Kolmogorov equation as a function of time, it is relatively easy to find the steady state solution to which the population settles after a long time. Let us denote the steady-state probability distribution by $p^*(x)$, which by definition must satisfy

$$\frac{\partial p^*(x)}{\partial t} = 0. \tag{1.60}$$

Therefore, setting the right-hand side of Eq. (1.42) to zero, we get

$$-\frac{\partial}{\partial x}\left[v(x)p^*(x)\right] + \frac{\partial^2}{\partial x^2}\left[D(x)p^*(x)\right] = 0. \tag{1.61}$$

The most general solution admits steady states in which there is an overall current and the integral over x of the last equation leads to a constant flow in probability. It is not clear how such a circumstance may arise in the context of population genetics, and we shall therefore focus on circumstances where there is no probability current, such that

$$-v(x)p^*(x) + \frac{\partial}{\partial x}(D(x)p^*(x)) = 0. \tag{1.62}$$

We can easily rearrange this equation to

$$\frac{1}{D(x)p^*}\frac{\partial}{\partial x}(D(x)p^*(x)) = \frac{\partial}{\partial x}\ln\left(D(x)p^*(x)\right) = \frac{v(x)}{D(x)}\,. \tag{1.63}$$

This equation can be integrated to

$$\ln D(x)p^*(x) = \int^x dx' \frac{v(x')}{D(x')} + \text{constant}, \tag{1.64}$$

such that

$$p^*(x) \propto \frac{1}{D(x)} \exp\left[\int^x \frac{v(x')}{D(x')}\right], \tag{1.65}$$

with the proportionality constant set by boundary conditions.

Let us examine the case of the dynamics of a fixed population, including mutations, and reproduction with selection. Adding the contributions in Eqs. (1.45), (1.46) and (1.59), we have

$$v(x) = \frac{s}{2}x(1-x) + \mu_1(1-x) - \mu_2 x\,, \tag{1.66}$$

while

$$D(x) = \frac{1}{4N}x(1-x) + \frac{\mu_1(1-x)+\mu_2 x}{2N} \approx \frac{1}{4N}x(1-x)\,. \tag{1.67}$$

The last approximation of ignoring the contribution from mutations to diffusion is common to population genetics, and well justified since typically the mutation rates are much less than unity. It enables a closed form solution to the steady state, as

$$\begin{aligned}
\log D(x)p^*(x) &= \int^x dx' \frac{v(x')}{D(x')} \\
&= 4N\int^x dx' \left[\frac{\mu_1}{x'} - \frac{\mu_2}{1-x'} + \frac{s}{2}\right] \\
&= 4N\left[\mu_1 \ln x + \mu_2 \ln(1-x) + \frac{s}{2}x\right] + \text{constant},
\end{aligned}$$

resulting in

$$p^*(x) \propto \frac{1}{x(1-x)} \times x^{4N\mu_1} \times (1-x)^{4N\mu_2} \times e^{2Nsx} \,. \tag{1.68}$$

In the special case of no selection, $s = 0$ and (for convenience) $\mu_1 = \mu_2 = \mu$, the steady-state solution (1.68) simplifies to

$$p^*(x) \propto [x(1-x)]^{4N\mu-1} \,. \tag{1.69}$$

The shape of the solution is determined by the parameter $4N\mu$. If $4N\mu > 1$, then the distribution has a peak at $x = 1/2$ and diminishes to the sides. On the other hand, if the population is small and $4N\mu < 1$, then $p^*(x)$ has peaks at either extreme—a situation where *genetic drift* is dominant.