

# 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*..., 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 (“null”) 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  $\{p_\alpha\}$ , where  $\alpha = 1, 2, \dots, m$ . (This choice is sometimes referred to as IID, for *identical, independently distributed* random variables.) Since the probabilities must be normalized, we require

$$\sum_{\alpha=1}^m p_\alpha = 1. \quad (1.1)$$

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

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

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

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

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

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

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 to 1. The terms within the sum on the right-hand side are known the *multinomial probabilities*

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

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 outcome (the mode  $\{N_\alpha^*\}$  of the multi-variate distribution) coincides with its average (mean  $\{\langle N_\alpha \rangle\}$ ), and is given by

$$N_\alpha^* = \langle N_\alpha \rangle = p_\alpha N. \quad (1.6)$$

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,  $N! \approx N \ln N - N$ , 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*

$$\mathcal{S}[\{p_{\alpha}\}] = - \sum_{\alpha} p_{\alpha} \log p_{\alpha}. \quad (1.7)$$

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

Let us illustrate the relations among entropy and information in the context of DNA. To transmit a sequence,  $ACTG \dots$ , 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  $\langle N_A \rangle \neq \langle N_G \rangle \neq \dots$ . 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

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

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 generations. 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

ATT CGC ATG ,

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

UAA GCG UAC .

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

Leu Ala Tyr .

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”

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

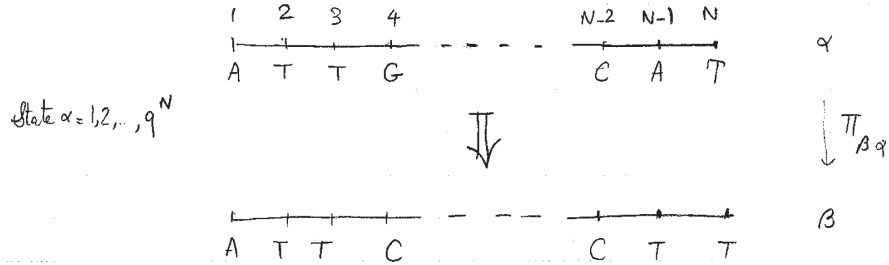
redundancy in the genetic code, there are also mutations that are *synonymous*, in that they do not change the amino acid which is added. Because these synonymous mutations do not affect the biological viability of the organism, we can find genes whose exact DNA varies from individual to individual. 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  $\alpha$  to  $\beta$  with a *transition probability*  $\pi_{\beta\alpha}$ . The  $m \times m$  such elements form the *transition probability matrix*  $\overleftrightarrow{\pi}$ . (Without the assumption that the sites evolve independently, we would have constructed a much larger  $(m^N \times m^N)$  matrix  $\overleftrightarrow{\Pi}$ . With the assumption of independence, this larger matrix is a direct product of transition matrices for individual sites, i.e.  $\overleftrightarrow{\Pi} = \overleftrightarrow{\pi}_1 \otimes \overleftrightarrow{\pi}_2 \otimes \cdots \otimes \overleftrightarrow{\pi}_N$ , with  $\overleftrightarrow{\pi}_i$  a  $m \times m$  matrix acting on site  $i$ .) With the transition probability matrix, we can track the evolution of the probabilities as

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

where the last identity is obtained by recursion, assuming that the transition probability matrix remains the same for all generations, i.e. does not change with time.



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

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

The last expression formalizes the statement that the 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

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

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

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

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

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

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),

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

We thus arrive at

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

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-eigenvector  $\overleftarrow{v}^* = (1, 1, \dots, 1)$  with eigenvalues of unity and zero respectively, i.e.

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

For each eigenvalue of a matrix 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

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

The elements of the vector  $\overrightarrow{p}^*$  represent the *steady state probabilities* for the process. These probabilities no longer change with time. In many cases, the Perron–Frobenius theorem ensures that the vector  $\overrightarrow{p}^*$  is unique, with positive elements as appropriate to probabilities.<sup>1</sup> Since the matrix  $\overleftrightarrow{\pi}$  is not symmetric, the remaining eigenvalues need not be real, but must occur in complex conjugate pairs. These remaining eigenvalues of the transition matrix have

---

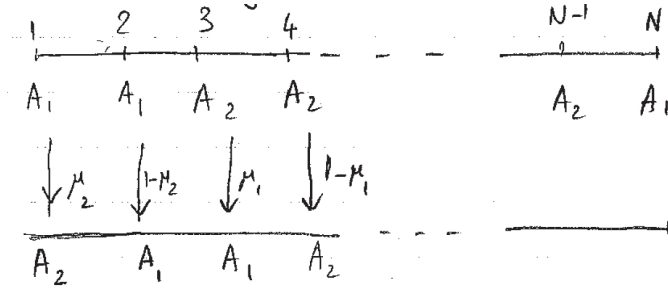
<sup>1</sup>A condition for the validity of the above is that it any state should be accessible (through a set of transitions) from any other states. Consequences of violation of this condition will be explored later in connection with genetic drift and absorbing states.

magnitude less than unity, and determine how an initial vector of probabilities approaches the steady state.

From Eq. (1.11) we observe that a particular steady state  $\vec{p}^*$  can be achieved through transition probabilities that satisfy the so-called condition of *detailed balance*,

$$\pi_{\alpha\beta}p_{\beta}^* = \pi_{\beta\alpha}p_{\alpha}^*. \quad (1.18)$$

This condition holds for systems in thermal equilibrium, where the steady-state probabilities are constrained to satisfy the Boltzmann distribution ( $p_{\alpha}^* \propto \exp(-\beta E_{\alpha})$ ). Equation (1.18) requires equal probability fluxes between any pair of states. More general steady states can be formed with the probability flux circulating along triplets and larger sets. Indeed, such circulation of flux can be used as indicator of non-equilibrium steady states.



As a simple example, consider a *binary* sequence (i.e.  $m = 2$ ) with independent states  $A_1$  or  $A_2$  at each site.<sup>2</sup> Let us assume that the state  $A_1$  can “mutate” to  $A_2$  at a rate  $\mu_2$ , while state  $A_2$  may change to  $A_1$  with a rate  $\mu_1$ . The probabilities  $p_1(t)$  and  $p_2(t)$  now evolve in time as

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

The above  $2 \times 2$  transition rate matrix has the following two eigenvectors

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

As anticipated, there is an eigenvector  $\vec{p}^*$  with eigenvalue of zero; the elements of this vector have been 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 demonstrate evolution of probabilities with time, 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

---

<sup>2</sup>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.

equation (1.19) is

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

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

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

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

#### 1.2.4 Mutating Population

The previous example of a binary sequence of length  $N$  can be recast and interpreted in terms of the evolution of a (haploid) 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  $n = 0, 1, \dots, N$  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  $\mu_2$ ,  $A_2$  to  $A_1$  at rate  $\mu_1$ , 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

$$R_{n,n+1} = \mu_2(n+1), \quad \text{and} \quad R_{n,n-1} = \mu_1(N-n+1), \quad (1.23)$$

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

$$\frac{dp(n, t)}{dt} = \mu_2(n+1)p(n+1) + \mu_1(N-n+1)p(n-1) - \mu_2np(n) - \mu_1(N-n)p(n), \quad (1.24)$$

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

$$\frac{dp(0, t)}{dt} = \mu_2p(1) - \mu_1Np(0), \quad \text{and} \quad \frac{dp(N, t)}{dt} = \mu_1p(N-1) - \mu_2Np(N). \quad (1.25)$$

The above equation looks complicated, but in fact admits a simple solution as justified in the different context described next.

### 1.2.5 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



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 so-called “mean-field” equation

$$\frac{dN_A}{dt} = -b'N_EN_A + a'N_EN_B \equiv -bN_A + a(N - N_A) \quad , \quad (1.27)$$

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 where  $N_A^*/N_B^* = a/b$ .

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. (1.24) 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 state solution also describes a chain of binary elements independently distributed with probabilities  $p_A^* = a/(a + b)$  and  $p_B^* = b/(a + b)$ . Hence, the steady state solution to the complicated looking set of equations (1.24) is simply

$$p^*(n) = \binom{N}{n} \frac{a^n b^{N-n}}{(a + b)^N} \quad . \quad (1.28)$$

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).