that it holds for j. Assuming that (4.10) is true for $j-1$, equation (4.6) gives

$$\frac{d}{dt}P_j(t) = \frac{\lambda e^{-\lambda t}(\lambda t)^{j-1}}{(j-1)!} - \lambda P_j(t).$$

From this,

$$e^{\lambda t}\left(\frac{d}{dt}P_j(t) + \lambda P_j(t)\right) = \frac{\lambda(\lambda t)^{j-1}}{(j-1)!}.$$

This equation may be rewritten as

$$\frac{d}{dt}\left(P_j(t)e^{\lambda t}\right) = \frac{\lambda(\lambda t)^{j-1}}{(j-1)!},$$

and integration of both sides of this equation gives

$$P_j(t)e^{\lambda t} = \frac{(\lambda t)^j}{j!} + C,$$

for some constant C. From (4.7) it follows that $C = 0$. Thus

$$P_j(t) = \frac{e^{-\lambda t}(\lambda t)^j}{j!}, \quad j = 0, 1, 2, \ldots. \tag{4.11}$$

This completes the induction, showing that at time t the random variable N has a Poisson distribution with parameter λt.

Conditions 1 and 2 are often taken as giving a mathematical definition of the concept of "randomness," and since many calculations in bioinformatics, some of which are described later in this book, make the randomness assumption, the Poisson distribution arises often.

4.2 The Poisson and the Binomial Distributions

An informal statement concerning the way in which the Poisson distribution arises as a limiting case of the binomial was made in Section 1.3.6. A more formally correct version of this statement is as follows. If in the binomial distribution (1.8) we let $n \to +\infty$, $p \to 0$, with the product np held constant at λ, then for any y, the binomial probability in (1.8) approaches the Poisson probability in (1.15). This may be proved by writing the binomial probability (1.8) as

$$\frac{1}{y!}(np)((n-1)p)\cdots((n-y+1)p)\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-y}. \tag{4.12}$$

Fix y and λ and write $p = \lambda/n$. Then as $n \to \infty$, each term in the above product has a finite limit as $n \to +\infty$: Terms of the form $(n-i)\lambda/n$ approach λ for any i, and

$$\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda},$$

(see (B.3)), and finally,

$$\left(1-\frac{\lambda}{n}\right)^{-y} \to 1.$$

Therefore, the expression in (4.12) approaches

$$\lambda^y e^{-\lambda}/y! \tag{4.13}$$

as $n \to +\infty$, and this is the Poisson probability (1.15).

4.3 The Poisson and the Gamma Distributions

There is an intimate connection, implied by equation (4.9), between the Poisson distribution and the exponential distribution. The (random) time until the first event occurs in a Poisson process with parameter λ is given by the exponential distribution with parameter λ. To see this, let $F(t)$ be the probability that the first event occurs before time t. Then the density function for the time until the first occurrence is the derivative $\frac{d}{dt}F(t)$. From (4.9), $F(t) = 1 - P_0(t) = 1 - e^{-\lambda t}$. Therefore, $\frac{d}{dt}F(t) = \lambda e^{-\lambda t}$. This is the exponential distribution (1.59), with notation changed from x to t.

It can also be shown that the distribution of the time between successive events is given by the exponential distribution. Thus the (random) time until the kth event occurs is the sum of k independent exponentially distributed times. The material surrounding (2.25) shows that this sum has the gamma distribution (1.68). Let t_0 be some fixed value of t. Then if the time until the kth event occurs exceeds t_0, the number of events occurring before time t_0 is less than k, and conversely. This means that the probability that $k-1$ or fewer events occur before time t_0 must be identical to the probability that the time until the kth event occurs exceeds t_0. In other words it must be true that

$$e^{-\lambda t_0}\left(1 + (\lambda t_0) + \frac{(\lambda t_0)^2}{2!} + \cdots + \frac{(\lambda t_0)^{k-1}}{(k-1)!}\right) = \frac{\lambda^k}{\Gamma(k)}\int_{t_0}^{+\infty} x^{k-1}\, e^{-\lambda x} dx. \tag{4.14}$$

This equation can also be established by repeated integration by parts of the right-hand side.

4.4 Introduction to Finite Markov Chains

In this section we give a brief outline of the theory of discrete-time finite Markov chains. The focus is on material needed to discuss the construction

of PAM matrices as described in Section 6.5.3. Further developments of Markov chain theory suitable for other applications, in particular for the evolutionary applications discussed in Chapter 13, are given in Chapter 10.

We introduce discrete-time finite Markov chains in abstract terms as follows. Consider some finite discrete set S of possible "states," labeled $\{E_1, E_2, \ldots, E_s\}$. At each of the unit time points $t = 1, 2, 3, \ldots$, a Markov chain process occupies one of these states. In each time step t to $t+1$, the process either stays in the same state or moves to some other state in S. Further, it does this in a probabilistic, or stochastic, way rather than in a deterministic way. That is, if at time t the process is in state E_j, then at time $t+1$ it either stays in this state or moves to some other state E_k according to some well-defined probabilistic rule described in more detail below. The process is called Markovian, and follows the requirements of a Markov chain if it has the following distinguishing Markov characteristics.

(i) The *memoryless* property. If at some time t the process is in state E_j, the probability that one time unit later it is in state E_k depends only on E_j, and not on the past history of the states it was in before time t. That is, the current state is all that matters in determining the probabilities for the states that the process will occupy in the future.

(ii) The *time homogeneity* property. Given that at time t the process is in state E_j, the probability that one time unit later it is in state E_k is independent of t.

More general Markov processes relax one or both properties.

The concept of "time" used above is appropriate if, for example, we consider the evolution through time of the nucleotide at a given site in some population. Aspects of this process are discussed later in this book. However, the concept of time is sometimes replaced by that of "space." As an example, we may consider a DNA sequence read from left to right. Here there would be a Markov dependence between nucleotides if the nucleotide type at some site depended in some way on the type at the site immediately to its left. Aspects of the Markov chains describing this process are also discussed later in this book. Because Markov chains are widely applicable to many different situations it is useful to describe the properties of these chains in abstract terms.

In many cases the Markov chain process describes the behavior of a random variable changing through time. For example, in reading a DNA sequence from left to right this random variable might be the excess of purines over pyrimidines so far observed at any point. Because of this it is often convenient to adopt a different terminology and to say that the value of the random variable is j rather than saying that the state occupied by the process is E_j. We use both forms of expression below, and also, when

no confusion should arise, we abuse terminology by using expressions like "the random variable is in state E_j."

4.5 Transition Probabilities and the Transition Probability Matrix

Suppose that at time t a Markovian random variable is in state E_j. We denote the probability that at time $t+1$ it is in state E_k by p_{jk}, called the *transition probability* from E_j to E_k. In writing this probability in this form we are already using the two Markov assumptions described above: First, no mention is made in the notation p_{jk} of the states that the random variable was in before time t (the memoryless property), and second, t does not occur in the notation p_{jk} (the time homogeneity property).

It is convenient to group the transition probabilities p_{jk} into the so-called *transition probability matrix*, or more simply the transition matrix, of the Markov chain. We denote this matrix by P, and write it as

$$P = \begin{array}{c} \\ (\text{from } E_1) \\ (\text{from } E_2) \\ \vdots \\ (\text{from } E_s) \end{array} \begin{array}{c} \begin{array}{cccccc} (\text{to } E_1) & (\text{to } E_2) & (\text{to } E_3) & \cdots & & (\text{to } E_s) \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1s} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{s1} & p_{s2} & p_{s3} & \cdots & p_{ss} \end{bmatrix} \end{array}. \tag{4.15}$$

The rows and columns of P are in correspondence with the states E_1, $E_2, \ldots, E_s$, so these states being understood, P is usually written in the simpler form

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1s} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{s1} & p_{s2} & p_{s3} & \cdots & p_{ss} \end{bmatrix}. \tag{4.16}$$

Any row in the matrix corresponds to the state *from* which the transition is made, and any column in the matrix corresponds to the state *to* which the transition is made. Thus the probabilities in any particular row in the transition matrix must sum to 1. However, the probabilities in any given column do not have to sum to anything in particular.

It is also assumed that there is some *initial* probability distribution for the various states in the Markov chain. That is, it is assumed that there is some probability π_i that at the initial time point the Markovian random variable is in state E_i. A particular case of such an initial distribution arises when it is known that the random variable starts in state E_i: In this case $\pi_i = 1$, while $\pi_j = 0$ for $j \neq i$. In principle the initial probability

distribution and the transition matrix P jointly determine the probability for any event of interest in the entire process.

Example. Random walks. Random walk theory underpins BLAST theory. The classic description of a random walk involves a gambler, with initial capital i dollars, $i > 0$, and his adversary, with initial capital $(s-i)$ dollars, $s-i > 0$. At unit time points a coin with probability p for heads is tossed, and if the coin lands heads up, the adversary gives the gambler 1 dollar, while if the coin lands tails up, the gambler gives his adversary 1 dollar. The game continues until either the gambler or his adversary has no money left. The random variable of interest is the current fortune of the gambler. This random variable undergoes what is called a simple random walk, taking at any time one of the values $0, 1, 2, \ldots, s$.

If at any time this random variable takes the value j $(1 \leq j \leq s-1)$, it moves one time unit later to the value $j+1$ with probability p or to the value $j-1$ with probability $q = 1-p$. This continues until the random variable reaches either 0 or s. When one of these values is reached no further changes in the gambler's fortune occur: His fortune is stopped at 0 or s.

The above implies that the evolution of the random variable is described by a Markov chain. The transition matrix P of this chain is

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & q & 0 & p & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4.17}$$

The appearance of the 1's in the top left and bottom right entries reflects the fact that the gambler's fortune remains unchanged when it reaches either 0 or s: The probability of a transition from 0 to 0 or from s to s is 1.

The probability that the Markov chain process moves from state E_i to state E_j after two steps can be found by matrix multiplication. It is this fact that makes much of Markov chain theory an application of linear algebra. The argument is as follows.

Let $p_{ij}^{(2)}$ be the probability that if the Markovian random variable is in state E_i at time t, then it is in state E_j at time $t+2$. We call this a *two-step* transition probability. Since the random variable must be in some state k at the intermediate time $t+1$, equation (1.83) gives

$$p_{ij}^{(2)} = \sum_k p_{ik} p_{kj}.$$

The right-hand side in this equation is the (i,j) element in the matrix P^2. Thus if we define the matrix $P^{(2)}$ as the matrix whose (i,j) element is $p_{ij}^{(2)}$, then the (i,j) element in $P^{(2)}$ is equal to the (i,j) element in P^2. From this we get the fundamental identity

$$P^{(2)} = P^2.$$

Extension of this argument to an arbitrary number n of steps gives

$$P^{(n)} = P^n. \tag{4.18}$$

That is, the "n-step" transition probabilities are given by the entries in the nth power of P.

4.6 Markov Chains with Absorbing States

The random walk example described by (4.17) is a case of a Markov chain with absorbing states. These can be recognized by the appearance of one or more 1's on the main diagonal of the transition matrix. If there are no 1's on the main diagonal, then there are no absorbing states. For the Markov chains with absorbing states that we consider, sooner or later some absorbing state will be entered, never thereafter to be left. The two most questions we are interested in for these Markov chains are:

(i) If there are two or more absorbing states, what is the probability that a specified absorbing state is the one eventually entered?

(ii) What is the mean time until one or another absorbing state is eventually entered?

We will address these questions in detail in Chapter 10. In the remainder of this chapter we discuss only certain aspects of the theory of Markov chains with no absorbing states, focusing on the theory needed for the construction of substitution matrices, to be discussed in more detail in Chapter 6.

4.7 Markov Chains with No Absorbing States

The questions of interest about a Markov chain with no absorbing state are quite different from those asked when there are absorbing states.

In order to simplify the discussion, we assume in the remainder of this chapter that all Markov chains discussed are *finite*, *aperiodic*, and *irreducible*.

Finiteness means that there is a finite number of possible states. The aperiodicity assumption is that there is no state such that a return to that state is possible only at t_0, $2t_0$, $3t_0, \ldots$ transitions later, where t_0 is an integer exceeding 1. If the transition matrix of a Markov chain with states E_1, E_2, E_3, E_4 is, for example,

$$P = \begin{bmatrix} 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.3 & 0.7 \\ 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \end{bmatrix}, \tag{4.19}$$

then the Markov chain is periodic. If the Markovian random variable starts (at time 0) in E_1, then at time 1 it must be either in E_3 or E_4, at time 2 it must be in either E_1 or E_2, and in general it can visit only E_1 at times $2, 4, 6, \ldots$. It is therefore periodic. The aperiodicity assumption holds for essentially all applications of Markov chains in bioinformatics, and we often take aperiodicity for granted without any explicit statement being made.

The irreducibility assumption implies that any state can eventually be reached from any other state, if not in one step then after several steps. Except for the case of Markov chains with absorbing states, the irreducibility assumption also holds for essentially all applications in bioinformatics.

4.7.1 Stationary Distributions

Suppose that a Markov chain has transition matrix P and that at time t the probability that the process is in state E_j is φ_j, $j = 1, 2, \ldots, s$. This implies that the probability that at time $t+1$ the process is in state j is $\sum_{k=1}^{s} \varphi_k p_{kj}$. Suppose that for every j these two probabilities are equal, so that

$$\varphi_j = \sum_{k=1}^{s} \varphi_k p_{kj}, \quad j = 1, 2, \ldots, s. \tag{4.20}$$

In this case we say that the probability distribution $(\varphi_1, \varphi_2, \ldots, \varphi_s)$ is *stationary*; that is, it has not changed between times t and $t+1$, and therefore will never change. It will be shown in Chapter 10 that for finite aperiodic irreducible Markov chains there is a unique such stationary distribution.

If the row vector $\boldsymbol{\varphi}'$ is defined by

$$\boldsymbol{\varphi}' = (\varphi_1, \varphi_2, \ldots, \varphi_s), \tag{4.21}$$

then in matrix and vector notation, the set of equations in (4.20) becomes

$$\boldsymbol{\varphi}' = \boldsymbol{\varphi}' P. \tag{4.22}$$

The prime here is used to indicate the transposition of the row vector into a column vector. The vector $(\varphi_1, \varphi_2, \ldots, \varphi_s)$ must also satisfy the equation $\sum_k \varphi_k = 1$. In vector notation, this is the equation

$$\boldsymbol{\varphi}' \mathbf{1} = 1, \tag{4.23}$$

where $\mathbf{1} = (1, 1, \ldots, 1)'$.

Equations (4.22) and (4.23) can be used to find the stationary distribution. An example is given in the next section.

In Chapter 10 we will show that if the Markov chain is finite, aperiodic, and irreducible, then as n increases, $P^{(n)}$ approaches the matrix

$$\begin{bmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_s \\ \varphi_1 & \varphi_2 & \cdots & \varphi_s \\ \varphi_1 & \varphi_2 & \cdots & \varphi_s \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1 & \varphi_2 & \cdots & \varphi_s \end{bmatrix}, \tag{4.24}$$

where $(\varphi_1, \varphi_2, \ldots, \varphi_s)$ is the stationary distribution of the Markov chain.

The form of this matrix shows that no matter what the starting state was, or what was the initial probability distribution of the starting state, the probability that n time units later the process is in state j is increasingly closely approximated, as $n \to \infty$, by the value φ_j.

There is another implication, relating to long-term averages, of the calculations above. That is, if a Markov chain is observed for a very long time, then the proportion of times that it is observed to be in state E_j is approximately φ_j, for all j.

4.7.2 Example

Consider the Markov chain with transition probability matrix given by

$$P = \begin{bmatrix} 0.6 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.5 & 0.1 \\ 0.1 & 0.3 & 0.1 & 0.5 \end{bmatrix}. \tag{4.25}$$

For this example the vector equation (4.22) consists of four separate linear equations in four unknowns. However, they form a redundant set of equations and any one of them can be discarded. The remaining three equations, together with (4.23), yield four linear equations in the four unknowns which can be solved for a unique solution. Discarding the last equation in (4.22), we get

$$\begin{aligned} 0.6\varphi_1 + 0.1\varphi_2 + 0.2\varphi_3 + 0.1\varphi_4 &= \varphi_1, \\ 0.1\varphi_1 + 0.7\varphi_2 + 0.2\varphi_3 + 0.3\varphi_4 &= \varphi_2, \\ 0.2\varphi_1 + 0.1\varphi_2 + 0.5\varphi_3 + 0.1\varphi_4 &= \varphi_3, \\ \varphi_1 + \varphi_2 + \varphi_3 + \varphi_4 &= 1. \end{aligned}$$

To four decimal place accuracy, these four simultaneous equations have the solution

$$\boldsymbol{\varphi}' = (0.2414, 0.3851, 0.2069, 0.1667). \tag{4.26}$$

This is the stationary distribution corresponding to the matrix P given in (4.25). In informal terms, from the point of view of long-term averages, over a long time period the random variable should spend about 24.14% of the time in state E_1, about 38.51% of the time in state E_2, and so on.

The rate at which the rows in $P^{(n)}$ approach this stationary distribution can be assessed from the following values:

$$P^{(2)} = \begin{bmatrix} 0.42 & 0.20 & 0.24 & 0.14 \\ 0.16 & 0.55 & 0.15 & 0.14 \\ 0.25 & 0.29 & 0.32 & 0.14 \\ 0.16 & 0.39 & 0.15 & 0.30 \end{bmatrix}, \tag{4.27}$$

$$P^{(4)} \approx \begin{bmatrix} 0.2908 & 0.3182 & 0.2286 & 0.1624 \\ 0.2151 & 0.4326 & 0.1899 & 0.1624 \\ 0.2538 & 0.3569 & 0.2269 & 0.1624 \\ 0.2151 & 0.4070 & 0.1899 & 0.1880 \end{bmatrix}, \tag{4.28}$$

$$P^{(8)} \approx \begin{bmatrix} 0.24596 & 0.37787 & 0.20961 & 0.16656 \\ 0.23873 & 0.38946 & 0.20525 & 0.16656 \\ 0.24309 & 0.38223 & 0.20812 & 0.16656 \\ 0.23873 & 0.38880 & 0.20525 & 0.16721 \end{bmatrix}, \tag{4.29}$$

$$P^{(16)} \approx \begin{bmatrix} 0.24142 & 0.38494 & 0.20692 & 0.16667 \\ 0.24135 & 0.38510 & 0.20688 & 0.16667 \\ 0.24140 & 0.38503 & 0.20691 & 0.16667 \\ 0.24135 & 0.38510 & 0.20688 & 0.16667 \end{bmatrix}. \tag{4.30}$$

After 16 time units, the stationary distribution has, for most purposes, been reached. The discussion in Chapter 10 shows how the rate at which this convergence occurs can be calculated in a more informative manner.

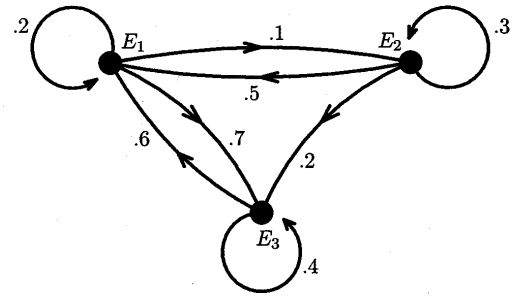
4.8 The Graphical Representation of a Markov Chain

It is often convenient to represent a Markov chain by a directed graph. A directed graph is a set of "nodes" and a set of "edges" connecting these nodes. The edges are "directed," that is, they are marked with arrows giving each edge an orientation from one node to another.

We represent a Markov chain by identifying the states with nodes and the transition probabilities with edges. Consider, for example, the Markov chain with states E_1, E_2, and E_3 and with probability transition matrix

$$\begin{bmatrix} .2 & .1 & .7 \\ .5 & .3 & .2 \\ .6 & 0 & .4 \end{bmatrix}.$$

This Markov chain is represented by the following graph:



Notice that we do not draw the edge if its corresponding transition probability is known to be zero, as is the case in this example with the transition from E_3 to E_2.

A graph helps us capture information at a glance that might not be so apparent from the transition matrix itself. Sometimes it is also convenient to include a *start state*; this is a dummy state that is visited only once, at the beginning. Therefore, all transition probabilities into the start state are zero. The transition probabilities out of the start state are given by the initial distribution of the Markov chain. If the Markov chain starts at time $t = 0$, we can think of the start state as being visited in time $t = -1$. We can further have an *end state*, which stops the Markov chain when visited.

We refer to the graph structure, without any probabilities, as the *topology* of the graph. Sometimes the topology of a model is known, but the various probabilities are unknown.

We will use these definitions when we discuss hidden Markov models in Chapter 11.

4.9 Modeling

There are many applications of the homogeneous Poisson process in bioinformatics. However, the two key assumptions made in the derivation of the Poisson distribution formula (4.10), namely homogeneity and independence, do not always hold in practice. Similarly, there are many applications of Markov chains in the literature, in particular in the evolutionary processes discussed in Chapter 13. Many of these applications also make assumptions, specifically the two Markov assumptions stated in Section 4.4. The modeling assumptions made in the evolutionary context are discussed further in Section 14.9.