obtained is then used instead of the conservative "unitary" matrix to construct a second, less conservative, set of blocks. A new substitution matrix is then obtained from these blocks. Then the process is repeated a third time. Henikoff and Henikoff derive the final family of BLOSUM matrices from this third set of blocks, and it is these whose use is suggested.

If the .85 similarity score criterion is adopted, the final matrix is called a BLOSUM85 matrix. In general if clusters with $X\%$ identity are used, then the resulting matrix is called BLOSUMX. The BLOSUM matrices currently available on the BLAST web page at NCBI (`www.ncbi.nlm.nih.gov/BLAST/`) are BLOSUM45, BLOSUM62, and BLOSUM80. Note that the larger-numbered matrices correspond to more recent divergence, and the smaller-numbered matrices correspond to more distantly related sequences.

One often has prior knowledge about the evolutionary distance between the sequences of interest that helps one choose which BLOSUM matrix to use. With no information, BLOSUM62 is often used. We explore the implications of the choice of various matrices in Section 10.2.4.

A central feature of the BLOSUM substitution matrix calculation is the use of (estimated) likelihood ratios. We see in the next section that the same is true of PAM matrices. In Section 9.2.1 it is shown that use of likelihood ratios has a statistical optimality property, and this optimality property explains in part their use in the construction of both BLOSUM and PAM matrices.

6.5.3 PAM Substitution Matrices

In this section we outline the Dayhoff et al. (1978) approach to deriving the so-called PAM substitution matrices. Two essential ingredients in the construction of these matrices, as with construction of BLOSUM matrices, are the calculation of an (estimated) likelihood ratio and the use of Markov chain theory as introduced in Section 4.8. We now describe this construction in more detail.

An "accepted point mutation" is a substitution of one amino acid of a protein by another that is "accepted" by evolution, in the sense that within some given species, the mutation has not only arisen but has, over time, spread to essentially the entire species. A PAM1 transition matrix is the Markov chain matrix applying for a time period over which we expect 1% of the amino acids to undergo accepted point mutations within the species of interest.

The construction of PAM matrices starts with ungapped multiple alignments of proteins into blocks for which all pairs of sequences in any block are, as in the BLOSUM procedure, "sufficiently close" to each other. In the original construction of Dayhoff et al. (1978), the requirement was that each sequence in any block be no more than 15% different from any other sequence. This requirement resulted, for their data, in 71 blocks of aligned

proteins. Imposing the requirement of close within-block similarity is aimed at minimizing the number of substitutions in the data that may have resulted from two or more consecutive substitutions at the same site. This is important because the initial goal is to create a Markov transition matrix for a short enough time period so that multiple substitutions are very unlikely to happen during this time period. We discuss later how to handle the multiple substitutions that are expected to arise over longer time periods.

The BLOSUM approach uses clustering to achieve two aims. One is to minimize biases in the databases that the sequences were taken from, since without clustering some closely related sequences may be overrepresented. The second is to account for evolutionary deviation of varying time periods. In the PAM approach, the first aim is approached by inferring a separate phylogenetic tree for the data in each aligned block of sequences, eventually using all the inferred trees in an aggregated manner to estimate a Markov chain transition matrix. The second aim is achieved by using Markov chain theory applied to this matrix.

The phylogenetic reconstruction method adopted for the data within any block in the database is the method of maximum parsimony, described in Chapter 14. This algorithm constructs trees with our sequences at the leaves, and with inferred sequences at the internal nodes, such that the total number of substitutions across the tree is minimal. Such a tree is called a *most parsimonious* tree. There are often several most parsimonious trees for any block, in which case all such trees are used and an averaging procedure is employed from the data in each tree, as described below. From now on, "tree" means one or other of the set of most parsimonious trees for one of the blocks.

For any column in any block, the data in each tree are used to obtain counts in the following manner. Suppose that two different but aligned amino acids A and B occur (in any order) in two nodes of a tree joined by a single edge. Then this edge contributes 1 to the "A–B" count. If the same amino acid A occurs aligned in two nodes of the tree joined by a single edge, then this edge contributes 2 to the "A–A" count. The counts for all "A–A" and all "A–B" amino acid pairs are then totaled over all edges of all trees in each block. If the block has n most parsimonious trees, these total counts are then divided by n. The sum of the resultant counts over all blocks is then calculated.

The following simple example demonstrates the calculations within any one block. Suppose that a given block of three sequences is

$$\begin{array}{l} AA \\ AB \\ BB \end{array} .$$

There are $n = 5$ most parsimonious trees leading to these three sequences at the leaves of the trees, as shown in Figure 6.3. Among these trees A is aligned with, and substituted for, B (or conversely) twice in each tree,

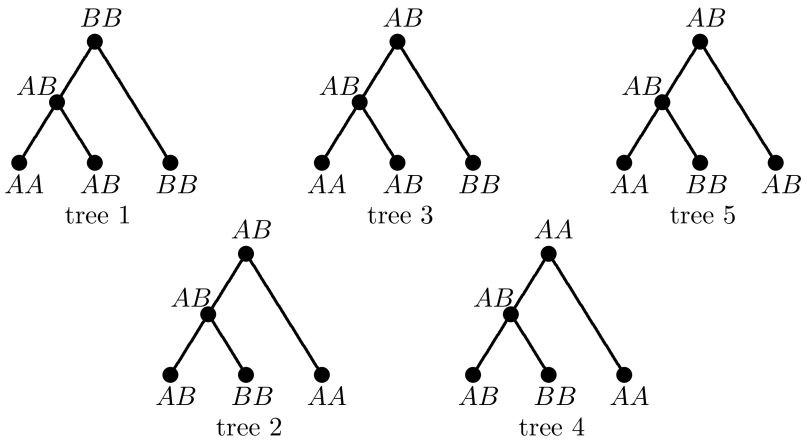

Figure 6.3.

leading to a total "A–B" count of 10. Division by the number of trees (5) in the block leads to a final contribution of 2 from this block to the A–B count.

Next, A is aligned with A two times in tree 1, three times in tree 2, three times in tree 3, four times in tree 4, and three times in tree 5, leading to a total of 15 A–A alignments. Each A–A alignment leads to a count of 2, so that the total A–A count is 30. Division by the number of trees for this block leads to a final block contribution of 6 to the overall A–A. Similar calculations show that that the contribution to the B–B count from this block is also 6.

If this were the only block in the data set, the final matrix of counts would then be

	A	B
A	6	
B	2	6

In general, there will be more than one block in the data set, and if so, as indicated above, we simply add the counts from the different blocks to obtain an overall count matrix.

Suppose that the amino acids are numbered from 1 to 20 and denote the (j,k)th entry in the overall count matrix by A_{jk}. The next task is to use this count matrix to construct an estimated Markov chain transition matrix. For any j and k (not necessarily distinct), define a_{jk} by

$$a_{jk} = \frac{A_{jk}}{\sum_m A_{jm}}. \tag{6.19}$$

For $j \neq k$, let

$$p_{jk} = ca_{jk}, \tag{6.20}$$

where c is a positive scaling constant (to be determined later), and let

$$p_{jj} = 1 - \sum_{k \neq j} c a_{jk}. \tag{6.21}$$

It follows from these definitions that $\sum_k p_{jk} = 1$. If c is chosen to be sufficiently small so that each p_{jj} is non-negative, the matrix $P = \{p_{jk}\}$ then has the properties of a Markov chain transition matrix. In this matrix smaller values of c imply larger diagonal entries relative to the nondiagonal entries; however, the relative sizes of the nondiagonal entries are independent of the choice of c. In practice it will always be the case that this matrix is irreducible and aperiodic, so that it has a well-defined stationary distribution.

Although the matrix P is derived from data, and thus any probability derived from it is an estimate, we assume that the data leading to P are sufficiently extensive so that no serious error is made by thinking of P not as an estimated transition matrix but as an actual transition matrix. We thus drop the word "estimated" below in discussing the probabilities deriving from this matrix.

The value of c is now chosen so that, after one step of the Markov chain defined by the transition matrix P, the weighted expected proportion of amino acid changes is 0.01, the weights being the various amino acid frequencies. A reasonable estimate for this set of frequencies is the observed distribution found from the data in the original blocks of aligned proteins. Let p_j be the observed such frequency for the jth amino acid. Then the expected proportion of amino acids that change after one step of the Markov chain defined by the transition matrix P is

$$\sum_j p_j \sum_{k \neq j} p_{jk} = c \sum_j \sum_{k \neq j} p_j a_{jk}. \tag{6.22}$$

This implies that if c is defined by the equation

$$c = \frac{.01}{\sum_j \sum_{k \neq j} p_j a_{jk}}, \tag{6.23}$$

the "expected proportion" requirement is satisfied, and the resulting transition matrix then corresponds to an evolutionary distance of 1 PAM. This matrix is often denoted by M_1, with typical element m_{jk}, and we follow this notation here. The matrix corresponding to an evolutionary distance of n PAMs is obtained by raising M_1 to the nth power, in line with the n-step transition probability formula in equation (4.23). This matrix is denoted here by M_n and is called the PAMn matrix. As n gets larger, the matrix M_n gets closer and closer to a matrix all of whose rows are identical to the stationary distribution corresponding to the matrix M_1 (see equation (4.29)). In practice, the element common to all positions in the jth column of this matrix is often close to the background frequency p_j.

The stationary distribution of the matrix M is independent of the choice of the scaling constant c (see Problem 6.7).

The next task is to construct a substitution matrix derived from the probability transition matrix M_n. Let $m_{jk}^{(n)}$ be the (j,k) entry in the matrix M_n, for some extrinsically chosen value of n. Then $m_{jk}^{(n)}$ is the probability, after n steps of the chain defined by the matrix M_1, that the kth amino acid occurs in some specified position, given that initially the jth amino acid occurred in that position. For reasons that will be developed in Section 10.2.4, the typical entry in a PAM substitution matrix is of the form

$$C \cdot \log\left(\frac{m_{jk}^{(n)}}{p_k}\right), \tag{6.24}$$

where C is a positive constant. The choice of C is not crucial; nevertheless, this also is discussed in Section 10.2.4.

A variant of the expression (6.24) is the following. Denote the joint probability that amino acid j occurs at some nominated position at time 0 and that amino acid k occurs at this position after n steps of the Markov chain whose one-step transition matrix is M_1 by $q(j,k)$. (Note that $q(j,k)$ is a function of n, but in accordance with common practice we suppress this dependence in the notation.) Then

$$q(j,k) = p_j m_{jk}^{(n)}, \tag{6.25}$$

and (6.24) can be written as

$$C \cdot \log\left(\frac{q(j,k)}{p_j p_k}\right). \tag{6.26}$$

The choice of the value n has so far not been discussed. This matter will be taken up in Section 10.6, where the effects of an incorrect choice will be evaluated.

The BLOSUM and PAM procedures differ in one interesting respect: the larger n is for a PAM matrix, the longer is the evolutionary distance, whereas for BLOSUM matrices *smaller* values of n correspond to longer evolutionary distance.

6.5.4 A Simple Symmetric Evolutionary Matrix

In order to elucidate some properties of PAM matrices and to assess the implications of the choice of n in these matrices, it is useful to discuss a simple symmetric example, which, while it does not correspond to any PAM matrix used in practice, has properties that are found easily and that illuminate properties of PAM matrices. The model we discuss is the discrete-time analogue of the simple (and unrealistic) model considered by Bishop and Friday (1985).

Suppose that all amino acids are equally frequent (so that $p_j = 0.05$), that all are equally likely to be substituted by some other amino acid in any given time, and that all substitutions are equally likely. Then the matrix M_1 is such that its elements $\{m_{jk}\}$ are given by

$$m_{jj} = 0.99, \quad m_{jk} = 0.01/19, \quad j \neq k. \tag{6.27}$$

The value 0.99 for m_{jj} derives from the fact that we wish to mimic a PAM1 matrix, that is, a matrix for which the probability of an amino acid change in unit time is 0.01. For this simple symmetric matrix it can be shown from the spectral theory of Appendix B.19 that

$$m_{jj}^{(n)} = 0.05 + 0.95(94/95)^n, \tag{6.28}$$

$$m_{jk}^{(n)} = 0.05 - 0.05(94/95)^n, \quad j \neq k. \tag{6.29}$$

These calculations, together with equation (6.24), imply that the typical diagonal entry and the typical off-diagonal entry in the substitution matrix are, respectively,

$$C \cdot \log(1 + 19(94/95)^n), \quad C \cdot \log(1 - (94/95)^n), \tag{6.30}$$

for some positive value of C. The ratio of these is independent of C, being

$$\frac{\log(1 + 19(94/95)^n)}{\log(1 - (94/95)^n)}. \tag{6.31}$$

This leads to a substitution matrix whose diagonal elements are all

$$-\frac{\log(1 + 19(94/95)^n)}{\log(1 - (94/95)^n)} \tag{6.32}$$

and whose off-diagonal elements are all -1. When $n = 259$ (more precisely, $n = 259.0675$), the expression in (6.32) is very close to 12, corresponding to a substitution matrix whose entries are

$$S(j, j) = 12, \quad S(j, k) = -1, (j \neq k). \tag{6.33}$$

For the case $n = 259$,

$$m_{jj}^{(259)} = 0.111251, \quad m_{jk}^{(259)} = 0.046776 \quad j \neq k. \tag{6.34}$$

The definition (6.25) of $q(j, k)$ implies that for this case

$$q(j, j) = 0.0055625, \quad q(j, k) = 0.0023388, \quad j \neq k. \tag{6.35}$$

With these values the probability $20q(j, j)$ of a match at any position is 0.111251, and the probability of a mismatch is 0.888749. The mean score in the substitution matrix is then

$$12(0.111251) - 0.888749 = 0.446. \tag{6.36}$$

We discuss this calculation further in Section 10.2.4, deriving the value 0.446 there by what appears initially to be a method different from that used here.