	A	*R*	*N*	*D*	*C*	*Q*	*E*	*G*	*H*	*I*	*L*	*K*	*M*	*F*	*P*	*S*	*T*	*W*	*Y*	*V*
A	4	−1	−2	−2	0	−1	−1	0	−2	−1	−1	−1	−1	−2	−1	1	0	−3	−2	0
R	−1	5	0	−2	−3	1	0	−2	0	−3	−2	2	−1	−3	−2	−1	−1	−3	−2	−3
N	−2	0	6	1	−3	0	0	0	1	−3	−3	0	−2	−3	−2	1	0	−4	−2	−3
D	−2	−2	1	6	−3	0	2	−1	−1	−3	−4	−1	−3	−3	−1	0	−1	−4	−3	−3
C	0	−3	−3	−3	9	−3	−4	−3	−3	−1	−1	−3	−1	−2	−3	−1	−1	−2	−2	−1
Q	−1	1	0	0	−3	5	2	−2	0	−3	−2	1	0	−3	−1	0	−1	−2	−1	−2
E	−1	0	0	2	−4	2	5	−2	0	−3	−3	1	−2	−3	−1	0	−1	−3	−2	−2
G	0	−2	0	−1	−3	−2	−2	6	−2	−4	−4	−2	−3	−3	−2	0	−2	−2	−3	−3
H	−2	0	1	−1	−3	0	0	−2	8	−3	−3	1	−2	−1	−2	−1	−2	−2	2	−3
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4	2	−3	1	0	−3	−2	−1	−3	−1	3
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4	−2	2	0	−3	−2	−1	−2	−1	1
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5	−1	−3	−1	0	−1	−3	−2	−2
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5	0	−2	−1	−1	−1	−1	1
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6	−4	−2	−2	1	3	−1
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7	−1	−1	−4	−3	−2
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4	1	−3	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5	−2	−2	0
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11	2	−3
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−3	−2	2	7	−1
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Table 6.7. The BLOSUM62 substitution matrix

et al. (1998) or Section 9.7 in Waterman (1995).

6.5 Protein Sequences and Substitution Matrices

6.5.1 Introduction

In the study of DNA sequences, simple scoring schemes are usually effective. For protein sequences, however, some substitutions are much more likely than others. The performance of any alignment algorithm is improved when it accounts for this difference. In all cases we consider, higher scores will represent more likely substitutions.

There are two frequently used approaches to finding substitution matrices. One leads to the PAM (Accepted Point Mutation) family of matrices, and the other to the BLOSUM (BLOcks SUbstitution Matrices) family. Table 6.7 gives an example of a typical BLOSUM substitution matrix (called the BLOSUM62 matrix). In this section we discuss how these substitution matrices are derived.

WWYIR	CASILRKIYIYGPV	GVSRLRTAYGGRK	NRG
WFYVR	CASILRHLYHRSPA	GVGSITKIYGGRK	RNG
WYYVR	AAAVARHIYLRKTV	GVGRLRKVHGSTK	NRG
WYFIR	AASICRHLYIRSPA	GIGSFEKIYGGRR	RRG
WYYTR	AASIARKIYLRQGI	GVGGFQKIYGGRQ	RNG
WFYKR	AASVARHIYMRKQV	GVGKLNKLYGGAK	SRG
WFYKR	AASVARHIYMRKQV	GVGKLNKLYGGAK	SRG
WYYVR	TASIARRLYVRSPT	GVDALRLVYGGSK	RRG
WYYVR	TASVARRLYIRSPT	GVGALRRVYGGNK	RRG
WFYTR	AASTARHLYLRGGA	GVGSMTKIYGGRQ	RNG
WFYTR	AASTARHLYLRGGA	GVGSMTKIYGGRQ	RNG
WWYVR	AAALLRRVYIDGPV	GVNSLRTHYGGKK	DRG

Table 6.8. A set of four blocks from the Blocks database

Any attempt to create a scoring matrix for amino acid substitutions must start from a set of data that can be trusted. The "trusted" data are then used to determine which substitutions are more or less likely. The matrix is then derived from these data, using (as we shall see) aspects of statistical hypothesis-testing theory.

Historically, the PAM matrices were developed first (in 1978), but since the derivation of BLOSUM matrices is somewhat simpler than that for PAM matrices, we start by considering BLOSUM matrices.

6.5.2 BLOSUM Substitution Matrices

The BLOSUM approach was introduced by Henikoff and Henikoff (1992). Henikoff and Henikoff started with a set of protein sequences from public databases that had been grouped into related families. From these sequences they obtained "blocks" of aligned sequences. A block is the *ungapped* alignment of a relatively highly conserved region of a family of proteins. Methods for producing such alignments are given in Section 6.6. These alignments provide the basic data for the BLOSUM approach to constructing substitution matrices. An example of such an alignment leading to four blocks is given in Table 6.8.

Since the algorithms used to construct the aligned blocks employ substitution matrices, there is a circularity involved in the procedure if the aligned blocks are subsequently used to find substitution matrices. Henikoff and Henikoff broke this circularity as follows. They started by using a simple "unitary" substitution matrix where the score is 1 for a match, 0 for a mismatch. Then, using data from suitable groups of proteins, they constructed only those blocks that they could obtain with this simple matrix. This procedure has the effect of generating a conservative set of blocks; that is, it tends to omit blocks with low sequence identity. While this restricted

the number of blocks derived, the blocks obtained were trustworthy and were not biased toward any specific scoring scheme.

Using the blocks so constructed, Henikoff and Henikoff then counted the number of occurrences of each amino acid and the number of occurrences of each pair of amino acids aligned in the same column. Consider a very simplified example, with only three amino acids, A, B, and C, and only one block:

$$\begin{array}{cccc} B & A & B & A \\ A & A & A & C \\ A & A & C & C \\ A & A & B & A \\ A & A & C & C \\ A & A & B & C \end{array} .$$

In this block there are 24 amino acids observed, of which 14 are A, 4 are B, and 6 are C. Thus the observed proportions are

amino acid	proportion of times observed
A	14/24
B	4/24
C	6/24

(6.14)

There are $4 \cdot \binom{6}{2} = 60$ *aligned pairs* of amino acids in the block. These 60 pairs occur with proportions as given in the following table:

aligned pair	proportion of times observed
A to A	26/60
A to B	8/60
A to C	10/60
B to B	3/60
B to C	6/60
C to C	7/60

(6.15)

We now compare these observed proportions to the *expected* proportion of times that each amino acid pair is aligned under a random assortment of the amino acids observed, given the observed amino acid frequencies (6.14). In other words, if we choose two sequences of the same length at random with these frequencies (6.14), and put them into alignment, then the expected proportion of pairs in which A is aligned with A is $\frac{14}{24} \cdot \frac{14}{24}$, the expected proportion of pairs in which A is aligned with B is $2 \cdot \frac{14}{24} \cdot \frac{4}{24}$, and so on. (The factor of 2 in the second calculation allows for the two cases where A is in the first sequence and B in the second, and that where B is in the first sequence and A in the second.)

These fractions are now used to calculate "estimated likelihood ratios" (see Section 3.6) as shown in the following table:

aligned pair	proportion observed	proportion expected	$2\log_2\left(\frac{\text{proportion observed}}{\text{proportion expected}}\right)$
A to A	26/60	196/576	0.70
A to B	8/60	112/576	−1.09
A to C	10/60	168/576	−1.61
B to B	3/60	16/576	1.70
B to C	6/60	48/576	0.53
C to C	7/60	36/576	1.80

(6.16)

For each row in this table the ratio of the entries in the second and third columns is an estimate, from the data, of the ratio of the proportion of times that each amino acid combination occurs in any column to the proportion expected under random allocation of amino acids into columns. With one important qualification, which we describe later, the respective elements in the BLOSUM substitution matrix are now found by calculating twice the logarithm (to the base 2) of this ratio (as shown in the final column of the above table), and then rounding the result to the nearest integer. In this simplified example, the substitution matrix would thus be

$$\begin{array}{c|ccc} & A & B & C \\ A & 1 & -1 & -2 \\ B & -1 & 2 & 1 \\ C & -2 & 1 & 2 \end{array}.$$

In general, the procedure is as follows. For each pair of amino acids x and y, first count the number of times we see x and y in the same column of an aligned block. We denote this number by n_{xy}. We then put

$$p_{xy} = \frac{n_{xy}}{\sum_{u \le v} n_{uv}},$$

where we take $u \le v$ to mean that the letter denoting u precedes the letter denoting v in the alphabet. This number p_{xy} is the estimate of the probability of a randomly chosen pair of amino acids chosen from one column of a block to be the pair x and y. Now, for each amino acid x, let p_x be the proportion of times x occurs somewhere in any block. Consider the quantity

$$e_{xy} = \begin{cases} \frac{2p_x p_y}{p_{xy}} & \text{if } x \neq y, \\ \frac{p_x p_y}{p_{xy}} & \text{if } x = y. \end{cases}$$

This quantity is the ratio of the likelihood that x and y are aligned by chance, given their frequencies of occurrence in the blocks, to the proportion of times we actually observe x and y aligned in the same column in the blocks. We convert this into a score by taking -2 times its logarithm to

the base 2, and rounding to the nearest integer. In this way pairs that are more likely than chance will have positive scores, and those less likely will have negative scores.

While this approach is still rudimentary, it does yield a more useful scoring scheme than the original one that merely scores 1 for a match and 0 for a mismatch. Its main shortcoming is that it overlooks an important factor that can bias the results. The substitution matrix derived will depend significantly on which sequences of each family happen to be in the database used to create the blocks. In particular, if there are many very closely related proteins in one block, and only a few others that are less closely related, then the contribution of that block will be biased toward closely related proteins. For example, suppose the data in one block are as follows:

$$\begin{array}{cccc} A & B & A & A \\ A & B & A & A \\ A & B & A & A \\ A & B & A & A \\ A & A & B & D \\ A & C & B & A \\ D & A & B & A \end{array}$$

The first four sequences possibly derive from closely related species and the last three from three more distant species. Since A occurs with high frequency in the first four sequences, the observed number of pairings of A with A will be higher than is appropriate if we are comparing more distantly related sequences. Ultimately, we would prefer to have sequences in each block such that any pair have roughly the same amount of "evolutionary distance" between them. The solution to this problem used by Henikoff and Henikoff is to group, or cluster, those sequences in each block that are "sufficiently close" to each other and, in effect, use the resulting cluster as a single sequence. This step requires a definition of "sufficiently close," and this is done by specifying a cut-off proportion, say 85%, and then grouping the sequences in each block into clusters in such a way that each sequence in any cluster has 85% or higher sequence identity to at least one other sequence in the cluster in that block.

We now describe how the counting is done in this case, and after the general method is described, we illustrate it with an example. The count of each amino acid is found by dividing each occurrence by the number of sequences in the cluster containing that occurrence, and summing over all occurrences. After this is done, we count aligned amino acid pairs. Here the rule we follow is that if in any block two sequences are in the same cluster, then in that block no counts are taken between amino acids in those two sequences. For any aligned amino acids in sequences in two different clusters in the same block, the count for any amino acid pair is divided by nm, where n and m are the sizes of the two clusters from which the amino acids are taken.

These weighted counts are then used in the same way as before. Consider a simple example with two blocks

$$\begin{array}{cccc} B & A & B & A \\ B & A & B & C \\ A & A & C & C \end{array}$$

and

$$\begin{array}{ccc} C & B & B \\ C & B & B \\ A & B & C \\ A & A & C \end{array}.$$

Suppose the identity for clustering is taken to be .75. Thus we cluster the first two sequences in each block together. The A's are counted as follows. The first column of the first block has one A, the second column contributes two A's, since the first two sequences are clustered it has $1 + \frac{1}{2} + \frac{1}{2} = 2$ A's. The fourth column contributes $\frac{1}{2}$ A. In the second block there are three A's, since each occurrence occurs in a cluster of size one. So in total there are 13/2 A's. Now to get the proportion of A's we must divide by 17, since each column of the first block contributes 2 to the counts of the symbols, and each column of the second block contributes 3 to the counts. So the proportion of A's is $(13/2)/17 = 13/34$. We record the proportions for all symbols in the following table:

amino acid	proportion of times observed
A	13/34
B	5/17
C	11/34

(6.17)

To count the A–B pairs, each occurrence in the first column of the first block contributes $\frac{1}{2}$, and in the second column of the second block the contribution is $\frac{1}{2} + \frac{1}{2} + 1$. So the total A–B count is 3. There are a total of 13 pairs in the blocks, four in the first block (each column contributes one pair, or more precisely, two half pairs) and nine in the second block. Thus the proportion of A–B pairs is 3/13. We record the proportions for all pairs of symbols in the following table:

aligned pair	proportion of times observed
A to A	2/13
A to B	3/13
A to C	5/26
B to B	1/13
B to C	3/13
C to C	3/26

(6.18)

The procedure is then carried out as before.

A further refinement was made by Henikoff and Henikoff (1992). After obtaining a BLOSUM substitution matrix as just described, the matrix

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