Figure 3.11. Example of local alignment of phage λ *cI* and phage P22 *c2* repressors by dynamic programming using the GCG GAP (Needleman-Wunsch algorithm) and BESTFIT (Smith-Waterman algorithm) programs. The log odds form of the PAM120 amino acid substitution matrix was used. PAM120 is optimal for proteins that are ~40% similar. The alignment reveals that the proteins are similar in the carboxy-terminal domain. The penalty for opening a gap in one of the sequences is 11 and for extending the gap 8; these were the default values assigned by the programs. Gaps at the unaligned ends of sequences were also weighted. In the program output, percent identity indicates the number of identical amino acids in the alignment, and percent similarity, the number of similar amino acids. Similar amino acids are defined by high-scoring matches between the amino acid pairs in the substitution matrix, and were defined at the time the program was run. The most similar pairs were indicated by a ':', less similar pairs by a '.' and unrelated pairs by a space, ' ', between the amino acid pairs. Although these dynamic programming programs provide a single optimal alignment, it is important to realize that a series of alignments are usually possible. Other programs, such as ALIGN in the FASTA set (Table 3.1 ALIGN-SITES), provide a user-specified number of alignments (see Fig. 3.12). Additionally, the alignments depend on the method used by the program to convert the traceback matrix into an alignment. GCG programs GAP and BESTFIT provide a method for printing two extremes of alignment, depending on whether gaps are favored in one sequence or the other. These options are called high road and low road.

USE OF SCORING MATRICES AND GAP PENALTIES IN SEQUENCE ALIGNMENTS

Amino Acid Substitution Matrices

Protein chemists discovered early on that certain amino acid substitutions commonly occur in related proteins from different species. Because the protein still functions with these substitutions, the substituted amino acids are compatible with protein structure and function. Often, these substitutions are to a chemically similar amino acid, but other changes also occur. Yet other substitutions are relatively rare. Knowing the types of changes that are most and least common in a large number of proteins can assist with predicting alignments for any set of protein sequences, as illustrated in Figure 3.13. If related

```
LALIGN finds the best local alignments between two sequences
 version 2.0u64 March 1998
Please cite:
 X. Huang and W. Miller (1991) Adv. Appl. Math. 12:373-381

 Comparison of:
(A) lamc1.pro  LAMC1  REFORMAT of: cipro.pro  from: 1 - 237
(B) p22c2.pro  P22C2  REFORMAT of: p22c2.pro  from: 1 - 216
 using matrix file: pam250.mat, gap penalties: -12/-2

  34.0% identity in 206 aa overlap; score:  338

          30        40        50        60        70        80
LAMC1  KKNELGLSQESVADKMGMGQSGVGALFNGINALNAYNAALLAKILKVSVEEFSPSIAREI
       ....: ..:.......:... ....  .. .. :. :   :.: :. : . .  .  ...
P22C2  RRKKLKIRQAALGKMVGVSNVAISQWERSETEPNGENLLALSKALQCSPDYLLKGDLSQT
            20        30        40        50        60        70

          90       100       110       120       130       140
LAMC1  YEMYEAVSMQPSLRSEYEYPVFSHVQAGMFSPELRTFTKGDAERWVSTTKKASDSAFWLE
         :..   ...     .::..: : :: .  .. .. : . :.: .:: . :...:::.
P22C2  NVAYHSRHEPRG-----SYPLISWVSAGQWMEAVEPYHKRAIENWHDTTVDCSEDSFWLD
            80              90       100       110       120

        150       160       170       180       190       200
LAMC1  VEGNSMTAPTGSKPSFPDGMLILVDPEQAVEPGDFCIARLGGD-EFTFKKLIRDSGQVFL
       :.:.:::::.:   :.:.::.:::::: .   :.. .:.:.:. : :::::..:.:. ::
P22C2  VQGDSMTAPAG--LSIPEGMIILVDPEVEPRNGKLVVAKLEGENEATFKKLVMDAGRKFL
       130       140       150       160       170       180

          210       220       230
LAMC1  QPLNPQYPMIPCNESCSVVGKVIASQ
       .:::::::::  :..:...: :....
P22C2  KPLNPQYPMIEINGNCKIIGVVVDAK
         190       200       210

----------

  17.8% identity in 90 aa overlap; score:  37

        20        30        40        50        60        70
LAMC1  RRLKAIYEKKKNELGLSQESVAD-KMGMGQSGVGALFNGINALNAYNAALLAKILKVSVE
       ..::   .   . .:.:. .... ... .... ..:..  .::.   . ::   :. .
P22C2  KKLKIRQAALGKMVGVSNVAISQWERSETEPNGENLLALSKALQCSPDYLLKGDLSQTNV
          20        30        40        50        60        70

          80        90       100
LAMC1  EF-SPSIAREIYEMYEAVSMQPSLRSEYEY
       .. :.  .:. : . . ::  . . .   :
P22C2  AYHSRHEPRGSYPLISWVSAGQWMEAVEPY
          80        90       100

----------

  40.0% identity in 15 aa overlap; score:  36

     220       230
LAMC1  SCSVVGKVIASQWPE
       :..... : :.:: :
P22C2  SYPLISWVSAGQWME
          90
```

Figure 3.12. Example of LALIGN program for finding multiple local alignments of two protein sequences. Three independent alignments of the phage λ and P22 repressors are shown. The amino acid substitution matrix used was the log odds form of the Dayhoff PAM250 matrix provided with the program, with a gap opening penalty of −12 and a gap extension penalty of −2.

protein sequences are quite similar, they are easy to align, and one can readily determine the single-step amino acid changes. If ancestor relationships among a group of proteins are assessed, the most likely amino acid changes that occurred during evolution can be predicted. This type of analysis was pioneered by Margaret Dayhoff (1978).

Amino acid substitution matrices or symbol comparison tables, as they are sometimes called, are used for such purposes. Although the most common use of such tables is for comparison of protein sequences, other tables of nucleic acid symbols are also used for comparison of nucleic acid sequences in order to accommodate ambiguous nucleotide

	Alignment			
sequence A	Tyr	Cys	Asp	Ala
sequence B	Phe	Met	Glu	Gly
BLOSUM62 matrix value	3	−1	2	0

Total score for alignment of sequence A with sequence B
= 3 − 1 + 2 + 0 = 4

Figure 3.13. Use of amino acid substitution matrix to evaluate an alignment of two protein sequences. The score for each amino acid pair (Tyr/Phe, etc.) is looked up in the BLOSUM62 matrix. Each value represents an odds score, the likelihood that the two amino acids will be aligned in alignments of similar proteins divided by the likelihood that they will be aligned by chance in an alignment of unrelated proteins. In a series of individual matches in an alignment, these odds scores are multiplied to give an overall odds score for the alignment itself. For convenience, odds scores are converted to log odds scores so that the values for amino acid pairs in an alignment may be summed to obtain the log odds score of the alignment. In this case, the logarithms are calculated to the base 2 and multiplied by 2 to give values designated as half-bits (a bit is the unit of an odds score that has been converted to a logarithm to the base 2). The value of 4 indicates that the 4 amino acid alignment is $2^{(4/2)}$= 4-fold more likely than expected by chance.

characters or models of expected sequence changes during different periods of evolutionary time that vary scoring of transitions and transversions.

In the amino acid substitution matrices, amino acids are listed both across the top of a matrix and down the side, and each matrix position is filled with a score that reflects how often one amino acid would have been paired with the other in an alignment of related protein sequences. The probability of changing amino acid A into B is always assumed to be identical to the reverse probability of changing B into A. This assumption is made because, for any two sequences, the ancestor amino acid in the phylogenetic tree is usually not known. Additionally, the likelihood of replacement should depend on the product of the frequency of occurrence of the two amino acids and on their chemical and physical similarities. A prediction of this model is that amino acid frequencies will not change over evolutionary time (Dayhoff 1978).

Dayhoff Amino Acid Substitution Matrices (Percent Accepted Mutation or PAM Matrices)

This family of matrices lists the likelihood of change from one amino acid to another in homologous protein sequences during evolution. There is presently no other type of scoring matrix that is based on such sound evolutionary principles as are these matrices. Even though they were originally based on a relatively small data set, the PAM matrices remain a useful tool for sequence alignment. Each matrix gives the changes expected for a given period of evolutionary time, evidenced by decreased sequence similarity as genes encoding the same protein diverge with increased evolutionary time. Thus, one matrix gives the changes expected in homologous proteins that have diverged only a small amount from each other in a relatively short period of time, so that they are still 50% or more similar. Another gives the changes expected of proteins that have diverged over a much longer period, leaving only 20% similarity. These predicted changes are used to produce optimal alignments between two protein sequences and to score the alignment. The assumption in this evolutionary model is that the amino acid substitutions observed over short periods of

evolutionary history can be extrapolated to longer distances. The BLOSUM matrices (see below) are based on scoring substitutions found over a range of evolutionary periods and reveal that substitutions are not always as predicted by the PAM model.

In deriving the PAM matrices, each change in the current amino acid at a particular site is assumed to be independent of previous mutational events at that site (Dayhoff 1978). Thus, the probability of change of any amino acid **a** to amino acid **b** is the same, regardless of the previous changes at that site and also regardless of the position of amino acid **a** in a protein sequence. Amino acid substitutions in a protein sequence are thus viewed as a Markov model (see also hidden Markov models in Chapter 4), characterized by a series of changes of state in a system such that a change from one state to another does not depend on the previous history of the state. Use of this model makes possible the extrapolation of amino acid substitutions observed over a relatively short period of evolutionary time to longer periods of evolutionary time.

To prepare the Dayhoff PAM matrices, amino acid substitutions that occur in a group of evolving proteins were estimated using 1572 changes in 71 groups of protein sequences that were at least 85% similar. Because these changes are observed in closely related proteins, they represent amino acid substitutions that do not significantly change the function of the protein. Hence they are called "accepted mutations," defined as amino acid changes "accepted" by natural selection. Similar sequences were first organized into a phylogenetic tree, as illustrated in Figure 1.1 in Chapter 1. The number of changes of each amino acid into every other amino acid was then counted. To make these numbers useful for sequence analysis, information on the relative amount of change for each amino acid was needed.

Relative mutabilities were evaluated by counting, in each group of related sequences, the number of changes of each amino acid and by dividing this number by a factor, called the exposure to mutation of the amino acid. This factor is the product of the frequency of occurrence of the amino acid in that group of sequences being analyzed and the total number of all amino acid changes that occurred in that group per 100 sites. This factor normalizes the data for variations in amino acid composition, mutation rate, and sequence length. The normalized frequencies were then summed for all sequence groups. By these scores, Asn, Ser, Asp, and Glu were the most mutable amino acids, and Cys and Trp were the least mutable.

The above amino acid exchange counts and mutability values were then used to generate a 20 × 20 mutation probability matrix representing all possible amino acid changes. Because amino acid change was modeled by a Markov model, the mutation at each site being independent of the previous mutations, the changes predicted for more distantly related proteins that have undergone N mutations could be calculated. By this model, the PAM1 matrix could be multiplied by itself N times, to give transition matrices for comparing sequences with lower and lower levels of similarity due to separation of longer periods of evolutionary history. Thus, the commonly used PAM250 matrix represents a level of 250% of change expected in 2500 my. Although this amount of change seems very large, sequences at this level of divergence still have about 20% similarity. For example, alanine will be matched with alanine 13% of the time and with another amino acid 87% of the time.

Do not confuse this mutation probability form of the PAM250 matrix with the log odds form of the matrix described below.

The percentage of remaining similarity for any PAM matrix can be calculated by summing the percentages for amino acids not changing (Ala versus Ala, etc.) after multiplying each by the frequency of that amino acid pair in the database (e.g., 0.089 for Ala) (Dayhoff 1978). The PAM120, PAM80, and PAM60 matrices should be used for aligning sequences that are 40%, 50%, and 60% similar, respectively. Simulations by George et al. (1990) have shown that, as predicted, the PAM250 matrix provides a better-scoring alignment than lower-numbered PAM matrices for distantly related proteins of 14–27% similarity.

PAM matrices are usually converted into another form, called log odds matrices. The odds score represents the ratio of the chance of amino acid substitution by two different hypotheses—one that the change actually represents an authentic evolutionary variation at that site (the numerator), and the other that the change occurred because of random sequence variation of no biological significance (the denominator). Odds ratios are converted to logarithms to give log odds scores for convenience in multiplying odds scores of amino acid pairs in an alignment by adding the logarithms (Fig. 3.13).

Example: Calculations for obtaining the log odds score for changes between Phe and Tyr at an evolutionary distance of 250 PAMs

1. Of 1572 observed amino acid changes, there were 260 changes between Phe and Tyr. These numbers were multiplied by (1) the relative mutability of Phe (see text), and (2) the fraction of Phe to Tyr changes over all changes of Phe to any other amino acid (since Phe to Tyr and Tyr to Phe changes are not distinguished in the original mutation counts, sums of changes are used to calculate the fraction) to obtain a mutation probability score of Phe to Tyr. A similar score was obtained for changes of Phe to each of the other 18 amino acids, and also for the calculated probability of not changing at all. The resulting 20 scores were summed and divided by a normalizing factor such that their sum represented a probability of change of 1%, as illustrated in Table 3.2.

   In this matrix, the score for changing Phe to Tyr was 0.0021, as opposed to a score of Phe not changing at all of 0.9946, as shown in Table 3.2. These calculations were repeated for Tyr changing to any other amino acid. The score for changing Tyr to Phe was 0.0028, and that of not changing Tyr was 0.9946 (not shown). These scores were placed in the PAM1 matrix, in which the overall probability of each amino acid changing to another is ~1%, and that of each not changing is ~99%.

2. The above PAM1 matrix was multiplied by itself 250 times to obtain the distribution of changes expected for 250 PAMs of evolutionary change. These changes can include both forward changes to another amino acid and reverse changes to a former one. At this distance, the probability of change of Phe to Tyr was 0.15 as opposed to a probability of 0.32 of no change in Phe. The corresponding probabilities for Tyr to Phe at 250 PAMs were 0.20 and 0.31 for no change.

3. The log odds values for changes between Phe and Tyr were then calculated. The Phe-Tyr score in the 250 PAM matrix, 0.15, was divided by the frequency of Phe in the sequence data, 0.040, to give the relative frequency of change. This ratio, $0.15/0.04 = 3.75$, was converted to a logarithm to the base 10 ($\log_{10}3.75 = 0.57$) and multiplied by 10 to remove fractional values ($0.57 \times 10 = 5.7$). Similarly, the Tyr to Phe score is $0.20/0.03 = 6.7$, and the logarithm of this number is $\log_{10}6.7 = 0.83$, and multiplied by 10 ($0.83 \times 10 = 8.3$). The average of 5.7 and 8.3 is 7, the number entered in the log odds table for changes between Phe and Tyr at 250 PAMs of evolutionary distance.

   The log odds from the PAM250 matrix, which is sometimes referred to as the mutation data matrix (MDM) at 250 PAMs and also as MDM_{78}, is shown in Figure 3.14. The log odds scores in this table lie within the range of -8 to $+17$. A value of 0 indicates that the frequency of the substitution between a matched pair of amino acids in related proteins is as expected by chance; a value less than 0 or greater than 0 indicates that the frequency is less than or greater than that expected by chance, respectively. Using such a matrix, a high positive score

between two amino acids means that the pair is more likely to be found aligned in sequences that are derived from a common ancestor, i.e., homologous, than in unrelated or nonhomologous sequences. The highest-scoring replacements are for amino acids whose side chains are chemically similar, as might be expected if the amino acid substitution is not to impede function. In the original data, the largest number of observed changes (83) was between Asp (D) and Glu (E). This number is reflected as a log odds score of +3 in the MDM. Many changes were not observed. For example, there were no changes between Gly (G) and Trp (W), resulting in a score of −7 in the table.

Table 3.2. *Normalized probability scores for changing Phe to any other amino acid (or of not changing) at PAM1 and PAM250 evolutionary distances*

Amino acid change	PAM1	PAM250
Phe to Ala	0.0002	0.04
Phe to Arg	0.0001	0.01
Phe to Asn	0.0001	0.02
Phe to Asp	0.0000	0.01
Phe to Cys	0.0000	0.01
Phe to Gln	0.0000	0.01
Phe to Glu	0.0000	0.01
Phe to Gly	0.0001	0.03
Phe to His	0.0002	0.02
Phe to Ile	0.0007	0.05
Phe to Leu	0.0013	0.13
Phe to Lys	0.0000	0.02
Phe to Met	0.0001	0.02
Phe to Phe	0.9946	0.32
Phe to Pro	0.0001	0.02
Phe to Ser	0.0003	0.03
Phe to Thr	0.0001	0.03
Phe to Trp	0.0001	0.01
Phe to Tyr	0.0021	0.15
Phe to Val	0.0001	0.05
SUM[a]	1.0000	1.00

[a]Approximate since scores are rounded off.

The multiplication of two PAM1 matrices to give a PAM2 matrix. Only three rows and columns are shown for illustrative purposes.

	aa1	aa2	aa3 →
aa1	a	b	c
aa2	d	e	f
aa3	g	h	i
↓			

×

	aa1	aa2	aa3 →
aa1	a	b	c
aa2	d	e	f
aa3	g	h	i
↓			

=

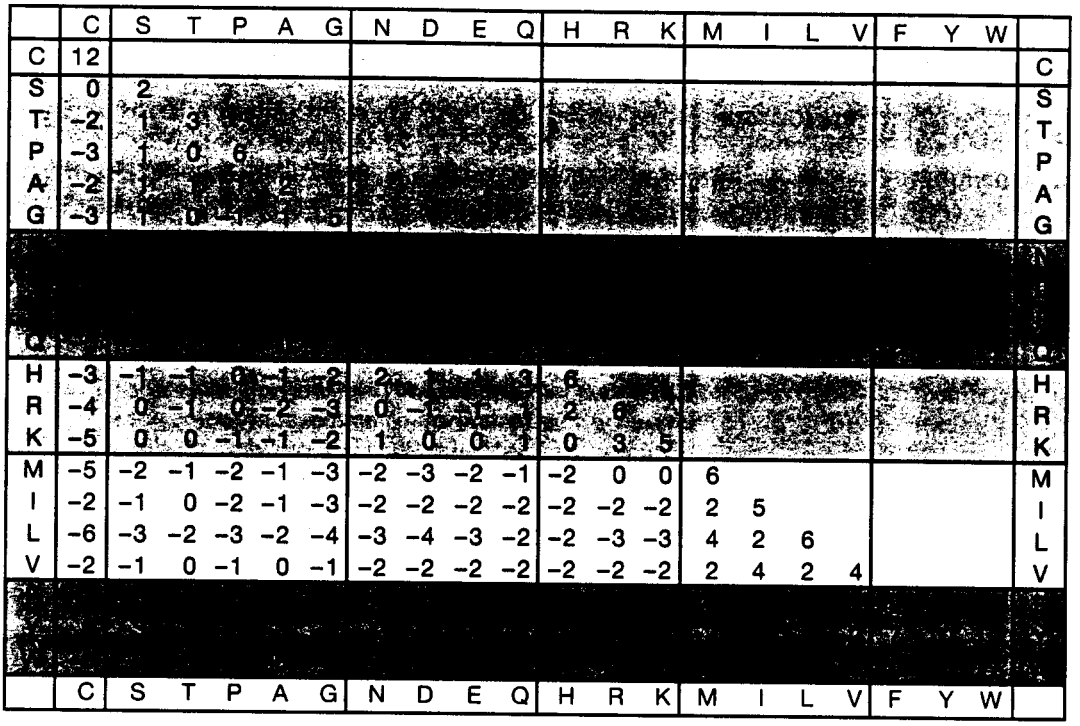
	aa1	aa2	aa3 →
aa1	A	B	C
aa2	D	E	F
aa3	G	H	I
↓			

$A = a^2 + bd + cg + \ldots$

$B = ab + be + ch + \ldots$

$C = ac + bf + ci + \ldots$

$D = da + ed + fg + \ldots$, etc.



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	-1	0	-1	1	2	2	4											Q
H	-3	-1	-1	-1	-1	-2	2	1	1	3	6										H
R	-4	0	-1	-1	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-4	-3	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	0	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Figure 3.14. The log odds form (the mutation data matrix or MDM) of the PAM250 scoring matrix. Amino acids are grouped according to the chemistry of the side group: (C) sulfhydryl, (STPAG) small hydrophilic, (NDEQ) acid, acid amide and hydrophilic, (HRK) basic, (MILV) small hydrophobic, and (FYW) aromatic. Each matrix value is calculated from an odds score, the probability that the amino acid pair will be found in alignments of homologous proteins divided by the probability that the pair will be found in alignments of unrelated proteins by random chance. The logarithm of these ODDS scores to the base 10 is multiplied by 10 and then used as the table value (see text for details). Thus, +10 means the ancestor probability is greater, 0 that the probabilities are equal, and −4 that the alignment is more often a chance one than due to an ancestor relationship. Because these numbers are logarithms, they may be added to give a combined probability of two or more amino acid pairs in an alignment. Thus, the probability of aligning two Ys in an alignment YY/YY is 10 + 10 = 20, a very significant score, whereas that of YY with TP is −2 −5 = − 7, a rare and unexpected alignment between homologous sequences.

At one time, the PAM250 scoring matrix was modified in an attempt to improve the alignment obtained. All scores for matching a particular amino acid were normalized to the same mean and standard deviation, and all amino acid identities were given the same score to provide an equal contribution for each amino acid in a sequence alignment (Gribskov and Burgess 1986). These modifications were included as the default matrices for the GCG sequence alignment programs in versions 8 and earlier and are optional in later versions. They are not recommended because they will not give an optimal alignment that is in accord with the evolutionary model.

Choosing the Best PAM Scoring Matrices for Detecting Sequence Similarity. The ability of PAM scoring matrices to distinguish statistically between chance and biologically meaningful alignments has been analyzed using a recently developed statistical theory for sequences (Altschul 1991) that is discussed later in this chapter. As discussed above, each PAM matrix is designed to score alignments between sequences that have diverged by a particular degree of evolutionary distance. Altschul (1991) has examined how well the PAM matrices actually can distinguish proteins that have diverged to a greater or lesser extent, when these proteins are subjected to a local alignment.

Initially, when using a scoring matrix to produce an alignment, the amount of similarity between sequences may not be known. However, the ungapped alignment scores obtained are maximal when the correct PAM matrix, i.e., the one corresponding to the degree of similarity in the target sequences, is used (Altschul 1991). Altschul (1991) has also examined the ability of PAM matrices to provide a reliable enough indication of an ungapped local alignment score between sequences on an initial attempt of alignment. For sequence alignments, the PAM200 matrix is able to detect a significant ungapped alignment of 16–62 amino acids whose score is within 87% of the optimal one. Alternatively, several combinations, such as PAM80 and PAM250 or PAM120 and PAM350, can also be used. Altschul (1993) has also proposed using a single matrix and adjusting a statistical parameter in the scoring system to reach more distantly related sequences, but this change would primarily be for database searches.

Scoring matrices are also used in database searches for similar sequences. The optimal matrices for these searches have also been determined (see book Web site and Chapter 7). It is important to remember that these predictions assume that the amino acid distributions in the set of protein families used to make the scoring matrix are representative of all families that are likely to be encountered. The original PAM matrices represent only a small number of families. Scoring matrices obtained more recently, such as the BLOSUM matrices, are based on a much larger number of protein families. BLOSUM matrices are not based on a PAM evolutionary model in which changes at large evolutionary distance are predicted by extrapolation of changes found at small distances. Matrix values are based on the observed frequency of change in a large set of diverse proteins. As is discussed on the book Web site, the BLOSUM scoring matrices (especially BLOSUM62) appear to capture more of the distant types of variations found in protein families.

In addition to the aforementioned differences among PAM scoring matrices for scoring alignments of more- or less-related proteins, the ability of each PAM matrix to discriminate real local alignments from chance alignments also varies. To calculate the ability of the entire matrix to discriminate related from unrelated sequences (H, the relative entropy), the score for each amino acid pair s_{ij} (in units of $\log_2$, called bits) is multiplied by the probability of occurrence of that pair in the original dataset, q_{ij} (Altschul 1991). This weighted score is then summed over all of the amino acid pairs to produce a score that represents the ability of the average amino acid pair in the matrix to discriminate actual from chance alignments.

$$H = \sum_{i=1}^{20} \sum_{j=1}^{i} q_{ij} \times s_{ij} \qquad (3)$$

In information theory, this score is called the average mutual information content per pair, and the sum over all pairs is the relative entropy of the matrix (termed H). The relative entropy will be a small positive number. For the PAM250 matrix the number is +0.36, for PAM120, +0.98, and for PAM160, +0.70. In general, all other factors being equal, the higher the value of H for a scoring matrix, the more likely it is to be able to distinguish real from chance alignments.

Analysis of the Dayhoff Model of Protein Evolution as Used in PAM Matrices. As outlined above, the Dayhoff model of protein evolution is a Markov process. In this model, each amino acid site in a protein can change at any time to any of the other 20 amino acids with probabilities given by the PAM table, and the changes that occur at each site are independent of the amino acids found at other sites in the protein and depend only on the cur-

rent amino acid at the site. The assumptions that underlie the method of constructing the Dayhoff scoring matrix have been challenged (for discussion, see George et al. 1990; States and Boguski 1991). First, it is assumed that each amino acid position is equally mutable, whereas, in fact, sites vary considerably in their degree of mutability. Mutagenesis hot spots are well known in molecular genetics, and variations in mutability of different amino acid sites in proteins are well known.

The more conserved amino acids in similar proteins from different species are ones that play an essential role in structure and function and the less conserved are in sites that can vary without having a significant effect on function. Thus, there are many factors that influence both the location and types of amino acid changes that occur in proteins. Wilbur (1985) has tested the Markov model of evolution (see box, below) and has shown that it can be valid if certain changes are made in the way that the PAM matrices are calculated.

Test of Markov Model of Evolution in Proteins

To test the model, Wilbur addressed a major criticism of the PAM scoring matrix, namely that the frequency of amino acid changes that require two nucleotide changes is higher than would be expected by chance. About 20% of the observed amino acid changes require more than a single mutation for the necessary codon changes. This fraction is far greater than would be expected by chance.

To correct for changes that require at least two mutations, Wilbur recalculated the PAM1 matrix using only amino acid substitution data from 150 amino acid pairs that are accountable by single mutations. To accomplish this calculation, he used a refined mathematical model that provided a more precise measure of the rate of substitution. He then estimated frequencies of the other 230 amino acid substitutions reachable only by at least two mutations, and compared these frequencies to the values calculated by Dayhoff, who had assumed these were single-step changes. If these numbers agreed, argued Wilbur, then the PAM model used to produce the Dayhoff matrix is a reliable one. In fact, the Dayhoff values exceeded the two-step model values by a factor of about 117. One source of discrepancy was the assumption that the two-step changes were a linear function of evolutionary time over short evolutionary periods of 1 PAM (average time of 1 PAM = 10 my), whereas, because two mutations are required to make the change, a quadratic function is expected. With this correction made to the Dayhoff calculations for amino acid substitutions requiring two mutations, agreement with the two-step model improved about 10-fold, leaving another 11.7-fold unaccounted for.

Wilbur analyzed the remainder by the covarion hypothesis (Fitch and Markowitz 1970; Miyamoto and Fitch 1995), in which it is assumed that only a certain fraction of amino acid sites in a protein are variable and that one site influences another. Thus, a change in one site may influence the variability of others. This model seems to be reasonable from many biological perspectives. The prediction of this hypothesis is that the frequency of two-step changes would be overestimated because we did not take into account the failure of many sites to be mutable. Using a reasonable estimate of 0.3 for the fraction of the sites that could change, the effect on the Dayhoff calculations for frequencies of two-step changes would be 3.3-fold. The remaining discrepancy in the 11.7-fold ratio between Dayhoff values and two-step values may be attributable to variations in mutation rates from site to site, or to the exclusion of certain amino acids at a particular site. In conclusion, Wilbur (1985) has shown that the Dayhoff model for protein evolution appears to give predictable and consistent

results, but that frequencies of change between amino acids that require two mutational steps must be calculated as a two-step process. Failure to do so generates errors due to variations in site-to-site mutability. George et al. (1990) have counterargued that it has never been demonstrated that two independent mutations must occur, each becoming established in a population before the next appears.

A further criticism of the PAM scoring matrices is that they are not more useful for sequence alignment than simpler matrices, such as one based on a chemical grouping of amino acid side chains. Although alignment of related proteins is straightforward and quite independent of the symbol comparison scoring scheme, alignments of less-related proteins are much more speculative (Feng et al. 1985). These matrices and the BLOSUM matrices have been very useful for finding more distantly related sequences (George et al. 1990). There have been recent changes in the way that members of protein families are identified (see Chapters 4 and 9). Once a family has been identified, family-specific scoring matrices can be produced, and there is no point in using these general matrices. As described in Chapter 4, a scoring matrix representing a section of aligned sequences with no gaps, or a matrix representing a section of aligned sequences with matches, mismatches, and gaps (a profile), are the best tools to search for more family members.

Another criticism of the PAM matrix is that constructing phylogenetic relationships prior to scoring mutations has limitations, due to the difficulty of determining ancestral relationships among sequences, a topic discussed in Chapter 6. Early on in the Dayhoff analysis, the evolutionary trees were estimated by a voting scheme for the branches in the tree, each node being estimated by the most abundant amino acid in distal parts of the tree. Once available, the PAM matrices were used to estimate the evolutionary distance between proteins, given the amount of sequence similarity. Such data can be used to produce a tree based on evolutionary distances (Chapter 6). This circular analysis of using alignments to score amino acid changes and then to use the matrices to produce new alignments has also been criticized. However, no method has yet been devised in any type of sequence analysis for completely circumventing this problem. Evidence that the values in the scoring matrix are insensitive to changes in the phylogenetic relationships has been provided (George et al. 1990).

Finally, the Dayhoff PAM matrices have been criticized because they are based on a small set of closely related proteins. The Dayhoff data set has been augmented to include the 1991 protein database (Gonnet et al. 1992; Jones et al. 1992). The ability of the Dayhoff matrices to identify homologous sequences has also been extensively compared to that of other scoring matrices. These comparisons are discussed on the book Web site.

Blocks Amino Acid Substitution Matrices (BLOSUM)

The BLOSUM62 substitution matrix (Henikoff and Henikoff 1992) is widely used for scoring protein sequence alignments. The matrix values are based on the observed amino acid substitutions in a large set of ~2000 conserved amino acid patterns, called blocks. These blocks have been found in a database of protein sequences representing more than 500 families of related proteins (Henikoff and Henikoff 1992) and act as signatures of these protein families. The BLOSUM matrices are thus based on an entirely different type of sequence analysis and a much larger data set than the Dayhoff PAM matrices.

These protein families were originally identified by Bairoch in the Prosite catalog. This catalog provides lists of proteins that are in the same family because they have a similar biochemical function. For each family, a pattern of amino acids that are characteristic of that function is provided. Henikoff and Henikoff (1991) examined each Prosite family for the presence of ungapped amino acid patterns (blocks) that were present in each family and that could be used to identify members of that family. To locate these patterns, the sequences of each protein family were searched for similar amino acid patterns by the MOTIF program of H. Smith (Smith et al. 1990), which can find patterns of the type aa1 d1 aa2 d2 aa3, where aa1 and aa2 are conserved amino acids and d1 and d2 are stretches of intervening sequence up to 24 amino acids long located in all sequences. These initial patterns were organized into larger ungapped patterns (blocks) between 3 and 60 amino acids long by the Henikoffs' PROTOMAT program (http://www.blocks.fhcrc.org). Because these blocks were present in all of the sequences in each family, they could be used to identify other members of the same family. Thus, the family collections were enlarged by searching the sequence databases for more proteins with these same conserved blocks.

The blocks that characterized each family provided a type of multiple sequence alignment for that family. The amino acid changes that were observed in each column of the alignment could then be counted. The types of substitutions were then scored for all aligned patterns in the database and used to prepare a scoring matrix, the BLOSUM matrix, indicating the frequency of each type of substitution. As previously described for the PAM matrices, BLOSUM matrix values were given as logarithms of odds scores of the ratio of the observed frequency of amino acid substitutions divided by the frequency expected by chance. An example of the calculations is shown in Figure 3.15.

This procedure of counting all of the amino acid changes in the blocks, however, can lead to an overrepresentation of amino acid substitutions that occur in the most closely related members of each family. To reduce this dominant contribution from the most alike sequences, these sequences were grouped together into one sequence before scoring the amino acid substitutions in the aligned blocks. The amino acid changes within these clustered sequences were then averaged. Patterns that were 60% identical were grouped together to make one substitution matrix called BLOSUM60, and those 80% alike to make another matrix called BLOSUM80, and so on. As with the PAM matrices, these matrices differ in the degree to which the more common amino acid pairs are scored relative to the less common pairs. Thus, when used for aligning protein sequences, they provide a greater or lesser distinction between the more common and less common amino acid pairs. The ability of these different BLOSUM matrices to distinguish real from chance alignments and to identify as many members as possible of a protein family has been determined (Henikoff and Henikoff 1992).

Two types of analyses were performed: (1) an information content analysis of each matrix, as was described above for the PAM matrices, and (2) an actual comparison of the ability of each matrix to find members of the same families in a database search, discussed below. As the clustering percentage was increased, the ability of the resulting matrix to distinguish actual from chance alignments, defined as the relative entropy of the matrix or the average information content per residue pair (see above), also increased. As clustering increased from 45% to 62%, the information content per residue increased from ~0.4 to 0.7 bits per residue, and was ~1.0 bits at 80% clustering. However, at the same time, the number of blocks that contributed information decreased by 25% between no clustering and 62% clustering. BLOSUM62 represents a balance between information content and data size. The BLOSUM62 matrix is shown in Figure 3.16.

...A...
...A...
...A...
...A...
...S...
...A...
...A...
...A...
...A...
...A...

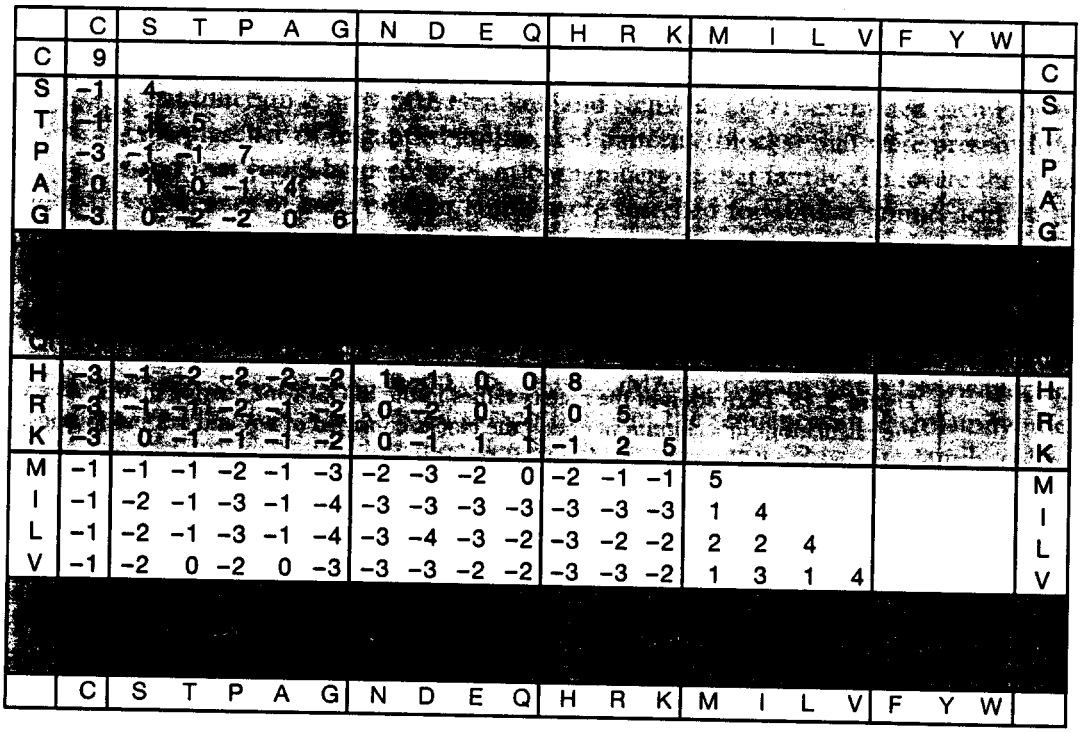
Figure 3.15. Derivation of the matrix values in the BLOSUM62 scoring matrix. As an example of the calculations, if a column in one of the blocks consisted of 9 A and 1 S amino acids, the following is true for this data set (see Henikoff and Henikoff 1992).

1. Since the original sequence from which the others were derived is not known, each column position has to be considered a possible ancestor of the other nine columns. Hence, there are $8+7+6 \ldots +1 = 36$ possible AA pairs (f_{AA}) and 9 possible AS pairs (f_{AS}) to be compared.
2. There are $20+19+18+ \ldots +1 = 210$ possible amino acid pairs.
3. The frequency of occurrence of an AA pair, $q_{AA} = f_{AA}/(f_{AA} + f_{AS}) = 36/(36+9) = 0.8$, and that of an AS pair, $q_{AS} = f_{AS}/(f_{AA} + f_{AS}) = 9/(36+9) = 0.2$.
4. The expected frequency of A being in a pair, $p_A = (q_{AA} + q_{AS}/2) = 0.8 + 0.2/2 = 0.9$, and that of $p_S = q_{AS}/2 = 0.1$.
5. The expected frequency of occurrence of AA pairs, $e_{AA} = p_A \times p_A = 0.9 \times 0.9 = 0.81$, and that of AS, $e_{AS} = 2 \times p_S \times p_A = 2 \times 0.9 \times 0.1 = 0.18$.
6. The matrix entry for AA will be calculated from the ratio of the occurrence frequency to the expected frequency. For AA, ratio $= q_{AA}/ e_{AA} = 0.8/0.81 = 0.99$, and for AS, ratio $= q_{AS}/ e_{AS} = 0.2/0.18 = 1.11$.
7. Both ratios are converted to logarithms to the base 2 and then multiplied by 2 (1/2 bit units). Matrix entry for AA, $s_{AA} = \log_2(q_{AA}/ e_{AA}) = -0.04$, and for AS, $s_{AS} = \log_2(q_{AS}/ e_{AS}) = 0.30$. These logarithms are both rounded to $1\ ^1/_2$ bit unit.

Henikoff and Henikoff (1993) have prepared a set of interval BLOSUM matrices that represent the changes observed between more closely related or more distantly related representatives of each block. Rather than representing the changes observed in very alike sequences up to sequences that were *n*% alike to give a BLOSUM-*n* matrix, the new BLOSUM-*nm* matrix represented the changes observed in sequences that were between *n*% alike and *m*% alike. The idea behind these matrices was to have a set of matrices corresponding to amino acid changes in sequence blocks that are separated by different evolutionary distances.

Comparison of the PAM and BLOSUM Amino Acid Substitution Matrices

There are several important differences in the ways that the PAM and BLOSUM scoring matrices were derived, and these differences should be appreciated in order to interpret the results of protein sequence alignments obtained with these matrices. First, the PAM matrices are based on a mutational model of evolution that assumes amino acid changes occur as a Markov process, each amino acid change at a site being independent of previous changes at that site. Changes are scored in sequences that are 85% similar after predicting



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

Figure 3.16. The BLOSUM62 amino acid substitution matrix. The amino acids in the table are grouped according to the chemistry of the side group: (C) sulfhydryl, (STPAG) small hydrophilic, (NDEQ) acid, acid amide, and hydrophilic, (HRK) basic, (MILV) small hydrophobic, and (FYW) aromatic. Each entry is the logarithm of the odds score, found by dividing the frequency of occurrence of the amino acid pair in the BLOCKS database (after sequences 62% or more in similarity have been clustered) by the likelihood of an alignment of the amino acids by random chance. The denominator in this ratio is calculated from the frequency of occurrence of each of the two individual amino acids in the BLOCKS database and provides a measure of a chance alignment of the two amino acids. The actual/expected ratio is expressed as a log odds score in so-called half-bit units, obtained by converting the odds ratio to a logarithm to the base 2, and then multiplying by 2. A zero score means that the frequency of the amino acid pair in the database is as expected by chance, a positive score that the pair is found more often than by chance, and a negative score that the pair is found less often than by chance. The accumulated score of an alignment of several amino acids in two sequences may be obtained by adding up the respective scores of each individual pair of amino acids. As with the PAM250-derived matrix, the highest-scoring matches are between amino acids that are in the same chemical group, and the very highest-scoring matches are for cysteine–cysteine matches and for matches among the aromatic amino acids. Compared to the PAM160 matrix, however, the BLOSUM62 matrix gives a more positive score to mismatches with the rare amino acids, e.g., cysteine, a more positive score to mismatches with hydrophobic amino acids, but a more negative score to mismatches with hydrophilic amino acids (Henikoff and Henikoff 1992).

a phylogenetic history of the changes in each family. Thus, the PAM matrices are based on prediction of the first changes that occur as proteins diverge from a common ancestor during evolution of a protein family. Matrices that may be used to compare more distantly related proteins are then derived by extrapolation from these short-term changes, assuming that these more distant changes are a reflection of the short-term changes occurring over and over again. For each longer evolutionary interval, each amino acid can change to any other with the same frequency as observed in the short term. In contrast, the BLOSUM matrices are not based on an explicit evolutionary model. They are derived from considering all amino acid changes observed in an aligned region from a related family of proteins, regardless of the overall degree of similarity between the protein sequences. However, these

proteins are known to be related biochemically and, hence, should share common ancestry. The evolutionary model implied in such a scheme is that the proteins in each family share a common origin, but closer versus distal relationships are ignored, as if they all were derived equally from the same ancestor, called a starburst model of protein evolution (see Chapter 6). Second, the PAM matrices are based on scoring all amino acid positions in related sequences, whereas the BLOSUM matrices are based on substitutions and conserved positions in blocks, which represent the most alike common regions in related sequences. Thus, the PAM model is designed to track the evolutionary origins of proteins, whereas the BLOSUM model is designed to find their conserved domains.

Other Amino Acid Scoring Matrices

In addition to the Dayhoff PAM, and related Gonnet et al. (1992), Benner et al. (1994), and Jones et al. (1992) matrices and the BLOSUM matrices, a number of other amino acid substitution matrices have been used for producing protein sequence alignments, and several representative ones are listed in Table 3.3. For a more complete list and comparison, see Vogt et al. (1995). These tables vary from a comparison of simple chemical properties of amino acids to a complex analysis of the substitutions found in secondary structural domains of proteins. Because most of these tables are designed to align proteins on the basis of some such feature of the amino acids, and not on an evolutionary model, they are not particularly suitable for evolutionary analysis. They can be very useful, however, for discovering structural and functional relationships, or family relationships among proteins. A sequence alignment program that uses a combination of these tables has been found to be particularly useful for detecting distant protein relationships (Argos 1987; Rechid et al. 1989). There have been extensive comparisons of the usefulness of various amino acid substitution matrices for aligning sequences, for finding similar sequences in a protein sequence database, or for aligning similar sequences based on structure that are described on the book Web site.

Table 3.3. *Criteria used in amino acid scoring matrices for sequence alignments*

1. Simple identity, which scores only identical amino acids as a match and all others as a mismatch.
2. Genetic code changes, which score the minimum number of nucleotide changes to change a codon for one amino acid into a codon for another, due to Fitch (1966), and also with added information based on structural similarity of amino acid side chains (Feng et al. 1985). A similar matrix based on the assumption that genetic code is the only factor influencing amino acid substitutions has been produced (Benner et al. 1994).
3. Matrices based on chemical similarity of amino acid side chains, molecular volume, and polarity and hydrophobicity of amino acid side chains (see Vogt et al. 1995).
4. Amino acid substitutions in structurally aligned three-dimensional structures (Risler et al. 1988; matrix JO93, Johnson and Overington 1993). A similar matrix was described by Henikoff and Henikoff (1993). Sander and Schneider (1991) prepared a similar matrix based on these same substitutions but augmented by substitutions found in proteins which are so similar to the structure-solved group that they undoubtedly have the same three-dimensional structure.
5. Gonnet et al. (1994) have prepared a 400 × 400 dipeptide substitution matrix for aligning proteins based on the possibility that amino acid substitutions at a particular site are influenced by neighboring amino acids, and thus that the environment of an amino acid plays a role in protein evolution.
6. Jones et al. (1994) have prepared a scoring matrix specifically for transmembrane proteins. This matrix was prepared using an analysis similar to that used for preparing the original Dayhoff PAM matrices, and therefore provides an estimate of evolutionary distances among members of this class of proteins.