comes out as

$$\begin{array}{cccccccccccccc} T & A & C & - & G & A & - & \mathbf{G} & T & A & C & - & G & A & - \\ - & A & C & T & G & A & C & \mathbf{G} & - & A & C & T & G & A & C \end{array}$$

Remark It turns out that the conditional optimal alignment shown above coincides with one of the two "unconditional" optimal global alignments of the given sequences. The second optimal global alignment (with the same score, 6) is given by

$$\begin{array}{cccccccccccccc} T & A & C & - & G & A & \mathbf{G} & T & A & C & - & G & A & - \\ - & A & C & T & G & A & C & \mathbf{G} & A & C & T & G & A & C \end{array}$$

In this alignment two **G**'s are not aligned. Still, the loss of the positive score for the match is compensated for by the decrease in the number of gaps. □

Problem 2.15 A substitution scoring matrix for alignment of nucleotide sequences is given as follows (with the log-odds scores defined in bits):

	T	C	A	G
T	1	0	−1	−1
C	0	1	−1	−1
A	−1	−1	1	0
G	−1	−1	0	1

(a) Determine the average score per nucleotide pair for DNA sequences described by the independence model with equal probabilities of nucleotides ($\frac{1}{4}$).

(b) Determine the "target frequencies" of nucleotide pairs this matrix is designed to search for in alignments of evolutionary related sequences.

Solution (a) In the ungapped alignment of two random nucleotide sequences the average score H per aligned pair is given by

$$H = \sum_{i,j} q_i q_j s_{ij}.$$

Here s_{ij} are the elements of the substitution scoring matrix, the sum is taken over all sixteen possible pairs of nucleotides, and q_i, q_j are the probabilities of nucleotides of types i and j under the uniform independence model. Therefore,

$$H = \frac{1}{16} \sum_{i,j} s_{ij} = -\frac{4}{16} = -0.25.$$

(b) By the definition of the substitution score as the log-odds score,

$$s_{ij} = \log_2 \frac{p_{ij}}{q_i q_j},$$

where p_{ij} are the "target frequencies" of the aligned pair (i, j) of nucleotides. Then the value of the target frequency $p_{ij} = (1/q_i q_j)2^{s_{ij}}$. For example, $p_{TT} = 1/16 \times 2 = 1/8$. All the "target frequencies" p_{ij} are shown in the following table:

	T	C	A	G
T	1/8	1/16	1/32	1/32
C	1/16	1/8	1/32	1/32
A	1/32	1/32	1/8	1/16
G	1/32	1/32	1/16	1/8

□

2.2.1 Derivation of the amino acid substitution matrices (PAM series)

Discussion of this method requires knowledge of the key notions of phylogenetic tree construction (Durbin *et al.* (1998), Chap. 7).

The substitution matrix is an important component of the scoring system for a pairwise sequence alignment. Dayhoff *et al.* (1978) offered a solid theoretical approach to defining the elements of substitution matrices, scaled by evolutionary distance, from counts of amino acid substitutions. These counts were calculated as frequencies of aligned residue pairs in carefully crafted alignments of closely related protein sequences from seventy-one families. Multiple alignments of these sequences were reduced to seventy-one ungapped alignment blocks (each sequence in a block had to be at least 85% identical to any other sequence in the block). The most parsimonious (with minimal number of substitutions along edges) phylogenetic tree, or several trees if parsimony was not unique, was constructed for sequences from each block. To illustrate the method, we use the artificial ungapped block considered by Dayhoff *et al.* (1978):

$$\begin{matrix} A & C & G & H \\ D & B & G & H \\ A & D & I & J \\ C & B & I & J \end{matrix} \tag{2.7}$$

The four most parsimonious trees, T_1, T_2, T_3, T_4, for sequences from block (2.7) are shown in Figure 2.2. For each pair of different amino acids (i, j) the total number a_{ij} of substitutions from i to j along the downward directed edges of trees T_k, $k = 1, \ldots, 4$, was calculated, and the matrix A of *accepted point mutations* with elements $A_{ij} = A_{ji} = a_{ij} + a_{ji}$, $i \neq j$, was produced (see Table 2.9).

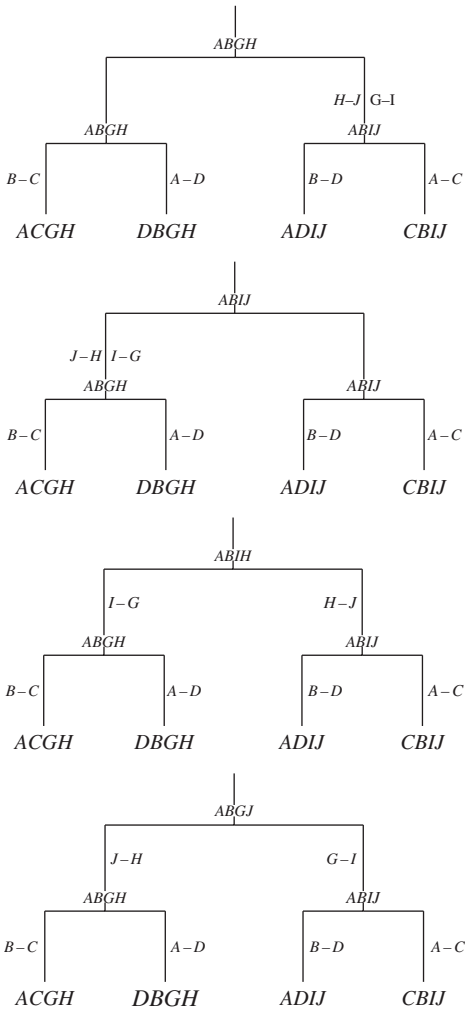

Figure 2.2. The most parsimonious trees T_1, T_2, T_3, and T_4 (in top down order) with the "observed" in block (2.7) amino acid sequences at the leaf nodes; each tree carries a total of six substitutions. The ancestor sequences (unobserved but inferred) are placed at the root and branch nodes; amino acid substitutions are indicated along the edges.

Note that we keep the term "accepted point mutation" and the two which will appear below, "relative mutability" and "mutation probability matrix," that were used by Dayhoff *et al.* (1978), although in contemporary literature the term "mutation" in this context has been replaced by the term "substitution."

Next, the relative mutability m_j for each amino acid j was determined as follows. An edge of a tree T_k, $k = 1, \ldots, 4$, is associated with the ungapped pairwise

Table 2.9. *The matrix A of accepted point mutation counts*

	A	B	C	D	G	H	I	J
A		0	4	4	0	0	0	0
B	0		4	4	0	0	0	0
C	4	4		0	0	0	0	0
D	4	4	0		0	0	0	0
G	0	0	0	0		0	4	0
H	0	0	0	0	0		0	4
I	0	0	0	0	4	0		0
J	0	0	0	0	0	4	0	

Table 2.10. *The relative amino acid mutability values m_j derived from the sequence alignment block (2.7)*

Amino acid	A	B	I	H	G	J	C	D
Changes (substitutions)	8	8	4	4	4	4	8	8
Frequency of occurrence	40	40	24	24	24	24	8	8
Relative mutability m	0.2	0.2	0.167	0.167	0.167	0.167	1	1

alignment of two sequences connected by this edge. Thus, any tree T_k in Figure 2.2 generates six alignments; for example, for T_1 they are as follows:

$$\begin{array}{cccc} A & B & G & H \\ A & B & G & H \end{array} \qquad \begin{array}{cccc} A & B & G & H \\ A & B & I & J \end{array} \qquad \begin{array}{cccc} A & B & G & H \\ A & C & G & H \end{array}$$

$$\begin{array}{cccc} A & B & G & H \\ D & B & G & H \end{array} \qquad \begin{array}{cccc} A & B & I & J \\ A & D & I & J \end{array} \qquad \begin{array}{cccc} A & B & I & J \\ C & B & I & J \end{array}$$

The relative mutability m_j is defined as the ratio of the total number of times that amino acid j has changed in all twenty-four pairwise alignments to the number of times that j has occurred in these alignments. The values of m_j are listed in Table 2.10.

Now we introduce the effective frequency f_j of an amino acid j. This notion takes into account the difference in variability of the primary structure conservation in proteins with different functional roles (thus, two alignment blocks corresponding to two different families may contribute differently to f_j, even if the number of occurrences of amino acid j in these blocks is the same). The effective frequency f_j is defined as

$$f_j = k \sum_{B_i} q_j^{B_i} N_i,$$

Table 2.11. *The amino acid effective frequencies f_j derived from the sequence alignment block (2.7)*

Amino acid	A	B	I	H	G	J	C	D
Frequencies f	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125

Table 2.12. *The amino acid effective frequencies f_i determined for the original alignment data (Dayhoff et al. (1978), Table 22)*

Gly (G)	*Ala (A)*	*Leu (L)*	*Lys (K)*	*Ser (S)*	*Val (V)*	*Thr (T)*
0.089	0.087	0.085	0.081	0.070	0.065	0.058
Pro (P)	*Glu (E)*	*Asp (D)*	*Arg (R)*	*Asn (N)*	*Phe (F)*	*Gln (Q)*
0.051	0.050	0.047	0.041	0.040	0.040	0.038
Ile (I)	*His (H)*	*Cys (C)*	*Tyr (Y)*	*Met (M)*	*Trp (W)*	
0.037	0.034	0.033	0.030	0.015	0.010	

where the sum is taken over all alignment blocks B_i, $q_j^{B_i}$ is the observed frequency of amino acid j in block B_i, N_i is the number of substitutions in a tree built for B_i, and coefficient k is chosen to ensure that the sum of the frequencies f_j is 1. In our example, with only the one block (2.7), the values of effective frequencies are equal to the values of compositional frequencies ($f_j = q_j$) and are shown in Table 2.11. The effective frequencies of twenty amino acids derived from seventy-one original alignment blocks (Dayhoff *et al.*, 1978), are given in Table 2.12.

The next step is to find elements of the mutation probability matrix $M = (M_{ij})$. The element M_{ij} defines the probability of an amino acid in column j having been substituted by an amino acid in row i over a given evolutionary time. The non-diagonal elements of M are defined by the following formula:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_k A_{kj}}, \tag{2.8}$$

where λ is a constant to be determined below, m_j is the relative mutability of amino acid j (Table 2.10), and A_{ij} is an element of the accepted point mutation matrix A (Table 2.9). For the diagonal elements of M we have

$$M_{jj} = 1 - \lambda m_j. \tag{2.9}$$

Note that M is a non-symmetric matrix if $m_i \neq m_j$ for some $i \neq j$.

The coefficient λ represents a degree of freedom that could be used to connect the matrix M with an evolutionary time scale. For instance, the coefficient λ could be adjusted to ensure that a specified (small) number of substitutions would occur

Table 2.13. *The example mutation probability matrix for evolutionary distance 1 PAM calculated from alignment block (2.7)*

The element M_{ij} gives the probability of an amino acid in column j having been substituted by an amino acid in row i over evolutionary time 1 PAM.

	A	B	C	D	G	H	I	J
A	0.9948	0	0.0131	0.0131	0	0	0	0
B	0	0.9948	0.0131	0.0131	0	0	0	0
C	0.0026	0.0026	0.9740	0	0	0	0	0
D	0.0026	0.0026	0	0.9740	0	0	0	0
G	0	0	0	0	0.9957	0	0.0043	0
H	0	0	0	0	0	0.9957	0	0.0043
I	0	0	0	0	0.0043	0	0.9957	0
J	0	0	0	0	0	0.0043	0	0.9957

on average per hundred residues. This adjustment of λ was done by Dayhoff *et al.* (1978) in the following way. The expected number of amino acids that will remain unchanged in a protein sequence one hundred amino acids long is given by the formula $100 \sum_j f_j M_{jj} = 100 \sum_j f_j (1 - \lambda m_j)$. If only one substitution per hundred residues is allowed, then λ is calculated from equation

$$100 \sum_j f_j (1 - \lambda m_j) = 99. \tag{2.10}$$

Subsequently, Equations (2.8) and (2.9) are used for the calculation of all elements of the matrix M. Such a mutation probability matrix is associated with an evolutionary time interval 1 PAM (one accepted point mutation per hundred amino acids), and is called the 1PAM matrix. From the sequence data in alignment block (2.7) we obtain $\lambda = 0.0261$ and the example matrix M (Table 2.13). The actual 1 PAM matrix M derived from the original data in Dayhoff *et al.* (1978) is shown in Table 2.14.

Further, it is assumed that the mutation probability matrix M serves as the matrix of transition probabilities for the stationary (homogeneous) Markov chain x_n, $n \in \mathbf{N}$, the model of the evolutionary change at each site of a protein sequence. Therefore, the stochastic 1 PAM matrix M associated with one unit (1 PAM) of evolutionary distance can be used for derivation of the matrices associated with larger evolutionary distances (multiples of 1 PAM). Then the mutation probability matrix for evolutionary distance n PAM will coincide with the matrix M^n of transition probabilities of the Markov chain x_n for n time units, the nth power of matrix M. (For how to derive a Markov process, which is a continuous-time version of Markov chain x_n, see Problem 8.18.) It is easy to check by direct calculation that M^3, the mutation probability matrix for time units 3 PAM, is strictly positive;

Table 2.14. *Mutation probability matrix M for the evolutionary distance 1 PAM (i.e., one accepted point mutation per hundred amino acids)*

The element of this matrix, M_{ij}, gives the probability of an amino acid in column j having been substituted by an amino acid in row i over the evolutionary time interval of 1 PAM. The actual transition probabilities are multiplied by 10 000 to simplify the matrix appearance (Dayhoff *et al.* (1978), Fig. 82).

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

therefore, the Markov chain with the matrix M of transition probabilities is a regular Markov chain. From the theory of stochastic matrices (Berman and Plemmons, 1979; Meyer, 2000) it is known that for a regular Markov chain there exists a unique strictly positive invariant (stationary) distribution π: $M\pi = \pi$, and for any initial distribution of probabilities of states (amino acids in our case) π_0,

$$\pi_n = M^n \pi_0 \to \pi \tag{2.11}$$

as $n \to +\infty$. It turns out that vector f of the effective frequencies $(f_1, f_2, \ldots, f_{20})$ (Table 2.12), with the order of its components corresponding to the order of rows (columns) of matrix M, satisfies the equation $Mf = f$ for the stationary distribution. Then convergence in (2.11) implies that f is the vector of equilibrium frequencies of M:

$$M^n \to \begin{pmatrix} f_A & f_A & \ldots & f_A \\ f_R & f_R & \ldots & f_R \\ \ldots & \ldots & & \ldots \\ f_V & f_V & \ldots & f_V \end{pmatrix} \tag{2.12}$$

as $n \to +\infty$. The convergence (2.12) was verified in Dayhoff *et al.* (1978) by direct calculation: it was shown that M^{2034} closely approximates the matrix of equilibrium frequencies.

The theory described above was used further to derive the matrices of the log-odds scores for amino acid substitutions. These matrices are critically important for algorithms of protein sequence alignment. Since we have a family of mutation probability matrices M^n, $n = 1, 2, \ldots$, we can derive a family of substitution matrices S_n, $n = 1, 2, \ldots$, with elements $s_n(i,j)$ of log-odds scores as follows. For a pair of amino acids (i,j), $i,j = 1, \ldots, 20$, the log-odds score $s_n(i,j)$ is defined by the following formula:

$$s_n(i,j) = \log \frac{M^n_{ji}}{f_j},$$

where $M^n_{ji} = P(x_n = j | x_0 = i)$ is an element of matrix M^n. The interpretation of the substitution score $s_n(i,j)$ from the standpoint of protein sequence evolution starts with the application of the properties of the Markov chain $\{x_k\}$, $k \in \mathbf{N}$:

$$\begin{aligned} s_n(i,j) = \log \frac{M^n_{ji}}{f_j} = \log \frac{P(x_n = j | x_0 = i)}{f_j} = \log \frac{f_i P(x_n = j | x_0 = i)}{f_i f_j} \\ = \log \frac{\sum_a f_i P(x_n = j | x_{n/2} = a) P(x_{n/2} = a | x_0 = i)}{f_i f_j}. \end{aligned} \tag{2.13}$$

Note that for large n we have $M^n_{ij} \approx f_i$, $M^n_{ji} \approx f_j$, since f is the vector of equilibrium frequencies for matrix M due to (2.12). Therefore, even though the Markov chain $\{x_k\}$ does not possess the reversibility property ($f_j M_{ij} \neq f_i M_{ji}$ for some $i \neq j$), for

sufficiently large n we can assume that $f_j M^n_{ij} \approx f_i M^n_{ji}$ for all i, j. Then Equation (2.13) becomes

$$\begin{aligned} s_n(i,j) &= \log \frac{\sum_a f_i P(x_n = j | x_{n/2} = a) P(x_{n/2} = a | x_0 = i)}{f_i f_j} \\ &\approx \log \frac{\sum_a f_a P(x_n = j | x_{n/2} = a) P(x_{n/2} = i | x_0 = a)}{f_i f_j} \\ &= \log \frac{\sum_a f_a P(x_{n/2} = j | x_0 = a) P(x_{n/2} = i | x_0 = a)}{f_i f_j} \\ &= \log \frac{P(x_{n/2} = i, x'_{n/2} = j | x_0 = x'_0)}{f_i f_j}, \end{aligned} \tag{2.14}$$

where $\{x'_k\}$ is an independent copy of the Markov chain $\{x_k\}$. The last expression in Equation (2.14) is the log-odds ratio that involves the Markovian evolutionary model with molecular clock property and the independence pair-sequence model R with parameters defined as the product of the effective $f_i f_j$, $i, j = 1, \ldots, 20$.

The numerator in (2.14) is the probability that two aligned protein sequences diverged from a common ancestor $n/2$ PAM time units ago would have at a given site amino acids i and j, assuming that substitutions in proteins are described by the Markov process $\{x_k\}$.

The term $f_i f_j$ in the denominator in (2.14) is the probability of observing amino acids i and j at a given site of two aligned protein sequences under the independence pair-sequence model R. Since the same arguments as in Equations (2.13) and (2.14) are true for the substitution score $s_n(j,i) = \log M^n_{ij}/f_i$, we have $s_n(i,j) \approx s_n(j,i)$. Thus, unlike the mutation probability matrix M^n, the substitution matrix S_n is a symmetric one. For practical convenience, log-odds values $s_n(i,j)$ are rescaled (by multiplying by ten, or by an other scaling factor) and then rounded to the nearest integer.

Among the PAM substitution matrices, the most frequently used is the 250 PAM matrix shown in Table 6.1.

With the rapid growth of the protein data, the update of the matrices of PAM series was undertaken by Jones, Taylor, and Thornton (1992). They used the same technique for counting amino acid substitutions as Dayhoff *et al.* (1978). Another empirical model of protein evolution which combines a parsimony-based counting and the maximum likelihood approach was derived by Whelan and Goldman (2001). Extended discussions of the Dayhoff model can be found in Wilbur (1985) and George, Barker, and Hunt (1990).

The important practical question is how to choose the optimal substitution matrix for the alignment of two given (homologous) protein sequences x and y. The accurate answer would include the estimation of the evolutionary distance (time)

Table 2.15. *Correspondence between the observed percent of amino acid differences d between two aligned homologous sequences and the evolutionary distance n (in PAM) between them*

As the evolutionary distance increases, the probability of multiple substitutions at the same site reversing initial changes becomes greater and results in a slower growth of observed percent difference. (Dayhoff *et al.* (1978), Table 23).

d	1	5	10	15	20	25	30	34	40	45
n	1	5	11	17	23	30	38	47	56	67
d	50	55	60	65	70	75	80	85		
n	80	94	112	133	159	195	246	328		

n between x and y, which requires the construction of an alignment of x and y to determine the percent difference d (the percentage of mismatches in aligned sites). The correspondence between the percent difference d and evolutionary distance n (in PAM) could be determined from the following equation:

$$100 \sum_j f_j M_{jj}^n = 100 - d,$$

where M_{jj}^n are elements of the mutation probability matrix M^n for the evolutionary time n PAM. For a given d, this equation allows us to choose the appropriate mutation probability matrix M^n and the substitution matrix S_n (Table 2.15 derived by Dayhoff *et al.* (1978) lists some pairs of corresponding values n and d). However, construction of an alignment of x and y to find d requires the substitution matrix in the first place! A possible, but cumbersome, way to break this circular logic is to use an iterative approach. Additional iterations, however, become prohibitively expensive for the database searches.

The problem of choosing an appropriate substitution matrix for local sequence alignment was studied by Altschul (1991) from the information theory perspective. It was shown that if a single matrix has to be selected then, for database searches, the 120 PAM matrix is the most appropriate one, while for comparing two specific proteins with suspected homology the best choice is the 200 PAM matrix.

Finally, to construct alignments with gaps, the scoring system has to be augmented by the gap scoring scheme. This issue was addressed, for example, by Vingron and Waterman (1994), Pearson (1996), Mott (1999), and Reese and Pearson (2002).

Problem 2.16 The original 250 PAM substitution matrix (Dayhoff *et al.*, 1978) scores a substitution of *Gly* by *Arg* by negative score -3 (decimal logarithms