and scaling factor 10 are used, with rounding to the nearest integer). The average frequency of *Arg* in the protein sequence database is 0.041. Use this information, as well as the method described in Section 2.2.1, to estimate the probability that *Gly* will be substituted by *Arg* after a 250 PAM time period.

Solution The element s_{ij} of the 250 PAM substitution matrix and the frequency of amino acid q_j in a protein sequence database (the data set from which the parameters of the background independence model are derived) are connected by the following formula:

$$s_{ij} = \left[10 \lg \frac{P(i \to j \text{ in } 250\,\text{PAM})}{q_j} \right].$$

Therefore, for the probability of substitution of *Gly* by *Arg* we have

$$P(Gly \to Arg \text{ in } 250\,\text{PAM}) = 0.041 \times 10^{-0.3} = 0.0205. \qquad \square$$

Problem 2.17 Matrices of the BLOSUM series are frequently used in protein sequence alignment algorithms. The method of derivation of the amino acid substitution scores used in the BLOSUM matrices was introduced by Henikoff and Henikoff (1992). From the following multiple alignment block clustered into three sections at the 75% threshold

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

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

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

define the BLOSUM-type substitution score matrix (3×3) using half-bit units.

Solution We define the matrix F of counts with elements $f_{ij} = f_{ji}$ equal to the number of (weighted) pairs of amino acids i and j over all columns of the block. We explain how to calculate f_{ij}, taking as an example the weighted count f_{AA}. First, note that every residue from the first cluster (sequences 1–3) has weight $1/3$, and that every residue from the second cluster (sequences 4, 5) and the third cluster (sequences 6, 7) has weight $1/2$, since the total weight of the residues from sequences of the same cluster (per column) must be equal to 1 (each cluster contributes into the alignment as a single sequence). In the first column there are three residues *A* from the first cluster (each with weight $1/3$) and one residue *A* from

the third cluster (with weight 1/3). Thus, the total weighted count (per column 1) of A-to-A substitutions between different clusters is $f_{AA}^1 = 1/3 \times 1/2 + 1/3 \times 1/2 + 1/3 \times 1/2 = 1/2$. There are no A-to-A substitutions between clusters in columns 2 and 4. The total count of A-to-A substitutions per column 3 is $f_{AA}^3 = 1/3 \times 1/2 = 1/3$. Finally, $f_{AA} = f_{AA}^1 + f_{AA}^2 + f_{AA}^3 + f_{AA}^4 = 1/3 + 0 + 1/2 + 0 = 5/6$. Repeating this counting scheme for each pair i and j, we fill in the following symmetric matrix F of pair counts:

	A	C	D
A	5/6	13/3	11/3
C	13/3	1	5/3
D	11/3	5/3	1/2

The observed frequency of occurrence of pair (i, j) is defined by the following formula:

$$q_{ij} = \frac{f_{ij}}{\sum_i \sum_{j=1}^{i} f_{ij}},$$

which produces the following values of q_{ij}:

	A	C	D
A	5/72	13/36	11/36
C		1/12	5/36
D			1/24

Next, the expected frequency e_{ij} of occurrence for the amino acid pair (i, j) is defined by $e_{ii} = p_i^2$ and $e_{ij} = 2p_i p_j$, $i \neq j$, where p_i is the probability of occurrence of amino acid i. The observed frequencies p_i are determined by the following formula:

$$p_i = q_{ii} + \frac{1}{2} \sum_{j \neq i} q_{ij}.$$

Thus, we have $p_A = 29/72$, $p_C = 19/72$, $p_D = 1/3$, and the expected frequencies of pairs e_{ij} come out as follows:

	A	C	D
A	0.1622	0.2683	0.2125
C		0.1108	0.1757
D			0.0696

Finally, calculating a log-odds ratios in half-bit units $s_{ij} = 2\log_2 q_{ij}/e_{ij}$ and rounding them to the nearest integer produces the (3×3) substitution score matrix:

	A	C	D
A	−2	1	1
C	1	−1	−1
D	1	−1	−1

□

2.2.2 Distributions of similarity scores

2.2.2.1 Theoretical introduction to Problems 2.19 and 2.20.

The development of comparative and evolutionary genomics would be impossible without efficient similarity search algorithms. However, as soon as a similarity characterized by a high score is found in a database search one needs to make sure that this high score has not occurred by chance. To establish confidence in identifying similarities, it is natural to study the alignments of random sequences and determine the statistically significant thresholds for similarity scores. Several definitions are necessary for an in-depth discussion.

Let a biological (protein) sequence be described by the independence model with probability q_a of occurrence of amino acid a at any position, $\sum_a q_a = 1$. The score of an ungapped pairwise alignment is determined as the sum of scores for amino acid pairs, while the score for amino acid pair (a, b) is defined by an element $s(a, b)$ of the substitution matrix **S** (such as BLOSUM or PAM matrix). Local ungapped pairwise alignments whose scores cannot be improved by extension or trimming are called *high-scoring segment pairs* (HSPs).

The elements of the substitution matrix and parameters of the independence model are supposed to satisfy the negative bias condition,

$$\sum_{a,b} q_a q_b s(a, b) < 0,$$

which means that the average score per alignment position is negative. This condition prevents the increase of a total score of a local alignment due to the mere increase of the length of the alignment. On the other hand, some elements of matrix **S** must be positive, otherwise all HSPs will have zero length. Note that the negative bias condition implies that the function $f(\lambda) = \sum_{a,b} q_a q_b e^{\lambda s(a,b)} - 1$ has only one positive root λ (see Problems 5.3–5.5).

Distributions of statistics of the HSP scores have been studied in numerous publications, for example Iglehart (1972); Lipman *et al.* (1984); Reich, Drabsch,

and Däumler (1984); Smith, Waterman, and Burks (1985); Karlin and Altschul (1990, 1993); Karlin, Dembo, and Kawabata (1990); Dembo, Karlin, and Zeitouni (1994a,b); Altschul and Gish (1996).

It was proved in Dembo, Karlin, and Zeitouni (1994b) that for unrelated (independent) sequences X and Y with sufficiently large lengths n and m the distribution of the number N_S of HSPs with scores greater than S can be closely approximated by the Poisson distribution with parameter $\Lambda = Knme^{-\lambda S}$. Here λ is the positive root of $f(\lambda)$, and the constant K depends only on $\{q_a\}$ and the scoring matrix $\mathbf{S}$. Therefore, the expected number of HSPs with scores greater than S is given by

$$\mathbf{E}N_S \approx \Lambda = Knme^{-\lambda S},$$

and the probability of observing a certain number z of such HSPs is given by

$$P(N_S = z) \approx e^{-\Lambda}\frac{\Lambda^z}{z!}.$$

The distribution of N_S can be used to derive the probability distribution of the maximum score $S_{\max}$ (the score of the optimal ungapped local alignment for sequences X and Y) as follows:

$$P(S_{\max} \geq S) = P(N_S > 0) = 1 - P(N_S = 0) \approx 1 - e^{-\mathbf{E}N_S}. \tag{2.15}$$

The value $\mathbf{E}N_S$ is called the *E-value* of the score S; the probability value $P(S_{\max} \geq S) = 1 - e^{-\mathbf{E}N_S}$ is called the *P-value* of the score S. The **E**-value is frequently used to characterize the statistical significance of the number of HSPs observed in database searches.

The normalized sum T_r of the scores $S_1,\ldots,S_r$ of r highest-scoring segment pairs,

$$T_r = \lambda\left(\sum_{k=1}^{r} S_k\right) - \ln K^r - \ln(mn)^r,$$

can be of interest as well. It was shown in Karlin and Altschul (1993) that for sufficiently large n and m the probability density function of T_r is approximated by

$$f(t) = \frac{e^{-t}}{r!(r-2)!}\int_0^\infty y^{r-2}\exp(-e^{(y-t)/r})\,dy.$$

The moments of this distribution can be calculated by the Laplace transform. In particular,

$$\mathbf{E}T_r \approx r\left(1 + \gamma - \sum_{k=1}^{r} 1/k\right) \approx r(1 - \ln r) - 1/2,$$

$$\mathbf{Var}\,T_r \approx r^2\left(\pi^2/6 - \sum_{k=1}^{r} 1/k^2\right) + r \approx 2r - 1/2,$$

where $\gamma \approx 0.577$ is the Euler constant.

Instead of ungapped alignments and HSPs, Neuhauser (1994) considered local alignments with insertions and deletions but no mismatches (so $x_i = y_j$ in a pair of aligned residues (x_i, y_j)). Neuhauser proposed an algorithm searching for local alignments of sequences X and Y with t matching pairs, k gaps, and a length of each gap at most l ((t,k,l)-alignments). It has been proved that, for independent sequences X and Y both described by the independence model and with sufficiently large lengths n and m, the number W of (t,k,l)-alignments found by this algorithm approximately has the Poisson distribution with parameter

$$\Lambda = C^k mn(1-p)\binom{t-1}{k} l^k p^{t-k},$$

where the probability of match p and the constant C depend only on the parameters of the independence model. Then for sufficiently large n and m the largest number S^* of matching pairs in any local alignment with k gaps (with length at most l each) has the following distribution:

$$P(S^* < t) \approx e^{-\Lambda}.$$

Finally, the formula for the P-value of an alignment score was derived for the general case of gapped local alignment (with mismatches in aligned pairs allowed) of two independent sequences and a scoring scheme based on the log-odds substitution matrix $\mathbf{S}$. The results were obtained heuristically by Mott (1999, 2000) and Mott and Tribe (1999) and analytically by Siegmund and Yakir (2000, 2003). We outline below the main results on the statistics of the gapped alignments scores (Siegmund and Yakir, 2000).

It is assumed, as it was for HSPs, that the negative bias condition holds for the elements of matrix $\mathbf{S}$ and parameters of the independence model, and that λ is the only positive root of function f. For the linear gap scoring function $\gamma(g) = -\delta g$, for sufficiently large n and m, the tail of the distribution of the maximum score $\mathbf{S}^j_{\max}$ of local alignment with at most j gaps can be approximated by the following formula:

$$P(\mathbf{S}^j_{\max} \geq S) \approx K' nme^{-\lambda S} \frac{(b\lambda S)^j}{j!} \sum_{l=j}^{\infty} a_l e^{-\lambda\delta l}.$$

Here constants K', b, and a_l depend only on $\{q_a\}$ and the scoring matrix $\mathbf{S}$, and the upper bound for a_l is $1/(\exp(\lambda\delta) - 1)$.

For the scoring function $\gamma(g) = -d - \delta g$ (the affine gap score), and sufficiently large n and m, the score $\mathbf{S}_{\max}$ of optimal local alignment (with any number of gaps) approximately has the following distribution:

$$P(\mathbf{S}_{\max} \geq S) \approx K' nme^{-\lambda S} \sum_{j=0}^{\infty} \frac{(b\lambda S e^{-d\lambda})^j}{j!} \sum_{l=j}^{\infty} a_l e^{-\lambda\delta l}. \tag{2.16}$$

In Equation (2.16) K', b, and a_l are the same constants as in the case of the linear gap scoring function, and it is assumed that $\lambda d = \ln S + C$ for some constant C. If, additionally, the rates of growth of n, m, and S satisfy the condition $nm \exp(-S) \to v$ for some finite positive v, then the formula for the P-value becomes

$$P(\mathbf{S}_{\max} \geq S) \approx 1 - \exp \Lambda, \tag{2.17}$$

where Λ is defined by the expression on the right-hand side of Equation (2.16). If in the expression for Λ we set $j = 0$ and, therefore, consider ungapped local alignments only, then Equation (2.17) for the P-value is reduced to Equation (2.15) derived by Dembo, Karlin, and Zeitouni (1994b). The P-values used in the original program, BLAST (Altscul *et al.*, 1990) for the similarity scores of ungapped local alignments have been calculated by formula (2.15). Notably, in the gapped version of the BLAST program Altschul *et al.* (1997) use the approximate formula (2.17) for P-values with the values of constants involved in the expression for Λ estimated numerically for each particular search. Schuler, Atschul, and Lipman (1991) used an analog of the distribution defined by Equation (2.15) for the assessment of the statistical significance of the sum-of-pair score of the ungapped block of several amino acid sequences. Further developments and discussions of problems related to the score distributions can be found in Waterman and Vingron (1994), Altschul *et al.* (2001), Webber and Barton (2001), Bailey and Gribskov (2002) and Grossman and Yakir (2004). Statistical methods for sequence analysis were reviewed by Karlin (2005).

The distributions of statistics of pairwise alignment scores $N_S, S_{\max}, T_r, W, S^*$, or $\mathbf{S}_{\max}$ provide a rigorous basis for testing hypotheses on the relatedness of sequences X and Y. The following two tests for the null hypothesis,

$$H_0 = \{X \text{ and } Y \text{ are unrelated (independent) sequences}\},$$

versus the alternative hypothesis,

$$H_a = \{X \text{ and } Y \text{ are (evolutionary) related sequences}\},$$

could be considered. In Test 1, based on statistic $S_{\max}$, H_0 is accepted (H_a is rejected) if $S_{\max} < S^\circ$; otherwise, H_0 is rejected (H_a is accepted). In Test 2, based on statistic T_r, H_0 is accepted (H_a is rejected), if $|T_r - \mathbf{E}T_r| < \varepsilon^*$; otherwise, H_0 is rejected (H_a is accepted). Note that the cut-off values S° and ε^* should be determined *a priori* (before the actual sequences X and Y are observed).

Problem 2.18 Test 1 and Test 2 are applied to establish relatedness of locally aligned protein sequences X and Y of lengths $n = 100$ and $m = 300$. It is assumed that $K = 0.1, \lambda = 0.7$.