6.4 Statistics of sequence alignments

Beneath the surface of the sequence alignment programs lie two important applications of statistics. First, statistics play a key role in the construction of the similarity score matrices. Second, the evaluation of the significance the "best" alignment found by any sequence alignment algorithm also depends on statistics.

The BLOSUM family of similarity score matrices

BLOSUM is the acronym for blocks substitution matrix. The name comes from the fact that the values of the matrices come from a large collection of blocks of biologically similar proteins. S. Henikoff and J.G. Henikoff (1991, Nucleic Acid Res. 19, 6565-6572) designed an automated system, PROTOMAT, for obtaining a set of blocks given a group of related proteins. This system was applied to catalog of several hundred protein groups, yielding a database of more than 2000 blocks.

Each block in this database consists of a number (say, d) of aligned amino acid sequences. Suppose there are w columns in the alignment. We say that the block has depth d and width w. From each column of d amino acids, one can form

$$1+2+...+(d-1) = d(d-1)/2$$

unordered pairs of amino acids. For example, if a column contains nine alanines and one serine, one can form 10(9)/2 = 45 pairs, 36 of them are [*A*, *A*] and 9 [*A*, *S*]. Gap letters in the alignment will be ignored and no pair will be formed with gaps.

Exercise If a column contains 2 *A*'s, 1 *S*, 1 *T* and 1 _, list the possible pairs formed and their frequencies.

When the procedure is repeated on every column of every protein block in the database, we obtain the frequency counts of all the 210 (i.e., 1 + 2 +...+ 20) amino acid pairs. For simplicity, we shall index the amino acids from 1 to 20 in some convenient order, (say, alphabetically by names) and denote the frequency counts by f_{ij}, where i=1, ..., 20, j = 1,..., i. These frequency counts will be used to calculate the score matrix.

First, we need to calculate an "odds ratio" which is defined to be the ratio of the observed relative frequencies to the expected relative frequencies of the amino acid pairs. The observed relative frequencies are calculated as

$$q_{ij} = f_{ij} / \sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij} .$$

Let us pretend that the entire database has only that column of 9 *A*'s and 1 *S* described before, where f_{AA} = 36 and f_{AS} = 9. Then q_{AA} = 36/45 = 0.8 and q_{AS} = 9/45 = 0.2.

The expected relative frequencies p_{ij} are calculated based on a rolling-die model where all the amino acids in the protein block were generated independently. In such a model, we can write $p_{ij} = p_i p_j$ where p_i and p_j are the probabilities of observing the two individual amino acids in the database. These probabilities are estimated by the observed relative

frequencies. So, $\hat{p}_i = q_{ii} + \sum_{j \neq i} q_{ij} / 2$. Hence the expected relative frequency of the pair is estimated by

$$\hat{p}_{ij} = \begin{cases} \hat{p}_i^2 & i = j \\ 2\hat{p}_i \hat{p}_j & i \neq j \end{cases}$$

In the example, the expected relative frequency for [*A*, *A*] is 0.9 x 0.9 = 0.81, that of [*A*, *S*] is 2 x 0.9 x 0.1 = 0.18, and that of [*S*, *S*] is 0.1 x 0.1 = 0.01.

The odds ratio is then calculated where each entry is $q_{ij} / \hat{p}_{ij}$. The base 2 logarithm, measured in number of bits, of this odds ratio is referred to as a "lod ratio"

$$lod = \log_2(q_{ij} / \hat{p}_{ij}) .$$

The lod ratio is positive, zero, or negative according to the amino acid pair occurs more frequently than expected, just as frequently as expected, or less frequently than expected. A positive lod ratio indicates that the pair of amino acids frequently substitute for each other in proteins with like functions. The pair usually have similar molecular structures and biochemical functions. The lod ratios are multiplied by a scaling factor of 2 and then rounded to the nearest integer value to produce the values in a BLOSUM matrix in half-bit units.

To reduce multiple contributions to amino acid pair frequencies from the most closely related members of a family, sequences are clustered within blocks and each cluster is weighted as a single sequence in counting pairs (Henikoff, S., Wallace, J.C., and Brown, J.P., 1990, *Methods Enzymol.* 183, 111-132.). This is done by specifying a clustering percentage in which sequence segments that are identical for at least that percentage of amino acids are grouped together. The BLOSUM matrix computed from this reduced block of proteins is associated with that percentage. That is why we have the BLOSUM62, BLOSUM80 matrices, etc.

The clustering procedure is best explained by the example given in Henikoff and Henikoff (1991). Suppose the clustering percentage is set at 80%, and sequence A is identical to sequence B at ≥ 80% of their aligned positions, then A and B are clustered and their contributions are averaged in calculating pair frequencies. If C is identical to either A or B at ≥ 80% of aligned positions, it is also clustered with them and the contributions of A, B, and C are averaged, even though C might not be identical to both A and B at ≥ 80% of the aligned positions. In the above example, if 8 of the 9 sequences with A residues in the 9A-1S column are clustered, then the contribution of this column to the frequency table is equivalent to that of a 2A-1S column, which contributes 2[A,S] pairs.