Amino Acid Substitution Matrices

Dannie Durand

BLOSUM Matrices

The BLOSUM (BLOck SUbstitution Matrices) matrices were derived by Steven and Jorja Henikoff in 1992 [1]. They were based on a much larger data set than the PAM matrices, and used conserved local alignments or "blocks," rather than global alignments of very closely related sequences. The *"trusted" alignments* used to construct the BLOSUM matrices consisted of roughly 2000 blocks of conserved regions representing 500+ groups of proteins.

Here, we discuss the procedure for constructing a substitution matrix in the BLOSUM framework from a single aligned block. In reality, the BLOSUM matrices were constructed from many blocks. See Ewens and Grant, Section 6.5.2, for a detailed treatment of the BLOSUM matrices, including a discussion of how pair frequencies from multiple blocks are combined. Their treatment includes a worked example with more than one block. Note that their notation is somewhat different from the notation we use in class.

BLOSUM matrix construction uses clustering rather than an explicit evolutionary model, to account for different degrees of sequence divergence. Clustering with different values of n, ranging from 45% to 90%, produces a parameterized set of matrices representing different degrees of sequence divergence. In order to construct a BLOSUMn matrix, the sequences in each block are first grouped into clusters, such that the percent identity of any pair of sequences from different clusters is less than n%. Next, for every pair of clusters, amino acids pairs consisting of one amino acid from each cluster are tabulated. Pairs of amino acids within the same cluster are ignored. Amino acid pair counts are normalized by cluster size so that all clusters contribute equally to the pair statistics.

The clustering step in BLOSUM matrix construction has two purposes: parameterizing evolutionary divergence and accounting for sample bias. First, since only amino acids pairs sampled from two different clusters are tabulated, the data used to construct the matrix consists of amino acid pairs observed in sequences with a particular divergence (i.e., $< n$%). Second, to control for sample bias, the contribution of each residue in a cluster is normalized by the number of sequences in that cluster. As a result, each cluster contributes the same amount of information to the estimation of amino acid pair frequencies, even though clusters may contain different numbers of sequences.

[1] *Amino acid substitution matrices from protein blocks*, PNAS, 1992 Nov 15;89(22):10915-9

The specific procedure for BLOSUM matrix construction is as follows:

Partitioning sequences into clusters with n% identity: The clustering step takes as input a block of k sequences of length L (no gaps) and generates C non-overlapping clusters. The ith cluster, C_i, has k_i sequences of length L, where $k = \sum k_i$. The sequences in the block are partitioned in such a way that every sequence in a cluster is at least $n\%$ identical to at least one other sequence in the cluster. One way to obtain such a clustering is to represent the block as a weighted graph, where the nodes correspond to sequences. The nodes for each pair of sequences are connected by an edge that is weighted by their percent identity. To get $n\%$ clusters, all edges with weights lower than $n\%$ are removed, resulting in one or more connected components. Each connected component corresponds to a cluster. If n is greater than the greatest edge weight, then each cluster will contain a single sequence. If n is smaller than the lowest edge weight, then all sequences will be in a single cluster. If this happens, it is not possible to construct a BLOSUM matrix for this value of n.

Amino acid pair counts: Following the clustering step, the *observed* frequency of amino acid x aligned with amino acid y is calculated as follows. For each pair of clusters, we determine the number of x, y and y, x pairs, where x and y are in the same column, but in different clusters. Let $N_l(C_i, x)$ be the number of times that residue x appears in the l^{th} column of cluster C_i. Then the total number of x, y pairs between C_i and C_j is given by

$$A^n_{xy} = \sum_{i=1}^{C} \sum_{j>i} \sum_{l=1}^{L} \frac{N_l(C_i, x) \cdot N_l(C_j, y) + N_l(C_i, y) \cdot N_l(C_j, x)}{k_i \cdot k_j}, \quad (1)$$

where $x \neq y$. We use the superscript n to indicate that these are pair counts for a BLOSUMn matrix, where n is the threshold used in the clustering. When $x = y$, the pairs are only counted in one direction:

$$A^n_{xx} = \sum_{i=1}^{C} \sum_{j>i} \sum_{l=1}^{L} \frac{N_l(C_i, x) \cdot N_l(C_j, x)}{k_i \cdot k_j} \quad (2)$$

In both cases, the (x, y) pair counts are normalized by the number of possible inter cluster pairs, $k_i \cdot k_j$.

Estimating substitution frequencies: The frequencies of amino acid pairs are derived from the pair counts by normalizing by the total number of possible pairs; that is, by the product of the number of sites in the block and the number of pairs of clusters:

$$q^N_{xy} = \frac{A^N_{xy}}{L \cdot \binom{C}{2}}.$$

Estimating the expected pair frequencies: The expected frequency of x aligned with y is the product of the *background* probabilities of observing x and y independently. In PAM matrix construction, the background frequency of an amino acid is assumed to be the frequency of that amino acid in typical proteins, for example, as tabulated by Robinson and Robinson[2]. In contrast, in BLOSUM matrix construction, the expected frequencies are estimated from the BLOCK data and adjusted for the current value of n.

In order to get the *expected* frequency of x aligned with y, we first estimate the frequencies of the individual residues in the current block, again using the clusters to correct for sample bias. As above, the counts from each cluster are "discounted" by a factor of $1/k_i$, and then normalized by the total number of elements, $L \cdot C$, to obtain the amino acid background frequency:

$$p_x = \frac{1}{L \cdot C} \sum_{i=1}^{C} \sum_{l=1}^{L} \frac{N_l(C_i, x)}{k_i}.$$

The expected pair frequencies are then obtained from the products of the background frequencies:

$$\begin{aligned} E_{xy} &= p_x p_y + p_y p_x \\ E_{xx} &= p_x^2. \end{aligned}$$

Finally, the *log odds scoring matrix* is calculated from the ratios of the observed and expected frequencies:

$$S^N[x, y] = 2 \log_2 \frac{q_{xy}^N}{E_{xy}}.$$

[2] *Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins*, PNAS, 1991 Oct;88:8880-4