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. In order to account for different degrees of sequence divergence, the Henikoffs used clustering rather than an explicit evolutionary model. Clustering with different values of n, ranging from 45% to 90%, produces a parameterized set of matrices representing different degrees of sequence divergence. The clustering procedure also addressed the issue of sample bias.

In class, we discussed 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 discussion of the BLOSUM matrices. 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.

1. The *"trusted" alignments* used to construct the BLOSUM matrices consisted of roughly 2000 blocks of conserved regions representing 500+ groups of proteins. In contrast to the PAM alignments, which were full length alignments of closely related sequences, the BLOSUM matrices are based on locally conserved regions (ungapped blocks) in multiple alignments of sequences that were not highly conserved, overall.

2. *Amino acid pair counts:* In the BLOSUM matrix construction process, amino acid pair counts are obtained directly from columns in the conserved blocks (no trees.) In order to construct a BLOSUMn matrix, the sequences in each block were first grouped into clusters, such that any pair of sequences that were sampled from two different clusters has less than n% identity.

   The clustering step in BLOSUM matrix construction has two purposes: parameterizing evolutionary divergence, via the percent identity parameter, n, and accounting for sample bias. The use of clusters in tabulating the incidence of amino acid pairs in the data contributes to these goals in several ways. Each cluster is treated as an "average sequence." For every pair of clusters, only amino acids pairs consisting of one amino acid from each cluster are tabulated. Since pairs of amino acids within the same cluster are ignored, only pairs found in sequences that are less than n% identical are counted. This limits the data used to construct the matrix to amino acid pairs observed in sequences with a particular divergence. To control for sample

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

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.

Specifically, 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 clique, where the nodes correspond to sequences. Each edge (s, t) is weighted by the percent identity between the sequences s and t. 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 largest 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.

Following the clustering step, the *observed* frequency of x aligned with y is calculated as follows. For each pair of clusters, we sum the number of x, y 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

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

where $x \neq y$. When $x = y$, the pairs are only counted in one direction:

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

In both cases, the counts from C_i (resp., C_j) are normalized by k_i (resp., k_j), the number of sequences in the cluster.

3. *Estimating substitution frequencies* The frequency 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:

$$A_{xy} = \frac{a_{xy}}{L \cdot \binom{C}{2}}.$$

In order to get the *expected* frequency of x aligned with y, we first estimate the background frequency of the individual residues. As above, each residue in the cluster is "discounted" by

a factor of $1/k_i$, and then normalized by the total number of elements:

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

The expected pair frequencies are then obtained from the background pair frequencies, as follows:

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

4. Finally, the *log odds scoring matrix* calculated from the observed and expected frequencies:

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

Comparing PAM and BLOSUM Matrices

	PAM	BLOSUM
Evolutionary model	Explicit evolutionary model	None
Data	Full length MSAs of closely related sequences.	Conserved blocks in protein
Bias correction	Trees	Clustering
Evolutionary distance	From Markov model of sequence evolution.	From clustering of sequences.
Matrices	Transition and log odds scoring matrices	Log odds scoring matrix only.
Parameter n	Distance increases with n	Distance decreases with n
Biophysical properties	Derived indirectly from data	Derived indirectly from data

The PAM and BLOSUM matrices were constructed from an evolutionary model and conserved blocks where amino acids are under selective constraints, respectively. Nevertheless, the matrices favor replacement of amino acids that share biochemical properties. Inspection of the BLOSUM62 matrix shows that alignments of residues in the same biochemical group tend to have positive log odds scores. These residues are more likely to be observed together in alignments of related sequences than by chance. Residues from different groups tend to have negative scores. These residues are less likely to be observed together in related sequences than in chance alignments. A score of zero means that this pair of residues is equally likely in related and chance alignments.

% Identity	PAM	BLOSUM
20	250	45
30	160	62
40	120	80
50	80	-
60	60	-