BLOSUM Lecture Notes

Dannie Durand

BLOSUM Matrices

See Ewens and Grant, 6.5.2. for a detailed discussion of how the BLOSUM matrices are computed. Note that their notation is slightly different.

Overview

- BLOSUM = BLOck SUbstitution Matrices, (Henikoff and Henikoff, 1992).
- Trusted Data
  - Much larger dataset ($\sim$ 2,000 blocks)
  - Local MSA's (ungapped blocks). Highly conserved due to selective pressure.
- Compute pairwise amino acid alignment counts
  - Count amino acid replacement frequencies directly from columns in blocks (no trees.)
  - General idea:
    - Cluster sequences that are n% identical.
    - Count amino acid pairs across clusters, treating clusters as an "average sequence". Normalize by the number of sequences in the cluster.
    - Do not count amino acid pairs within a cluster.

Specifics of calculating the Blosum$n$ log odds matrix

- Input: B blocks of sequences. Each block b contains k_b sequences of length n_b (no gaps).
- Cluster sequences such that within each cluster, each sequence is at least n% identical to one other sequence in the cluster.
- Let C_b be the number of clusters in block b following the clustering step, where the ith cluster C_{b_i} has k_{b_i} sequences ($k_b = \sum k_{b_i}$).

1) The *observed* frequency of *x* aligned with *y* is calculated as follows:

$$A_{xy}^{b} = \frac{\sum_{i=1}^{C_b}\sum_{j>i}\sum_{l=1}^{n_b}(\#\,x\text{'s at site } l \text{ in } C_{b_i})\cdot(\#\,y\text{'s at site } l \text{ in } C_{b_j}) + (\#\,y\text{'s at site } l \text{ in } C_{b_i})\cdot(\#\,x\text{'s at site } l \text{ in } C_{b_j})}{k_{b_i}\cdot k_{b_j}}$$

$$A_{xy} = \frac{\sum_{b=1}^{B} a_{xy}^{b}}{\sum_{b=1}^{B} n_b \cdot \binom{C_b}{2}}$$

2) The *expected* frequency of *x* aligned with *y* is calculated as follows:

$$\begin{aligned} p_x^b &= \sum_{i=1}^{C_b}\sum_{l=1}^{n_b}\frac{(\#\,x\text{'s at site } l \text{ in } C_{b_i})}{k_{b_i}} \\ p_x &= \frac{\sum_{b=1}^{B} p_x^b}{\sum_{b=1}^{B} n_b \cdot C_b} \\ E_{xy} &= p_x p_y + p_y p_x \\ E_{xx} &= p_x^2 \end{aligned}$$

3) Calculate the log odds matrix from the observed and expected frequencies:

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

Blosum*x* matrices:

- Sequences that are *n*% identical were clustered during the construction of the matrix.
- Corrects for sample bias.
- Parameter of evolutionary divergence.

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 which share biochemical properties. Inspection of the BLOSUM 62 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 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.

Seq Indentity	PAM	BLOSUM
20	250	45
30	160	62
40	120	80
50	80	-
60	60	-