PAM and BLOSUM Matrices

Prepared by: *Jason Banich and Chris Hoover*

Background

As DNA sequences change and evolve, certain amino acids are more likely to mutate into other amino acids.

PAM and BLOSUM are substitution matrices, which means that they describe the rate at which one character in a sequence is replaced by another character. This is applicable to amino acids and nucleotides, since they can be represented by a single character.

PAM was published in 1966 by Margaret Dayhoff. BLOSUM is more recent, it has been around since it was published in 1992 by Henikoff.

PAM

PAM stands for **P**oint **A**ccepted **M**utation. PAM is a scoring matrix for sequence alignment, calculated from very similar sequences of DNA by measuring their differences.

Sequences are defined as 'one PAM unit diverged' if the series of accepted mutations converted S_1 and S_2 with an average of one point mutation per 100 amino acids.

PAM matrices are represented by a number, such as PAM 250. This number means that the series have diverged by 250 PAM units [1].

Calculation for PAM

In order to produce a n PAM matrix, many distinct pairs of sequences must be collected that are known to diverge by n PAM units. The sequences must then be aligned in pairs, and then for each amino acid pair (i and j), count the number of times that they align opposite each other, and divide by the total number of pairs in the aligned data, where $f(i, j)$ is the resulting frequency. Also let $f(i)$ be the number of times that i appears in all of the sequences. The

entry for that i, j entry is given by the following equation:

$$log \frac{f(i,j)}{f(i)f(j)}$$

Doing so normalizes the replacement frequency by the frequency that is expected due to chance alone. The table will now contain entirely probabilities.

This is only for the ideal case though, where it's very easy to determine where insertions and deletions occurred. In order to solve this, Dayhoff used extremely similar sequences that were believed to have diverged from a common ancestor very little. In doing so, she was able to construct a PAM1 matrix. To convert that to a PAM 250 matrix, you simply multiply the matrix by itself 250 times. PAM 250 is more accurately written as PAM^{250} [1].

Usage of PAM

Different levels of PAM are used for different purposes. PAM 250 is known for being good when doing a database search, whereas PAM 40 is known to be good for a nucleotide sequence. PAM 250 happens to correspond to sequences being about 20% identical, having diverged 250 mutations per 100 amino acids of the sequence [4].

BLOSUM

BLOSUM is short for **BLO**cks **SU**bstitution **M**atrix. Like PAM matrices, BLOSUM matrices are scoring matrices for sequence alignment. However, unlike PAM matrices, BLOSUM matrices are calculated from databases of local alignments, using alignments that have been observed already. [2]

Constructing a BLOSUM matrix

This section's content is taken from the construction procedure given in [2].

Outline of algorithm

1. Cluster similar sequences into one sequence.
2. Compute the probability of observing each amino acid pair (X, Y) in our sample.
3. Compute the expected probability of observing a pair (X, Y) based on the probabilities of observing X and Y individually.
4. Use the ratio of the probabilities from the previous two steps to derive a score for each pair (X, Y). These scores are used as the BLOSUM matrix's entries.

Detailed procedure

A BLOSUM matrix is calculated based on a database of amino acid alignment blocks. See Figure 1.

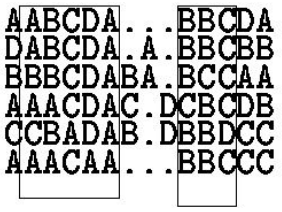

Figure 1: Blocks in amino acid segments

1. Given a percentage x, within blocks, cluster segments at least x% identical into a single sequence. This is done to avoid multiple contributions from very similar sequences. This can be done by either removing sequences from the block or by representing the cluster with a new sequence. The parameter x is reflected in the name of the BLOSUM matrix computed using that value. For example, if x = 62, the resulting BLOSUM matrix is called BLOSUM-62.

2. Compute observed probability of a pair (i, j) in our database where i and j are amino acids. If the number of (i, j) pairs in our database is f_{ij}, the probability of observing the $pair_{i,j}$ is:

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

   In other words, this probability is computed as the number of i, j pairs in our database divided by the total number of amino acid pairs in our database.

3. Compute the expected probability of observing a pair (i, j) in our database. As an intermediate calculation, we compute the probability that amino acid i will occur in a pair i, j. This is given by pi, where:

$$p_i = q_{ii} + \sum_{j \neq 1} q_{ij}/2$$

   Either ordering of the acids i and j will represent the same alignment, so i,j and j, i are considered equivalent. Thus, when i != j, the expected probability of a pair i, j, e_{ij}, is:

$$p_i p_j + p_j p_i = 2 p_i p_j$$

   When i = j, e_{ij} is given by:

$$p_i p_j$$

4. Compute a logarithm of odds (lod) matrix. For each pair i, j, we use the previously calculated probabilities to obtain the ratio shown in

$$q_{ij}/eij$$

   Then, for each such ratio, we compute the lod matrix's entries as:

$$s_{ij} = log_2(q_{ij}/e_{ij})$$

These values are then multiplied by 2 and rounded to the nearest integer. The resulting matrix is a BLOSUM matrix.

BLOSUM Performance Compared to PAM

This section's discussion, data and figures are taken from [3].

Experiment 1: Comparing results to alignments seen in 3D structures

MUTALIN, a hierarchical multiple-alignment program, was tested using several PAM and BLOSUM matrices, along with a simple +6/-1 matrix where a match is assigned a +6 score and a mismatch is assigned a -1 score. BLOSUM matrices produced the most accurate alignments, as can be seen in Figure 2.

Matrix aligned	Program	Residue positions missed*	
		All positions	Side chains
	MSA	12	6
PAM 120	MULTALIN	31	22
PAM 160	MULTALIN	30	22
PAM 250	MULTALIN	30	22
+6/−1	MULTALIN	34	26
BLOSUM 45	MULTALIN	9	5
BLOSUM 62	MULTALIN	6	4
BLOSUM 80	MULTALIN	9	6

Figure 2: Results of multiple alignment experiment with PAM and BLOSUM matrices

Experiment 2: Database searches

BLAST, FASTA, and Smith-Waterman programs were tested, using both PAM and BLOSUM matrices, against a set of sample queries. As in the previous experiment, BLOSUM matrices demonstrated superior performance. (See Figure 3)

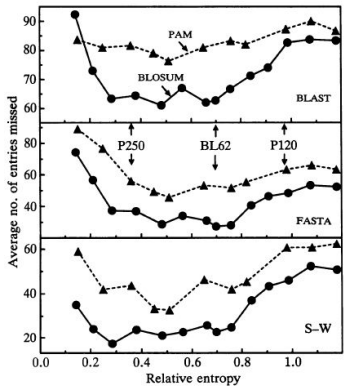

Figure 3: Database query accuracy using PAM and BLOSUM matrices.

References

[1] http://cs124.cs.ucdavis.edu/lectures/scoringmatrices.pdf

[2] http://www.bioinfo.rpi.edu/bystrc/courses/biol4540/lecture5.pdf

[3] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992;89:1091510919.

[4] http://www.ebi.ac.uk/help/matrix.html