Lecture No. 7
Blast and Substitution Matrix

February 3, 2003

- BLAST: Basic Local Alignment Search Tool
  A set of similarity search programs for proteins or DNA sequences (BLASTN, BLASTP,–) developed at NCBI.

1). Searches databases for related sequences.

2). Seeking local alignment is able to detect relationships among sequences that share only isolated regions of similarity.

- Biological implication of greater than 30% sequence identity:

1). Structural similarity (usually functional similarity).

2). A common ancestor.

Statistics of Align Scores

To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone. In this context, "chance" can mean the comparison of

(i) real but non-homologous sequences;

(ii) real sequences that are shuffled to preserve compositional properties or

(iii) sequences that are generated randomly based upon a DNA or protein sequence model.

A local alignment without gaps consists simply of a pair of equal length segments, one from each of the two sequences being compared. It is possible to express the score of interest in terms of standard deviations from the mean; it is a mistake to assume that the relevant distribution is normal and convert the Z-value into a p-value; the tail behavior of global alignment scores is unknown. The most one can say reliably is that if 100 random alignments have score inferior to the alignment of interest, the p-value in question is likely less than 0.01.

Distribution of Scores:

If scores of segment pairs can not be improved by extension or trimming. These are called high-scoring segment pairs or HSPs. The maximum of a large number of independent identically distributed random variables tends to an extreme value distribution.
$P(S < x) = exp[-exp(-x)]$

$P(S \geq x) = 1 - exp[-exp(-x)]$

In the limit of sufficiently large sequence lengths m and n, the statistics of HSP scores are characterized by two parameters, K and lambda.

$$p(S > x) = 1 - \exp\left(-kmn \exp\left(-\lambda x\right)\right)$$

This comes from extreme value distribution and S is a given threshold score. k and λ are constants that can be calculated from substitution matrix.

- Most simply, the expected number of HSPs with score at least S is given by the E-value for the score S:

$$E = Kmn \exp(-\lambda S)$$

- Doubling the length of either sequence should double the number of HSPs attaining a given score.

- For an HSPs to attain the score 2 times it must attain the score times twice in a row, so one expects E to decrease exponentially with score.

- p-value: The number of random HSPs with score $\geq S$ is described by a poisson distribution. This means that the probability of finding exactly a HSPs with scores $\geq S$ is given by:

$$p(HSPs = a) = \frac{e^{-E} E^a}{a!}$$

where E is the E-value of S.

- The chance of finding zero HSPs with score $\geq S$ is e^{-E}, so the probability of finding at least one such HSP is

  $$p = 1 - e^{-E}$$

  This is the p-value associated with the score S.

- For example, if one expects to find three HSPs with scores $\geq S$, the probability of finding at least one is 0.95. The BLAST programs report E-value rather than p-values because it is easier to understand the difference between, e.g., E-value 5 and 10 than p-values of 0.993 and 0.99995. However, when $E < 0.01$, p-values and E-value are nearly identical.

- The E-values we were studying was related to the comparison of two sequences of length m and n. How does one assess the significance of an alignment that arises from the comparison of a sequence of length m to a database containing many different sequences of varying lengths? There are two views:

- All sequences in a database are a priori equally likely to be related to the query. This implies that a low E-value for an alignment involving a short database sequence should carry the same weight as a low E-value for an alignment involving a long database sequence. To calculate a database search E-value, simply multiply the pairwise-comparison E-value by the number of sequences in the database.

Recent version of FASTA protein comparison programs take this approach.

- A query is a priori more likely to be related to a long than a short sequence, because long sequences are often composed of multiple distinct domains. If one assumes that a priori chance of relatedness is proportional to sequence length, then the pairwise E-value involving a database sequence of length n should be multiplied by N/n, where N is the total length of the database in residues. This can be done simply by treating the database as a single long sequence of length N.
  The BLAST programs take this approach to calculating database E-value.

- A key element in evaluating the quality of a pairwise sequence alignment is the Substitution Matrix, which assign a score for aligning any possible pair of residues. Two popular set of matrices:

1. PAM Matrices (Dayhoff et al. 1978).

2. BLOSUM Matrices (Henikoff and Henikoff 1992).

   Both try to capture the relative substitutability of sequence pairs in the context of evolution.

- Substitution Matrix Motivation:

1. Consider simplest ungapped global alignment of two sequences x and y of length n.

2. In scoring this alignment one would like to assess

$$\frac{p(x, y \mid M)}{p(x, y \mid R)} = \frac{p(x, y \mid \text{ seq. has comm. ancestor})}{p(x, y \mid \text{ aligning by chance})}$$

Substitution matrix is constructed by estimating this ratio.

- Substitution Matrix: Basic Ideas:

1. Let q_a be the frequency of amino acid a.

2. Consider the case where alignment of x and y is random

$$p(x, y \mid R) = \prod_i q_{x_i} \prod_i q_{y_i}$$

3. Let p_{ab} be the probability that a and b derived from a common ancestor.

4. Then the case where the alignment is due to common ancestry is

$$p(x, y \mid M) = \prod_i p_{x_i y_i}$$

Joint probability with independence.

5. The odds ratio of these two alternatives is given by:
$\frac{p(x,y|M)}{p(x,y|R)} = \frac{\prod_{i=1} p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}}$

6. Taking the log, we get

$$\begin{aligned} \log \frac{p(x, y \mid M)}{p(x, y \mid R)} &= \log \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} q_{y_i}} \\ &= \sum_i \log \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \end{aligned}$$

7. The score from alignment is given by :

$$\begin{aligned} S &= \sum_i S(x_i, y_i) \\ &= \log \frac{p(x, y \mid M)}{p(x, y \mid R)} \end{aligned}$$

8. The substitution matrix score for a pair a and b in then given by:

$$S(a, b) = \log \frac{p_{ab}}{q_a q_b}$$

1. QUESTION: How do we get values for p_{ab} (probability that a and b arose from a common ancestor)?

2. It depends on how long ago sequences diverged. If diverged recently then

$$p_{ab} = 0 \text{ for } a \neq b.$$

If diverged long ago :

$$p_{ab} = p_a p_b$$

There are two approaches PAM approach and BLOSUM approach.

- PAM approach —Percent Accepted Mutation

1 PAM is the time over which one amino acid replacement is expected to occur in a 100 amino acid polypeptide chain of average composition.

1. Observed

A C G H D B G H A D I J C B I J

1A—A 1B—B 1A—A 1B—B

1G—G 1G—G 1I—I 1I—I

1H—H 1H—H 1J—J 1J—J

1C—B 1A—D 1D—B 1C—A

inferred

A B G H A B I J

1A—A 1A—A

1B—B 1B—B

1G—G 1I—G

1H—H 1J—H

inferred

A B G H

$$\begin{pmatrix} & A & B & C & D & G & H & I & J \\ A & 8 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ B & 0 & 8 & 1 & 1 & 0 & 0 & 0 & 0 \\ C & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ D & 1 & 1 & 0 & 6 & 0 & 0 & 0 & 0 \\ G & 0 & 0 & 0 & 0 & 6 & 0 & 1 & 0 \\ H & 0 & 0 & 0 & 0 & 0 & 0 & 4 & \\ I & & & & & & & & 4 \\ J & & & & & & & & \end{pmatrix}$$

A_{ij} = number of times amino acid j mutates to amino acid i.
A mutation could go in both directions, therefore tally of mutation $I|J$ enters both A_{IJ} and A_{JI} enteries, while the tally of conservation $A|A$ enters A_{AA} enter twice.

2. Relative Mutability of Amino Acid j:

$$m_j = \frac{\sum_{i=1,20, i\neq j} A_{ij}}{\sum_{i=1,20} A_{ij}}$$

m_j is the probability that amino acid j will change in a given evolutionary interval. The absolute value of m_j depends on how similar the sequences used in tally A_{ij} are, but the relative values do not.

m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8
$\frac{2}{10}$	$\frac{2}{10}$	0	0	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{5}$	$\frac{1}{5}$

3. PAM-1 Probability Transition Matrix:

M_{ij} = probability of amino acid j changing into i in the evolutionary period (1 PAM).

$$\begin{aligned} M_{ij} &= 1 - \lambda m_j \\ &= \frac{A_{ij}}{\sum_{i=20} A_{ij}} \lambda \end{aligned}$$

λ is scaling constant to make sure that the total mutation rate is 1%.

$$\lambda \sum_{j=1,20} p_j m_j = 1\%$$

p_j = probability of random occurance of amino acid j

$$p_j = \frac{\sum_{i=1,20} A_{ij}}{\sum_i \sum_j A_{ij}}$$

where $\sum_i \sum_j A_{ij} = 48$

j	1	2	3	4	5	6	7	8
p_j	$\frac{10}{48}$	$\frac{10}{48}$	$\frac{2}{28}$	$\frac{2}{48}$	$\frac{7}{48}$	$\frac{7}{48}$	$\frac{5}{48}$	$\frac{5}{48}$

Now find λ and then M matrix is:

$$\begin{vmatrix} .048 & 0 & .03 & .03 & 0 & 0 & 0 & 0 \\ 0 & .048 & .03 & .03 & 0 & 0 & 0 & 0 \\ .03 & .03 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{36}{7} & 0 & \frac{.06}{5} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{.36}{7} & 0 & \frac{.06}{5} \\ 0 & 0 & 0 & 0 & \frac{.06}{7} & 0 & \frac{.24}{5} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{.06}{7} & 0 & \frac{.24}{5} \end{vmatrix}$$

4. PAM log-odds score matrix for k units of evolution:

$$S_{ij} = 10 \log_{10} \frac{(M^k)_{ij}}{p_i}$$

5. The score from an alignment is thus given by

$$S = \sum_i S_{ij}$$

BLOSUM Matrices: Blocks substitution matrices

Local alignments can be represented as ungapped blocks, where rows are the protein segments within protein sequences and columns are the aligned positions.

1. For a block of width w amino acids, and depth s sequences, there are $s(s-1)/2$ pairs of amino acids for each column and $ws(s-)/2$ pairs overall.

2. Let f_{ij} be the amino acid types i,j. The proportion of i,j pairs is:

$$q_{ij} = \frac{f_{ij}}{\sum_{i,j} f_{ij}}$$

3. The expected proportion under independence is:

$$e_{ij} = \begin{cases} p_i^2 & i = j \\ 2p_i p_j & i \neq j \end{cases}$$

where

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

4. The i, j^{th} element of the BLOSUM matrix is defined as:

$$s_{ij} = 2 \log_2(q_{ij}/e_{ij})$$

5. The BLOSUM62 matrix is derived from a database of blocks in which sequence segments that are identical for at least 60% of aligned positions are clustered.

There are two views:

- To calculate a database search E-value, simply multiply the pairwise-comparison E-value by the number of sequences in the database.
  Recent version of FASTA protein comparison programs take this approach.

- If one assumes that a priori chance of relatedness is proportional to sequence length, then the pairwise E-value involving a database sequence of length n should be multiplied by N/n, where N is the total length of the database in residues. This can be done simply by treating the database as a single long sequence of length N.