Shortened Glossary of Terms from the BLAST NCBI Web-Site.

Bioinformatics

The merger of biotechnology and information technology with the goal of revealing new insights and principles in biology.

Proteomics

Systematic analysis of protein expression of normal and diseased tissues that involves the separation, identification and characterization of all of the proteins in an organism.

****

Alignment:

The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

Bit score

The value S' is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.

BLAST

Basic Local Alignment Search Tool. (Altschul et al.) A sequence comparison algorithm optimized for speed used to search sequence databases for optimal local alignments to a query. The initial search is done for a word of length "W" that scores at least "T" when compared to the query using a substitution matrix. Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S". The "T" parameter dictates the speed and sensitivity of the search. For additional details, see one of the BLAST tutorials (Query or BLAST) or the narrative guide to BLAST.

BLOSUM

Blocks Substitution Matrix. A substitution matrix in which scores for each position are derived from observations of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62% identity. Sequences more identical than 62% are represented by a single sequence in the alignment so as to avoid over-weighting closely related family members. (Henikoff and Henikoff)

Domain

A discrete portion of a protein assumed to fold independently of the rest of the protein and possessing its own function.

E value

Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

FASTA

The first widely used algorithm for database similarity searching. The program looks for optimal local alignments by scanning the sequence for small matches called "words". Initially, the scores of segments in which there are multiple word hits are calculated ("init1"). Later the scores of several segments may be summed to generate an "initn" score. An optimized alignment that includes gaps is shown in the output as "opt". The sensitivity and speed of the search are inversely related and controlled by the "k-tup" variable which specifies the size of a "word". (Pearson and Lipman)

gap

A space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.

Homology

Similarity attributed to descent from a common ancestor.

HSP

High-scoring segment pair. Local alignments with no gaps that achieve one of the top alignment scores in a given search.

Identity

The extent to which two (nucleotide or amino acid) sequences are invariant.

Motif

A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domains.

Multiple Sequence Alignment

An alignment of three or more sequences with gaps inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column. **Clustal W** is one of the most widely used multiple sequence alignment programs

Orthologous

Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.

P value

The probability of an alignment occurring with the score in question or better. The p value is calculated by relating the observed alignment score, S, to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant P values will be those close to 0. P values and E values are different ways of representing the significance of the alignment.

PAM

Percent Accepted Mutation. A unit introduced by Dayhoff et al. to quantify the amount of evolutionary change in a protein sequence. 1.0 PAM unit, is the amount of evolution which will change, on average, 1% of amino acids in a protein sequence. A PAM(x) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have experienced a certain amount (x) of evolutionary divergence.

Paralogous

Homologous sequences within a single species that arose by gene duplication.

PSI-BLAST

Position-Specific Iterative BLAST. An iterative search using the BLAST algorithm. A profile is built after the initial search, which is then used in subsequent searches. The process may be repeated, if desired with new sequences found in each cycle used to refine the profile. Details can be found in this discussion of PSI-BLAST. (Altschul et al.)

Substitution Matrix

A substitution matrix containing values proportional to the probability that amino acid i mutates into amino acid j for all pairs of amino acids. such matrices are constructed by assembling a large and diverse sample of verified pairwise alignments of amino acids. If the sample is large enough to be statistically significant, the resulting matrices should reflect the true probabilities of mutations occuring through a period of evolution.