**Course: Advanced Bioinformatics**

**Module title: Blast**

**Module no. : 34**

**Basic Local Alignment Search Tool**

Blast has supplanted FASTA as the most commonly used database search tool. BLAST was developed as an improvement in speed from the FASTA suite without a sacrifice in sensitivity.

The first step of the BLAST algorithm is to locate common words or k-tuples in the query sequence and the target database sequences. However, BLAST does not search for every possible k-tuple, it only considers those that are most significant. For the NCBI BLAST program, the word length is fixed at 3 for proteins and 11 for nucleic acids. This k-tuple is referred to as the word-length, and is the minimum length needed to achieve a word score that is high enough to be significant but not so long as to miss short but significant patterns.

MSP – Maximal Segment Pair: The highest scoring pair of identical length segments chosen from two sequences. The boundaries of an MSP are chosen to maximize its score, so an MSP can be of any length.

The number of MSP scores with a score greater than a cutoff score S are reported.

BLAST minimizes the time spent on sequence regions where the score is unlikely to exceed this cutoff score.

The main strategy of BLAST is to seek only segment pairs that contain a word pair with a score of at least T. Any such hit is extended to determine if it is contained within a segment pair whose score is greater than or equal to the cutoff score S.

The scanning phase of BLAST locates the words within the sequences in linear time. One method is to map each possible word to an integer so that it can be used as an index into an array. For instance, if the word size was 4, and amino acids were used, there are 204 = 160,000 entries in the array. The second approach was the use of a deterministic finite state automaton.

**Hit Extension**

Initial hits are then examined, and extended in either direction until they fall below a certain score threshold.

In order to get around the problem of using uninformative hits, BLAST stores a list of words that are found much more often than expected by random. Hits to these words are discarded from consideration

**Steps Used by BLAST**

1. The sequence is optimally filtered to remove low-complexity regions that will not lead to meaningful sequence alignments.
2. A list of words of the predefined word length (3 for amino acids; 11 for DNA sequences) in the query sequence is made.
3. The query words are evaluated for an exact match with a word in any database sequence, using substitution scores for amino acids, and +5,-4 scoring scheme for DNA.
4. A cutoff score, called neighborhood word score threshold (T) is selected to reduce the number of possible matches to the word to be the most significant ones. This pares down the list of possible matching words to those resulting in the most significant alignments.
5. The procedure is repeated for each word in the query sequence.
6. The remaining high-scoring words for each possible match to a word are organized into an efficient search tree.
7. Each database sequence is scanned for an exact match to one of words in the search tree, one position at a time. If a match is found, it is used to seed a possible un-gapped alignment between the query and database sequences.
8. (UNGAPPED BLAST – VERSION 1.0) In the original BLAST suite of programs, an attempt is made to extend an alignment from the matching words in each direction along the sequences, as long as the score does not drop below a certain threshold. At this point, a larger stretch of sequence (called the HSP (high-scoring segment pair) which has a larger score than the original word may have been found.

(GAPPED BLAST – VERSION 2.0) In the newer version of BLAST, the neighborhood word threshold T is reduced in order to find shorter matching word hits that can be aligned along the same diagonal

1. The score of each HSP is compared against a cutoff score S, which is empirically determined.
2. The statistical significance for each HSP is calculated using the Karlin-Altschul statistics and the extreme value distribution, as previously discussed with sequence alignments. Recall that the probability, p, of observing a score S greater than or equal to x is given by the equation:



where



and m’ and n’ are the effective lengths of the query and database, such that





where H is the average expected score per aligned pairs of residues in an alignment of two random sequences; m and n are the length of the query and database; K and lambda are parameters calculated based on the sequences and the scoring scheme.

These effective, or reduced, lengths are used as a correction factor in order to allow alignments starting near the end of one of the sequences to be detected.

The expectation, E, of seeing a score S >= x in a database of D sequences is approximately given by the Poisson distribution,



1. Two or more HSP regions may be combined to a longer alignment region, even though the individual HSPs may result in a lower score.
2. Smith-Waterman type alignments are shown for the query sequence with each of the matched sequences in the database. BLAST-2 can produce alignments with gaps, while BLAST-1 cannot.
3. When the expected score for a given database sequence satisfies the threshold for E, the match score is reported.