**BLAST (Basic Local Alignment Search Tool)**

**The BLAST algorithm**

* **One of the most well known and commonly used tools for searching sequence databases.**
* **Introduced by S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman in the early 1990s**
* **The original BLAST algorithm searches a sequence database for maximal ungapped local alignments.**
* **BLAST finds subsequences from the database that are similar to subsequences in the query sequence.**
* **Several variations of BLAST are available**
* **BLASTP algorithm – searching for protein sequences matches using PAM (Point Accepted Mutation) or BLOSUM matrices to score the ungapped alignments.**
* **The most recent versions of BLAST insert gaps to optimize the alignment**
* **PSI – BLAST - summarize the results of sequence searches into position-specific scoring matrices, which are useful for protein modeling and structure prediction**

The BLAST algorithm is a heuristic search method that seeks words of length W (default = 3 or 4 in BLASTP) that score at least T (threshold) when aligned with the query and scored with a substitution matrix. Words in the database that score T or greater are extended in both directions in an attempt to find a locally optimal un-gapped alignment or HSP (high scoring pair) with a score of at least S or an E value lower than the specified threshold. HSPs that meet these criteria will be reported by BLAST, provided they do not exceed the cutoff value specified for number of descriptions and/or alignments to report.

Example:

Input Query Sequence: AILVPTV

1. Break the query sequence into words, or subsequences of a fixed length (4 is a default word length). All possible words in the query sequence are calculated:

AILV

ILVP

LVPT

VPTV

1. Words composed mostly of common amino acids will be discarded
2. The sequences in the database are then searched for occurrences of the search words

AILV

MVQGWALYDFLKCRAILVGTVIAML…….

1. Each time a word match is found in the database, the match is extended in both directions from the matching word until the alignment score falls below a given threshold (T). Since the alignment is un-gapped, the extension only involves adding additional residues to the matching region and recalculating the score according to the scoring matrix.

5. AILVPTV

MVQGWALYDFLKCRAILVGTVIAML…….

The choice of the threshold is an important search parameter, because it determines how likely the resulting sequences are to be biologically relevant homologous of the query sequence.

Key ideas in BLAST searches:

* BLAST computes "similarity", not "alignment".
* Given two protein sequences:
  + Find all substrings of length *k* (typical: *k = 4*) that occur in both strings.
  + Build score based on matches.
  + Extend substrings to see if match score can be increased.
  + Compute total score when no more extensions are possible.
* Compute BLAST score against all proteins in database.
* Rank order search results by score.

Practice:

1. BLAST one sequence from HIV lab
2. Use Specialized BLAST to compare 2 sequences from the same subject from HIV lab
3. Human Actin Gene (FLNA), filamin-A isoform 2, use FASTA to get a sequence, BLASTn (DNA sequence), Distance Tree, Cladogram
4. Use sequence in shared files named BLAST.txt to perform BLASTp (protein sequence)
5. Use sequence in shared files named BLAST1.txt to perform BLASTn (nucleotide sequence)
6. Use sequence in shared files named BLAST2.txt to perform BLASTx (nucleotide sequence to protein sequence search)