Basic Local Alignment Search Tool

Diploma in Bioinformatics

Jain Institute of Vocational and Advanced Studies

May 16, 2007

Outline

About BLAST

- A computer program that is a widely used homology search tool
  - Computer program = Implementation of an algorithm.
  - An Algorithm is a set of rules for solving a problem in a finite number of steps.
    - Series of well stated instructions to be performed by the computer.
    - These instructions are stated in a programming language.
    - Machine and Language independent
- BLAST is a heuristic approach based on Smith Waterman algorithm
  - Heuristic = solution quality or the running time improved using approximations.
- Finds best local alignments
- Provides statistical significance of the results

About BLAST

Reference

- The BLAST programs are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query
- The original BLAST algorithm[1] was written balancing speed and increased sensitivity for distant sequence relationships
- Speed is a critical issue since databases are increasing in size at exponential rates

[1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers and David J. Lipman. *Basic local alignment search tool*. J Mol Biol 1990 5:403-10

About BLAST

Why Local Alignments?

- Instead of relying on global alignments, BLAST emphasizes regions of local alignment to detect relationships among sequences which share only isolated regions of similarity

Example

Domains are shared between proteins, amino acids surrounding the active site are conserved compared to the other sites..

Comparison Methods Available

- *blastp* - Protein - Protein BLAST
  - compares a protein query with a protein database
- *blastn* - Nucleotide - Nucleotide BLAST
  - compares a nucleotide query with a nucleotide database
- *blastx* - Translated BLAST
  - Compares a nucleotide query sequence translated in all reading frames against a protein sequence database
  - To find potential translation products of an unknown nucleotide sequence
  - This also helps to identify sequencing errors that change the ORF

Comparison Methods Available

- *tblastn* - Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames

Application

If you have predicted the ORF for your sequence, you can search the nucleotide database to determine if any sequence encodes a similar protein

Comparison Methods Available

- *tblastx* - Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Application

If you aren't sure of the ORFs particularly in a genomic sequence, you can search a nucleotide database dynamically translated to see if a similar protein is encoded in other known sequence.

Databases Searched

1. Proteins
    1. non-redundant - nr (GenBank CDS translations + PDB + SwissProt + PIR + PRF)
    2. month(last 30 days), swissprot, patents, pdb
2. Nucleotides
    1. non-redundant - nr(GenBank + EMBL + DDBJ)
    2. month, dbest,dbsts,yeast,E.coli, patents,mito,vector,gss,htgs etc.,

Search Procedure

Summary

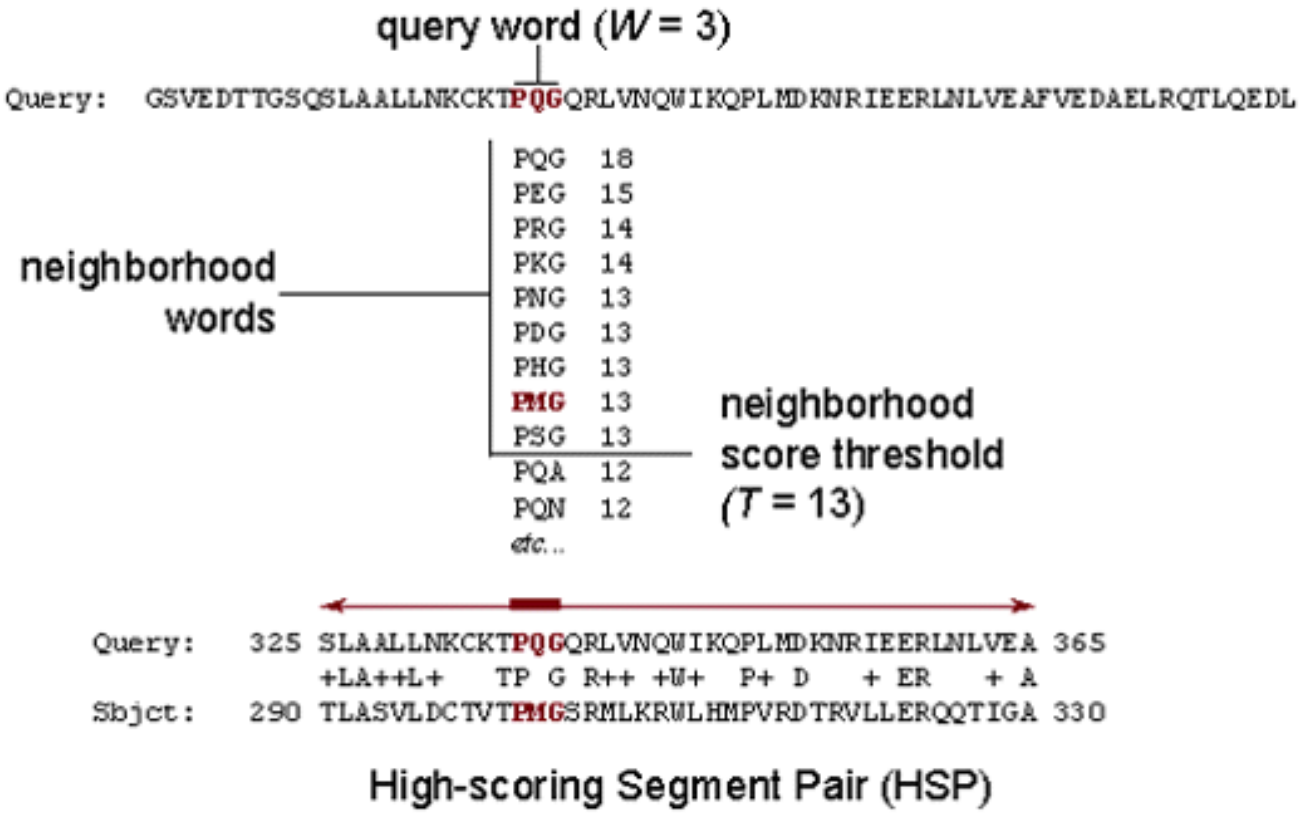
- The initial scanning phase identifies matching [query:database] fragments.
- A match is determined by the sum alignment score for a region (defined as a "word") of the query sequence.
- The alignment for each base in the word is scored
- if a nucleotide in the query word exactly matches a nucleotide at the same position in the database word (e.g. A with A), then a positive score is awarded.
- If two nucleotides do not match, a negative score is awared. The sum score is used to determine the degree of similarity.
- Sequences with a high score are referred to as high-scoring segment pairs (HSPs).

Search Procedure

Summary

- The program tries to extend the best HSP (those with the highest score; the best matches) by extending the alignment in both directions.
- The alignment extension is continued until the sequence ends, or the alignment becomes non-biologically significant.
- Substitution matrices are used during both scanning and extension.
- The reported sequences are those with the overall highest scores (maximal-scoring segment pair, MSP).

The BLAST Search Algorithm



Scoring

- The quality of each pair-wise alignment is represented as a score and the scores are ranked.
- Scoring matrices are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein).
- A unitary matrix is used for DNA pairs because each position can be given a score of +1 if it matches and a score of zero if it does not.
- Substitution matrices are used for amino acid alignments.
  - These are matrices in which each possible residue substitution is given a score reflecting the probability that it is related to the corresponding residue in the query.

Scoring

- The alignment score will be the sum of the scores for each position.
- Various matrix scoring systems (e.g. PAM, BLOSUM and PSSM) for quantifying the relationships between residues have been used.
- Positions at which a letter is paired with a null are called gaps.
  - Gap scores are negative.
  - Since a single mutational event may cause the insertion or deletion of more than one residue, the presence of a gap is frequently ascribed more significance than the length of the gap.
  - Hence the gap is penalized heavily, whereas a lesser penalty is assigned to each subsequent residue in the gap.
  - There is no widely accepted theory for selecting gap costs.
  - It is rarely necessary to change gap values from the default.

Significance Of The Results

- The significance of each alignment is computed as a probability (P value) or an expected frequency (E value).
- The p-value relates the score returned for an alignment to the likelihood of it having arisen by chance
  - In general, the closer the value approaches zero the greater the confidence that the match is real = 4e-45 (4×10^{-45}), 4e-16 (4×10^{-16}), or 1e-05 (1×10^{-05})
    - 1e-05 (1×10^{-5}) indicates that there is 1 in 100,000 possiblity that the match is due to chance
    - Note that 0.0 is identity
    - The nearer the value is to unity (1), the greater the chance that the match is spurious = 0.34 or 0.08
    - Between these ranges the estimation of the significance is difficult

References

- Basic Local Alignment and Search Tool, JMB Research Article
- Gapped BLAST and PSI-BLAST, JMB Research Article
- NCBI
  - BLAST Query tutorial
  - BLAST guide - NCBI Handbook
- Bioinformatics and Molecular Evolution
  - Chapters: Sequence Alignment Algorithms, Searching Sequence Databases