

BLAST

Bioinformatics Computing; September 29

1. Heuristics in BLAST
 - a. Search for common words (k-tuples) in query and database sequences to initiate extensions (11 for nucleotides and 3 for peptides)
 - b. (Optionally) Mask regions that may cause artificially high scores
 - i. Low complexity regions
 - ii. Human repetitive elements (LINEs and SINEs)
2. WU-BLAST vs. BLAST: Different statistical methods to evaluate sequence similarity scores
3. BLAST Parameters
 - a. Databases: nr, est, refseq, swissprot, month, pdb, environmental sequences, etc.
 - b. Advanced Options: Limit by entrez query, limit by organism, word size, custom scoring options
4. Main Types of BLAST
 - a. Nucleotide
 - i. blastn: nucleotide – nucleotide
 - ii. MEGABLAST: greedy algorithm; results in an algorithm that more efficiently finds long alignments between very similar sequences
 - b. Protein
 - i. blastp: protein – protein
 - ii. rpsblast: search for conserved domains
 - iii. cdart: find conserved domains and identify other proteins with similar domain architectures
 - c. Translated
 - i. blastx: nucleotide (translated) – protein; search for proteins encoded in your query DNA sequence

- ii. tblastn: protein – nucleotide (translated); search for new genes encoding a protein
 - iii. tblastx: nucleotide (translated) – nucleotide (translated); search for new genes
- d. Specialized BLASTs
- i. GEO BLAST – sequences with microarray information
 - ii. IgBlast – matches to immunoglobulin sequences
 - iii. SNP BLAST – matches to reference SNPs

5. Applications

- a. Identify the query sequence
- b. Find similar sequences
- c. Search for conserved domains
- d. Screen for vector contamination
- e. Search for homologs in specific organisms

6. References

- a. DW Mount: pages 248-258
- b. Course site: <http://www.nd.edu/~gmadey/bio05>
- c. NCBI BLAST:
<http://www.ncbi.nlm.nih.gov/blast/index.shtml>
- d. BLAST FAQ:
http://www.ncbi.nlm.nih.gov/blast/blast_FAQs.shtml
- e. Toolkit: <ftp://ftp.ncbi.nih.gov/toolbox/>