BLAST

Basic Local Alignment Search Tool
(Altschul et al. 1990)
What is blast?

BLAST ⇔ DNA/protein databases
GOOGLE ⇔ internet

D. Fristrom, Boston U.
Uses for Blast

- Identifying unknown sequence(s)
- Sequence comparison
- Identifying repeated sequences ("self" blast)
- Finding conserved motifs (PSI-blast)
## Main Blast flavors

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>blastp</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>blastx</td>
<td>translated nucleotide</td>
<td>protein</td>
</tr>
<tr>
<td>tblastn</td>
<td>protein</td>
<td>translated nucleotide</td>
</tr>
<tr>
<td>tblastx</td>
<td>translated nucleotide</td>
<td>translated nucleotide</td>
</tr>
</tbody>
</table>
Blast Applications

• blastn (megablast, discontinuous megablast)
  – within species (genus) comparisons

• blastp, blastx
  – cross-species comparisons (protein sequences evolve less rapidly than nucleotide sequences, due to redundant genetic code & functional constraints)
  – BUT... protein databases contain only those proteins that have been identified to date

• tblastn, tblastx
  – finds similarity to pseudogenes and “cryptic” genes
How Blast Works

• The query is broken up into short words (11–mers)
• These short words are used to search the database
• When a match is identified, blast tries to extend the match using gap or non-gapped alignments

• When performing whole genome analyses, Blast can miss significant matches
Interpreting BLAST outputs

• The top hit is not always the “correct” hit
• Alignment score is highly dependent on alignment length
• Longer alignments with multiple mismatches “incorrect” matches can be ranked higher than short, perfect alignments
• Depends on database “completeness”
Web-based Blast

BLAST: Basic Local Alignment Search Tool

NCBI/BLAST Home

BLAST finds regions of similarity between biological sequences. more...

NEW DELTA-BLAST, a more sensitive protein-protein search

BLAST Assembled RefSeq Genomes
Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- Arabidopsis thaliana
- Oryza sativa
- Bos taurus
- Danio rerio
- Drosophila melanogaster
- Gallus gallus
- Pan troglodytes
- Microbes
-Apis mellifera

Basic BLAST
Choose a BLAST program to run.

- nucleotide blast
- protein blast
- blast
- tblast
- Search a nucleotide database using a nucleotide query
  Algorithms: blastn, megablast, discontiguous megablast
- Search protein database using a protein query
  Algorithms: blastp, psi-blast, phiblast, delta-blast
- Search protein database using a translated nucleotide query
- Search translated nucleotide database using a protein query
- Search translated nucleotide database using a translated nucleotide query

Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (odat)
Disadvantages to Web Blast

- Limited options
- Results are stored on the NCBI server
  - Must be downloaded
- Not possible to batch process
- SEARCHES ARE LIMITED TO THE SEQUENCES THAT ARE AVAILABLE IN GENBANK
  - What about your own sequences?
Local Blast

• Impressive selection of search/output options
• Results saved directly to local machine
• Possible to search NCBI and local databases
• Creation of new local databases is very easy
• Amenable to batch processing
This morning’s exercise

• Learn how to find an executable on a website and download it via ftp
• Install the blast executables in the blast directory
• Run a blast search locally but query a remote database (at NCBI)
• Format a sequence to make a local blast database BLAST
• Search the local database
• Play with different output formats