BLAST

Slides adapted & edited from a set by Cheryl A. Kerfeld (UC Berkeley/JGI) & Kathleen M. Scott (U South Florida)


Starts with a Query Sequence in FASTA Format

Amino acid sequence:

>ribosomal protein L7/L12 [Thiomicrospira crunogena XCL-2]
MAITKDDILEAVANMSVMEVVELVEAMEEKFGVSAAAAVAVAPGDAGAA
GEEQTEFDVVLGAGDNKVAIAAGVRGATGLGLKEAKSAVESAPFTLKEG
VSKEEAETLANELKEAGIEEVK

Nucleotide sequence:

>gi|118139508:333094-333465 Thiomicrospira crunogena XCL-2
ATGGCAATTACAAAAGACGATATTTAGAAGCAGTTGCTAACATGTCAGTAATGGAAG
TTGTGAACTTGGGAAAGGAGAGGAGTTGTGCTGCAAGCAGCTTGCC
GGTGGAACTTGGGAAGAAGAAAGTTGGTCTTGTGCTGCTGCTGGCAGTGCAGTTGAC
GTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGT GG

Note the description line
Starts with “>”, ends with carriage return
Not read as sequence data

Kerfeld and Scott, PLoS Biology 2011
NCBI BLAST Interface
(blastp: for protein-protein alignments)

NCBI BLAST Results Page:
Potential homologs retrieved from database
Overview of BLAST

1. Segment the query sequence into short “words”
2. Use the query sequence segments to scan the database for matching sequences
3. Extend the matched segments in either direction to find local alignments.
4. Create a list of hits & alignments, with best matches first

BLAST Phase 1: Segment the query sequence and identify words that could form potential alignments

- Segment the query sequence into pieces ("words")
  - Default word length: 3 amino acids or 11 nucleic acids
- Create a list of synonyms and their scores for comparing query words to target words
  - Uses scoring matrix to calculate scores for synonyms that might be found in the database
- Save the scores (and synonyms) exceeding a given threshold T
**BLAST Phase 2: Using the query sequence word list, scan the database for synonyms (hits)**

- Scan the database for matches to the word list with acceptable T values
- Require two matches ("hits") within the target sequence
- Set aside sequences with matches above T for further analysis

```
.........SWITEASFSPPGIM....
```

Possible match from the database

**Words**

SWI  PGI

**BLAST Phase 3: Extending the hits**

- Search 5' and 3' of the word hit on both the query and target sequence
- Add up the score for sequence identity or similarity until value exceeds S
- Alignment is dropped from subsequent analyses if value never exceeds S

```
RHQGILTSWVSQASFTPPGIMLAIPGFDAYGLAGQNK
......................................................

Word match from Step 1  Extension until score drops
```
So, to summarize:

- BLAST segments query sequence into “words” and scores potential word matches
- Scans this list for alignments that meet a threshold score $T$
  - uses a scoring matrix to calculate this (e.g., BLOSUM62)
- Uses this list of ‘synonyms’ to scan the database
- Extends the alignments to see if they meet a cutoff score $S$
  - uses a scoring matrix to calculate this
- Reports the alignments that exceed $S$

PAM and BLOSUM Matrices

- Scoring matrices are calibrated to capture different degrees of sequence similarity
- In practice, this means choosing a matrix appropriate to the suspected degree of sequence identity between the query and its hits
- PAM: empirically derived for close relatives
- BLOSUM: empirically derived for distant relatives

<table>
<thead>
<tr>
<th>More divergent</th>
<th>Less divergent</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOSUM45</td>
<td>PAM240</td>
</tr>
<tr>
<td>BLOSUM62</td>
<td>PAM180</td>
</tr>
<tr>
<td>BLOSUM80</td>
<td>PAM120</td>
</tr>
<tr>
<td>BLOSUM90</td>
<td>PAM30</td>
</tr>
</tbody>
</table>
Raw Scores (S values) from an Alignment

\[ S = (\sum M_{ij}) - cO - dG, \]

where

- \( M \) = score from a similarity matrix for a particular pair of amino acids (ij)
- \( c \) = number of gaps
- \( O \) = penalty for the existence of a gap
- \( d \) = total length of gaps
- \( G \) = per-residue penalty for extending the gap

Limitations of Raw Scores

- S values depend on the substitution matrix, gap penalties
- Impossible to compare S values from hits retrieved from BLAST searches when different matrices and gap penalties are used
Going from Raw Scores to Bit Scores

\[ S' = \frac{\lambda S - \ln(K)}{\ln(2)} \]

where

- \( S' \) = bit score
- \( \lambda \) and \( K \) = normalizing parameters of the specific matrices and search spaces

- Larger raw scores result in larger bit scores
- Allows user to compare scores obtained by using different matrices and search spaces

Limitations of Bit Scores

- How high does a bit score have to be to suggest common ancestry?
  - Hard to evaluate hits as homologs or not, based solely on bit scores
E-value

- Number of distinct alignments with scores greater than or equal to a given value expected to occur in a search against a database of known size, based solely on chance, not homology.
  - Large E-values suggest that the query sequence and retrieved sequence similarities are due to chance
  - Small E-values suggest that the sequence similarities are due to shared ancestry (or potentially convergent evolution)

Calculating E-values

\[ E = \frac{n \times m}{2^S} \]

where

- \( m \) = effective length of the query sequence
- \( n \) = effective length of the database sequence

\( m = \text{length of query sequence} - \text{average length of alignments} \)

(Controls for fewer alignments occurring at the ends of the query sequence)

The value of \( E \) decreases exponentially with increasing \( S \)
**BLAST Parameters**

- Expect
- Word size
- Matrix
- Gap costs
- Filter
- Mask

**E value Threshold**

- Alignments will be reported with E-values less than or equal to the expect values threshold
  - Setting a larger E threshold will result in more reported hits
  - Setting a smaller E threshold will result in fewer reported hits
Filter and Mask

- **Filter: Low complexity**
  - Replaces the following with N (nucleotides) or X (amino acids)
    - Dinucleotide repeats
    - Amino acid repeats
    - Leader sequences
    - Stretches of hydrophobic residues

- **Mask: Lower case**
  - Replaces lowercase letters in sequence with N or X
    - Lowercase letters typically indicate base or amino acid not known with certainty

Parameter Summary is Found at the Bottom of the Output.....
Evaluating BLAST Results

<table>
<thead>
<tr>
<th>Score</th>
<th>Total Score</th>
<th>Query Coverage</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;40</td>
<td>279</td>
<td>100%</td>
<td>1e-101</td>
</tr>
<tr>
<td>30-40</td>
<td>199</td>
<td>100%</td>
<td>2e-40</td>
</tr>
<tr>
<td>20-30</td>
<td>149</td>
<td>100%</td>
<td>5e-35</td>
</tr>
<tr>
<td>10-20</td>
<td>138</td>
<td>100%</td>
<td>5e-40</td>
</tr>
</tbody>
</table>

Does it cover the whole length of both the query and subject sequences?

Examine the BLAST Alignment

Does it cover the whole length of both the query and subject sequences?
High E-value: Discovery of a Distant Homolog or Garbage?

- Take another look at the target (subject) sequence(s) that have high E-values
  - Similar length?
  - Recurring motifs?
  - Similar biological functions?
- Use target sequences as query sequences for another BLAST search
  - Does the original query sequence come up in report?