BLAST

Anders Gorm Pedersen
&
Rasmus Wernersson
Database searching

Using pairwise alignments to search databases for similar sequences

Query sequence

Database
Database searching

Most common use of pairwise sequence alignments is to search databases for related sequences. For instance: find probable function of newly isolated protein by identifying similar proteins with known function.

Most often, local alignment ("Smith-Waterman") is used for database searching: you are interested in finding out if ANY domain in your protein looks like something that is known.

Often, full Smith-Waterman is too time-consuming for searching large databases, so heuristic methods are used (fasta, BLAST).
# Database searching: heuristic search algorithms

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Uses heuristics to avoid calculating the full dynamic programming matrix</td>
<td>Uses rapid word lookup methods to completely skip most of the database entries</td>
</tr>
<tr>
<td>Speed up searches by an order of magnitude compared to full Smith-Waterman</td>
<td>Extremely fast</td>
</tr>
<tr>
<td>The statistical side of FASTA is still stronger than BLAST</td>
<td>One order of magnitude faster than FASTA</td>
</tr>
<tr>
<td>Two orders of magnitude faster than Smith-Waterman</td>
<td>Almost as sensitive as FASTA</td>
</tr>
<tr>
<td>BLAST flavors</td>
<td>TBLASTN</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>BLASTN</td>
<td>Protein query sequence</td>
</tr>
<tr>
<td>Nucleotide query sequence</td>
<td>Nucleotide database</td>
</tr>
<tr>
<td></td>
<td>&quot;On the fly&quot; six frame translation of database</td>
</tr>
<tr>
<td>BLASTP</td>
<td>Protein query sequence</td>
</tr>
<tr>
<td>Protein query sequence</td>
<td>Protein database</td>
</tr>
<tr>
<td>BLASTX</td>
<td>Nucleotide query sequence</td>
</tr>
<tr>
<td>Nucleotide query sequence</td>
<td>Nucleotide database</td>
</tr>
<tr>
<td></td>
<td>Compares all reading frames of query with all reading frames of the database</td>
</tr>
</tbody>
</table>
Searching on the web: BLAST at NCBI

Very fast computers dedicated to running BLAST searches

Many databases that are always up to date (e.g. NR and Human Genome)

Nice simple web interface

But you still need knowledge about BLAST to use it properly
When is a database hit significant?

• **Problem:**
  
  – Even *unrelated* sequences can be aligned (yielding a low score)
  
  – How do we know if a database hit is *meaningful*?
  
  – When is an *alignment score* sufficiently high?

• **Solution:**
  
  – Determine the range of alignment scores you would expect to get for *random reasons* (i.e., when aligning unrelated sequences).
  
  – Compare actual scores to the *distribution of random scores*.
  
  – Is the real score much higher than you’d *expect by chance*?
Distribution of random alignment scores

- **Software simulation**
Significance of alignment score expressed as E-value

Searching a database of unrelated sequences results in scores following an extreme value distribution.

The exact shape and location of the distribution depends on the exact nature of the database and the query sequence.

**E-value**: the number of random hits to expect for any given score.

Want E-values below 1 (the lower the better).
Significance of alignment score expressed as E-value

**E-value / Expect-value:**
Number of unrelated hits with an equal or better alignment score to expect due to strictly stochastic reasons.

**Example:**
Alignment score = 110
E-value = 8.7

Alignment score = 135
E-value = 0.0001
BLAST heuristics

• BLAST speeds up the search >100x by pre-screening the database sequences and only performing the full Dynamic Programming on “promising” sequences.

• Promising sequences: database sequences that have sub-strings (“words”) which also occur in the query sequence (found rapidly using a so-called “suffix-tree”)

• BLASTN and BLASTP use different criteria for overlap required for a sequence to be deemed promising
• Heuristics:
  – Perfect match “word” of at least size: 7, 11 (default) or 15.

• Alignment matrix:
  – Match: 1
  – Mismatch: -3

• Notice: All mismatches are equally penalized:
  – E.g. A:G == A:C == A:T
  – More advanced models for DNA evolution does exist.
• **Heuristics:**
  – 2 x “Near match” within a window.
  – Default word length: 3 aa
  – Default window length: 40 aa

• **Alignment matrix:**
  – PAM and BLOSUM-series (default: BLOSUM 62)

• **Notice:** These alignment matrices incorporate knowledge about protein evolution.