Alignments 3: BLAST

Sequence Analysis

Part of the presentation courtesy of Radek Szklarczyk (radek@cs.vu.nl)
Sequence searching – challenges

- Exponential growth of databases
Sequence searching – definition

- Task:
  - Query: short, new sequence (~1000b)
  - Database (searching space): very many sequences
  - Goal: find seqs related to query

- We want:
  - fast tool
  - primarily a filter: most sequences will be unrelated to the query
  - fine-tune the alignment later
What is BLAST

• Basic Local Alignment Search Tool
• Bad news: it is only a heuristics

• Basic idea:
  – High scoring segments have well conserved (almost identical) part
  – As well conserved part are identified, extend it to the real alignment
What means *well conserved* for BLAST?

- BLAST works with *k-words* (words of length $k$)
  - $k$ is a parameter
  - different for DNA (>10) and proteins (2..4)
- word $w_1$ is *T-similar* to $w_2$ if the sum of pair scores is at least $T$ (e.g. $T=12$)

**Similar 3-words**

<table>
<thead>
<tr>
<th>$W_1$</th>
<th>R</th>
<th>K</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_2$</td>
<td>R</td>
<td>R</td>
<td>P</td>
</tr>
</tbody>
</table>

**Score:** 9 -1 7 $\sum = 15$
BLAST algorithm
3 basic steps

1) Preprocess the query sequence:
   extract all the *k-words*

2) Scan for *T-similar* matches in database

3) Extend them to alignments
BLAST, Step 1: Preprocess the query

- Take the query (e.g. LVNRKPVVP)
- Chop it into overlapping k-words (k=3 in this case)

<table>
<thead>
<tr>
<th>Query:</th>
<th>LVNRKPVVP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word1:</td>
<td>LVN</td>
</tr>
<tr>
<td>Word2:</td>
<td>VNR</td>
</tr>
<tr>
<td>Word3:</td>
<td>NRK</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

- For each word find all similar words (scoring at least T)
- E.g. for RKP the following 3-words are similar:
  - QKP
  - KKP
  - RQP
  - REP
  - RRP
  - RKP
BLAST, Step 2: Find "exact" matches with scanning

- Select sequences from the database which match with at least one of the extracted k-words. So all selected sequences have at least one well conserved part in common with the query sequence.

...VLQKPLKKPPLVKRQPCCEVVVRKPLVKVIRCLA...
BLAST, Step 3: Extending "exact" matches

- Having the list of exact matches we extend alignment in both directions

Query: L V N R K P V V P

T-similar: R R P

Subject: G V C R R P L K C

Score: -3 4 -3 5 2 7 1 -2 -3

...till the sum of scores drops below some level X from the best known
Gapped BLAST (now standard)

- Original Blast program did not allow gapped alignments.
- Observation: Well conserved parts in sequences are usually much longer than $k$.
- Gapped Blast uses this observation by only extending the alignment if there are two close hits on the same diagonal.
  - This reduces the number of extensions (90% time is spent on extensions), so it is much faster.
Seeds

- More general form of k-word is a seed
- The seed
  \[CT\cdot GT\cdot AT\]
- gives “hits” with both sequences
  \[\ldots CTCGT\text{TATA}\ldots\]
  \[\ldots CTAGTAATG\ldots\]
Profiles

- A profile is a generalized form of a sequence.
- Matrix with probabilities instead of sequence with letters.
- Scoring with profiles:

\[
\text{score}(i, A) = \sum_B p(i, B) \times \text{score}_{\text{blosum62}}(A, B)
\]

position  letter  probability of some letter being at position i
Constructing a profile

- Take significant BLAST results
- Make an alignment
- Assign weights to sequences
- Construct profile
PSI-BLAST

- Position-Specific Iterated BLAST
- A profile is derived from the result of the first search
- Database is searched against the profile (instead of a sequence)
- Up to 3 iterations
BLAST flavours

- **blastp**: protein query, protein db
- **blastn**: DNA query, DNA db
- **blastx**: DNA query, protein db
  - in all reading frames. Used to find potential translation products of an unknown nucleotide sequence.
- **tblastn**: protein query, DNA db
  - database dynamically translated in all reading frames.
- **tblastx**: DNA query, DNA db
  - all translations of query against all translations of db
Dot-plots

Can be a bit messy, though...

Filter:
6/10 residues have to match...
Dot-plots, what about...

- Insertions
- Duplications (tandem repeats)
- Inversions
Dot-plots, self-comparison

Direct repeat

Inverted repeat

Tandem repeat
Take-home message

- BLAST method
- Gapped Blast
- PSI-BLAST
- dot-plots