Week 4, Lecture 8

István Albert
Bioinformatics Consulting Center
Penn State

Lecture Data

• Download and unpack the data for this lecture, it contains quite a few files

• Nucleotide sequences and protein sequences

• Single sequence others and multiple sequences (records)

Blast Bit Scores (S)

A bit score corresponds to alignment quality

• the higher the score, the better the alignment

• the formula uses a scoring matrix

  similar/identical residues \(\rightarrow\) increase the score
  gaps, non-similar residues \(\rightarrow\) decrease the score

\[(\text{the bit score is also normalized from the alignment score})\]

Blast E (expect) -value

• The number of hits one can "expect" to see by chance when searching a database of a particular size.

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

\(m, n\) are the sequence sizes, \(S\) is the bitscore
Reminder on DNA to Protein translation

Frame of translation usually leads to different proteins

blastn - nucleotide vs nucleotide

- EST: An expressed sequence tag or EST is a short sub-sequence of a cDNA sequence
- Find the best matches of a single EST nucleotide sequence (read) against a nucleotide based EST collection
- You may use different search tasks (optimizations) within blastn. Named (confusingly) as blastn and megablast

List of BLAST+ programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Input-Output</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>nucleotide-nucleotide</td>
<td>This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.</td>
</tr>
<tr>
<td>blastp</td>
<td>protein-protein</td>
<td>This program, given a protein query, returns the most similar protein sequences from the protein database that the user specifies.</td>
</tr>
<tr>
<td>blastx</td>
<td>nucleotide-6-frame translation-protein</td>
<td>This program compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.</td>
</tr>
<tr>
<td>blastx6</td>
<td>nucleotide-6-frame translation</td>
<td>This program is the slowest of the BLAST family. It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. The purpose of blastx is to find very distant relationships between nucleotide sequences.</td>
</tr>
<tr>
<td>tblastx</td>
<td>protein-nucleotide-6-frame translation</td>
<td>This program compares a protein query against the all six reading frames of a nucleotide sequence database.</td>
</tr>
</tbody>
</table>

Orientation

> AI109564 AI109564 Caenorhabditis elegans cDNA clone:yk7887
  S' end, single read.
  Length=326
  Score = 26.5 bits (28), Expect = 2.0
  Identities = 20/24 (83%), Gaps = 0/24 (0%)
  Strand=Plus/Minus
  Query 109 AAATTTTCTACGAGAGAAAT 132
  Objet 141 AAATTTTCTACGAGAGAAAT 118

> AI109529 AI109529 Caenorhabditis elegans cDNA clone:yk7886
  S' end, single read.
  Length=325
  Score = 26.5 bits (28), Expect = 2.0
  Identities = 15/16 (94%), Gaps = 0/16 (0%)
  Strand=Plus/Plus
  Query 1 TOGGCTACTGGAAAAA 16
  Objet 1 TOGGCTACTGGAAAAA 16
Understanding a blast report

We are accessing a sequence reported in the blast report with the blastdbcmd command.

Fine tuning blastn searches

Four different tasks are supported in blastn:

1. **megablast** - for very similar sequences (e.g., sequencing errors)
2. **dc-megablast**, typically used for inter-species comparisons
3. **blastn** - the traditional program used for inter-species comparisons
4. **blastn-short** - optimized for sequences less than 30 nucleotides.

blastp – protein vs protein

- Find the best alignments of the gamma2 globin protein against a list of globin proteins

blastx – nucleotide vs protein

Looking for matches on both strands and 3 reading frames → 6
Align the nucleotide sequence of the fugu globin against the globin list.
Align the protein sequence of the gamma 2 globin against the EST database

>nd|AD109149 AD109149 Caenorhabditis elegans cDNA clone yk703p6 : 5' end, single read.
Score = 26.2 bits (56), Expect = 0.11, Method: Compositional matrix adjust.
Identities = 11/23 (48%), Positives = 16/23 (69%), Gaps = 0/23 (0%)
Frame = -1
Query  99 REILIDLSEFQQLSHELBCDKLEHV
Subject 314 RLILGSSATTFQYLRQQCHEYV  246

Compare coding regions between more distant organisms
- Example: having only nucleotide sequence compare the coding regions of a chicken with those of a fugu

Pairwise alignments
- Previously called **bl2seq**
- Folded into the individual tools under the parameters: **-subject, -query**

```bash
# compare the results of the following three alignments
# blastx, megablast
/bin/blastx -query data/chicken_glabin.nuc.fa -subject data/fugu_glabin.nuc.fa
# blastx, blastx
/bin/blastx -query data/chicken_glabin.nuc.fa -subject data/fugu_glabin.nuc.fa -task blastx
# tblastx
/bin/tblastx -query data/chicken_glabin.nuc.fa -subject data/fugu_glabin.nuc.fa
```

Homework 8
1. You may use the data provided with this lecture
2. Align a nucleotide sequence against a protein database. Extract the nucleotide sequence of the query that corresponds to the worst local alignment. (hint: construct a blast database for the query, then use blastdbcmd)
3. Extract and compare the nucleotide sequences that correspond to the subject and the query of a local alignment obtained via **tblastx**