Data Retrieval from GenBank

Peter J. Myler

Bioinformatics of Intracellular Pathogens

JNU, Feb 7-0, 2009
• ENTREZ – Browsing – Text search – Accession number/GID
• ORF Finder
• BLAST – Sequence search

Accessing sequence data at GenBank
Kinetoplastid genomes
L. major

Friedlin
L. major Friedlin chr18
Entrez gene entries
Only a single hit
• originally genetic locus name
• now just a random number
• Unique identifier
• Letters can denote sequence type
  e.g. RefSeq (curated, non-redundant sequence standard)
  NC_ (chromosomes)
  NM_ (mRNAs)
  NP_ (proteins)
  NT_ (constructed genomic contigs)
  NG_ (genomic regions)
• Accession number version

• GI (GenInfo Identifier)
unique identification number
changes when sequence changes
Publications describing sequence
• Authors
• Title
• Journal
• Link to PubMed
• Information about genes and gene products and regions of biological significance
• mRNA
• CDS – coding sequence (ORF)
• Translation – link to Protein DB
• gene – genetic definition
• Sequence
• Can view in fasta format for sequence only
## ORF Analysis

<table>
<thead>
<tr>
<th>ORF</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF1</td>
<td>Description1</td>
</tr>
<tr>
<td>ORF2</td>
<td>Description2</td>
</tr>
<tr>
<td>ORF3</td>
<td>Description3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ORF</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF4</td>
<td>Description4</td>
</tr>
<tr>
<td>ORF5</td>
<td>Description5</td>
</tr>
<tr>
<td>ORF6</td>
<td>Description6</td>
</tr>
</tbody>
</table>
BLAST searching at GenBank
Different types of BLAST
Nucleotide BLASTs
How does BLAST work?

• Step 1 - Indexing
• Step 2 - Initial searching
• Step 3 - Extension
• Step 4 - Gap insertion
• Step 5 - Score reporting
Step 1 - Indexing

- The program accesses a pre-constructed index of every oligomer, or word of defined size ($W$), in the database. For DNA sequences:
  - Default $W=26$ for megablast
  - Default $W=11$ or $12$ for discontinuous megablast
  - Default $W=11$ for blastn
- For protein sequences the default $W=3$
- When you run the program it indexes your probe
GTGTCAGCTAAC
TGTCAGCTAACG
GTCAGCTAACGG
TCAGCTAACCGC
e tc...

Probe sequence
TCATATCACGGCCCTTCGGACCTGAGG

Words
TCATATCACGG
CATATCACCGC
ATATCACCGCC
e tc...

Database sequence (N = 46 bases)
GTGTCAGCTAACGGCCGTTACGATGCTAAAGCTATACGATTAGCG
Step 2 - Initial Searching

- Each word in the probe sequence index is compared to the database index and residue pairs are scored.
- For DNA sequences, a match is +1, a mismatch is -3.
- For protein sequences, scores for matches and mismatches are based on a substitution matrix.
<table>
<thead>
<tr>
<th>A</th>
<th>R</th>
<th>N</th>
<th>D</th>
<th>C</th>
<th>Q</th>
<th>E</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>L</th>
<th>K</th>
<th>M</th>
<th>F</th>
<th>P</th>
<th>S</th>
<th>T</th>
<th>W</th>
<th>Y</th>
<th>V</th>
<th>B</th>
<th>Z</th>
<th>X</th>
<th>*</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>-4</td>
</tr>
<tr>
<td>R</td>
<td>-1</td>
<td>5</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>N</td>
<td>-2</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>-4</td>
<td>-2</td>
<td>-3</td>
<td>3</td>
<td>0</td>
<td>-1</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>6</td>
<td>-3</td>
<td>0</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>-4</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>4</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>9</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-4</td>
</tr>
<tr>
<td>Q</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>5</td>
<td>2</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>3</td>
<td>-1</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>-4</td>
<td>2</td>
<td>5</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>1</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>4</td>
<td>-1</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>2</td>
<td>-2</td>
<td>6</td>
<td>-2</td>
<td>-4</td>
<td>-4</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>8</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-4</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>4</td>
<td>2</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
</tr>
<tr>
<td>L</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>2</td>
<td>4</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-4</td>
<td>-3</td>
<td>-1</td>
<td>-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>1</td>
<td>1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>3</td>
<td>-1</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>3</td>
<td>3</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-4</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>3</td>
<td>-3</td>
<td>-2</td>
<td>2</td>
<td>2</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>3</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-2</td>
<td>-1</td>
<td>3</td>
<td>4</td>
<td>-3</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>-3</td>
<td>-4</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-3</td>
<td>3</td>
<td>4</td>
<td>-2</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
<td></td>
</tr>
</tbody>
</table>
The score for each word pair is the sum of the scores for each pair of residues.

Matching words scoring above a threshold (T) are retained for further analysis.

- DNA: T = 0
- Protein: T = 11

Stretches of similarity are called HSPs (high scoring segment pairs).
Probe sequence word

TCAGATCACGG

Database sequence

TGTCAGCTAACGG

Database sequence words

TGTCAGCTAAC
GTCAGCTAACG
TCAGCTAACGG

Alignments

\[
\begin{array}{c|c|c|c}
\text{Probe} & \text{Database} & \text{Misalignment} \\
\hline
TCAGATCACGG & TGTCAGCTAAC & (3 \times 1) + (8 \times -3) = -21 \\
TCAGATCACGG & GTCAGCTAACG & (2 \times 1) + (9 \times -3) = -25 \\
TCAGATCACGG & TCAGCTAACGG & (9 \times 1) + (2 \times -3) = 3 \\
\end{array}
\]
Step 3 - Extension

• The program tries to extend matching segments out in both directions by adding pairs of residues.
• Residues will be added until the incremental score drops below a threshold.
• Identical or extremely conserved sequences will align over their entire length.
TCAGATCACGGCCCAACGGACCTGAGG

TCAGCTAACGGCCGTTACGATGCTAAA (11 - 39) = -28

Probe sequence

TCATATCACGGCCCTTCGGACCTGAGG

Database sequence

TGTCAGCTAACGGCCGTTACGATGCTAAAGCTATACGATTAGCG

Initial Alignment

TCAGATCACGGC (9 * 1) + (2 * -3) = 3

Probe sequence

TCATATCACGGCCCTTCGGACCTGAGG

Database sequence

TGTCAGCTAACGGCCGTTACGATGCTAAAGCTATACGATTAGCG

Extension

TCAGATCACGGCC (10 * 1) + (2 * -3) = 4

TCAGATCACGGCCCAACGGACCTGAGG

TCAGCTAACGGCCGTTACGATGCTAAA (11 * 1) + (2 * -3) = 5
Step 4 - Gap Insertion

• Gaps are added to connect adjacent HSPs

• BLAST uses what are called “affine gap costs”
  – Different costs for opening vs extending a gap
  – For Blosum 62 the defaults are 11 and 1
Step 5 - Score Reporting

- At each step of the process, the statistical significance of matches are evaluated. Only the most significant are kept.

- The program employs Karlin-Altschul statistics, which relates the score of an HSP to the expected frequency of its occurrence.

  - The frequency depends on the lengths of the query and database sequences, and the size of the sequence alphabet.
Scores are reported in various ways:

- Raw values based on the specific scoring matrix employed
- As bits, which are matrix-independent normalized values

The statistical significance of a score is represented by the E parameter.
A match will only be reported if its E value falls below the threshold set. The default value for E is 10, which means that 10 matches with scores this high are expected to be found by chance. Lower EXPECT thresholds are more stringent, and report fewer matches. The EXPECT (E) threshold is used to control score reporting.
Probabilities reported are summations of the probabilities of multiple HSPs. For HSPs to be included in a sum statistic or gapped alignment they must exhibit consistency—They must be in the same orientation, in a consistent order, and do not overlap. Repeated motifs will result in multiple, independent alignments between query and subject sequences.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Yes</strong></td>
<td><strong>No</strong></td>
</tr>
<tr>
<td><strong>Maybe</strong></td>
<td></td>
</tr>
</tbody>
</table>