Sequence Analysis '17: lecture 6

BLAST
You have seen....

**Dynamic programming:**

- Global alignment
- Global/local alignment (no end gaps. 3 ways to do it.)
- Local alignment
- Linear gap penalty
- Affine gap penalty

**Substitution matrices:**
Asymmetric substitution matrices?

If two different species have different overall *amino acid compositions*, then the substitutions between those species are asymmetric, $S[i \rightarrow j] \neq S[j \rightarrow i]$.

For example, if tetanus has *more L* than tuberculosis does, then $S[X_{tetan} \rightarrow L_{tuber}] > S[L_{tetan} \rightarrow X_{tuber}]$ (where $X$ is any amino acid).

Solving crime using NCBI BLAST

• Open “mystery sequence” on the course website.*
• Paste into a text file in UGENE
• right-click/Analyze/Query NCBI Blast database
• Select blastn, short, megablast. Expectation value=1. Max hits=20, nr
• Wait for results.
• Read annotations.

*Provided by “crime scene investigators”
NCBI is phasing out sequence GIs - use Accession.Version instead!
Wednesday, March 2, 2016
As of September 2016, the integer sequence identifiers known as "GIs" will no longer be included in the GenBank, GenPept, and FASTA formats supported by NCBI for sequence records. The FASTA header will be further simplified to report only the sequence accession.version and record title for accessions managed by the International Sequence Database Collaboration (INSDC) and NCBI’s Reference Sequence (RefSeq) project. As NCBI makes this transition, we encourage any users who have workflows that depend on GI's to begin planning to use accession.version identifiers instead. After September 2016, any processes solely dependent on GIs will no longer function as expected.
Why do a database search?

**Mol. Bio:** Determination of gene function. Primer design.

**Pathology, epidemiology, ecology:** Determination of species, strain, lineage, phylogeny.

**Biophysics:** Prediction of RNA or protein structure, effect of mutation.
Searching millions of sequences

Given a protein or DNA sequence, we want to find all of the sequences in GenBank (> 255 billion bases!!) that have a good alignment score.

Each alignment score should be the *optimal* score (or a close approximation).

How do we do it?
Searching using Dynamic Programming

**SSEARCH** Smith & Waterman

Dynamic Programming (DP) returns the **optimal alignment**, given the scoring function (usually *affine gap local* alignment).

Relatively slow, but more *sensitive*, and more *selective*, than FASTA and BLAST

**Optimal.**

<table>
<thead>
<tr>
<th></th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>T+</td>
<td>F-</td>
</tr>
<tr>
<td>False</td>
<td>F+</td>
<td>T-</td>
</tr>
</tbody>
</table>

Sensitivity (or coverage) is the fraction of all Trues that are Positive

\[
\text{sensitivity} = \frac{T^+}{T^+ + F^-}
\]

Selectivity (or accuracy) is the fraction of all Positives that are True

\[
\text{selectivity} = \frac{T^+}{T^+ + F^+}
\]
Searching using word matches

**FASTA**  
W. Pearson, 1988

First searches for *k-tuples*, then links them. Results are similar to a **dot plot**. Finally, diagonals are scored using a substitution matrix, and the highest-scoring diagonals are joined.

High-scoring alignments are re-calculated using DP (local/affine).

At least 50-times faster than SSEARCH. Not as sensitive. Final DP step makes it more *sensitive*, but less *selective*.

**FASTA** is a Heuristic alignment method, not Optimal.
Finding identity matches is very fast.

If two \textit{k-tuples} are separated by exactly the same amount in both sequence, draw a diagonal. A gapless alignment.
Find all gapless alignments

Score them using BLOSUM, keep the best

Connect them using simple affine gap. (gap ext. = 0)

If this alignment one of the best scores in the database search, go back and realign it to the query using DP.
Searching using lookup tables

BLAST  S. Altschul et al.

First make a set of lookup tables for all 3-letter (protein) or 11-letter (DNA) matches.

Make another lookup table: the locations of all 3-letter words in the database.

Start with a match, extend to the left and right until the score no longer increases.

Very fast. Selective, but not as sensitive as SSEARCH. Good statistics.

Heuristic.
BLAST, precalculations

All 8000 possible 3-tuples

Each 3-tuple is scored against all 8000 possible 3-tuples using BLOSUM. The top scoring 50 are kept as that 3-tuple’s “neighborhood words”
For every 3-residue window, we get the set of 50 nearest neighbors. Use each word to get identity matches (seeds). Then extend the seed alignments as long as the score increases.
The best extended seeds are called HSPs (high scoring pairs). The top scoring HSP is picked first, then the second (as long as it falls "northwest" or "southeast" of the first.), and so on.
» Local Dynamic Programming (DP) alignment is applied to only the sequences that pass the FASTA score cutoff.

» DP scores are converted to e-values.

» Local alignments are output for the top hits.

» Optionally, multiple sequence alignment output ("star" alignment)
Protein Databases available for BLAST search

On BLAST search page, select a database to search and then select ? to learn a little about that database.
## Protein Databases available for BLAST search

On BLAST search page, select a database to search and then select ? to learn a little about that database.

<table>
<thead>
<tr>
<th>Database</th>
<th>Title</th>
<th>Molecule Type</th>
<th>Update date</th>
<th>Number of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProtKB/Swiss-Prot(swissprot)</td>
<td>Non-redundant UniProtKB/SwissProt sequences.</td>
<td>Protein</td>
<td>2017/09/20</td>
<td></td>
</tr>
<tr>
<td>Patented protein sequences(pat)</td>
<td>Protein sequences derived from the Patent division of GenBank</td>
<td>Protein</td>
<td>2017/09/20</td>
<td>2054527</td>
</tr>
<tr>
<td>Protein Data Bank proteins(pdb)</td>
<td>PDB protein database</td>
<td>Protein</td>
<td>2017/09/20</td>
<td>93500</td>
</tr>
</tbody>
</table>
Protein Databases available for BLAST search

On BLAST search page, select a database to search and then select ? to learn a little about that database.

Metagenomic proteins (env_nr)

Title: Proteins from WGS metagenomic projects (env_nr).
Molecule Type: Protein
Update date: 2017/09/20
Number of sequences: 7003668

Transcriptome Shotgun Assembly proteins (tsa_nr)

Title: Transcriptome Shotgun Assembly (TSA) sequences
Description: The Transcriptome Shotgun Assembly proteins are produced from CDS features on mRNA sequences in the Transcriptome Shotgun Assembly sequences. See http://www.ncbi.nlm.nih.gov/genbank/TSA.html for details.
Molecule Type: Protein
Update date: 2013/03/31
Number of sequences: 2382761
BLAST -- Filters

You can restrict the search by Taxonomy
You can enter a Entrez search query to restrict the search.
(Test your Entrez query first)
(Learn about Entrez here:

Watch this!

https://www.youtube.com/watch?v=t8fKz9rvuOk&feature=youtu.be
Limit by Entrez Query

A BLAST search can be limited to the result of an Entrez query against the database chosen. This restricts the search to a subset of entries from that database fitting the requirement of the Entrez query. Terms normally accepted by Entrez nucleotide or protein searches are accepted here. Examples are given below.

protease NOT hiv1[organism]
This will limit a BLAST search to all proteases, except those in HIV 1.

1000:2000[slen]
This limits the search to entries with lengths between 1000 to 2000 bases for nucleotide entries, or 1000 to 2000 residues for protein entries.

Mus musculus[organism] AND biomol_mrna[properties]
This limits the search to mouse mRNA entries in the database. For common organisms, one can also select from the pulldown menu.

10000:100000[mlwt]
This is yet another example usage, which limits the search to protein sequences with calculated molecular weight between 10 kD to 100 kD.

src specimen voucher[properties]
This limits the search to entries that are annotated with a /specimen_voucher qualifier on the source feature.

all[filter] NOT enviromnental sample[filter] NOT metagenomes[orgn]
This excludes sequences from metagenome studies and uncultured sequences from anonymous environmental sample studies.
## Other forms of BLAST

<table>
<thead>
<tr>
<th>BLAST</th>
<th>query</th>
<th>database</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>blastp</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>tblastn</td>
<td>protein</td>
<td>translated DNA</td>
</tr>
<tr>
<td>blastx</td>
<td>translated DNA</td>
<td>protein</td>
</tr>
<tr>
<td>tblastx</td>
<td>translated DNA</td>
<td>translated DNA</td>
</tr>
<tr>
<td>psi-blast</td>
<td>protein, profile</td>
<td>protein</td>
</tr>
<tr>
<td>phi-blast</td>
<td>pattern</td>
<td>protein</td>
</tr>
<tr>
<td>transitive blast*</td>
<td>any</td>
<td>any</td>
</tr>
</tbody>
</table>

*not really a blast. Just a way of using blast.*
Psi-BLAST searches the database \textit{iteratively}.

(Cycle 1) Normal BLAST (with gaps)

(Cycle 2) (a) Construct a \textbf{profile} from the results of \textbf{Cycle 1}.

(b) Search the database using the profile.

(Cycle 3) (a) Construct a \textbf{profile} from the results of \textbf{Cycle 2}.

(b) Search the database using the profile.

And So On... (user sets the number of cycles)

Psi-BLAST is much more \textit{sensitive} than BLAST.

Also more vulnerable to \textit{low-complexity}. 

PHI-BLAST --
Patterned Hit Initiated BLAST

Table 1. Detection of subtle protein sequence relationships using PHI-BLAST

<table>
<thead>
<tr>
<th>Conserved domain or motif under investigation</th>
<th>Pattern$^a$</th>
<th>GenBank (30) accession no. of query</th>
<th>Top non-trivial relevant hit found by PHI-BLAST Accession no.</th>
<th>$E$-value</th>
<th>Top non-trivial relevant hit found by BLAST Accession no.</th>
<th>$E$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. P-loop ATPase domain in apoptosis regulators and plant stress response proteins</td>
<td>[GA]xxxxGK[ST]</td>
<td>231729</td>
<td>2213598</td>
<td>0.038</td>
<td>2961373</td>
<td>4.7</td>
</tr>
<tr>
<td>B. ATPase domain in mismatch repair protein MutL, type II topoisomerases, histidine kinases, and HS90 molecular chaperones</td>
<td>hxhxDxGxG</td>
<td>127552</td>
<td>488200</td>
<td>0.017</td>
<td>2495364</td>
<td>1.8</td>
</tr>
<tr>
<td>C. Nucleotidyltransferase domain in archaeal tRNA nucleotidyltransferases</td>
<td>DhDhhh</td>
<td>2826366</td>
<td>2650333</td>
<td>0.061</td>
<td>2650333</td>
<td>8.6</td>
</tr>
<tr>
<td>D. Motif VI of superfamily II helicases in archaeal homologs of bacterial DNA primases</td>
<td>QxxGRx[GAR]</td>
<td>2128723</td>
<td>2499099</td>
<td>0.54</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The problem with...

Transitive Blast searches

- Transitive-BLAST = using the hits of a BLAST search to do additional BLAST searches.
- If $A$ detects $B$, and $B$ detects $C$. Therefore, $A$ is homologous to $C$ Right?
- Problem: most genes have multiple domains. Any domain can give a hit. $XYZ$ hits $XYAB$ hits $ABC$, but $XYZ$ is not a homolog of $ABC$!
- Transitive Blast is "legit" if all queries are bonafide single domains.
Exercise 5 -- write pseudocode for BLAST

due Mon Sep 25

Write pseudocode for the BLAST database search algorithm. Input is
1) a query sequence Q
2) a database DB containing M sequences
3) an index (F) relating each 3-tuple to locations L in DB sequence S.
4) a neighborhood words list N relating each 3-tuple to 50 similar 3-tuples.

Other data structures are created by the program.

Steps: (a) identity matches of 3-tuples to neighborhood words, (b) HSP generation, (c) FASTA alignment and triaging, (d) DP, (e) e-value