Sequence Analysis '18: 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)

Asymmetric subs matrix is the full square, not just half.

Yu YK, Wootton JC, Altschul SF.
The compositional adjustment of amino acid substitution matrices.
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”
Databases of NCBI

- **nr** (non-redundant)
  Includes all of GenBank, curated or not. *This is what you pick to be sure you have searched everything.*

- **RefSeq**
  The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. As of May 2015: *RefSeq hits have functional annotation keywords PROVISIONAL, PREDICTED, or INFERRED, if the function has not been determined directly!* *This is what you pick if you want confident functional annotations.*

- **RefSeq genomes** are copies of selected assembled genomes available in GenBank.

- **PDB** (Protein Data Bank)
  Sequences of known protein structures (NMR, Xray, cryo-EM). *This is what you pick if you want to know the structure of your sequence.*
GenBank database @ NCBI

» **Accession Identifiers:**

» **Uploading sequences to GenBank using BankIt**

---

NCBI phases 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.
Database searching

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

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.

![Confusion Matrix](image)

- **True**
  - T+
  - F-

- **False**
  - F+
  - T-

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

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

Selectivity (or accuracy) is the fraction of all Positive 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 $k$-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

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.
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.
Random access database

» Speed is gained because BLAST reads only the sequences that pass the FASTA cutoff.
BLAST -- last steps

» 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.

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Title</th>
<th>Molecule Type</th>
<th>Update Date</th>
<th>Number of Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-redundant protein sequences (nr)</td>
<td>All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects</td>
<td>Protein</td>
<td>2017/09/20</td>
<td>131807364</td>
</tr>
<tr>
<td>Reference proteins (refseq_protein)</td>
<td>NCBI Protein Reference Sequences</td>
<td>Protein</td>
<td>2017/09/20</td>
<td>92512280</td>
</tr>
<tr>
<td>Model Organisms (landmark)</td>
<td>Landmark database for SmartBLAST</td>
<td>Protein</td>
<td>2015/09/09</td>
<td>414146</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.

- **UniProtKB/Swiss-Prot**
  - Title: Non-redundant UniProtKB/SwissProt sequences.
  - Molecule Type: Protein
  - Update date: 2017/09/20

- **Patented protein sequences**
  - Title: Protein sequences derived from the Patent division of GenBank
  - Molecule Type: Protein
  - Update date: 2017/09/20
  - Number of sequences: 2054527

- **Protein Data Bank proteins**
  - Title: PDB protein database
  - Description: This database consists of sequences from the Protein Data Bank (PDB), which contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies.
  - Molecule Type: Protein
  - Update date: 2017/09/20
  - Number of sequences: 93500
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
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 enviromntal 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: Blast with profiles

Psi-BLAST searches the database iteratively.

(Cycle 1) Normal BLAST (with gaps)

(Cycle 2) (a) Construct a profile from the results of Cycle 1.
   (b) Search the database using the profile.

(Cycle 3) (a) Construct a profile from the results of Cycle 2.
   (b) Search the database using the profile.

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

Psi-BLAST is much more sensitive than BLAST.

Also more vulnerable to 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>Patterna</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>hxxDxGxG</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.
NEW Exercise 5 -- write pseudocode for BLAST
due Mon Sep 24

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

Write pseudocode for the FASTA algorithm
input: a matrix of HSPs (high scoring pairs, or extended seeds), each with start,end pairs ((i,j), (k,l)) and a score H.
output: An alignment consisting of a sequential set of HSPs.
Algorithm: "Greedy" with sequentiality constraint. No new HSP can have a point "north-east" or "south-west" of any accepted HSP.