Sequence Database Searching
(Basic Tools and Advanced Methods)

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB

What’s database searching
- Goal: find similar (homologous) sequences of a query sequence in a sequence of database
- Input: query sequence & database
- Output: hits (pairwise alignments)

Basics of database searching
- Core: pairwise alignment algorithm
- Speed (fast sequence comparison)
- Relevance of the search results (statistical tests)
- Recovering all information of interest
  - The results depend of the search parameters like gap penalty, scoring matrix.
  - Sometimes searches with more than one matrix should be preformed
- Specificity and sensitivity

The different strategies for database searching

<table>
<thead>
<tr>
<th>Query</th>
<th>Alignment methods</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence</td>
<td>BLAST, FASTA, SW</td>
<td>Sequence</td>
</tr>
<tr>
<td>MSA</td>
<td>PSI-BLAST, Sequence</td>
<td>MSA</td>
</tr>
<tr>
<td>MSA</td>
<td>Profile alignment</td>
<td>MSA</td>
</tr>
<tr>
<td>Sequence</td>
<td>hmmmer, HMM models</td>
<td></td>
</tr>
</tbody>
</table>

What program to use for searching?
- BLAST is fastest and easily accessed on the Web
  - A suite; BLASTP, BLASTN, BLASTX
  - lag behind the development of sequencing techniques
- FASTA
  - more sensitive for DNA-DNA comparisons
  - FASTX and TFASTX can find similarities in sequences with frameshifts
- Smith-Waterman is slower, but more sensitive
  - known as a “rigorous” or “exhaustive” search
  - SSEARCH in GCG and standalone FASTA
- Other tools: BLAT and RAPSearch (RAPSearch2)

FASTA
- Derived from logic of the dot plot
  - compute best diagonals from all frames of alignment
- Word method looks for exact matches between words in query and test sequence
  - hash tables (fast computer technique)
  - DNA words are usually 6 bases
  - protein words are 1 or 2 amino acids
  - only searches for diagonals in region of word matches = faster searching
FASTA algorithm

- Makes longest diagonal
  - after all diagonals found, tries to join diagonals by adding gaps
  - computes alignments in regions of best diagonals

FASTA alignments

- BLAST
  - [BLAST = Basic Local Alignment Search Tool]
  - The central idea of the BLAST algorithm is that a statistically significant alignment is likely to contain a high-scoring pair of aligned words.
  - Uses word matching like FASTA
    - Similarity matching of words (3 aa’s, 11 bases)
    - Does not require identical words.
  - If no words are similar, then no alignment
    - won’t find matches for very short sequences
  - Original BLAST does not handle gaps well; the “gapped” blast is better

The BLAST Search Algorithm

<table>
<thead>
<tr>
<th>Query word (w = 3)</th>
<th>Neighborhood words</th>
<th>Word match</th>
<th>Extend hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAAVKEEISVEDEAVDKNI</td>
<td>EEA, EAA, AVK, KEE, ESI, ISV, EIS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BLAST: word matching

- Break query into words:
- Break database sequences into words:
- Compare word lists by Hashing (allow near matches)
BLAST: extend hits

For each word match, extend the alignment in both directions to find alignments that score greater than a threshold of value 3.

Use two word matches as anchors to build an alignment between the query and a database sequence.

Gapped BLAST algorithm

- The NCBI’s BLAST website now both use “gapped BLAST”
- This algorithm is more complex than the original BLAST
- It requires two word matches close to each other on a pair of sequences (i.e. with a gap) before it creates an alignment allow gaps (using Smith-Waterman algorithm constrained by a band width).

Statistical tests

- Evaluate the probability of an event taking place by chance.
- Very little is known about the random distribution of optimal global alignment scores
- Statistics for the scores of local alignments, are well understood (particularly true for local alignments lacking gaps)
- P-value
  - Randomized data
  - Distribution under the same setup

BLAST statistics

- E-value is equivalent to standard P value (based on Karlin-Altschul theorem)
  - Karlin-Alschul equation is probably one of the most recognized equation in bioinformatics.
  - E-value represents the likelihood that the observed alignment is due to chance alone
  - E-value: the number of alignments expected by chance (E) during a sequence database search. E = 1 indicates that an alignment this good would happen by chance with any random sequence searched against the same database.

\[ E = K \cdot e^{-\frac{S}{m}} \]

How to compute BLAST E-value

\[ S' = (S - \ln k)/\ln 2 \]

\[ E = m \cdot e^{-S} \]

\[ E = -\ln(1 - P) \]

P (the probability of getting at least one HSP)
**Which BLAST?**

<table>
<thead>
<tr>
<th>Engine</th>
<th>Query</th>
<th>Database</th>
<th>Comparison</th>
<th>Reference use</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>DNA</td>
<td>DNA</td>
<td>DNA level</td>
<td>Search for genes in non-coding DNA</td>
</tr>
<tr>
<td>blastp</td>
<td>Protein</td>
<td>Protein</td>
<td>Protein level</td>
<td>Search for homologous proteins</td>
</tr>
<tr>
<td>blastx</td>
<td>DNA</td>
<td>Protein</td>
<td>Protein level</td>
<td>Analyze new DNA to find genes and seek homologous proteins</td>
</tr>
<tr>
<td>tblastn</td>
<td>Protein</td>
<td>DNA</td>
<td>Protein level</td>
<td>Search for genes in non-coding DNA</td>
</tr>
<tr>
<td>tblastx</td>
<td>DNA</td>
<td>DNA</td>
<td>Protein level</td>
<td>Discover gene structure</td>
</tr>
</tbody>
</table>

BLAST+ package has different user interface

**Which database?**

- For blastp (blastx)
  - Non-redundant protein sequences (nr)
  - Reference proteins (refseq_protein)
  - Swissprot protein sequences (swissprot)
  - Patented protein sequences (pat)
  - Protein Data Bank proteins (pdb)
  - Environmental samples (env_nr)
- For blastn (tblastx)
  - Human genomic plus transcript (not for tblastx)
  - Mouse genomic plus transcript (not available for tblastx)
  - Nucleotide collection (nr/nt)
  - etc

**BLAST is approximate**

- BLAST makes similarity searches very quickly because it takes shortcuts.
  - Looks for short, nearly identical “words” (11 bases)
- It also makes errors
  - Misses some important similarities
  - Makes many incorrect matches
    - Easily fooled by repeats or skewed composition

**Interpretation of BLAST hits**

- Very low E values (e-100) are homologs
- Moderate E values are related genes
- Long list of gradually declining of E values indicates a large gene family
- Long regions of moderate similarity are more significant than short regions of high identity

**Is the hit with smallest E-value the closest sequence to the query?**

- Not necessarily
- Some people argue that more strict phylogeny analysis is needed for further conclusion
  - Why? Different evolutionary rates, etc

**Biological relevance**

- It is up to you, the biologist to scrutinize these alignments and determine if they are significant.
- Were you looking for a short region of nearly identical sequence or a larger region of general similarity?
- Are the mismatches conservative ones?
- Are the matching regions important structural components of the genes or just introns and flanking regions?
20 tips to improve your BLAST searches
- “Don’t Use the Default Parameters” (but what parameters?)
- “Treat BLAST Searches as Scientific Experiments”
- “View BLAST Reports Graphically” (depends!)
- “Be Skeptical of Hypothetical Proteins”
- “Look for Stop Codons and Frame-Shifts to find Pseudo-Genes”
- “Parse BLAST Reports with Bioperl” (your own python, or perl, program)
- “How to Lie with BLAST Statistics” (“Lies, dammed lies, and statistics”)
- ...

Borderline similarity
- What to do with matches with E values that are not so impressive? (e.g., E values > 0.01, smaller numbers are more significant)
- this is the “Twilight Zone”
- retest these sequences and look for related hits (not just your original query sequence)
- similarity is (is not) transitive: if A~B and B~C, then A~C

Distant homology detection
- Use more data (MSA)
- Use advanced approaches (like HMM)

PSI-BLAST
- Position Specific Iterated BLAST (PSI-BLAST)
- Basic idea
  - Use results from BLAST query to construct a profile matrix (or PSSM)
  - Search database with profile instead of query sequence
- Iterate
- Position-specific scoring matrices are an extension of substitution scoring matrices

PSSM—Position Specific Scoring Matrix

Representing a profile as a logo
- Logos are used to show the residue preferences or conservation at particular positions
- Based on information theory
Profile-profile alignment
(alignment of alignments)

- Inputs: two MSAs (profiles)
- Alignment of two MSAs (similar to pairwise sequence alignment)
- Many existing programs
- Scoring functions vary
  - Dot product (FFAS)
  - FFAS scoring function
    - Dot product of two amino acid "frequency" vectors

FFAS scoring function

$$C_{m,n} = \sum_{a=1}^{20} f_{m,a} f_{n,a}$$

Dot product of two amino acid "frequency" vectors

BLAT—the blast-like alignment tool

- Question: "What are the differences between Blat and Blast?"
- Response: Blat is an alignment tool like BLAST, but it is structured differently. On DNA, Blat works by keeping an index of an entire genome in memory. Thus, the target database of BLAT is not a set of GenBank sequences, but instead an index derived from the assembly of the entire genome. Blat of DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments.
- On proteins, Blat uses 4-mers rather than 11-mers, finding protein sequences of 80% and greater similarity to the query of length 20+ amino acids. The protein index requires slightly more than 2 gigabytes of RAM. In practice — due to sequence divergence rates over evolutionary time — DNA Blat works well within humans and primates, while protein Blat continues to find good matches within terrestrial vertebrates and even earlier organisms for conserved proteins. Within humans, protein Blat gives a much better picture of gene families (paralogs) than DNA Blat. However, BLAST and psi-BLAST at NCBI can find much more remote matches.

Ref: http://genome.ucsc.edu/FAQ/FAQblat.html

RAPSearch: fast protein similarity search using long seeds with reduced alphabets

RAPSearch algorithm:
- seed + extension (ungapped + gapped extension)
- borrow statistical evaluation method from BLAST
- utilize reduced amino acid alphabet and flexible seed
- identify seeds using suffix array (RAPSearch) -> hash table (RAPSearch2)

Selection of reduced alphabet

High coverage
High specificity
Sequence identity scoring zones

- >25-30%: homology zone
- 15-25%: twilight zone
- <15%: midnight zone

Weak sequence similarity detection is still not solved!

And sequence similarity $\neq$ structural similarity (for proteins) $\neq$ functional similarity

Readings

- Chapter 5 (Pairwise Sequence Alignment and Database Searching)

- Box 5.1 – saving space by throwing away the intermediate calculations
- “Time can be saved with a loss of rigor by not calculating the whole matrix” (Figure 5.18)