Lecture 2: BLAST search.

Sanjay Tiwari

Starting a BLAST search

- Go to the NCBI BLAST website:
  http://www.ncbi.nlm.nih.gov/BLAST/
- In our example: click on protein-protein BLAST (blastp).
- Open electronic file containing your sequence of interest.
- Copy sequence and then paste it in the “Search” box.

Starting a BLAST search (cont.)

- Thus query sequence(s) to be used for a BLAST search should be pasted in the **'Search'** text area.
- It accepts a number of different types of input and automatically determines the format or the input.
- Accepted input types are (i) bare sequence, (ii) FASTA, or (iii) sequence identifiers .

BLAST input formats

- **Bare sequence** (for our protein, L1 metallo-beta-lactamase):
  MRSTLLAFALAVALPAAHTSAAEVPLP
  QLRAYTVDASWLQPMAPLQIADHTWQI
  GTEDLTALLVQTPDGAVLLDGGMPQM
  ASHLLDNMKARGVTPRDLRLILLSHAH
  ADHAGPVAELKRRTGAKVAANAESAVL
  LARGGSDDLHFGDGITYPPANADRIV
  MDGEVITVGGIVFTAHFMAGHTPGSTA
  WTWTDTRNGKPVRIAYADSLSAPGYQ
  LQGNPRYPHLIEDYRRSFATVRALPCD
  VLLTPHPGASNWDYAAGARAGAKALT
  CKAYADAAEQ KFDGQLAKETAGAR

Note: Accepted Amino Acid Codes

- the accepted amino acid codes are: A alanine P proline B aspartate/asparagine Q glutamine C cystine R arginine D aspartate S serine E glutamate T threonine F phenylalanine *U selenocysteine* G glycine V valine H histidine W tryptophan I isoleucine Y tyrosine K lysine Z glutamate/glutamine L leucine X any M methionine * translation stop N asparagine

  - gap of indeterminate length

BLAST input formats (cont.)

- **Bare sequence**. The format can also be sequence interspersed with numbers and/or spaces, such as the sequence portion of a GenBank/GenPept flatfile report:
- 1 mrstllafal avalpaahts aaevplpqlr aytvdaswlq pmaplqiadh twqigtedlt
- 61 allvqtpdga vlldggmpqm ashlldnmka rgvtprdlrl illshahadh agpvaelkrr
- 121 tgakvaanae savllarggs ddlhfgdgit yppanadriv mdgevitvgg ivftahfmag
- 181 htpgstawtw tdtrngkpvr iayadslsap gyqlqgnpry phliedyrrs fatvralpcd
- 241 vlltphpgas nwdyaagara gakaltckay adaaeqkfdg qlaketagar

BLAST input formats (cont.)

- **FASTA format**. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length. Our protein sequence in FASTA format looks like:

BLAST input formats (cont.)

- >gi|525299|emb|CAA52968.1| L1 metallo-beta-lactamase [Stenotrophomonas maltophilia]
MRSTLLAFALAVALPAAHTSAAEVPLPQLRAYTVDASWLQPMAPLQIADHTWQIGTEDLTALLVQTPDGA
VLLDGGMPQMASHLLDNMKARGVTPRDLRLILLSHAHADHAGPVAELKRRTGAKVAANAESAVLLARGGS
DDLHFGDGITYPPANADRIVMDGEVITVGGIVFTAHFMAGHTPGSTAWTWTDTRNGKPVRIAYADSLSAP
GYQLQGNPRYPHLIEDYRRSFATVRALPCDVLLTPHPGASNWDYAAGARAGAKALTCKAYADAAEQKFDG
- QLAKETAGAR

BLAST input formats (cont.)

- **Identifiers or accession numbers**. For our sequence:
- P52700 (Swiss-Prot)
- 525299 (gi number)

Any of these numbers can be typed in the search box.

Continuing a BLAST search

- To speed the process up a little, uncheck the box that says, “Do CD Search.”
- Also (for speed-up), check the box in lower portion of screen that says, “Mask for lookup table only.”
- Now click: BLAST!

Continuing the BLAST search (cont.).

- The resulting screen, titled “formatting BLAST,” gives two important pieces of information:
  - The request ID. Make note of this, in case you need to go back to search.
  - Expected time in seconds for the search.
- To see the search results, scroll down to bottom of the page and click “Format.”

Continuing a BLAST search (cont.)

- (Aside) If you have noted the request ID number and need to refer back to the search later, do the following:
  - Go the BLAST **Home Page**
  - At bottom of page, click on link (under “meta”): “Retrieve results.”
  - This will bring up a screen with empty request ID number.
  - Paste request ID number in that box.
  - Now click, “Format” and wait for BLAST display.

Understanding the BLAST results

- A new window, titled “results of BLAST,” will open. It will consist of three features:
  - Graphic overview
  - Descriptions
  - Alignments
- **Graphical Overview** : is an overview of the database sequences aligned to the query sequence. The score of each alignment is indicated by one of five different colors, which divides the range of scores into five groups. Multiple alignments on the same database sequence are connected by a striped line. Mousing over a hit sequence causes the definition and score to be shown in the window at the top, clicking on a hit sequence takes the user to the associated alignments.

Understanding the BLAST results (cont.)

- Scroll down the “results” page until you see the section titled, “sequences producing significant alignments”
- Each sequence will have identifiers:
  - gi number (a unique identifier for a file, whatever database it resides in).
  - The database name in which the sequence was found.
  - accession number (varies with database)
    .

Understanding the BLAST results (cont.)

- In the middle of the BLAST result is a brief description of the file.
- Next in the display is the **bit score**. The higher the bit score, the better the alignment.
- Bit score is also a link to the actual alignment (lower down the page) with the “query” sequence.
- At the far right is the “E-value”

E-value

- Def: The E-value is the proportion of such matches in the combined (non-redundant) databases that are expected by chance alone (i.e. the proportion of purely random matches that one an expect)
- The smaller the E-value, the higher the chances that the similarity is “real,” i.e. reflects common descent and not the result of sheer chance.

What sequences to include?

- It is essential (for reconstructing phylogeny) that we only include sequences that are truly homologous (i.e. have descended from a common ancestral sequence).
- **Rule of thumb**: sequences with E-value less that 1/100000, i.e. e-5, are homologs of query sequence.

What sequences to include? (cont.)

- First sequence could be the query sequence itself (if it is found in a database).
- The second sequence has an “S” to the right of it. This means it is a protein structure file of the same query sequence, so we can eliminate it.

Examining the alignments

- Scroll down to the alignments themselves.
- Will decide which sequences to include in phylogeny.
- Will also select sequences for later downloading.

Examining the alignments (cont.)

- Now consider sequence 3, click on its bit score to automatically scroll down to the alignment.
- Note that sequences 3 and 4 are 92% and 88% identical to the query sequence. They are from the same species and represent “within species variation”.
- If interested in every homologous sequence, choose both; otherwise, choose only one.

Examining the alignments (cont.)

- Will stop when E-value is less than e-10.
- The decision about which sequences to keep and which to eliminate **cannot be reduced to an algorithm.**
- It depends on what you intend to accomplish with your phylogenetic tree.
- If you want as complete a tree as possible, keep everything that is a true homolog.
- If you only want to show representations of major groups, be more selective

Downloading the selected sequences.

- Assuming you have been checking the boxes for the selected sequences, scroll up to the beginning of the alignments and click on: **Get selected sequences**.
- In the textbook example, ten sequences were selected and they will all be displayed in the new page.
- Select all ten sequences.
- Change display choice from “Summary” to “FASTA” and the “Send to” choice to “File”.
- Click Display.

Distance Tree in BLAST

- BLAST has a new feature: a distance tree. Note that it is a simple tree and it is based on pairwise alignment and not multiple alignment.
- This feature can be found, at the end of the graphic overview (in the “results of BLAST” page) and just before the beginning of the sequence descriptions.
- Click on “Distance tree of results.”

Distance tree in BLAST (cont.)

- **Remark:** BLAST computes a pairwise alignment between a query and the database sequences searched. It does not explicitly compute an alignment between the different database sequences (i.e., does not perform a multiple alignment). For purposes of this sequence tree presentation an implicit alignment between the database sequences is constructed, based upon the alignment of those (database) sequences to the query. It may often occur that two database sequences align to different parts of the query, so that they barely overlap each other or do not overlap at all. In that case it is not possible to calculate a distance between these two sequences and only the higher scoring sequence is included in the tree.

Distance tree in BLAST (cont.)

- The BLAST distance tree allows you to choose two methods:
  - Fast minimum evolution (default)
  - Neighbor joining.
- It has five different display formats:
  - Rectangle (default)
  - Slanted
  - Radial
  - Force

Distance tree in BLAST (cont.)

- It allows you to mouse over an internal node and choose:
  - the subtree
  - the multiple alignment.
- In the default view, the tree is scaled according to distance.

Distance tree in BLAST (cont.)

- The BLAST distance tree is a kind of rough and ready attempt at displaying the relationship between all the related sequences showing up in a BLAST search.
- It should not be confused with a true Phylogenetic tree.