fasta.bioch.virginia.edu/biol4559/blast_demo.html

Biol4559 - Similarity Searching Exercises

These exercises use programs on the FASTA WWW Search page[pgm] and the BLAST WWW Search page [pgm].

In the links below, [pgm] indicates a link with most of the information filled in; e.g. the program name, query, and library. [seq] links go to the NCBI, for more information about the sequence. In general, you should click [pgm] links, but not [seq] links.

Identifying homologs and non-homologs; using domain annotations

Use the FASTA search page [pgm] to compare Honey bee glutathione transferase D1 NP_001171499/ H9KLY5_APIME [seq] (gi|295842263) to the PIR1 Annotated protein sequence database. Be sure to press Search Database , not Compare Sequences .

1. Take a look at the output.
   a. How long is the query sequence?
   b. How many sequences are in the PIR1 database?
   c. What scoring matrix was used?
   d. What were the gap penalties? (what is the penalty for a one-residue gap? two residues?)
   e. What are each of the numbers after the description of the library sequence? Which one is best for inferring homology?
   f. How similar is the highest scoring sequence? What is the difference between %_id and %_sim? Why is there no 100% identity match?
   g. Looking at an alignment, where are the boundaries of the alignment (the best local region)? How many gaps are in the best alignment? The second best?

2. Homology (and non-homology) can also be inferred from domain relationships. There are three parts to the domain display, the domain structure of the query (top) sequence (if available), the domain structure of the library (bottom) sequence, and the domain alignment boundaries in the middle (inside the alignment box). The boundaries and color of the alignment domain coloring match the `Region:` sub-alignment scores.
   a. Ignoring the domain annotations, what is the highest scoring non-homologous sequence (the highest scoring -- lowest E()-value -- sequence that is unlikely to share any structural similarity with the honey bee query sequence?
   b. Test your candidate non-homolog(s) by comparing it to SwissProt using the `General re-search` link. Does it show significant similarity to any glutathione transferases?
   c. Is your testing by re-searching consistent with the domain annotation coloring?
   d. Note that the alignment of Honey bee `GSTD1` and `SSPA_ECO57` includes portions of both the N-terminal and C-terminal domains, but neither domain is completely aligned. Why do might the alignments not include the complete domains?
   e. Is your explanation for the partial domain alignment consistent the the argument that domains have a characteristic length? How might you test whether a complete domain is present?
3. Repeat the GSTD1 search [pgm] using the BLASTP62/-11/-1 scoring matrix BlastP62 (30%) ⇕ that BLAST uses. Re-examine the `GSTD1:SSPA_ECO57` alignment. Are both Glutathione transferase domains present? Look at the alignments to the homologs above and below `SSPA_ECO57`. Based on those aligments, do you think the Glutathione-S-Trfase C-like domain is really missing? Why did the alignment become so much shorter?

4. One of the candidate non-homologs is `sp|Q9SI20|EF1D2_ARATH`, with an E()-value of 0.11.
   a. Does the domain structure of `EF1D2_ARATH` suggest that it could be a glutatione transferase homolog?
   b. Use the General Research to explore the domains contained in `EF1D2_ARATH` homologs found in SwissProt.
   c. Does this second search support homology or non-homology?

5. Compare the Drosophila glutathione transferase GSTT1_DROME [pgm] sequence to the sequences with solved 3D structures (**Library:** PDB structures (NCBI) ⇕ making certain that the **Annotations:** CATH structural domains ⇕ is set.)
   a. Are the structural domains in GSTT1_DROME consistent with the InterPro domains?
   b. Looking `pdb|2DSAA` or `pdb|1EV4A`, how many "Glutaredoxin" domains appear to be present?
   c. Looking at `pdb|3NIVA`, what do you think the "correct" structural domain annotation should be? How could you confirm you hypothesis?

`blastp` and `fasta36` on the command line

1. Login (ssh) to `franklin.achs.virginia.edu`

2. Download the Honey Bee sequence from the NCBI using `curl`:

   ```
   curl "http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&id=295842263&rettype=fasta&retmode=text" > honeybee_gst.aa
   ```

   Or copy the `/data/slib/biol4559/data/honeybee_gst.aa` file into your own homework (`biol4559/hwk2`) directory.

3. To run the `blast` and `fasta` programs, you need to have the `/data/slib/bin` directory in your path, and some additional environment variables set. The easiest way to do this is to copy two files from the `/data/slib/biol4559/bash` directory:

   ```
   cd    # make sure you are in your home directory
   cp /data/slib/biol4559/bash/.bash_profile .
   cp /data/slib/biol4559/bash/.bashrc .
   source ./.bashrc
   echo $PATH
   ```

   At this point, your `$PATH` should include `/seqprg/bin`, which is where the `blast` and `fasta` reside.

4. The following databases are available for searching with both the `fasta36 / ssearch36` and `blastp` programs. All `blast` names must be preceeded by `/data/slib/bl_dbs/`, e.g. `pir1` is referred to as `/data/slib/bl_dbs/pir1`.

Database	Blast name	FASTA abbreviation	# sequences

PIR1	`pir1`	a	~13,0001
SwissProt	`swissprot`	q	~500,000
RefSeq	`refseq_protein`	r	~40,000,0001

You can see the names of some of the `blastp` databases by looking in the `/data/slib/bl_dbs` directory: (but pir1 does not have this format)

```
ls /data/slib/bl_dbs/*[^0-9].pal  # protein databases
ls /data/slib/bl_dbs/*[^0-9].nal  # DNA databases
```

(The `[^0-9]` excludes the many files of the form swissprot.##.pal, which are not typically referred to directly.)

You can see the list of libraries that `fasta36` knows about by running the program in the *interactive* (`fasta36 -I`) mode. After you enter the name of your query sequence (`honeybee_gst.aa`), you will be given a list of sequence libraries, and their abbreviations. (`-I` interactive mode is useful for learning about available sequence libraries, but normally you should run in non-interactive mode.)

5. Running **`fasta36`** and **`blastp`** non-interactively.

   a. Look at the help commands. To see how to run the programs, and the various command line options, type:

   ```
   fasta36 -h
   fasta36 -help > fasta.help
   ```

   Likewise:

   ```
   blastp -h
   blastp -help > blastp.help
   ```

   b. Do the same FASTA search you did on the web site by typing:

   ```
   fasta36 honeybee_gst.aa a > honeybee_v_pir1.fa_result
   ```

   Look at the `honeybee_v_pir1.fa_result` file. Does the library hit summary look exactly the same?

   c. Do the search with `blastp`:

   ```
   blastp -query honeybee_gst.aa -db /data/slib/bl_dbs/pir1 > honeybee_v_pir1.bp_result
   ```

   Note that with `blastp`, the query and library sequence files are specified with arguments (`-query`, `-db`). With `fasta36` (and the other `FASTA` programs), they are specified as the first and second non-option commands. With `fasta`, all options *must* precede the query and library file names.

   d. Both `fasta36` and `blastp` can display search results in different formats. The simplest format to parse with other programs is `tabular` format:

   ```
   fasta36 -m 8 query.file library > fasta_output.tab
   blastp -outfmt 6 -query query.file library.file > blast_output.tab
   ```

   Run the `fasta36` and `blastp` `honeybee_gst.aa` vs PIR1 searches producing tabular output and saving it to a file.

   You will parse these `tabular` results files for this weekend's homework.

Where to get the FASTA package: faculty.virginia.edu/wrpearson/fasta/

The "normal" FASTA WWW site:

Contact Bill Pearson: wrp@virginia.edu
