BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 3)

Advanced Database Searching

https://bioboot.github.io/bimm143_S18/lectures/#3

Dr. Barry Grant

Overview: Searching in databases for homologues of known proteins is a central theme in bioinformatics. The core goals are:

- High ***sensitivity*** - that is, detecting even very distant relationships, and
- High ***selectivity*** - namely, minimizing the number of reported 'hits' that are not true homologues.

All database search methods involve a trade-off between *sensitivity*, *selectivity* and *performance*. Important questions to ask include does the method find all or most of the examples that are actually present, or does it miss a large fraction? Conversely, how many of the 'hits' that it reports are incorrect? Finally does the approach scale to the tractable analysis of large datasets?

In this hands-on session we will explore the detection limits of conventional BLAST and introduce more sensitive (but often more time consuming) approaches including **Profiles**, **PSI-BLAST** and **Hidden Markov Models** (HMMs).

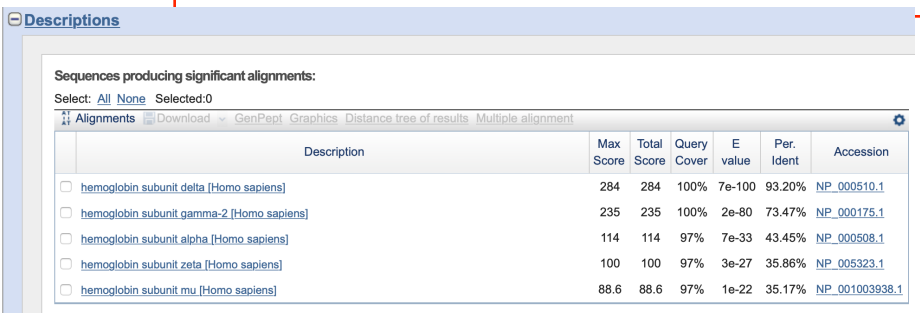
Section 1: The limits of using BLAST for remote homologue detection

Let's return to the HBB protein that we explored in a previous class and see if we can find distantly related myoglobin and neuroglobin using this as a BLAST query.

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

After selecting **blastp** and entering the sequence, be sure to change the search database to "**refseq-protein**" and restrict our search organism to only **humans** (taxid: 9605). This will help focus our results to highlight distant homologs in humans.

Q1. What homologs did you find with this simple blastp search? Note their *precent identities*, *coverage* and *E-values*.



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
hemoglobin subunit delta [Homo sapiens]	284	284	100%	7e-100	93.20%	NP_000510.1
hemoglobin subunit gamma-2 [Homo sapiens]	235	235	100%	2e-80	73.47%	NP_000175.1
hemoglobin subunit alpha [Homo sapiens]	114	114	97%	7e-33	43.45%	NP_000508.1
hemoglobin subunit zeta [Homo sapiens]	100	100	97%	3e-27	35.86%	NP_005323.1
hemoglobin subunit mu [Homo sapiens]	88.6	88.6	97%	1e-22	35.17%	NP_001003938.1

Now we could try changing the **Algorithm parameters** on the submission page to increase the number of hits reported. To do this you can click on the **Edit and Resubmit** link at the top left of your results page.

Q2. Try increasing the **Expect threshold** for your blasts search. What new hits were reported? What about their alignment statistics? Do you trust these matches?

Many useful 'rules of thumb' are expressed in terms of precent identity. If two proteins have more than 45% identical residues in their optimal alignment they typically have very similar structures and are likely to have a similar function. If two proteins have more than 25% identical residues (but less than 45% identity), they are likely to have a similar general folding pattern. Note that we will expand on the basis of this important *sequence > structure > function* relationship in a subsequent class unit.

Observations of a lower degree of sequence similarity cannot however rule out homology. Our very own Russ Doolittle (http://biology.ucsd.edu/research/faculty/rdoolittle) defined the region between 18-25% sequence identity as the "**twilight zone**" in which the suggestion of homology is tantalizing but dangerous. Below the twilight zone is a region where pairwise sequence alignments tell us very little - sometimes called the "midnight zone".

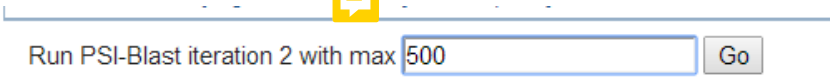
Section 2: Using PSI-BLAST

Although the twilight zone is a treacherous region, we are not entirely helpless. In deciding whether there is a genuine relationship, the '*texture*' of the alignment is important - essentially are the similar residues isolated and scattered throughout the sequences, or are there characteristic 'icebergs' - local regions of high similarity seen in many distant sequences that may correspond to a shared active site or other functional motif?

Lets return to your previous BLAST submission page with the HBB example from before. This time select the **PSI-BLAST** algorithm from the 'Program Selection' options section. Other settings should be as before (remember to reset your Expect threshold to default if you changed this previously) and use **refseq_protein** and search only in humans again.

Q3. The first iteration should be similar to your previous blastp search. Did you find any new potential homologs that you did not see previously?

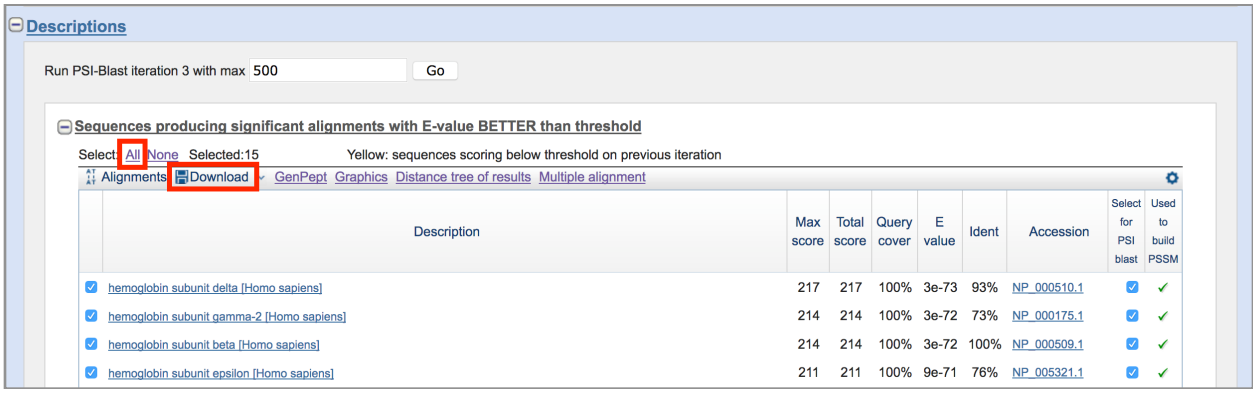
Q4. Now, we'd like to search for more distant homology, using another iteration of PSI-BLAST. Were you able to find any other proteins? If so, what were they and what function do they perform?



Q5. Perform a third iteration. Did the algorithm find any other proteins? Did we find myoglobin and neuroglobin? What could we change if we wanted to find even more distantly related sequences?

It can be difficult to visually identify conserved regions in the regular online NCBI BLAST alignment display. Selecting alternative display formats can be helpful. At the very top of the results page is a '**Formatting options**' link. Using the available options for '*Alignment View*' try the alternative 'query-anchored' display formats.

At the top of the '**Descriptions**' sub-section of the results page find the '**Downloads**' link, make sure all sequences are selected, and then chose "**FASTA (complete sequences)**".



Next paste or upload your FASTA sequences to **MUSCLE** (http://www.ebi.ac.uk/Tools/msa/muscle/) and use either the **Jalviewer** link (under the Result Summary tab) or select "**Download Alignment File**" and view the resulting alignment in the standalone **Seaview** program (http://doua.prabi.fr/software/seaview).

Optional: From Seaview you can export a FASTA format alignment file (Using **File** > **Save As…** and selecting **FASTA format**) and then use this file with alternate viewers to more clearly highlight conserved positions/columns in your alignment (e.g. http://www.bioinformatics.org/strap/aa/)

Q6. List the identities and alignment position of several invariant residues in your alignment, what role might these play in the protein?
HINT: We will return to this in Q8 after further analysis.

Section 3: Using HMMER

HMMER is an alternative sequence search and alignment method that employees probabilistic models called profile hidden Markov models (HMMs). HMMER aims to be significantly more accurate and more able to detect remote homologs than BLAST because of the strength of its underlying mathematical models. In the past, this strength came at significant computational expense, but in the new HMMER3 project, HMMER is now essentially as fast as BLAST.

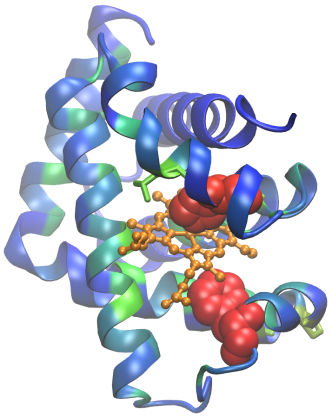
Lets use the new HMMER3 online @ http://www.ebi.ac.uk/Tools/hmmer/search/phmmer to examine how results compare to those obtained from BLAST and PSI-BLAST in the last section.

Q7. Performing a HMMER (phmmer) search with our HBB sequence above against the **SwissProt** database and setting the **"Restrict by Taxonomy"** to **9606**, how do your results compare to those from regular BLAST and PSI-BLAST?

Q8. Did you find myoglobin and neuroglobin? Are there any neuroglobin PDB structures available? If so take a record of their PDB codes for later.

Q9. How long did your search take?

HMMER is at the forefront of sequence-only based methods for detecting distant relatives. This tool is used to construct the **PFAM** (protein families) database. Find the link to the PFAM entry for the **Globin** family from your HMMER search results. Click on the HMM Logo link and determine the most conserved residues in this family.



Q10. Inspect the **HMM Logo** link for the PFAM Globin family and determine the most conserved residues in this family. What role might these residues play in these proteins?

In the molecular figure of beta globin here we have colored each residue position by the level of conservation in the alignment obtained from HMMER (blue - least conserved, red - most conserved). This information should help you answer Q8.

Section 4: Divergence of protein sequence and protein structure during evolution

In this case, as in many other examples in the twilight zone, protein structure can yield important insights. This is primarily because protein structure similarities remain robust as sequence similarities fade during the course of evolution. If protein structures are available for your tentative homologues it is advisable to examine their structural similarity and the overlap of conserved sequence regions at potentially functional sites. We will cover this important topic in more detail in Lecture 10. For now we will use the FATCAT **pairwise structural alignment** server to examine the similarities of our beta globin and neuroglobin proteins.

Visit: http://fatcat.sanfordburnham.org select **Pairwise alignment** and enter the *PDB code* **2HBS** *chain* **B** for the first structure. Then enter one PDB code for neuroglobin you found from answering Q8 previously (see below for an example).

Get the 1st structure (please use only one method from the 3 methods below):

Enter a name for your structure: pdb1 (optional)

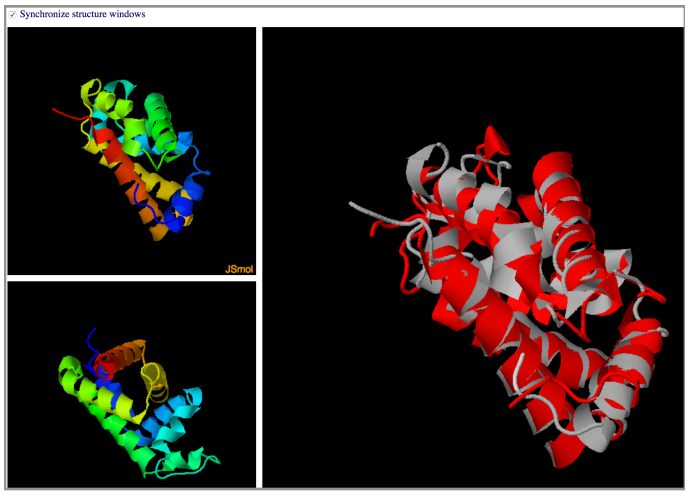
- Upload file (in PDB format): Choose File no file selected
- **or** Provide PDB code: 2HBS Chain: B
- **or** Provide SCOP domain code:

Get the 2nd structure (please use only one method from the 3 methods below):

Enter a name for your structure: pdb2 (optional)

- Upload file (in PDB format): Choose File no file selected
- **or** Provide PDB code: 4MPM Chain: B
- **or** Provide SCOP domain code:

Run the calculation and view the resulting structure **superposition** (basically a fit of one structure onto the other) online with JSmol.



Note how similar in structure these two distant homologues are.

Unfortunately, we wont always have a structure available for the system under investigation but when we do they can provide invaluable insight into evolutionary and functional mechanisms.

Q11. What one part of this hands-on session or associated lecture material is still confusing? Please answer using the following anonymous form:

Muddy Point Assesment