Student Name: Joshua Cheung UCSD PID #: A15441585

Class 3 Lab

Advanced Database Searching

http://thegrantlab.org

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 *percent identities*, *coverage* and *E-values*.

Hemoglobin subunit beta, Per. Ident: 100.00%, Cover: 100%, E-value: 2e-106
Hemoglobin subunit delta, Per. Ident: 93.20%, Cover: 100%, E-value: 7e-100
Hemoglobin subunit epsilon, Per. Ident: 75.51%, Cover: 100%, E-value: 2e-82
Hemoglobin subunit gamma-2, Per. Ident: 73.47%, Cover: 100%, E-value: 2e-80
Hemoglobin subunit gamma-1, Per. Ident: 72.79%, Cover: 100%, E-value: 3e-79
Hemoglobin subunit alpha, Per. Ident: 43.45%, Cover: 97%, E-value: 7e-33
Hemoglobin subunit zeta, Per. Ident: 35.86%, Cover: 97%, E-value: 3e-27
Hemoglobin subunit theta, Per. Ident: 39.31%, Cover: 97%, E-value: 6e-26
Hemoglobin subunit mu, Per. Ident: 35.17%, Cover: 97%, E-value: 1e-22

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?

After increasing the Expect threshold for the blasts search, there are many new hits reported such as 'zinc finger CCCH domain-containing protein isoform 6' and 'inhibitor of bruton tyrosine kinase isoform 1'. The alignment statistics of the new results are quite poor as the percent coverange of the new results range from 9% to 16%. Also the percent identities of the new results range from 20% to 60%. Additionally, the E-values of the new results are very high (ranging from 42 to 9094). Thus, these new matches should not be trusted.

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 amino-acids 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 (see image below). 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?

The first iteration of the search is indeed similar to the previous blastp search. There are three new potential homologs in the search. These homologs are cytoglobin, cytoglobin isoform X1, and cytoglobin isoform X2

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

We are able to to find three new proteins, these are neuroglobin, myoglobin isoform 1, and myoglobin isoform 2. Both myoglobin isoform 1 and 2 is involved in enabling heme binding, enabling metal ion binding, enabling oxygen binding, and enabling oxygen carrier activity. Neuroglobin is involved in enabling heme binding, enabling metal ion binding, enabling oxygen binding, enabling oxygen carrier activity, and enabling protein binding

Q5. Perform a third iteration. Did the algorithm find any other proteins? Did we find myoglobin and neuroglobin?

No we did not find any other proteins in the third iteration. We did indeed find myoglobin and neuroglobin during the second iteration

Section 3: Examining conservation patterns and evolutionary relationships

It can be difficult to visually identify conserved regions in the regular online NCBI BLAST alignment display. Selecting alternative display formats can be helpful. Toward the top of your results page under "**Other reports**" click the "**Multiple alignment**" option (see image below).

This will submit your identified (or selected subset) of sequences for multiple alignment. On the resulting page scroll down past the "*Graphical Overview*" and "*Descriptions*" to the "*Alignments*" section and note the coloring by conservation. Change this to **Conservation Setting**: **Identity**.

Q6. Are there any invariant amino acid positions across all the globins that you have identified? If so how many positions, and what amino acids are these in relation to beta globin? **Hint**: A common way to write these results in terms of the one-letter amino acid code and position number in the sequence you care most about e.g. H64 for Histidine in position 64.

There are indeed invariant amino acid positions across the goblins we have identified. There are 3 positions. The amino acids and their positions are H64, H96, and F23

Q7. What do you think these invariant amino acid residues might do in all these globins?

Both the H amino acid is hydrophillic while the F amino acids is hydrophobic so it is possible that these residues are involved in facillitating and coordinating heme binding.

At the very top of the page you can find a **Phylogenetic Tree** and **Download** link for your results. A common format to download is "Fasta plus gaps". You can then open this downloaded file in a program such as Seaview (that we used in lab 1) or input to **R** as we will use next day.

Feel free to examine the "**Phylogenetic Tree**" link and discuss with your neighbors, IAs and Barry whether this makes sense based on what we now know about relationships between these globins.

Section 4: Using HMMER (**OPTIONAL**: Note *server can be very slow - If so skip to section 5*)
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.

Q8. 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?

The result of the the HMMER include nearly all the results of the BLAST and all of the results of the PSI-BLAST search as well

Q9. Did you find myoglobin and neuroglobin? Are there any neuroglobin PDB structures available? If so take a record of their PDB codes for later.

There is indeed both a myoglobin and neuroglobin result from the search. There are neuroglobin PDB structures available. Its PDB code is 4MPM

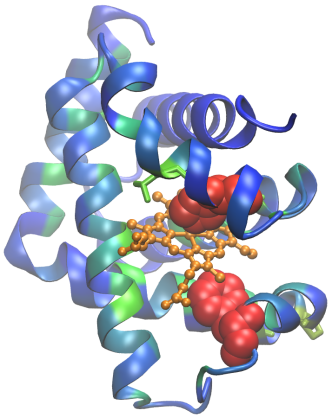
Q10. How long did your search take? **Was the web server accessible and responsive?**

The web server was relatively accessible and responsive. The search took less than 30 seconds.

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.

Q11. Inspect the **HMM Logo** link for the PFAM Globin family and determine the most conserved residues in this family. Again the key question is what role might these residues play in these proteins?

The most conserved regions are phenylalanine and histidine. A potential role these residues play in these proteins is to facilitate heme binding.



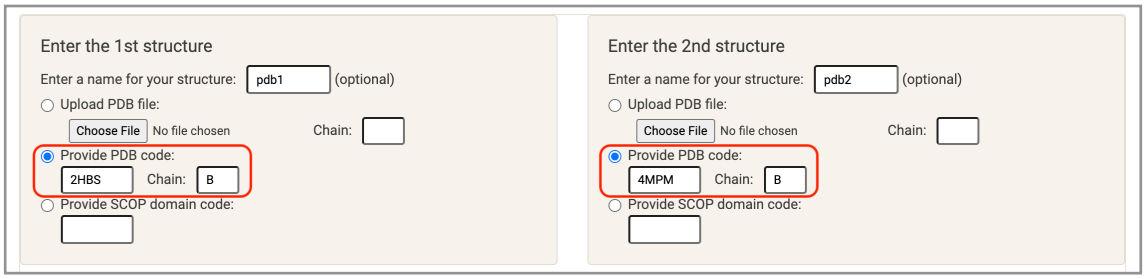
In the molecular figure of beta globin above I have colored each residue position by the level of conservation in the alignment obtained from HMMER (blue - least conserved, red - most conserved). This information should help you answer Q11.

Note: If the HMMER web server was unresponsive you can search PFAM directly @ https://pfam.xfam.org to help answer Q11.

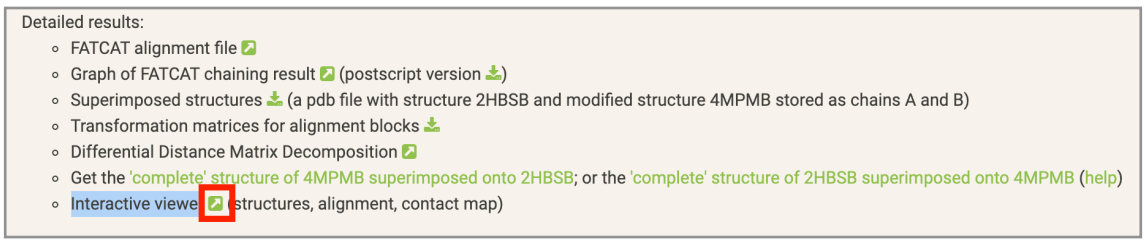
Section 5: Divergence of protein sequence and protein structure during evolution

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 s later class. 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.godziklab.org/fatcat/fatcat_pair.html and enter the *PDB code* **2HBS** *chain* **B** for the first structure. Then enter one PDB code for neuroglobin you found from answering **Q17** previously (see below for an example).

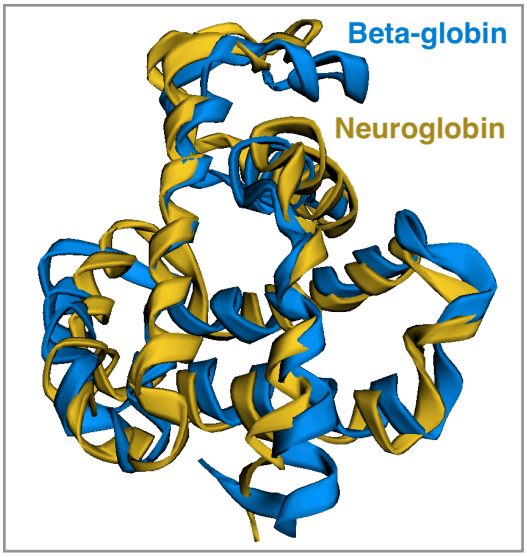


Click *SUBMIT* to run the calculation and view the resulting structure *superposition* (basically a fit of one structure onto the other) online in their "**Interactive viewer**" by clicking the green arrow (see below):



Note how similar in structure these two distant homologues are.

Explore the different display options on this page. For the image here I have selected *Render as*: **cartoon** and *Color by*: **chain**. This has the effect of having the first chain colored blue and (that is our beta globin) and the second (neuroglobin) dark yellow.



Q12. Can you find the most divergent in structure regions? Where are they located in the structure (interior/exterior in secondary structure elements or loops)?

The most divergent in structure regions are located on surface exposed loop regions.

Take home: 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.

Q13. What one part of this lab or associated lecture material is still confusing? Please answer in the following anonymous form: https://forms.gle/FEbKxnq4X7nUMhcn8