# Class 3 Lab*

## Advanced Database Searching

Barry Grant

Version 221003

**Instructions**

Save this document to your computer and open it in a PDF viewer such as Preview (available on every mac) or Adobe Acrobat Reader (free for PC and Linux). Be sure to add your name and UC San Diego personal identification number (PID) and email below before answering all questions in the space provided.

| Student Name | UCSD PID | UCSD Email |
|---|---|---|
| Charlie Rezanka | A15837296 | crezanka@ucsd.edu |

## 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?

*http://thegrantlab.org/teaching/

In this hands-on lab 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.

Delta globin: 100% coverage, 93.2% identity, 1e-100
Epsilon globin: 100% coverage, 75.5% identity, 1e-83
gamma-2 globin: 100% coverage, 73.5% identity, 4e-81

% identity goes from 100% to 35%, coverage goes from 100% to 97%, e values go from 1e-100 to 2e-23

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 (e.g. to 2000). What new hits were reported? What about their alignment statistics? Do you trust these matches? Did you find myoglobin?

> Tip
>
> Remember that you can use the *E-value* together with percent *identity* and *coverage* values to help judge your alignment results.

% coverage goes from 100% to 6%
% identity goes from 93% to 23%
E values go from 1e-100 to 1808

Some new matches include Hox-A6, SSX1, and 52M1. I don't really trust any of the new matches, many false positives.

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 late Russ Doolittle 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 Figure 1). 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.

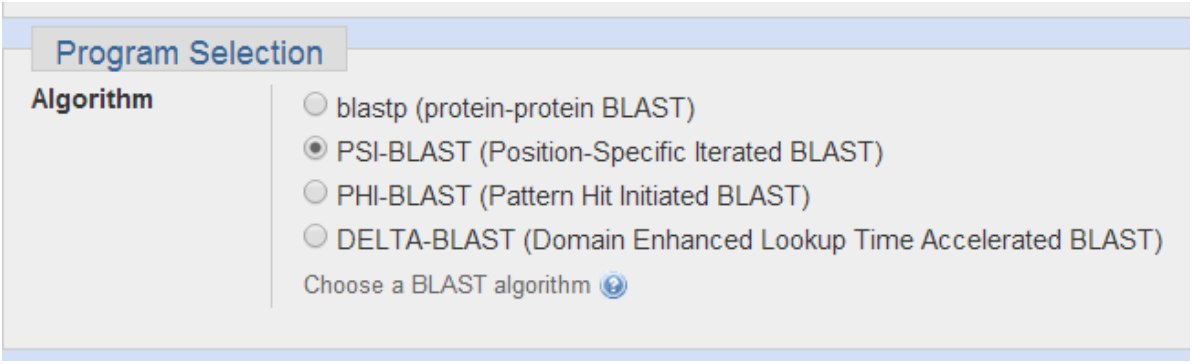



Figure 1: Selecting PSI-BLAST on the NCBI Protein BLAST submission page

**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?

Two new potential homologues are shown:
Cytoglobin (NP_599030.1)
Neuroglobin (NP_067080.1)

**Q4** Now, we'd like to search for more distant homology, using another iteration of PSI-BLAST (click the "Run" button, see Figure 2 below). Were you able to find any other proteins? If so, what were they and what function do they perform?

> Tip
>
> You can use the link-outs to the NCBI Gene database to help answer this question if you need to.

The only new protein found was myoglobin, which is found primarily in skeletal or cardiac muscles, and facilitates oxygen transfer from the cell membrane to mitochondria.

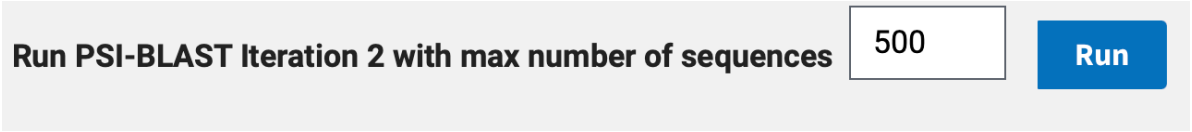



Figure 2: You must manually inspect results before each PSI-BLAST itteration

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

> Tip
>
> Remember that you can use the *E-value* together with precent identity and coverage to help judge your alignment results.

The algorithm did not find any new proteins, but myoglobin (3e-50) and neuroglobin (9e-42) are reflected. E values for both are well under .05

## 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 Figure 3 below).

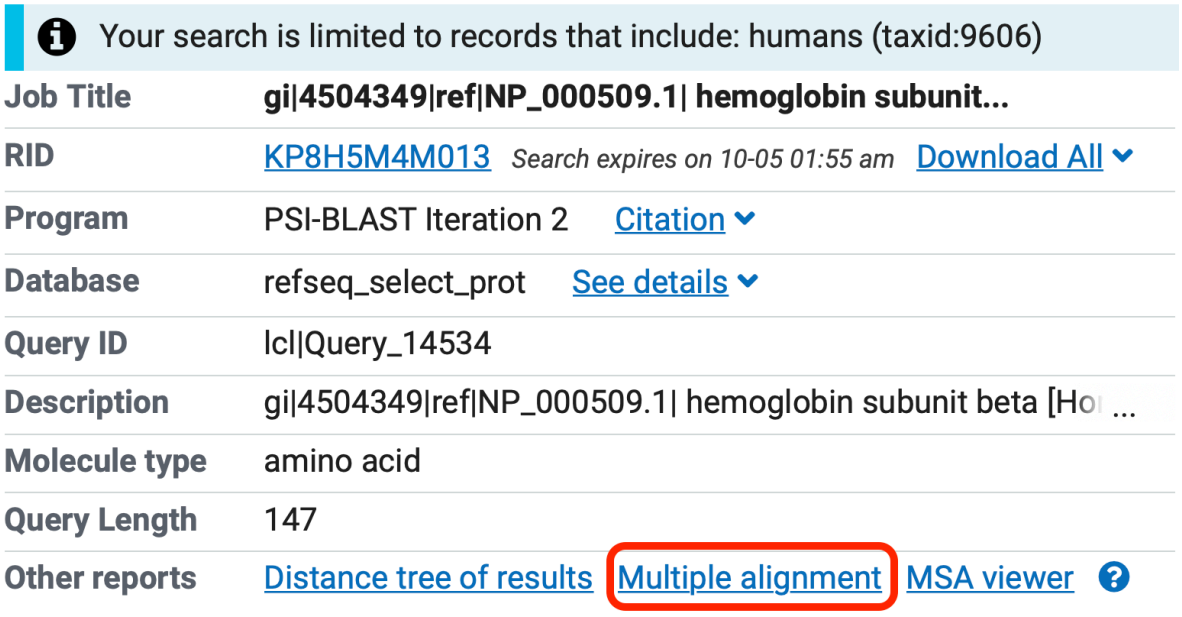



Figure 3: Other result report formats are available. The most useful of these is often *"Multiple alignment"*

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** (see Figure 4).

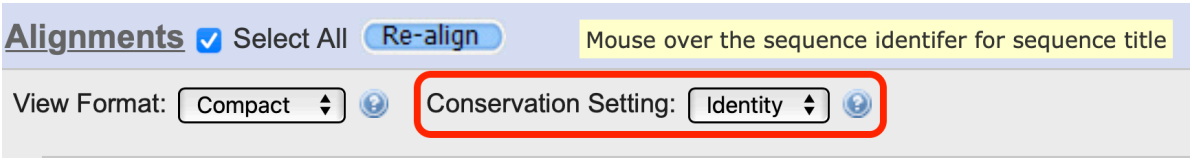



Figure 4: Color by *Identity* to highlight invarent amino acids accross your set of aligned sequences

**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?

> Tip
>
> 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.

7 invariant residues: W16, R33, P39, F46, H64, H93, F122

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

> Tip
>
> Do you think they are doing different things or the same thing in these distinct but clearly related proteins?

Each invariant residue likely has the same role in each homologue that is crucial to the function of the protein.

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" (see Figure 5 below). 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.

Figure 5: Be sure to download the FASTA format aligned (i.e. plus gaps) globin sequences as we will use these in the next lab

## Section 4: Using HMMER (OPTIONAL)

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?

> Tip
>
> This EBI server can be very slow if there are multiple users. You may need to skip this question - if so just note "server unresponsive" as your anser here.

Found all homoglobins (alpha, beta) and more distantly related globins including neurolgobin and myoglobin

**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.

> Tip
>
> The HMMER server results detail little colored icons, one of which lists PDB accession codes for identified structures. Alternatively, you can search the PDB database using (PSI)BLAST or HMMER just like we did back in lab 1.

Both were found.
neuroglobin ID codes: 4MPM & 1OJ6

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

> **Tip**
>
> If you find yourself waiting along time please make a note of this here.

The website was accaptibly responsive, roughly 5 minutes to gather all the info needed for these questions.

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. Or visit the page directly here. Click on the **HMM Logo** link and determine the most conserved residues in this family.

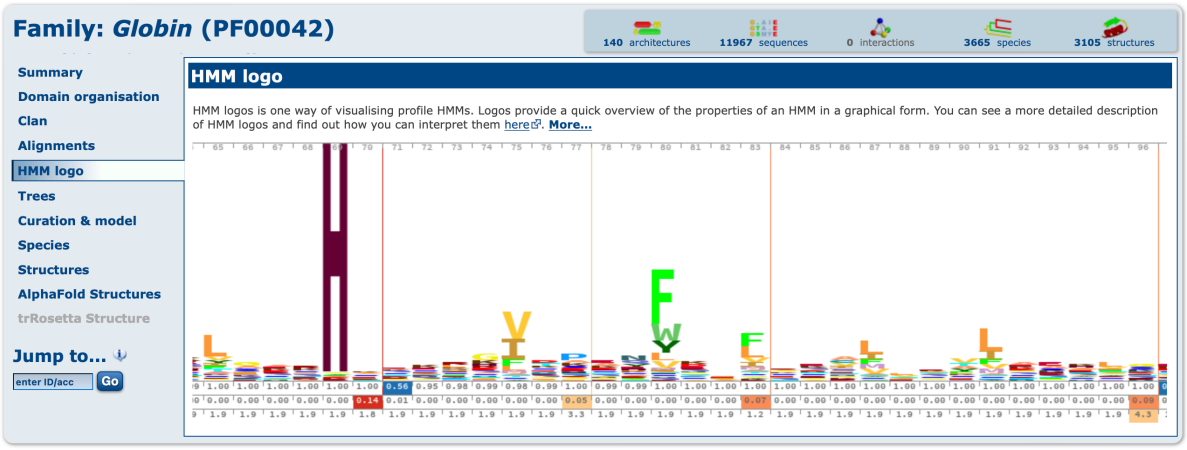



Figure 6: Sequence logos are one way of visualising sequence conservation patterns. Logos provide a quick overview of a given profile in a graphical form.

**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?

Tip

At this stage we will often turn to protein structure inspection just like we did back in our first lab. To help you here I include a molecular figure of beta globin (Figure 7) with the most conserved amino acids in red spheres and other conserved positions in green (unconserved positions in blue).

The most conserved regions are two histidine residues found in alpha helices flanking the heme group.

**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 **Q9** previously (see Figure 8 for an example where we use the neuroglobin structure **4MPM**, chain B).

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):

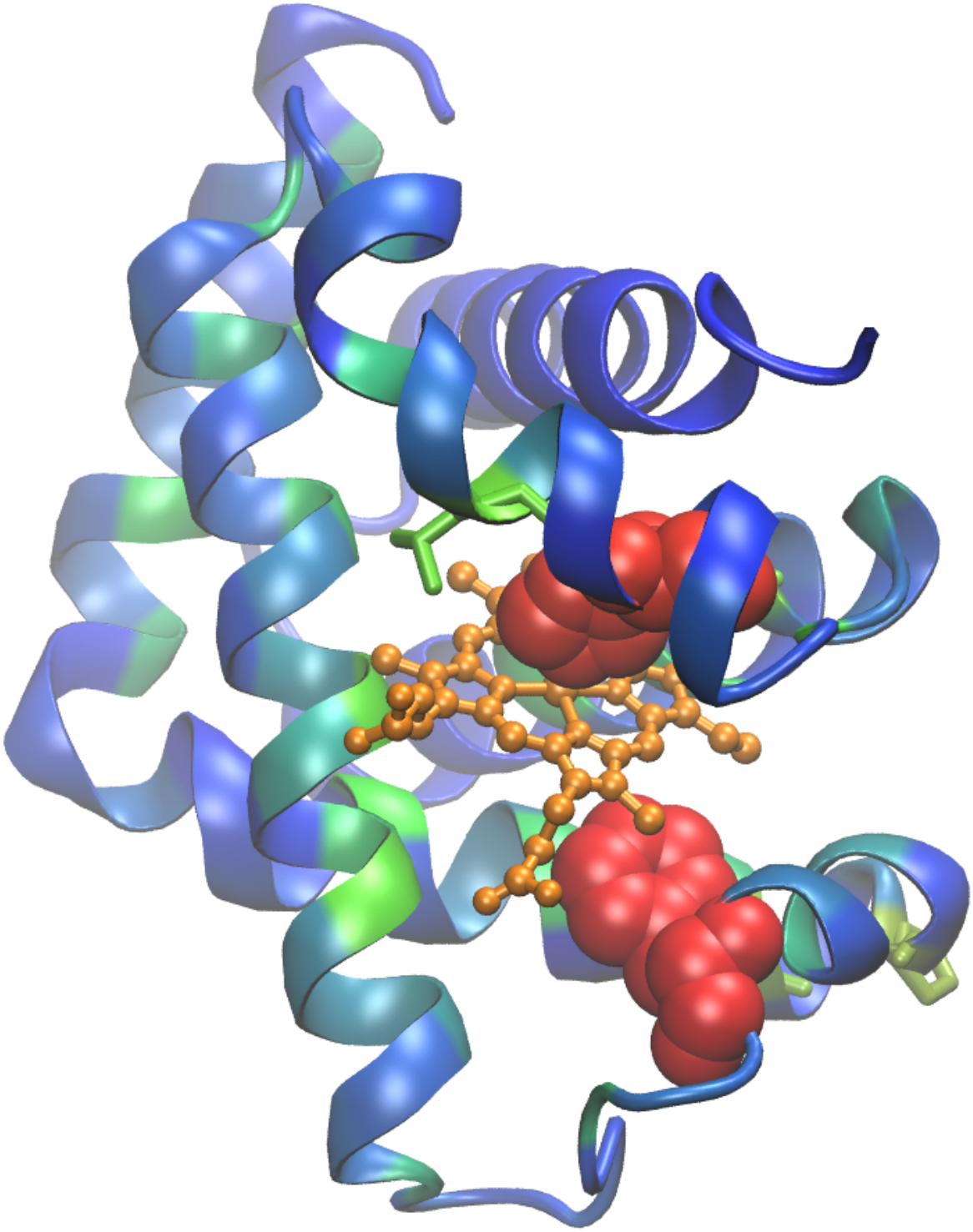


Figure 7: Molecular figure of beta globin where each residue position is colored by the level of conservation in the alignment obtained from HMMER (blue - least conserved, red - most conserved). This information should help you answer Q11.

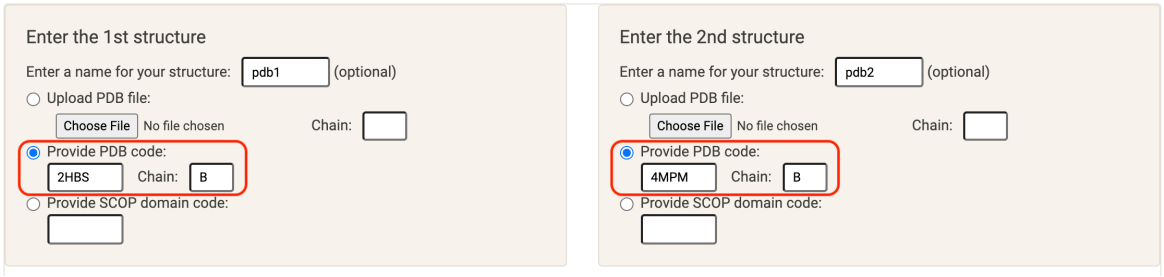
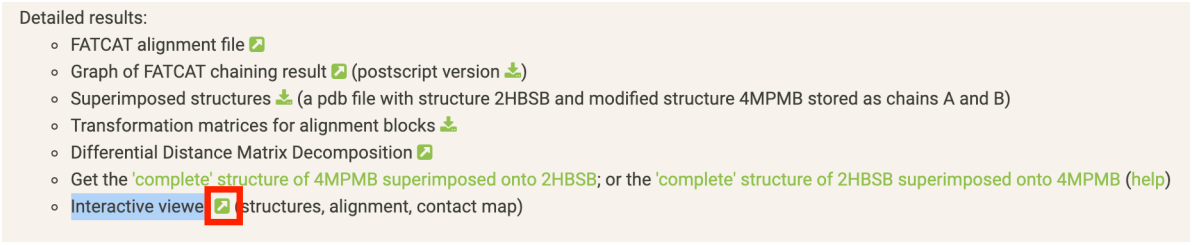
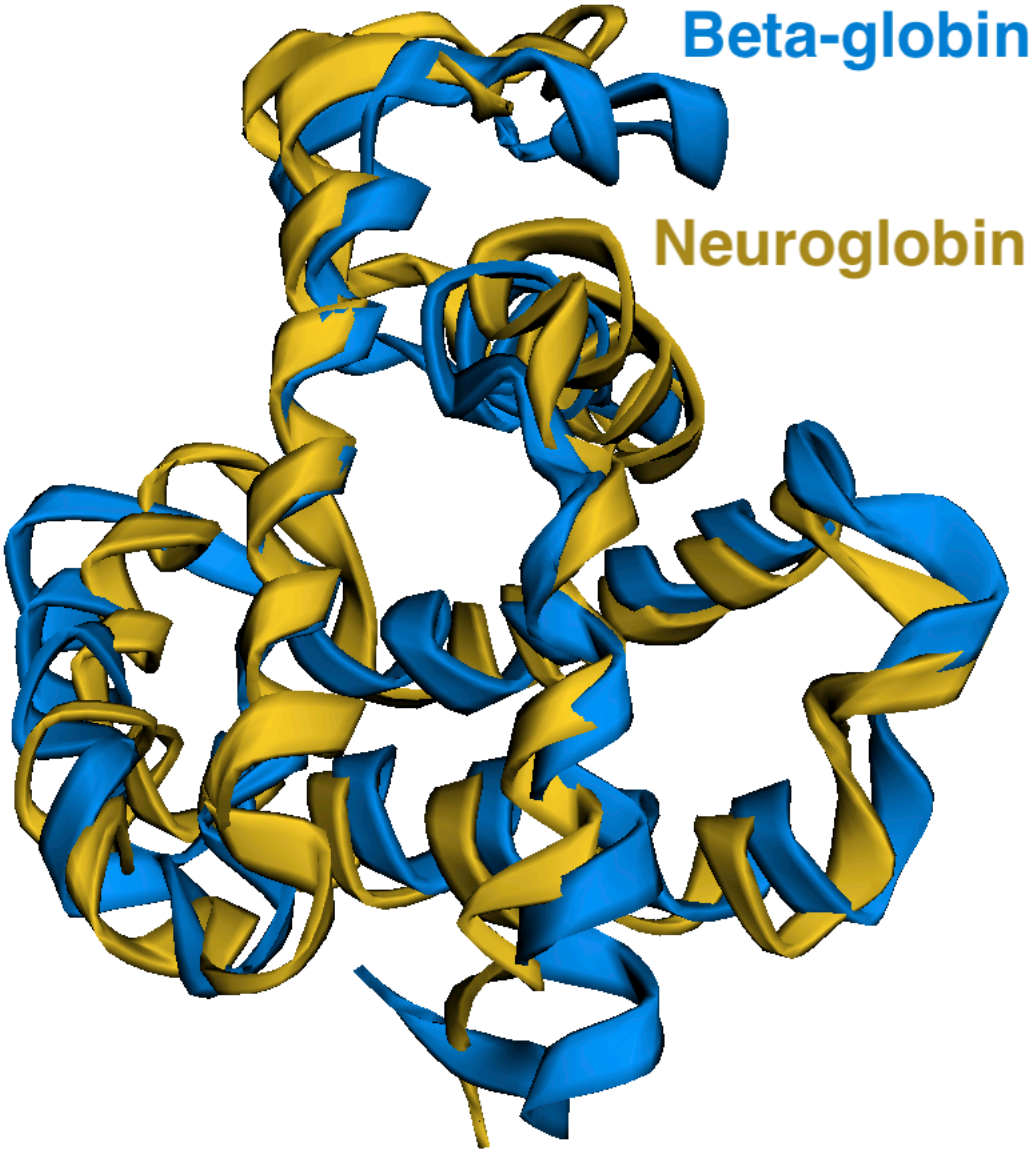



Figure 8: Performing structureal alignment and superposition of beta globin and neuroglobin




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)?

Tip

Try the different coloring and display options to help answer this question.

The regions with the most variance are in the extremities and outer loops of each protein. In particular, a loop in the vicinity of R47.

## Discussion

Structural alignment approaches like those employed in the last section highlight how protein structure similarities can remain robust even as sequence similarities fade below our conventional sequence based detection limits during the course of evolution.

> **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 that are dificult and often imposible to find from sequence alone. One of the major advances in molecular biology and bioinformatics of the last deceade, **AlphaFold** released last summer, is rapidly changing this narative. We will discuss and learn how to use use AlphaFold in a later lab after we learn some bioinformatics codding. Learning codding together will take us to the next level and enable us to do advanced bioinformatics analysis like this.

**Q13**: What one part of this exercise or associated lecture material is still confusing? If appropriate please also indicate the question number from this document and answer the question in the following anonymous form: **Mudy_Point_Assesment_Form** Your comments will let us know which material needs to be further clarified and will help us gain stronger control of the material in this course. Thank you!