

Class 3 Lab*

Advanced Database Searching

Barry Grant

2026-01-12

i 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

Learning Objectives

By the end of this lab, you will be able to:

1. **Recognize** the limitations of conventional BLAST for detecting distantly related (remote) homologs;
2. **Build** position-specific scoring matrices (PSSMs) from aligned sequences and use these to score the similarity of new sequences;
3. **Use** PSI-BLAST and HMMER to detect remote homologs that standard BLAST misses;
4. **Identify** conserved and invariant amino acid positions from multiple sequence alignments and relate these positions to their potential functional roles;

*<http://thegrantlab.org/teaching/>

5. **Perform** pairwise structural alignments to reveal evolutionary relationships that are undetectable by sequence comparison alone;
6. **Appreciate** that protein structure is more conserved than protein sequence during evolution.

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 lab we will explore the detection limits of conventional BLAST ([Section 1](#)) and introduce more sensitive, but often more time consuming, approaches including **Profiles** ([Section 2](#)), **PSI-BLAST** ([Section 3](#)) and **Hidden Markov Models** (HMMs) ([Section 4](#)). Finally for today we turn to time consuming **structural alignment** approaches ([Section 5](#)). Collectively, these sections 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.

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]
MVHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPWTRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLNLDLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHFGKEFTPPVQAAYQKVVAGVAN
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.

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.

Many useful ‘rules of thumb’ are expressed in terms of percent 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?

Let’s 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.



The image shows a screenshot of the 'Program Selection' section on the NCBI Protein BLAST submission page. The section is titled 'Program Selection' and contains a list of algorithms with radio buttons next to them. The selected algorithm is 'PSI-BLAST (Position-Specific Iterated BLAST)'. Below the list is a link that says 'Choose a BLAST algorithm' with a small icon.

Program Selection	
Algorithm	<input type="radio"/> blastp (protein-protein BLAST)
	<input checked="" type="radio"/> PSI-BLAST (Position-Specific Iterated BLAST)
	<input type="radio"/> PHI-BLAST (Pattern Hit Initiated BLAST)
	<input type="radio"/> DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
	Choose a BLAST algorithm

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?

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.

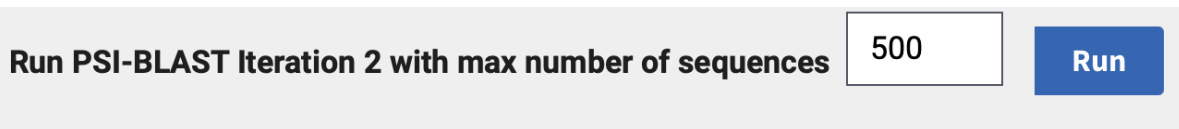


Figure 2: You must manually inspect results before each PSI-BLAST iteration

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 percent identity and coverage to help judge your alignment results.

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).

i Your search is limited to records that include: humans (taxid:9606)

Job Title	gi 4504349 ref NP_000509.1 hemoglobin subunit...
RID	KP8H5M4M013 Search expires on 10-05 01:55 am Download All ▼
Program	PSI-BLAST Iteration 2 Citation ▼
Database	refseq_select_prot See details ▼
Query ID	lcl Query_14534
Description	gi 4504349 ref NP_000509.1 hemoglobin subunit beta [Ho ...
Molecule type	amino acid
Query Length	147
Other reports	Distance tree of results Multiple alignment MSA viewer ?

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).

Alignments Select All [Re-align](#) Mouse over the sequence identifier for sequence title

View Format: [?](#) **Conservation Setting:** [?](#)

Figure 4: Color by *Identity* to highlight invariant amino acids across 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.

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? Later we will have a look at their proximity to each other and the heme group in example structures.

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.

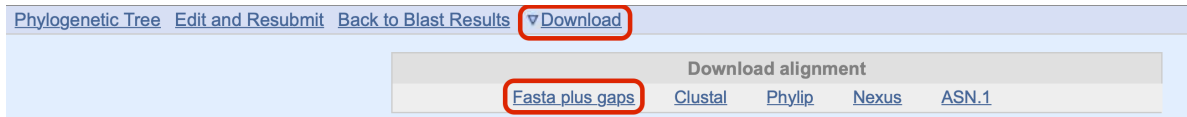


Figure 5: Be sure to download the FASTA format aligned (i.e. plus gaps) globin sequences as we will use these in the next lab

Section 4: Using HMMER (OPTIONAL)

HMMER is an alternative sequence search and alignment method that employs 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.

Let's 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, if available, setting the “**Restrict by Taxonomy**” to **9606**, how do your results compare to those from regular BLAST and PSI-BLAST?

Tip

The EBI HMMER server can be very slow if there are multiple users. You may need to skip and come back to this question - if so just note “server unresponsive” as your temporary answer here.

Q9 Did you find human myoglobin and human neuroglobin? If so list their main “Target” accession IDs and Secondary Accession IDs (e.g. *P68871* and *HBB_HUMAN*).

Tip

These will likely be towards the end of your results list.

Q10 Are there any human neuroglobin PDB structures available? If so take a record of their PDB codes (PDB database accession number) and corresponding chain ID for later.

Tip

Clicking on the target ID will take you to the corresponding **UniProt** entry where you can scroll down (or click) to see associated “**Structure**” entries. Alternatively, you can search the PDB database using (PSI)BLAST or HMMER just like we did back in lab 1.

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: [PF00042](#). Click on the **Profile HMM Logo** link and determine the most conserved residues in this family.

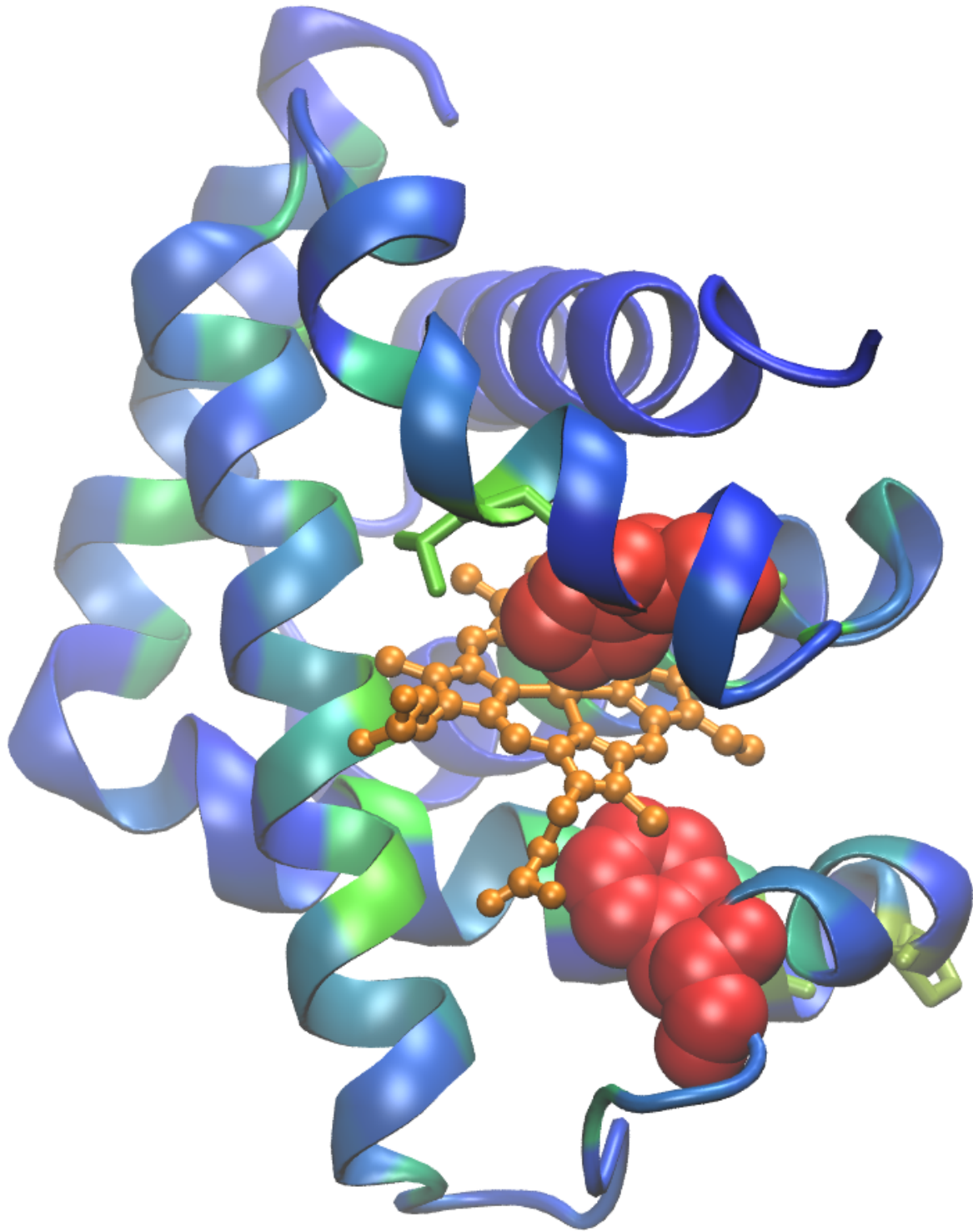


Figure 7: Molecular figure of beta globin where each residue position is colored by the level of conservation in the alignment obtained from HMMER (blue - least conserved, red - most conserved). This information should help you answer Q11.

Note: If the HMMER web server was unresponsive you can search PFAM directly @ <https://www.ebi.ac.uk/interpro/entry/pfam/> 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 a later class. For now we will use the PDB **pairwise structural alignment** server to examine the similarities of our beta globin and neuroglobin proteins.

Visit: <https://www.rcsb.org/alignment> and enter the *PDB code* **2HBS chain B** for the first structure. Then enter one PDB code for neuroglobin you found from answering **Q10** previously (see Figure 8 for an example where we use the neuroglobin structure **4MPM**, chain B).

The screenshot shows the RCSB PDB Pairwise Structure Alignment tool interface. The top navigation bar includes links for Deposit, Search, Visualize, Analyze, Download, Learn, About, Careers, and COVID-19. Below the navigation bar are logos for PDB-101, PDB, EMDataResource, NAKB, wwPDB Foundation, and PDB-IHM. The main heading is "Pairwise Structure Alignment" with a "Help" link. Below the heading is a description: "This tool allows the selection of protein 3D structures for alignment. Use an existing PDB or Computed Structure Model entry ID, upload a local file with atomic coordinates, or enter a URL of a file on the web." There is an "Alignment API" button. The main form area is titled "Compare Protein Structures" and contains two rows of input fields. The first row is for PDB entry 2HBS, chain B, with a range from residue 1 to 146. The second row is for PDB entry 4MPM, chain B, with a range from residue 1 to 151. Below these rows is an "Alignment Method" dropdown menu set to "TM-align". At the bottom of the form are "Compare" and "Clear" buttons.

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

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

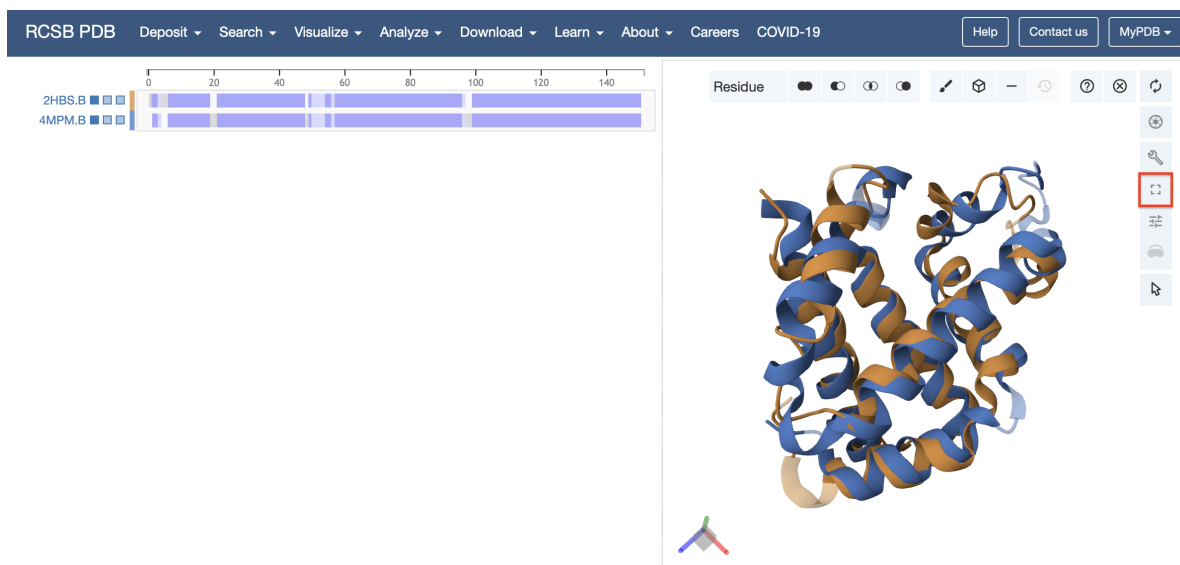


Figure 9: Structural alignment results for beta globin and neuroglobin. Clicking the full screen button (red highlight box) in Mol-star will give you full control of the molecular viewer.

Note how similar in structure these two distant homologues are and explore the different display options in the full screen Mol-star viewer (Figure 9). For the image here I displayed only the aligned chains and displayed the two conserved HIS amino-acids as spacefill (Figure 10):

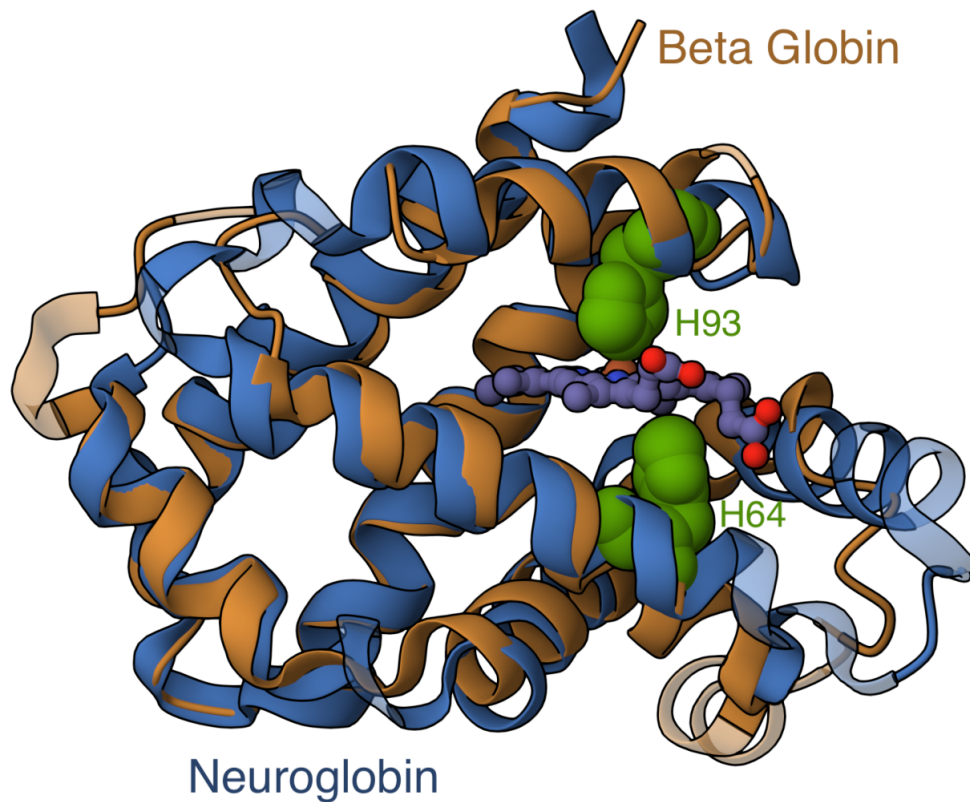


Figure 10: Superposed globins with conserved H64 and H93 displayed in green

Q12 From your explorations in the 3D viewer 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.

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 won't 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 difficult and often impossible to find from sequence alone. One of the major advances in molecular biology and bioinformatics of the last decade, [AlphaFold3](#) released in the summer of 2024, is rapidly changing this narrative. We will discuss and learn how to use AlphaFold in a later lab after we learn some bioinformatics coding. Learning coding 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: [Muddy_Point_Assessment_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!