BIMM-143: INTRODUCTION TO BIOINFORMATICS (week 3)

Sequence Alignment & Database Searching
http://thegrantlab.org/bimm143/
Dr. Barry Grant

Overview: Aligning novel sequences with previously characterized genes or proteins provides important insights into their common attributes and evolutionary origins.

In **sections 1** & **2** of this hands-on session we will first explore the principles and methods underlying the computational comparison and alignment of biomolecular sequences.

In **section 3** we explore how these methods are used to search databases to identify homologues sequences (*i.e.* finding evolutionary related genes or proteins that are descended from a common ancestor).

With the optional extension exercises in **section 4** to **6** we highlight the detection limits of conventional BLAST. We then introduce more sensitive (but often more time consuming) approaches including Profiles, PSI-BLAST and Hidden Markov Models (HMMs).

Finally, in **section 7** we apply "gold standard" structural alignment approaches to highlight how protein structure similarities can remain robust even as sequence similarities fade below or detection limits during the course of evolution.

Section 1: Dot Plot Parameters
Dot plots are a simple graphical approach for the visual comparison of two sequences. They have a long history (see Maizel and Lenk 1981 and references therein) and entail placing one sequence on the vertical axis of a 2D grid (or matrix) and the other on the horizontal.

In its simplest form, a dot is placed where the horizontal and vertical sequence values match. More elaborate forms use '*sliding windows*' composed of multiple characters and a threshold value, or '*match stringency*' for two windows to be considered as matched.
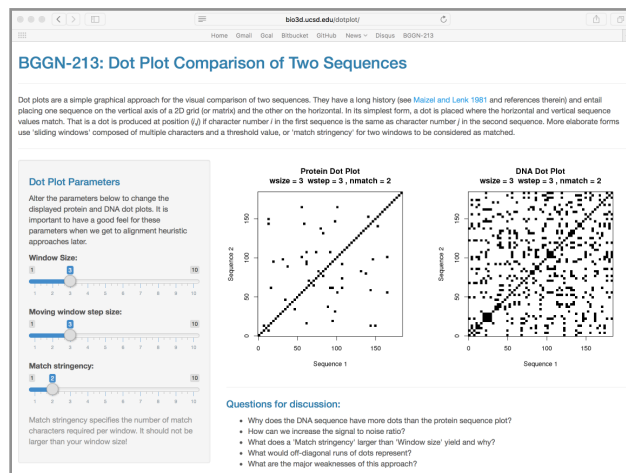
Visit our very own simple dot plot web-app (http://bio3d.ucsd.edu/dotplot/ or it's mirror https://bioboot.shinyapps.io/dotplot/) and get a feel for how altering these major dot plot parameters change the displayed protein and DNA dot plots.

N.B. Note the questions listed on the web page (also found below) and add your answers in space provided on the next page.

Q1. Why does the DNA sequence have more dots than the protein sequence plot? HINT: what do you know about DNA composition vs protein composition?

Q2. How can we increase the signal to noise ratio?

Q3. What does a 'Match stringency' larger than 'Window size' yield and why?

Q4. What are the major weaknesses of this approach? HINT: is your inner nerd happy with this approach? How would you use it to determine if a second set of sequences was more similar to each other than a first set of sequences?

Section 2: Needleman-Wunsch Alignment

Sequence alignment methods often use something called a 'dynamic programming' algorithm that can be usefully considered as an extension of the dot plot approach. Here we have two sample sequences, and we'd like to use the **Needleman-Wunsch algorithm** discussed in class to align them.

Sequence 1: **ATTGC**
Sequence 2: **AGTTC**

		A	G	T	T	C
	0					
A						
T						
T						
G						
C						

Q5. Using a **match score of +2**, a **mismatch score of -1**, and a **gap score of -2**. Fill in the table below and translate it into a alignment. What is the optimal score for this alignment? Is there one unique alignment with this score?

Practice makes perfect. Again use the **Needleman-Wunsch algorithm** discussed in class to align the following sequences:

Sequence 1: **TATAG**
Sequence 2: **GTTAC**

		G	T	T	A	C
	0					
T						
A						
T						
A						
G						

Q6. Using a **match score of 2**, a **mismatch score of -1**, and a **gap score of -2**. Write out your alignment matrix (table), fill in the values and translate your results into all optimal alignments. What is the optimal alignment score for these sequences? Write out all alignments consistent with this score?

Section 3: Finding homologous sequence

Your collaborators found a protein while working on a fly species and have asked you to see if there are any human homologs.

```
>fly_protein
MDNHSSVPWASAASVTCLSLDAKCHSSSSSSSSKSAASSISAIPQEETQTMRHIAHTQRCLSRLTSLVAL
LLIVLPMVFSPAHSCGPGRGLGRHRARNLYPLVLKQTIPNLSEYTNSASGPLEGVIRRDSPKFKDLVPNY
NRDILFRDEEGTGADRLMSKRCKEKLNVLAYSVMNEWPGIRLLVTESWDEDYHHGQESLHYEGRAVTIAT
SDRDQSKYGMLARLAVEAGFDWVSYVSRRHIYCSVKSDSSISSHVHGCFTPESTALLESGVRKPLGELSI
GDRVLSMTANGQAVYSEVILFMDRNLEQMQNFVQLHTDGGAVLTVTPAHLVSVWQPESQKLTFVFADRIE
EKNQVLVRDVETGELRPQRVVKVGSVRSKGVVAPLTREGTIVVNSVAASCYAVINSQSLAHWGLAPMRLL
STLEAWLPAKEQLHSSPKVVSSAQQQNGIHWYANALYKVKDYVLPQSWRHD
```

Q7. Using the default settings for NCBI BLAST, can you find any homologs for this protein in Humans? HINT: try using the *LIMITS* and *FILTERING* options we covered in the last lab.

Q8. Try changing the database to **refseq_protein**. From the results, select a few proteins and find the common name for the species. What trend do you notice as you move down the results list? HINT: search google for the species name.

Q9. Finally, try also limiting the search to only *H. Sapiens*. HINT: you can simply type the Taxon ID **9606** in the "**Organism**" box. What function do these proteins have?

Q10. What function do you think this protein performs for your collaborators' organism?

EXTENSION SECTIONS

The remaining sections of this worksheet are optional. These sections delve deeper into more advanced topics that will be of interest to motivated students.

Section 4: 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.

> **Q11.** What homologs did you find with this simple blastp search? Note their *precent 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.

> **Q12.** Try increasing the **Expect threshold** for your blasts search. What new hits were reported? What about their alignment statistics? Do you trust these matches?
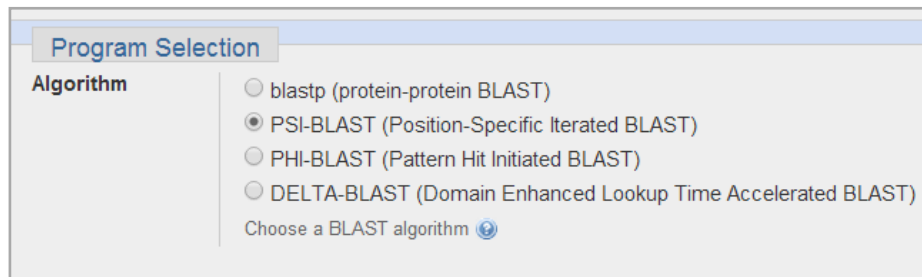
Many useful 'rules of thumb' are expressed in terms of precent identity. If two proteins have more than 45% identical residues in their optimal alignment they typically have very similar structures and are likely to have a similar function. If two proteins have more than 25% identical residues (but less 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 5: 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. 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.

Q13. The first iteration should be similar to your previous blastp search. Did you find any new potential homologs that you did not see previously?

Q14. 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?

Run PSI-Blast iteration 2 with max | 500 | | Go |

Q15. Perform a third iteration. Did the algorithm find any other proteins? Did we find myoglobin and neuroglobin?

Section 6: Using HMMER (**OPTIONAL**: Note *server can be very slow!*)

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.

Q16. 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?

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.

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 Q10.
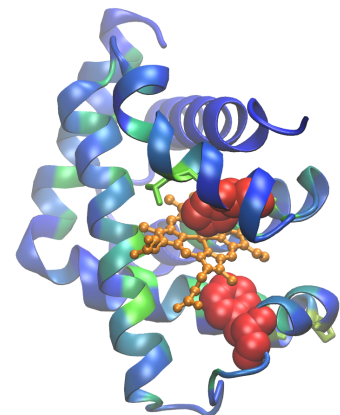
Note: If the HMMER web server was unresponsive you can search PFAM directly @ https://pfam.xfam.org to help answer Q10.

Section 7: 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.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).

 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.
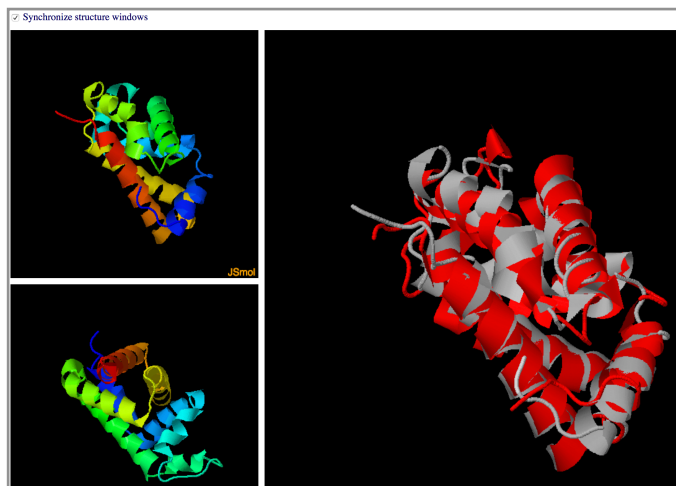
Q20. What one part of this lab or associated lecture material is still confusing?
If appropriate please also indicate the question number from this lab instruction pdf and answer the question in the following anonymous form:

https://forms.gle/FEbKxnq4X7nUMhcn8