BIOINF525: INTRODUCTION TO BIOINFORMATICS LAB SESSION 2

Sequence Alignment

http://bioboot.github.io/bioinf525_w17/module1/#1.2

Dr. Barry Grant

Jan 2017

Overview: Aligning novel sequences with previously characterized genes or proteins provides important insights into their common attributes and evolutionary origins. In this lab session we will explore the principles underlying the computational tools that can be used to compute and evaluate sequence alignments.

Section 1: Dot Plots

There are a number of tools available that attempt to aid the visual comparison of two sequences; here we will be using **YASS** (http://bioinfo.lifl.fr/yass/yass.php) to generate dot plot comparisons. Below are the mRNA sequences for α and β globin.

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1)
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTC
AAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCC
TGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGG
CCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTG
TCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACT
GCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAA
GTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTT
CTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAA
AGTCTGAGTGGGCGGC

>gi|28302128|ref|NM_000518.4| Homo sapiens hemoglobin, beta (HBB)
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATG
CTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGC
TCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGAT
CCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGC
```

Copy these sequences into the two boxes, be sure to click the "**select**" button beside each pasted sequence, and then click "**run YASS**". After it finishes running, select the **simple dotplot** view from the results section.

Q1: List the positions (in terms of first and last nucleotide number) of the first major segment of alpha globin that appears to have a similar region in beta globin? Also note down the “score” for this matching segment. HINT: You can click on the plot to view details of the corresponding sequence segment.

Go back to the sequence submission page and choose an *alternate scoring scheme* and change the ‘*Gap costs (opening,extension)*’ values in the **Parameters** section of the submission page.

Q2: How does changing the parameters, specifically lowering the ‘gap costs’, change the dot plot’s overall appearance and the “score” for the segment you noted in Q1?

Section 2: 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
```

Q3. 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.

Q4. 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.

Q5. 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?

Q6. Are the proteins you found examples of paralogs or orthologs? Given your answer to Q6, what function do you think this protein performs for your collaborators' organism?

Section 3: Finding Distant Relationships

Your collaborators found a transcript while working with a *Drosophila* species, and were wondering if similar sequences had been found outside of *Drosophila*.

```
>fly_mRNA
AGAAGCTCAACCAGGAGAACGAACAGTCGGCAAACAAGGAGAACGACTGCGCTAAGACGGTAATTTCGCCATCCTCC
AGCGGCCGTTCCATGAGTGACAACGAGGCCAGCTCCCAGGAAATGTCCACCAACCTCAGGGTGCGCTACGAACTAAA
GATCAACGAGCAGGAGGAGAAGATCAAGCAGTTGCAGACGGAAGTAAAGAAGAAGACGGCGAATCTGCAAAATCTGG
TCAACAAGGAGCTATGGGAGAAAAATCGTGAGGTGGAGCGCCTCACTAAGCTGCTGGCTAACCAACAGAAGACGTTG
CCACAGATAAGTGAGGAATCCGCCGGAGAAGCAGATCTGCAGCAATCCTTCACGGAGGCGGAGTACATGAGGGCATT
GGAGCGAAACAAGCTGCTGCAGCGAAAGGTGGATGTGCTCTTCCAGCGCCTGGCAGACGATCAACAGAACAGCGCTG
TGATTGGGCAGTTGCGTTTGGAACTTCAACAAGCTCGCACGGAAGTCGAGACGGCGGATAAGTGGCGTCTTGAATGC
GTCGATGTCTGCAGTGTGCTGACAAACCGATTGGAAGAGCTGGCTGGTTTCCTCAACTCTCTGCTGAAGCACAAAGA
TGTTCTTGGCGTGTTGGCCGCTGATCGACGCAATGCCATGCGTAAGGCGGTGGATCGCAGCTTGGATCTTTCCAAGA
GTCTTAATATGACTCTGAATATAACAGCTACATCCTTGGCTGATCAAAGCCTCGCTCAGCTGTGCAATCTATCCGAG
ATCTTGTACACCGAAGGTGATGCAAGCCACAAAACTTTCAATTCCCACGAAGAGCTGCACGCCGCTACTTCGATGGC
TCCGACTGTAGAGAACTTAAAGGCCGAGAATAAGGCTCTTAAAAAGGAGTTGGAAAAGCGACGCAGCTCAGAAGGAC
AGAGGAAAGAGCGCCGCTCCTTACCGCTGCCCTCCCAGCAGTTCGATAACCAGAGCGAGTCAGAGGCCTGGTCAGAG
CCTGACCGCAAGGTTTCCTTGGCACGCATTGGCCTGGACGAAACCTCCAACAGTTTGGCAGCGCCTGAGCAGGCGAT
CAGCGAGTCGGAGAGCGAGGGA
```

Q7. Using standard blastn, were you able to find similar sequences outside of *Drosophila*? If so, in which species?

Q8. Now, use blastx instead of blastn. What species did you find this time?

Q9. Among the more distantly related sequences, are there any regions that seem to be more conserved than others? If so, what do these regions correspond to?

Q10. Return to the search page for blastx, and limit the search to only humans. Did you find any related human sequences?

We can increase the sensitivity of our search by changing some of the algorithm parameters. By changing the **substitution matrix**, we can change how the alignment scoring is performed and potentially find more distant evolutionary relationships. We can also change the **Expect Threshold** to return alignments with higher e-values.

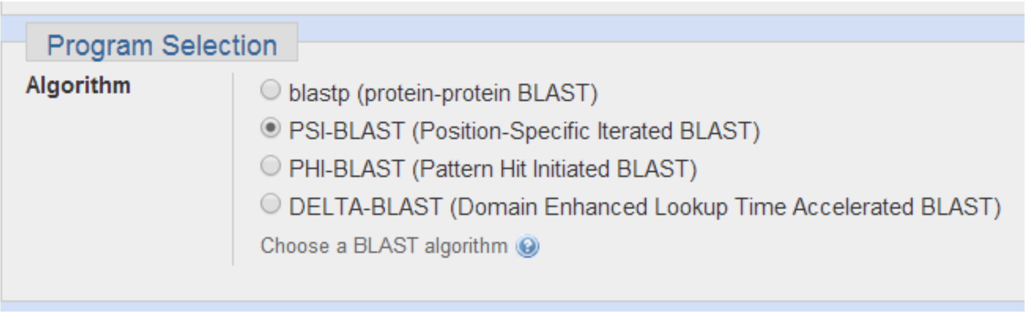
Q11. Change your matrix and set the Expect Threshold to 100 while searching only within humans. Did you find anything? If so, is it meaningful? HINT: Large numbers with PAM suggest larger evolutionary distance, but large numbers with BLOSUM suggest closer relationships.)

Section 4: Using PSI-BLAST

Let's return to the HBB protein that we explored last week 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
```

We'd like to find more distant homologs for this protein by using PSI-BLAST. After selecting **blastp** and entering the sequence, select the **PSI-BLAST** algorithm from the 'Program Selection' options section. Also, in order to focus our results, let's use **refseq_protein** and search only in humans again.



Program Selection

Algorithm

○ blastp (protein-protein BLAST)

● PSI-BLAST (Position-Specific Iterated BLAST)

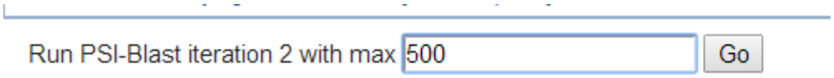
○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Q12. The first iteration should be similar to a simple blastp search. What homologs did you find with this?

Q13. 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

Q14. Perform a third iteration. Did the algorithm find any other proteins? 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 results page is a '**formatting options**' link. Using the available options try 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 the **Jalviewer** link (under the Result Summary tab) to display the resulting alignment along with per-residue conservation scores.

Q15. 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 Q17 after further analysis.

Section 5: Using HMMER

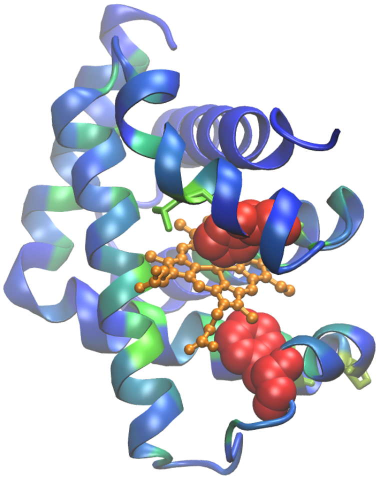
HMMER is an alternative sequence search and alignment method that employees probabilistic models called profile hidden Markov models (HMMs). HMMER aims to be significantly more accurate and more able to detect remote homologs than BLAST because of the strength of its underlying mathematical models. In 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.

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 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.

Q17. 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 Q17.

(Optional) Section 6: 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						

Q18. Using a **match score of 2**, a **mismatch score of -1**, and a **gap score of -2**. Fill in the table and translate it into a alignment. What is the optimal score for this alignment? Is there one unique alignment with this score?

Q19. 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:

http://tinyurl.com/bioinf525-lab1-2