Class 2 Lab*

Sequence Alignment & Database Searching (Pt. 1)

Barry Grant

Version 220926

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

Nhi To A18053310 tuto@ucsd.edu

Overview:

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

In sections 1, 2 and 3 of this hands-on session we will first explore the principles and methods underlying the computational comparison and alignment of biomolecular sequences.

In section 4 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).

In section 5 we highlight the detection limits of conventional BLAST. This sets the scene for introducing more sensitive (but often more time consuming) approaches including Profiles, PSI-BLAST and Hidden Markov Models (HMMs).

^{*}http://thegrantlab.org/teaching/

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 (Link1 or it's mirror Link2) 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 the space provided on the next page.



Figure 1: Interactive dot plot web app to explore fundemental concepts of biomolecular sequence comparasion

Q1 Why does the DNA sequence have more dots than the protein sequence plot?



What do you know about DNA composition vs protein composition?

The DNA sequence has more dots than the protein sequence plot because there is a higher chance for the sequence code to match by nucleotide than by codon. DNA sequence is by each individual nucleotide, while protein sequence is by each codon (3 nucleotides). Finding a match between 4 options (A,C,T,G) is more likely than finding a match between 20 codon options.

Q2 How can we increase the signal to noise ratio?



? Tip

Signal in this case means correct matches that we actually want to highlight and noise means spurious matches that we don't want.

To increase the signal to noise ratio, one should increase the 'window size' and 'match stringency' because the data plot become more clear, and there's less noise. We are increasing the amount of signal there is while decreasing the amount of noise.

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

A 'Match stringency' larger than 'Window size' yields an error because if the number of match characters required per window is larger than the window size, it is out of the parameters. There cannot be more matches per window more than the actual window size.

Q4 What are the major weaknesses of this approach?



? Tip

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?

The major weakness of this approach is that one can only compare two sequences and that it is difficult to find the best 'window size' and 'match stringency'. In addition, it is hard to see if there is a more complex relationship other than character matching. This approach is a visual appraoch, and thus similarities and differences are hard to distinguish without including numbers into the analyzation.

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. Feel free to use the clasroom white-boards and/or pen and paper and attach a photo to this PDF for gradescope.

Q5 Using a match score of +2, a mismatch score of -1, and a gap score of -2. Fill in the table below (or use pen and paper) for the following two sequences:

Sequence 1: ATTGC Sequence 2: AGTTC

From your completed table what is the optimal score and corresponding alignment (with one sequence above the other)?

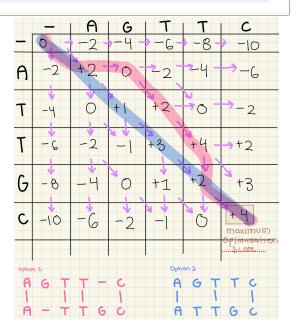
? Tip

It can be hard to store the all important progress arrows in the PDF version of this document and thus you may prefer to use your own paper (or white-board) version that you can take a photo off for upload to gradescope.

The optimal score is +4 and the corresponding alignment would be

Sequence 1: A G T T - C Sequence 2: A G T T G C

		A	G	Т	Τ	С
	0					
A						
Т						
Т						
G						
С						



Section 3: Practice makes perfect

Again use the Needleman-Wunsch algorithm discussed in class to align the following sequences:

Q6 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?

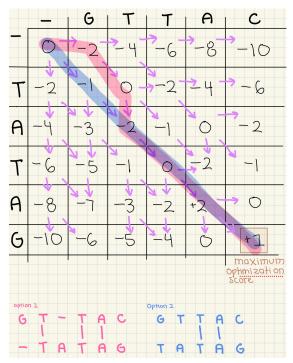
Sequence 1: TATAG

Sequence 2: GTTAC

The optimal score is +1 and the corresponding alignment would be

Sequence 1: G T - T A C Sequence 2: - T A T A G

		G	Т	Т	A	С
	0					
Т						
A						
Т						
A						
G						



Section 4: 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?



? Tip

Try using the LIMITS and FILTERING options we covered in the last lab.

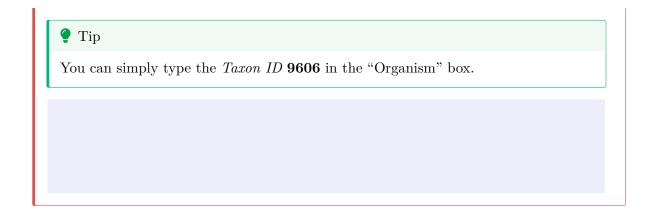
Q8 Try changing the database to refseq_protein. From the results, select a few proteins and find the common name for the species. What trend do you notice as you move down the results list?



Tip

Search google for the species name and use the taxonomy tab on your NCBI BLAST results page.

Q9 Finally, try also limiting the search to only H. Sapiens. What function do these proteins have?



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

OPTIONAL EXTENSION

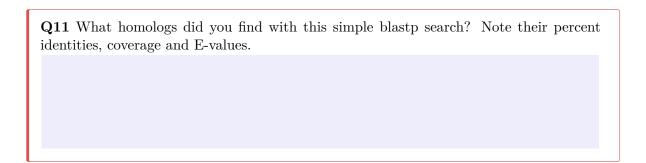
We will revisit this problem and introduce approaches with greater sensitivity (i.e. ability to find more remote homolouges) in the next lab.

Section 5: 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 HBB 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.



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

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_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!

Discussion

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 the next lab.

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".

Our next class will introduce more advanced topics including profile and structure based approaches that can delve deeper into these important, but often hard to detect, sequence-structure-function relationships.