Class 2 Lab*

Sequence Alignment & Database Searching (Pt. 1)

Barry Grant

Version 220926

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
Shreyas Sankaranarayanan	A17077452	shsankaranarayanan@ucs

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.

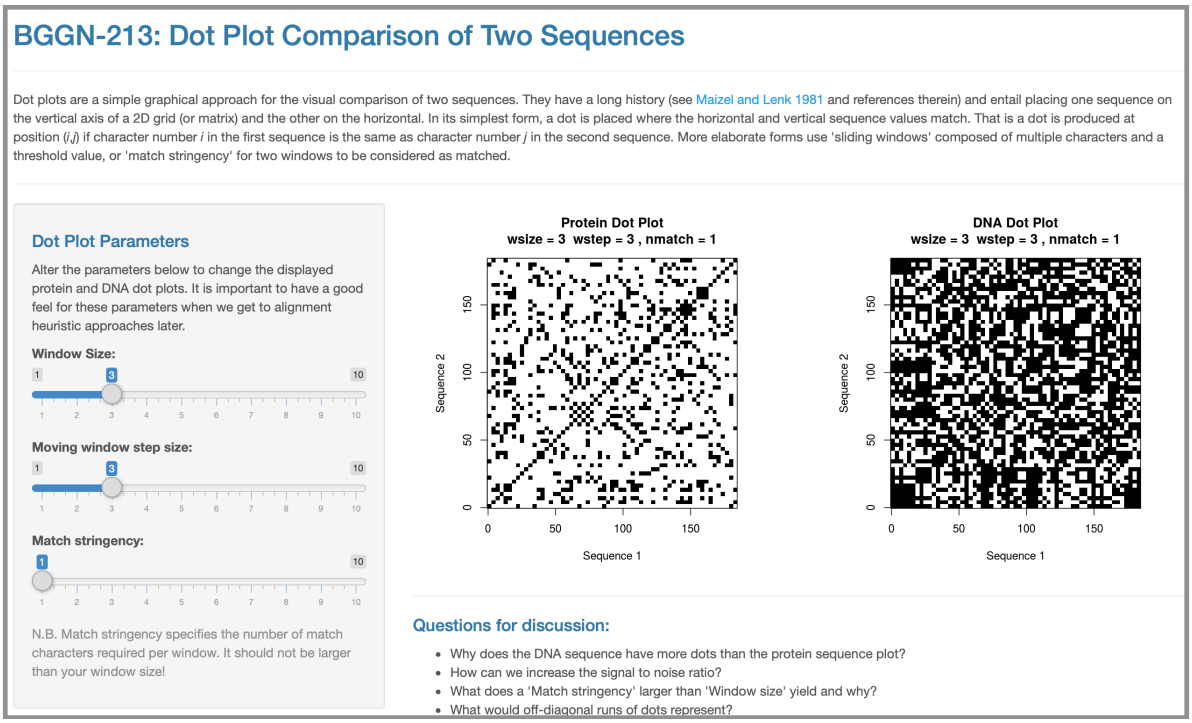

Figure 1: Interactive dot plot web app to explore fundemental concepts of biomolecular sequence comparasion

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

> Tip
>
> What do you know about DNA composition vs protein composition?

DNA is composed of single nucleotides so in order for a match to occur in the window size of 3 above with a match stringency of 1 only a single nucleotide (out of 4) has to be repeated. This is not the same for protein which is consistent of amino acids that are created by codon of 3 nucleotides which requires more specificity. This is comprcomparable to increasing match stringency by 3 for the DNA plot. Also, there are 20 different amino acids meaning there is less likelv to be a match.

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.

Reducing window noise and increasing match strignency. This is because a larger window size allows for greater potential of matches and increasing match stringency reduces the number of matches due to the need for consecutive matches. We could also see less noise by increasing step size as this would mean less cross-checks meaning less matches. Increasing alphabet size.

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

An error occurs as the required number of matches is higher than the number of elements checked.

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?

Does not show an actual comparative alignment, but rather only provides a detection of matches. We cannot determine how to optimally align 2 sequences compared to each other. Lots of unnecessary noise, and we don't get a clearly outlined idea of mutation and indel locations.

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 equal to 4 and the corresponding alignment is:
match, gap, match, match,gap,match

```
AG TT - C
| | || | |
A - TTGC
```

		A	G	T	T	C
	0	-2	-4	-6	-8	-1
A	-2	2	0	-2	-4	-6
T	-4	0	1	2	0	-2
T	-6	-2	-1	3	4	2
G	-8	-4	0	1	2	3-
C	-1	-6	-2	-1	0	4

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 is:
gap, match, gap, match, match, mismatch

GT-TAC
| ||| ||
- TATAG

		G	T	T	A	C
	0	-2	-4	-6	-8	-1
T	-2	-1	0	-2	-4	-6
A	-4	-3	-2	-1	0	-2
T	-6	-5	-1	0	-2	-1
A	-8	-7	-3	-2	2	0
G	-1	-6	-5	-4	0	1

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.

The top 4 hits are:
1. NP_066382.1, 50.13%, 2e-122, 81% (Accession #, % Identity, E-Value, Query Coverage)
2. BAG36160.1, 50.13%, 3e-122, 81% (Accession #, % Identity, E-Value, Query Coverage)
3. AEF38366.1, 49.87%, 5e-120, 81% (Accession #, % Identity, E-Value, Query Coverage)
4. NP_002172.2, 46.97%, 2e-110, 78% (Accession #, % Identity, E-Value, Query Coverage)

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.

These genes are present heavily in flies and mosquitoes and the E-Value gets larger and the percent identity gets lower. Some hits include:

1. NP_001034065.1, Drosophila Melanogaster (Common Fruit Fly)
2. XP_030370903.1, Scaptodrosophila Lebanonensis (Vinegar Flies)
3. XP_017471220.1, Rhagoletis Zephyria (Fruit Fly)

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

Tip

You can simply type the *Taxon ID* **9606** in the "Organism" box.

The top 4 hits are:
1. NP_066382.1, 50.13%, 7e-123, 81% (Accession #, % Identity, E-Value, Query Coverage)
2. NP_002172.2, 46.97%, 7e-111, 78% (Accession #, % Identity, E-Value, Query Coverage)
3. NP_000184.1, 49.86%, 1e-108, 71% (Accession #, % Identity, E-Value, Query Coverage)
4. XP_016874869.1, 48.39%, 5e-89, 64% (Accession #, % Identity, E-Value, Query Coverage)

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

The functions of this protein include:

- enabling calcium ion binding
- enables cholesterol-protein transferase activity
- enables morphogen activity
- enables patched binding

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.

Q11 What homologs did you find with this simple blastp search? Note their percent identities, coverage and E-values.

The top 4 hits include
1. NP_000509.1, 100.00%, 3e-106, 100% (Accession #, % Identity, E-Value, Query Coverage)
2. NP_000510.1, 93.20%, 1e-99, 100%(Accession #, % Identity, E-Value, Query Coverage)
3. NP_005321.1, 75.51%, 2e-82, 100%(Accession #, % Identity, E-Value, Query Coverage)
4. NP_000175.1, 73.47%, 3e-80, 100% (Accession #, % Identity, E-Value, Query Coverage)

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?

The new hits reported include
1. NP_001006940.3
2. XP_016883210.1
3. NP_001001410.1
4. NP_001356276.1

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!

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.