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
Rachel Field	A69042948	rafield@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.

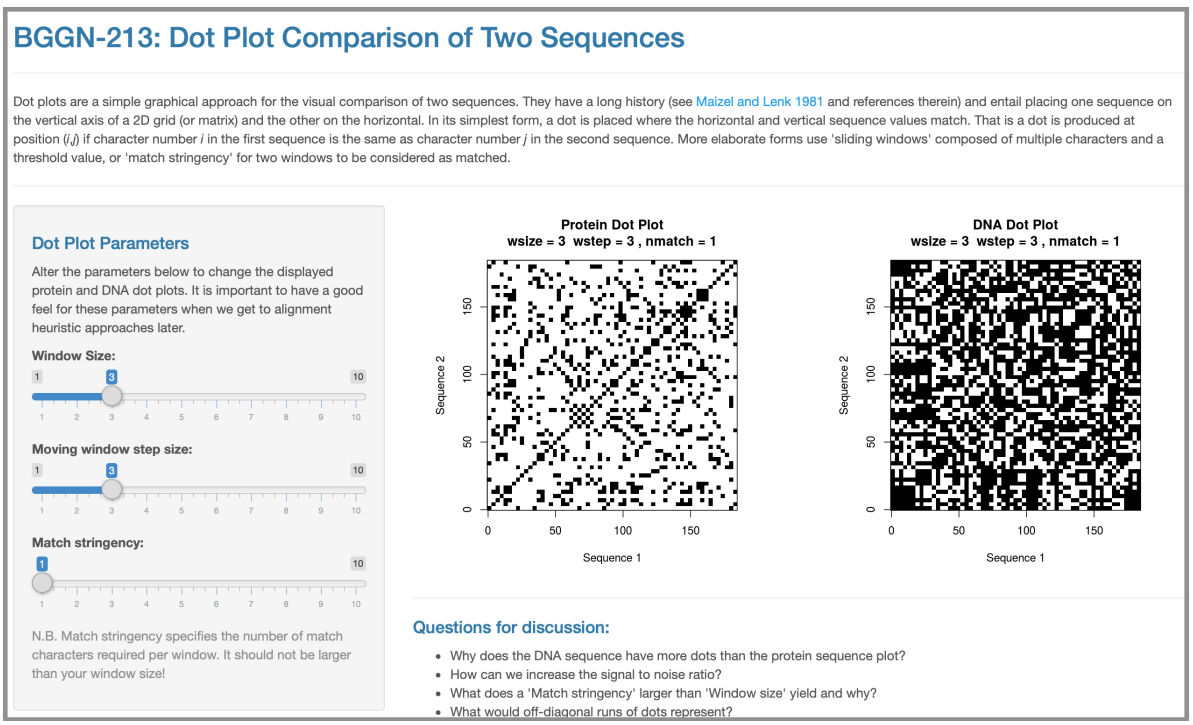

Figure 1: Interactive dot plot web app to explore fundemental concepts of biomolecular sequence comparasion

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

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

There are 4 options for the identity of each nucleotide, versus 20 options for the identity of each amino acid, so with 4 options versus 20, there will be more matches between two DNA sequences than two protein sequences.

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.

We can translate the nucleotide sequence to an amino acid sequence, since protein sequences have less noise due to having a lower chance of randomly matching. You can also decrease the window size for a nucleotide sequence; this increase the signal to noise ratio because there are less random matches with less nucleotides included. Additionally, you can increase the stringency; since it is less likely that, for example, 3 nucleotides in a given window will match than 2 nucleotides, increasing the stringency from 2 to 3 reduces the

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

Match stringency is the number of match characters required per window. If this number is larger than the window size, you cannot reach this required number, and you get an error.

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?

You can see it but you can't quantify it! So you could "see" if one set of sequences is more similar than another, but there is no number to back this up.

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.
Optimal alignment:
AGTTC
ATTGC

AGTT-C

		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

see Excel & photo, included in submission

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

Optimal score: 1
Optimal alignment:
TATAG
GTTAC

		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	

see Excel, included in submission

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.

No

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.

No; there is no human - it's all Drosophila

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.

desert hedgehog protein preproprotein [Homo sapiens]
indian hedgehog protein preproprotein [Homo sapiens]
sonic hedgehog protein isoform 1 preproprotein [Homo sapiens]

Function: It is involved in cell fate specification and cell-cell signaling, by functioning in the hedgehog signaling pathway. It enables calcium binding, cholesterol-protein transferase

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

It functions in the hedgehog signaling pathway in fly, so it's involved in developmental morphogenesis, such as limb development.

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.

hemoglobin subunit beta [Homo sapiens] - 100% identity, 100% query cover, E = 3 e -106
hemoglobin subunit delta [Homo sapiens] - 93.20% identity, 100% query cover, E = 1 e -99
hemoglobin subunit epsilon [Homo sapiens] - 75.51% identity, 100% query cover, E = 2 e -82
hemoglobin subunit gamma-2 [Homo sapiens] - 73.47% identity, 100% query cover, E = 3 e -80
hemoglobin subunit gamma-1 [Homo sapiens] - 72.79% identity, 100% query cover, E = 4 e

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_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.

		A	G	T	T	C
	0	-2	-4	-6	-8	-10
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	-10	-6	-2	-1	0	4 ♡

	_	A	G	T	T	C
_	0	-2	-4	-6	-8	-10
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	-10	-6	-2	-1	0	4

jagged path					*use gap and G bc it's vertical movement
hor seq	A	G	T	T	_	C
vert seq	A	_	T	T	G	C
		*use G and gap bc it's horizontal movement				
all-diagonal path						
hor seq	A	G	T	T	C	
vert seq	A	T	T	G	C	

match score of +2, a mismatch score of -1, and a gap score of -2.

for each square, you have 3 options; try all of them and see which gives the

*for horizontal and vertical, always use a gap score; for diagonal, use a ma

*do the top and left-most rows using horizontal and vertical at first

choose your best path (there are 2 here)

along the best path, fill the hor/vert spots with gaps, fill the diagonal with matches or mismatch

top row of photo is the result for the hor sequence, bottom row is the vert sequence

& the top-written sequence in black is for the jagged path, bottom-written sequence is for the all-diagonal path

C3		C5		D7		F3		G3	
diagonal	**=0+(2) = 2**	diagc	= -4+-1 =	diagc	= -4 + -1 = -5	diagonal	= -6 + -1 = -7	diagona	= -8 + -1 = =-9
horizontal right	=-2 + (-2) = -4	horiz	= -6 + -2 =	horiz	= -6 + -2 = -8	**horizontal right**	**= -2 + -2 = -4**	**horizon**	**= -4 + -2 = -6**
vertical down	=-2 + (-2) = -4	**verti**	**= 0 + -2 =**	**verti**	**= 0 + -2 = -2**	vertical down	= -8 + -2 = -10	vertical	= -10 + -2 = -12
C4		C6		E4		F4		G4	
diagonal	=-2 + (-1) = -3	diagc	=-6 + -1 =	**diagc**	**= 0 + 2 = 2**	diagonal	-2 + 1 = -1	diagona	= -4 + -1 = -5
horizontal right	=-4 + (-2) = -6	horiz	= -8 + -2 =	horiz	= 1 + -2 = -1	**horizontal right**	**=2 + -2 = 0**	**horizon**	**= 0 + -2 = -2**
vertical down	**=-2 + (-2) = 0**	**verti**	**= -2 + -2 =**	vertic	= -2 + -2 = -4	vertical down	= -4 + -2 = -6	vertical	= -6 + -2 = -8
D3		C7		E5		F5		G5	
diagonal	=-2 + (-1) = -3	diagc	=-8 + -1 =	**diagc**	**1 + 2 = 3**	**diagonal**	**= 2 + 2 = 4**	diagona	0 + -1 = -1
horizontal right	**=2 + (-2) = 0**	horiz	= -10 + -2	horiz	= -1 + -2 = -3	horizontal right	= 3 + -2 = 1	**horizon**	**4 + -2 = 2**
vertical down	= -4 + (-2) = -6	**verti**	**= -4 + -2 =**	vertic	= 2 + -2 = 0	vertical down	= 0 + -2 = -2	vertical	-2 + -2 = -4
example:		D4		E6		F6		G6	
diagonal	**=2 + (-1) = 1**	**diagc**	**= 0 + -1 =**	diagc	-1 + -2 = -3	**diagonal**	**= 3 + -1 = 2**	**diagona**	**4 + -1 = 3**
horizontal right	=0 + (-2) = -2	horiz	= -2 + -2 =	horiz	= 0 + -2 = -2	horizontal right	= 1 + -2 = -1	horizon	= 2 + -2 = 0
vertical down	0 + (-2) = -2	**verti**	**1+ -2 = -1**	**verti**	**= 3 + -2 = 1**	**vertical down**	**= 4 + -2 = 2**	vertical	= 2 + -2 = 0
E3		D6		E7		F7		G7	
diagonal	-4 +-1 = -5	**diagc**	**-2 + 2 = 0**	**diagc**	**= 0 + -1 = -1**	**diagonal**	**= 1 + -1 = 0**	**diagona**	**= 2 + 2 = 4**
horizontal right	**0 + (-2) = -2**	horiz	-4 +-2 = -6	horiz	= -2 + -2 = -4	horizontal right	= -1 +-2 = -3	horizon	= 0 + -2 = -2
vertical down	-6 + (-2) = -8	vertic	-1 + -2 = -	vertic	= 1 + -2 = -1	**vertical down**	**= 2 + -2 = 0**	vertical	= 3 + -2 = 1

_	G	T	T	A	C
0	-2	-4	-6	-8	-10
-2	-1	0	-2	-4	-6
-4	-3	-2	-1	0	-2
-6	-5	-1	0	-2	-1
-8	-7	-3	-2	2	0
-10	-6	-5	-4	0	1

match score of +2, a mismatch score of -1, and a gap score of -2.

diagonal
horizontal right
vertical down

Calculations:

C3		D3		E3		F3		G3	
diagonal	**=0+(-1)=-1**	**diagonal**	**=-2+2=0**	**diagonal**	**=-4+2=-2**	diagonal	=-6+(-1)=-7	diagonal	=-8+(-1)=-9
horizontal right	=-2+(-2)=-4	horizontal right	=-1+(-2)=-3	**horizontal right**	**=0+(-2)=-2**	**horizontal right**	**=-2+(-2)=-4**	**horizontal right**	**=-4+(-2)=-6**
vertical down	=-2+(-2)=-4	vertical down	=-4+(-2)=-6	vertical down	=-6+(-2)=-8	vertical down	=-8+(-2_=-10	vertical down	=-10+(-2)=-12

C4		D4		E4		F4		G4	
diagonal	**=-2+(-1)=-3**	**diagonal**	**=-1+(-1)=-2**	**diagonal**	**=0+(-1)=-1**	**diagonal**	**=-2+2=0**	diagonal	=-4+(-1)=-5
horizontal right	=-4+(-2)=-6	horizontal right	=(-3)+(-2)=-5	horizontal right	=-2+(-2)=-4	horizontal right	=-1+(-2)=-3	**horizontal right**	**=0+(-2)=-2**
vertical down	**=-1+(-2)=-3**	**vertical down**	**=0+(-2)=-2**	vertical down	=-2+(-2)=-4	vertical down	=-4+(-2)=-6	vertical down	=-6+(-2)=-8
						`			

C5		D5		E5		F5		G5	
diagonal	**=-4+(-1)=-5**	**diagonal**	**=-3+(2)=-1**	**diagonal**	**=-2+2=0**	**diagonal**	**=-1+(-1)=-2**	**diagonal**	**=0+(-1)=-1**
horizontal right	=-6+(-2)=-8	horizontal right	=-5+(-2)=-7	horizontal right	=-1+(-2)=-3	**horizontal right**	**=0+(-2)=-2**	horizontal right	=-2+(-2)=-4
vertical down	**=-3+(-2)=-5**	vertical down	=-2+(-2)=-4	vertical down	=-1+(-2)=-3	**vertical down**	**=0+(-2)=-2**	vertical down	=-2+(-2)=-4

C6		D6		E6		F6		G6	
diagonal	**=-6+(-1)=-7**	diagonal	=-5+(-1)=-6	**diagonal**	**=-1+(-1)=-2**	**diagonal**	**=0+2=2**	diagonal	=-2+(-1)=-3
horizontal right	=-8+(-2)=-10	horizontal right	=-7+(-2)=-9	horizontal right	=-3+(-2)=-5	horizontal right	=-2+(-2)=-4	**horizontal right**	**=2+(-2)=0**
vertical down	**=-5+(-2)=-7**	**vertical down**	**=-1+(-2)=-3**	**vertical down**	**=0+(-2)=-2**	vertical down	=-2+(-2)=-4	vertical down	=-1+(-2)=-3

C7		D7		E7		F7		G7	
diagonal	**=-8+2=-6**	diagonal	=-7+(-1)=-8	**diagonal**	**=-3+(-1)=-4**	diagonal	=-2+(-1)=-3	**diagonal**	**=2+(-1)=1**
horizontal right	=-10+(-2)=-12	horizontal right	=-6+(-2)=-8	horizontal right	=-5+(-2)=-7	horizontal right	=-4+(-2)=-6	horizontal right	=0+(-2)=-2
vertical down	=-7+(-2)=-9	**vertical down**	**=(-3)+(-2)=-5**	**vertical down**	**=-2+(-2)=-4**	**vertical down**	**=2+(-2)=0**	vertical down	=0+(-2)=-2

Best path

	_	G	T	T	A	C
_	0	-2	-4	-6	-8	-10
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	-10	-6	-5	-4	0	1

	_	G	T	T	A	C
_	0	-2	-4	-6	-8	-10
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	-10	-6	-5	-4	0	1

alignment yellow

G	T	T	A	C
T	A	T	A	G

alignment pink

G	T	_	T	A	C
_	T	A	T	A	G