# BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 2)

## Sequence Alignment

http://thegrantlab.org/bggn213/

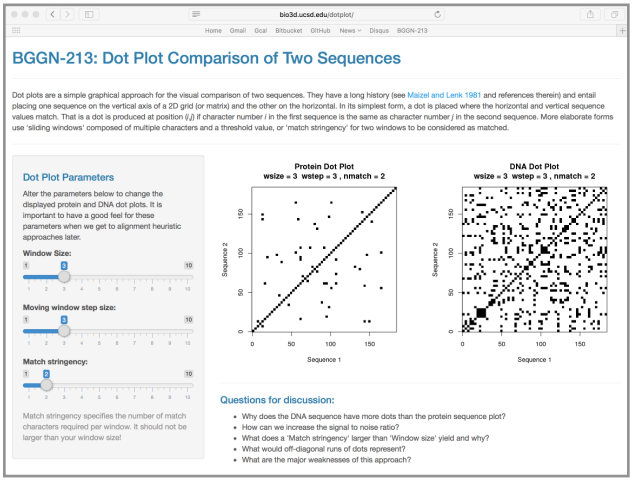
Dr. Barry Grant

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

## 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.** It is important to have a good feel for these parameters when we explore alignment heuristic approaches later.

Now discuss with your neighbor the answers the following questions:

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

**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 would off-diagonal runs of dots represent?

**Q5.** What are the major weaknesses of this approach?

**Dot Plots Revisited [OPTIONAL EXTENSION]**

There are a number of more useful 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.

> **Q6**: 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.

> **Q7**: 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 Q6?

## 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.

| | | A | G | T | T | C |
|---|---|---|---|---|---|---|
| | 0 | | | | | |
| A | | | | | | |
| T | | | | | | |
| T | | | | | | |
| G | | | | | | |
| C | | | | | | |

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

**Q8.** 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?

## Section 3: [OPTIONAL EXTENSION]

Use the **Needleman-Wunsch algorithm** discussed in class to align the following sequences:

Sequence 1: **TATAGC**
Sequence 2: **GTTATC**

**Q9.** 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 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
```

**Q10.** 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.

**Q11.** 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.

**Q12.** 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?

**Q13.** What function do you think this protein performs for your collaborators' organism?

## Section 5: Finding Distant Relationships [OPTIONAL EXTENSION]

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
```

**Q14.** Using standard blastn, were you able to find similar sequences outside of *Drosophila*? If so, in which species?

**Q15.** Now, use blastx instead of blastn. What species did you find this time?

**Q16.** 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?

**Q17.** 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.

**Q18.** 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.)

**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:

Muddy_Point_Assesment_Link