

8 - 12 April

Readings and Synopsis

- Should have read chapter 2, Chapter 3 intro: pp 53-58
- Reading for this week:
 - Ch 3: pp 59-64 for Today
 - Ch 3: pp 65-76 for Wednesday
 - Ch 3: 77-121 for Friday
 - Ch 3:119-124 for next week
- This week:
 - Homology
 - Sequence Analysis/Comparison
 - Dot matrix analysis
 - Introduction to alignments
 - Evaluation of Alignments
 - Scoring Systems
 - Quiz Friday, discussion session



BIMM 140

Introduction to Bioinformatics 1

Homology

What is Homology?

- Nothing in biology makes sense except in the light of evolution.
 - Theodosius Dobzhansky (1900-1975)
 - ...without that light it becomes a pile of sundry facts some of them interesting or curious but making no meaningful picture as a whole.
- homology - the presence of a similar feature because of descent from a common ancestor
- homoplasy - the presence of a similar feature because of convergence
 - Homology cannot be observed. We can't actually see the ancestral organisms/molecules and trace descent.
 - Homology is an inference, a conclusion we draw based on observed similarity.
 - Homology is an all-or-none relationship



BIMM 140

Introduction to Bioinformatics 2

Homology

Why is homology Important?

- Homology strongly suggests that the molecules have similar structure and function
- There are (very) many ways to fold a polypeptide to place specific chemical groups at specific locations. There is no reason, *a priori*, why proteins with a specific function should have similar 3-D structures.
- Therefore, there is no reason, *a priori*, why unrelated sequences should have any detectable similarity in sequence. Significantly similar molecular sequences are very unlikely to arise by chance - i.e. homoplasy on the molecular level is very unlikely.
- When we see significant similarity, we infer that the sequences/structures are homologous, i.e. at some point in the past they share an identical structure.
- The only thing that keeps the sequences tied to each other is the commonality of structure and function arising from homology.



BIMM 140

Introduction to Bioinformatics 3

Sequence Comparison

Alignments and DotPlots

- We want to match two sequences so that we can see the evolutionary similarity between them
 - Which functional domains correspond
 - Which functional residues correspond
- The matching acts as a map for applying information known about one molecule to the other
 - What activities can one infer
- Two important depictions of sequence matchings
 - Dot matrix plots
 - Alignments



BIMM 140

Introduction to Bioinformatics 4

Sequence Comparison

Dot Matrix Plots

- Simplest method - put a dot wherever sequences are identical
- A little better - use a scoring table, put a dot wherever the residues have better than a certain score
- Or, put a dot wherever you get at least n matches in a row (identity matching, compare/word)
- Even better - filter the plot

Sequence Comparison

Dot Matrix Plots

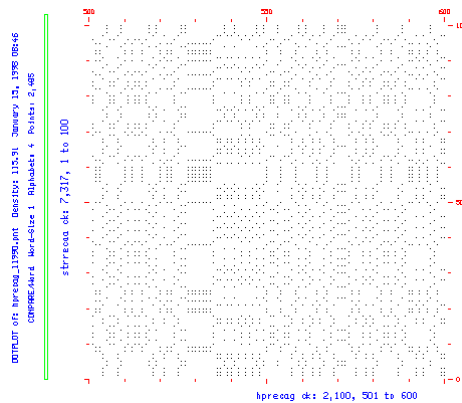
- Windowed scores
 - Calculate a score within a window
 - Move the window over one

```
A C C T T G T C C T C T T T A C C T G C C G A A
A C G T T G A C C T G T A A C C T G C C G A T T
```

Window Length = Segment = Span = 6
Every pair of letters, AA, AC, AG, ... TT, must have a defined score

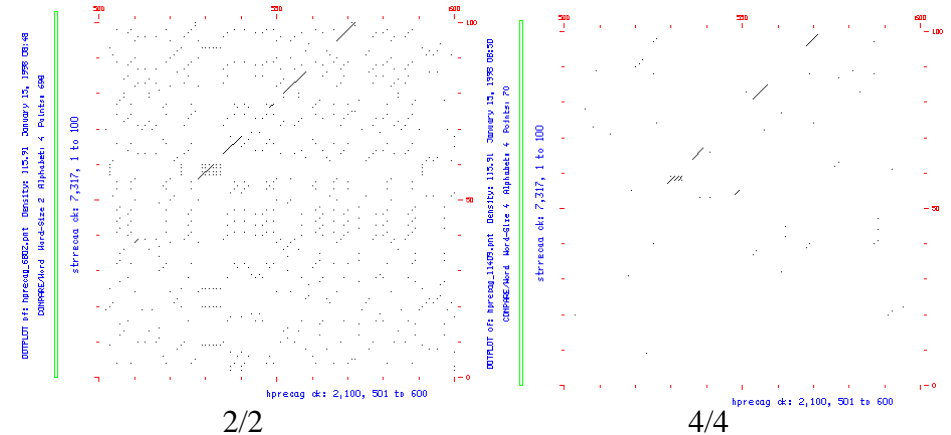
Sequence Comparison

RecA DNA sequence from *Helicobacter pylori* and *Streptococcus mutans*, window=1 match=1



Sequence Comparison

RecA DNA sequence from *Helicobacter pylori* and *Streptococcus mutans*, window/ match shown below figure



Sequence Comparison

Dot Matrix Plots

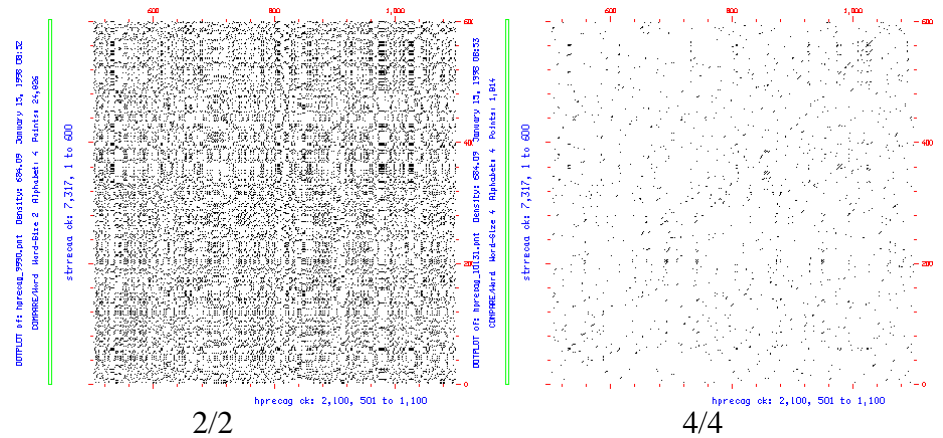
•Filtering dotplots

- Proportional matching COMPARE (GCG)
- Sum the score over a window, plot a point if over a threshold (stringency in GCG speak)
- Similar to GC content – but it is as if we only plot the part of the line over a specified threshold



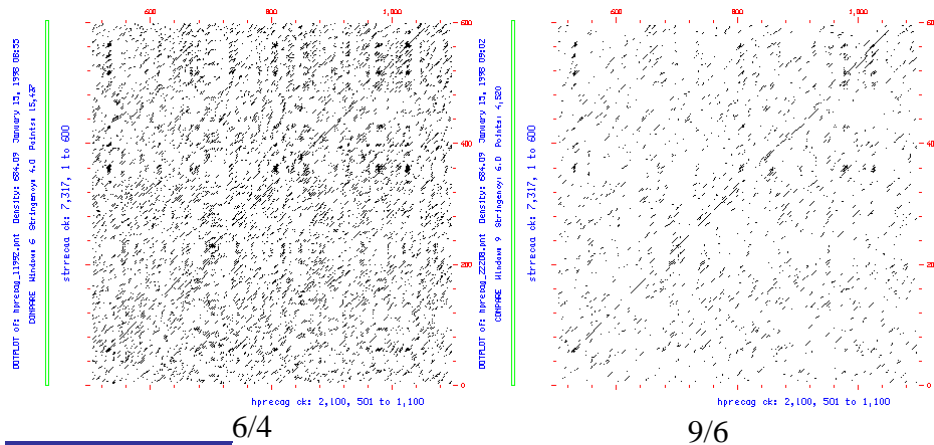
Sequence Comparison

RecA DNA sequence from *Helicobacter pylori* and *Streptococcus mutans*, window/ match shown below figure



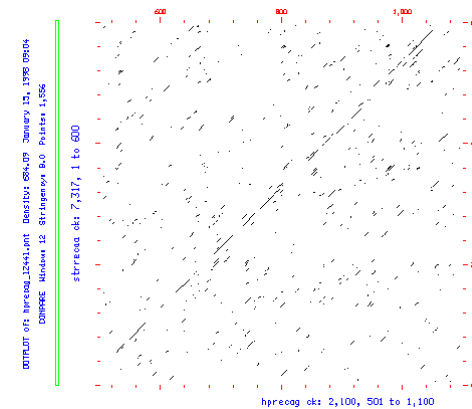
Sequence Comparison

RecA DNA sequence from *Helicobacter pylori* and *Streptococcus mutans*, window/ match shown below figure



Sequence Comparison

RecA DNA sequence from *Helicobacter pylori* and *Streptococcus mutans*, window/ match = 12/8



Sequence Comparison

Dot Matrix Plots

• What can you see in dotplots?

- Similar regions
- Repeated sequences
- Rearrangements
- RNA structures



Sequence Comparison

Dot Matrix Plots

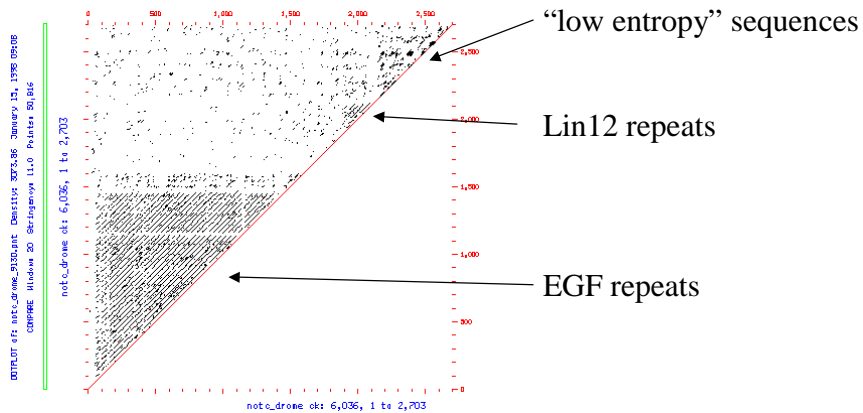
• Dotplots - rules of thumb

- Don't get too many points, about 3-5 times the length of the sequence is about right (1-2%)
- Window size about 20 for distant proteins 12 for nucleic acid (try stringency =10 for GCG proteins)
- Check sequence vs. itself
 - Finds internal repeats
- Check sequence vs. sequence
 - Finds repeats and rearrangements



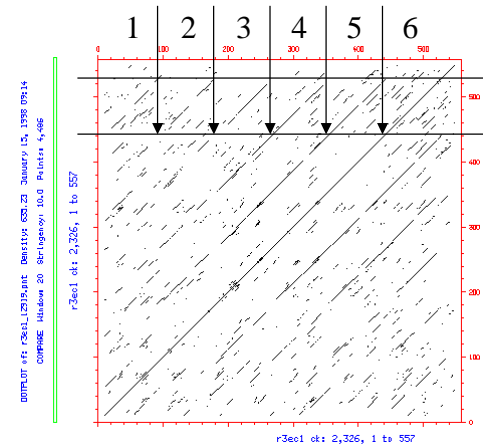
Sequence Comparison

Drosophila Notch protein



Sequence Comparison

Repeated sequence in E. coli ribosomal protein S1



Statistics: Part II

Is there homology?

- We can answer, “yes,” if the result is surprising when compared to unrelated sequences

What do we mean by surprising?

- We are surprised when an event is very unlikely to happen by chance. In this case, we are surprised when the observed level of similarity is very unlikely between unrelated sequences

This requires a model for unrelated sequences

- Most common choice is a random sequence model

Statistics: Part II

Dotplots

- How many windows do you expect at or above a certain score, x ?
- $\text{Exp_Num_Windows}(\text{Score} \geq x) = \text{Total_Windows} \times P(\text{Score} \geq x)$
- $\text{Total_Windows} = N_Windows_Seq1 \times N_Windows_Seq2$
 - $P(\text{Score} \geq 140) = 2.63 \times 10^{-5}$
 - $\text{Total_Windows} = (141-11+1) \times (154-11+1) = 18,864$
- $\text{Exp_Num_Windows}(140) = 2.63 \times 10^{-5} \times 18,864 = 2$

Statistics: Part II

Models

- Models allow us to answer the question, "How surprised am I?"
- Models tell us what we expect to see
 - If we know what to expect, we can tell if we should be surprised
 - Usually predict behavior of unrelated sequences
 - Models allow us to be quantitative
 - Statistics are a formal and quantitative way of measuring surprise
 - Models are nearly always a simplification
 - In some situations, the simplification may not be appropriate!
 - If you understand the model you are less likely to be fooled

Statistics: Part II

Common models

- Three kinds models commonly used in molecular biology
- Random sequence model
 - Assumes unrelated sequences behave as random or "scrambled" sequences
 - GAP, BESTFIT
 - often evaluated by Monte Carlo approach
- Unrelated sequence model
 - Assumes you can actually tell which ones are unrelated
 - FASTA, Profilesearch
- Theoretical models
 - Many possibilities, many assumptions
 - Extreme value theory/BLAST

Statistics: Part II

Random Sequence Model

- Assumes that unrelated sequences act like random sequences
- A random sequence is typically created by sampling residues or bases a random according to the frequencies in the database
- How are random sequences unrealistic
 - Lengths?
 - Composition?
 - Patterns?

Sequence Comparison

Review

- Homology is the key concept that relates sequence similarity to inferences about structure and function
- The inference of homology relies on determining that similarity is surprising, i.e., too high for unrelated sequences
- Dotplots are a qualitative way to examine sequence similarity
- Dot plots detect
 - Similar regions
 - Repeated sequences
 - Rearrangements