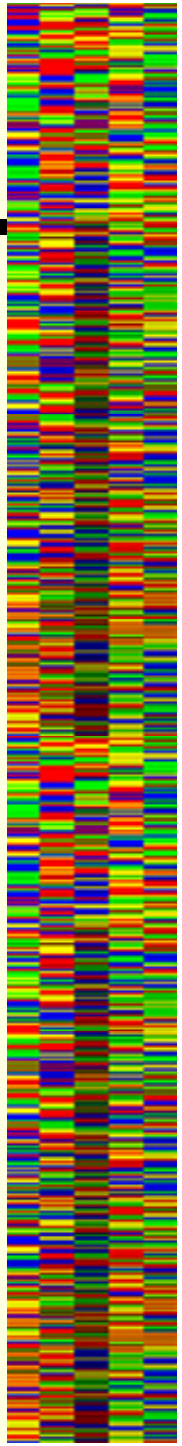
Biol478/595

29 August

#	Day	Inst.	Topic	Hwk	Reading
			August		
1	M 25	MG	Introduction		
2	W 27	MG	Sequences and Evolution		Handouts
3	F 29	MG	Sequences and Evolution		
			September		
	M 1		Labor Day		
4	W 3	MG	Database Searching		Ch. 6
5	F 5	MG	Database Searching	Hw1	

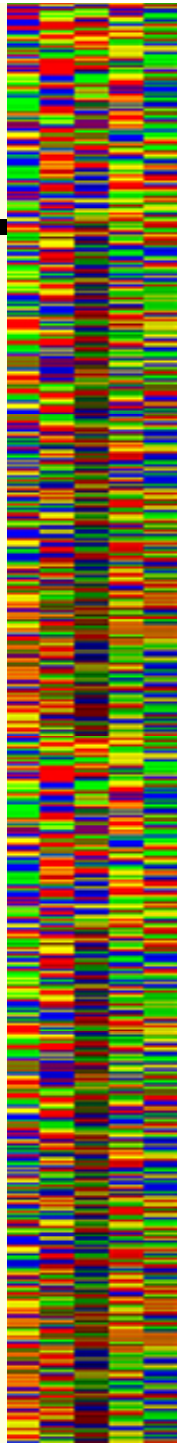
- ***Reading***
  - Handouts
  - Mount: Chapter 6 – Sequence database searching for similar sequences
- ***Homework***
  - Usually posted Wednesday, due following week Friday



Looking at similar molecules can tell us

- ***where they came from (history)***
- ***how they work (mapping knowledge)***

- ***The key starting point is the knowledge that you are looking at molecules that are ancestrally related. This is called <u>homology</u>***



Genomics

What is Homology?

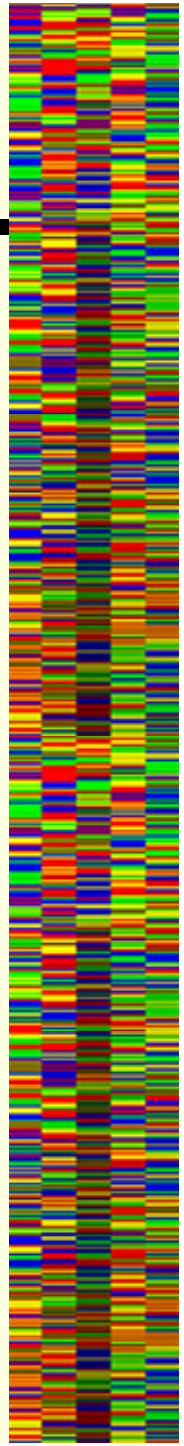
Seen in the light of evolution, biology is, perhaps, intellectually the most satisfying and inspiring science. Without that light it becomes a pile of sundry facts some of them interesting or curious but making no meaningful picture as a whole.

Nothing in biology makes sense except in the light of evolution.
Theodosius Dobzhansky (1900-1975)

homology - the presence of a similar feature because of descent from a common ancestor

- ***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 – no partial homology***

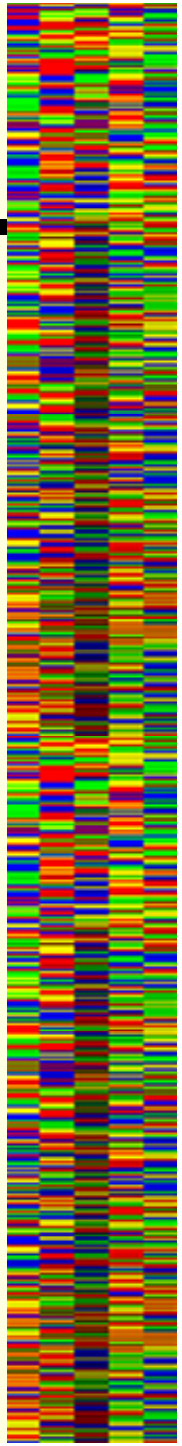
homoplasy - the presence of a similar feature because of convergence



Genomics

Why is homology Important?

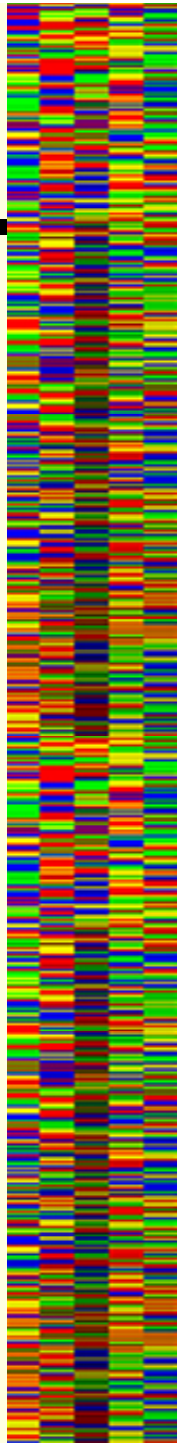
- ***Homology strongly suggests that the molecules have similar structure and function***
- ***Some time in the past, the molecules had identical structure and function***
- ***Biology is conservative - small accumulations of mutations lead to small changes in function, but not to radical changes***
- ***If you can prove homology, you have a strong basis for predicting similar structure and function***
- ***Known information about related molecules can be "mapped" onto unknown molecules***



Genomics

Proving homology

- *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.*



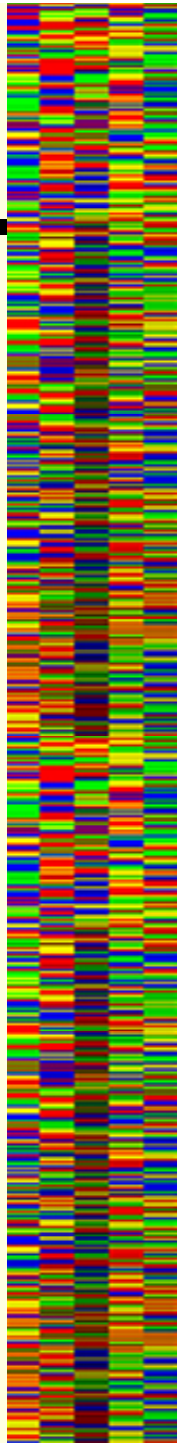
Genomics

We can only make comparisons when we know we are comparing the "same" things. More precisely

- ***homologous genes***
- ***homologous proteins***

- ***Argument: This gene in mouse is the same as a gene in humans, therefore it does about the same thing***

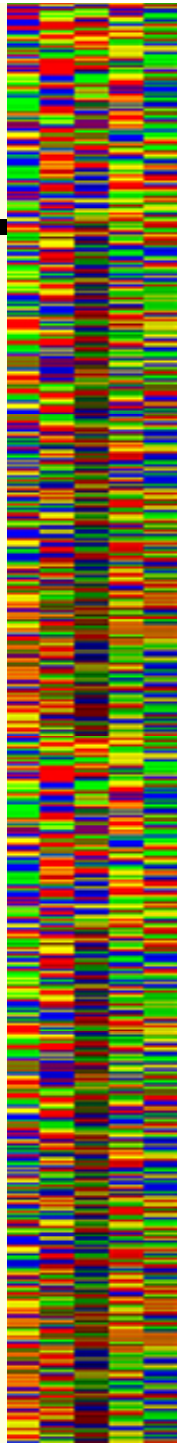
- ***What do you…***
  - … find the same gene in two genomes?
  - … find the same protein in two proteomes?
  - … guess the function of a new gene?



Genomics

Homology

- ***Sequences alignments and database searches let us***
  - Find homologous sequences (genes/proteins)
  - Map information from known systems to new ones
    - Gene identification
    - Gene function
    - Metabolic and regulatory systems
- ***Two common classes of homologs***
  - Orthologs – genes separated by a speciation event, i.e. the same gene in two species
  - Paralogs – genes separated by a duplication events, originally the same but now diverged with possibly different functions



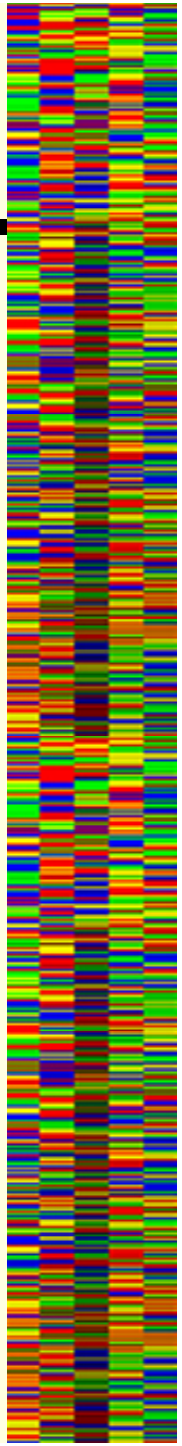
Genomics

How is homology determined?

- ***Because molecular homoplasy is unlikely, significant sequence similarity strongly indicates homology***

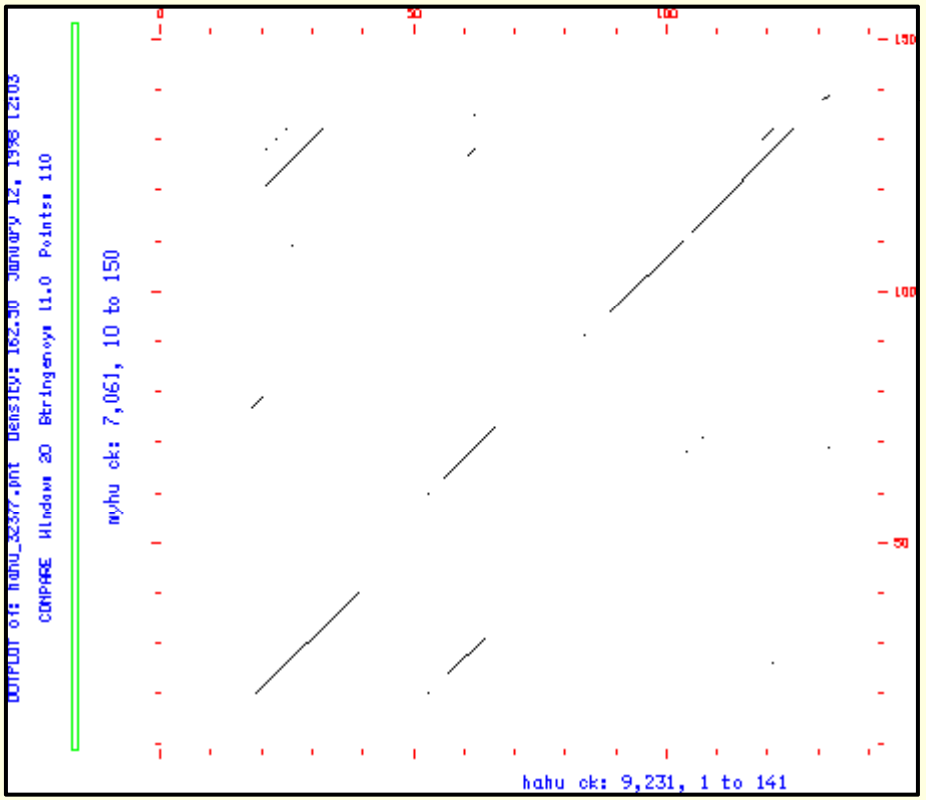
Similarity ≈ homology

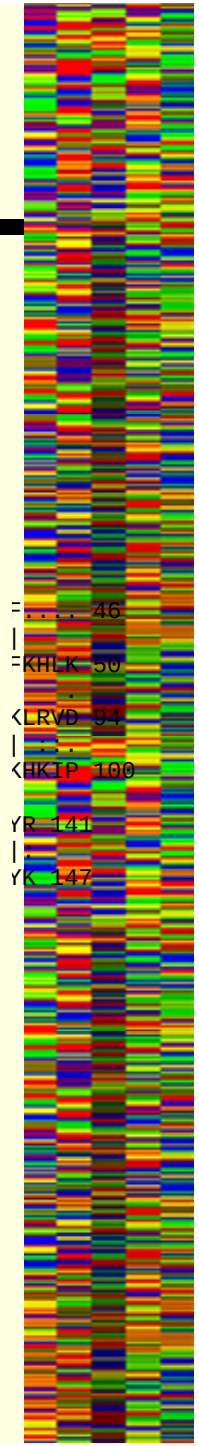
- ***Similarity is determined by sequence matching or comparison, more commonly called sequence alignment***
- ***Approaches***
  - Dotplots
  - Alignments
  - Database searches



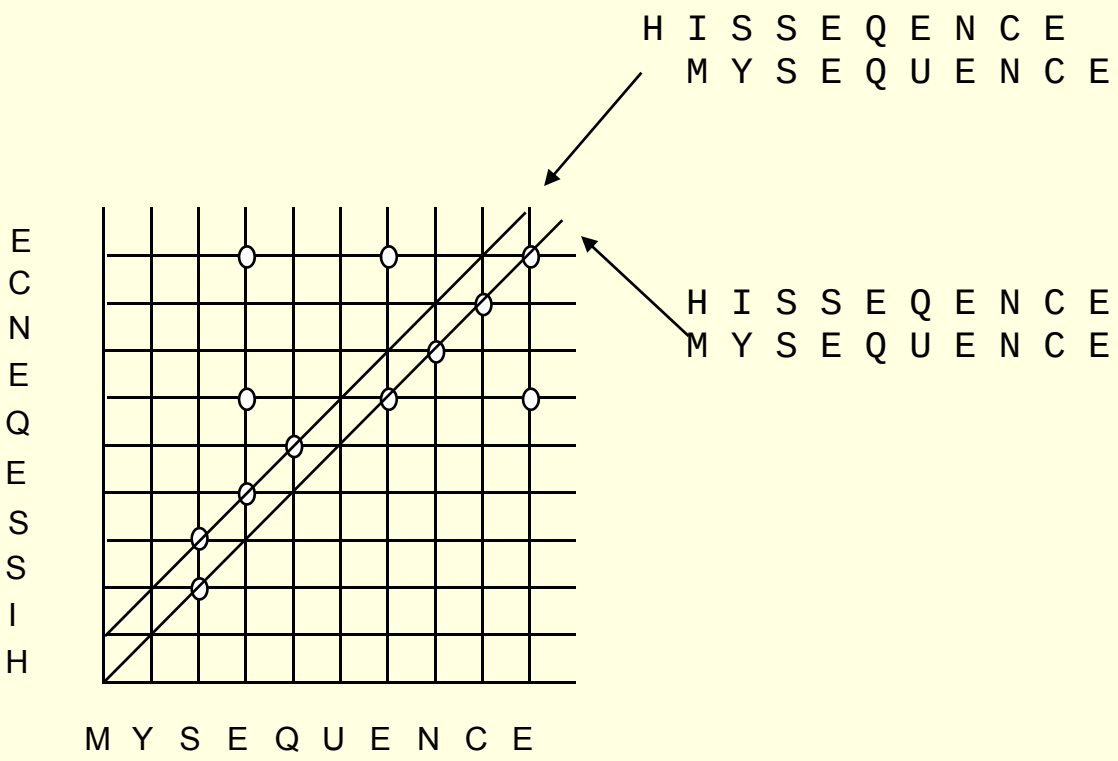
Genomics

Dotplots – a simple way to compare sequences





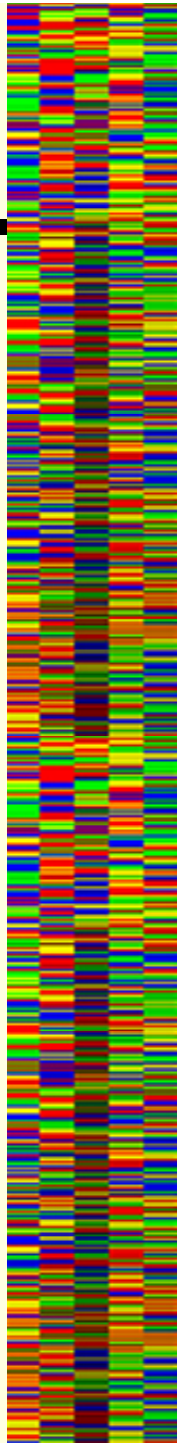
Genomics


H I S S E Q E N C E
M Y S E Q U E N C E
H I S S E Q E N C E
M Y S E Q U E N C E
E C N E Q E S S I H
M Y S E Q U E N C E

Genomics

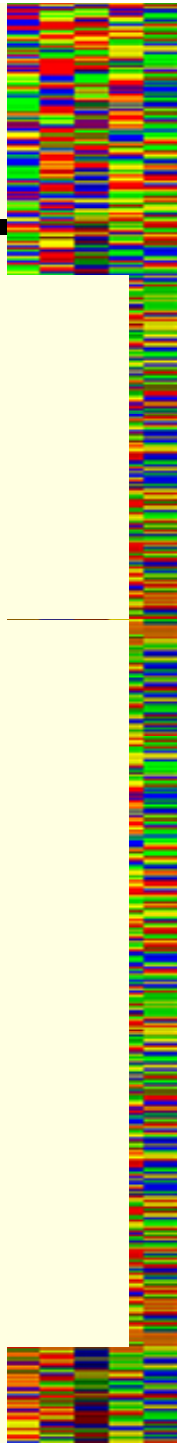
Dotplots

- *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*

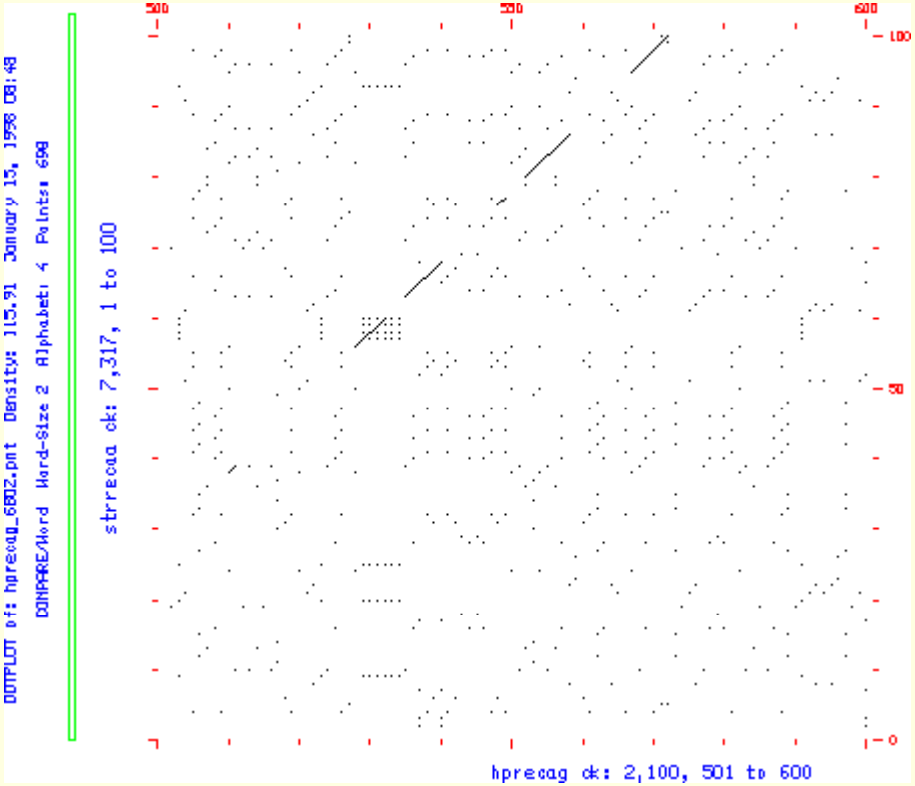


Genomics

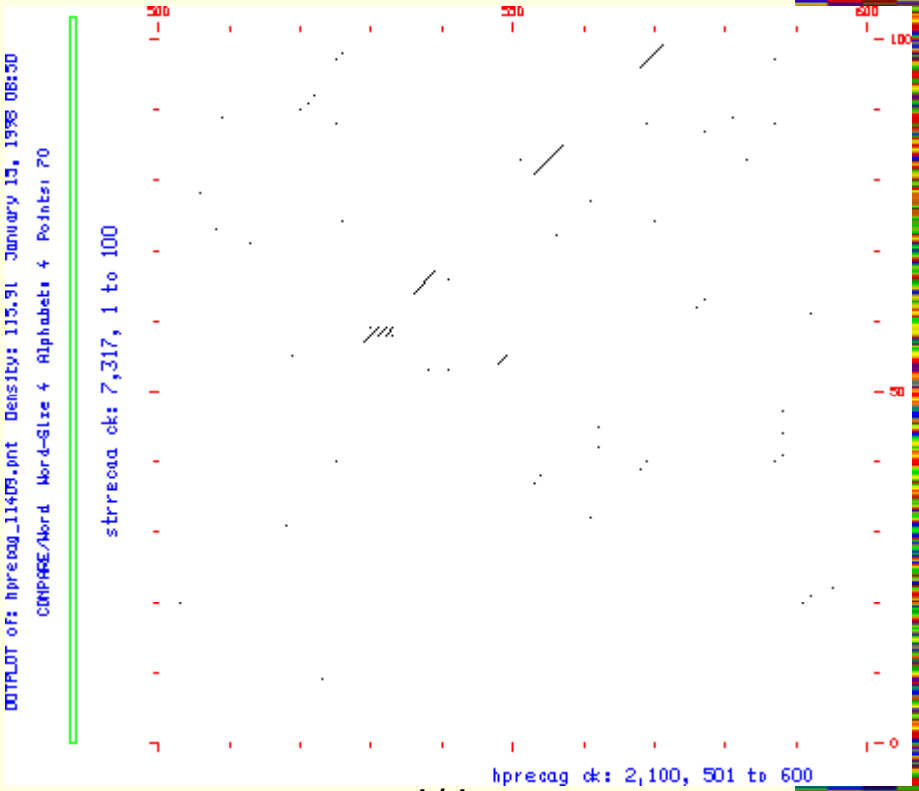
Dotplots



Genomics



2/2



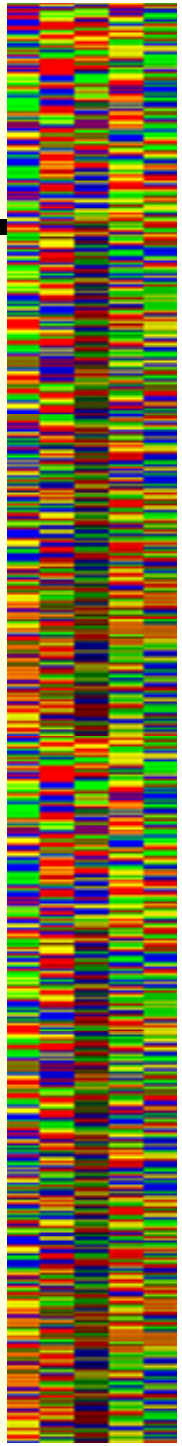
4/4

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

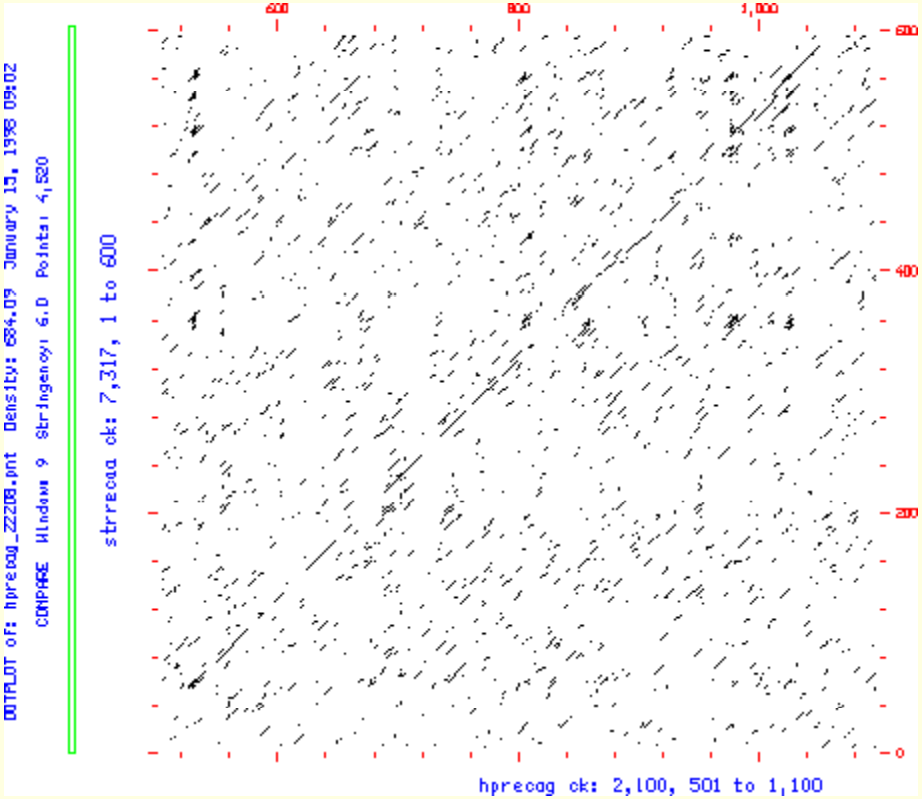
Genomics

Dotplots

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



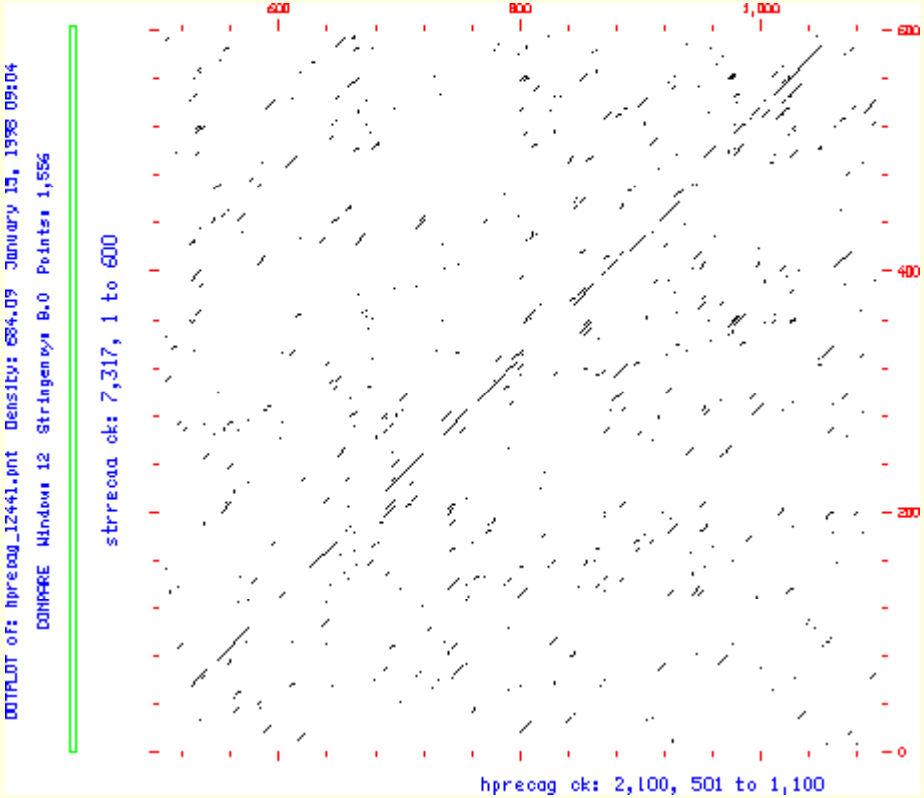
Genomics



9/6

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

Genomics

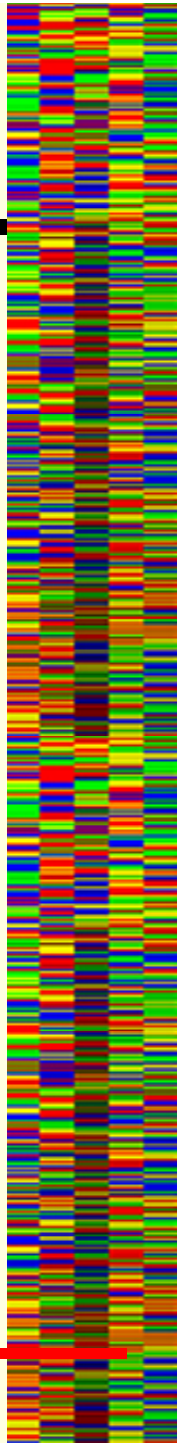


RecA DNA sequence from *Helicobacter pylori* and *Streptococcus mutans,* window/ match = 12/8

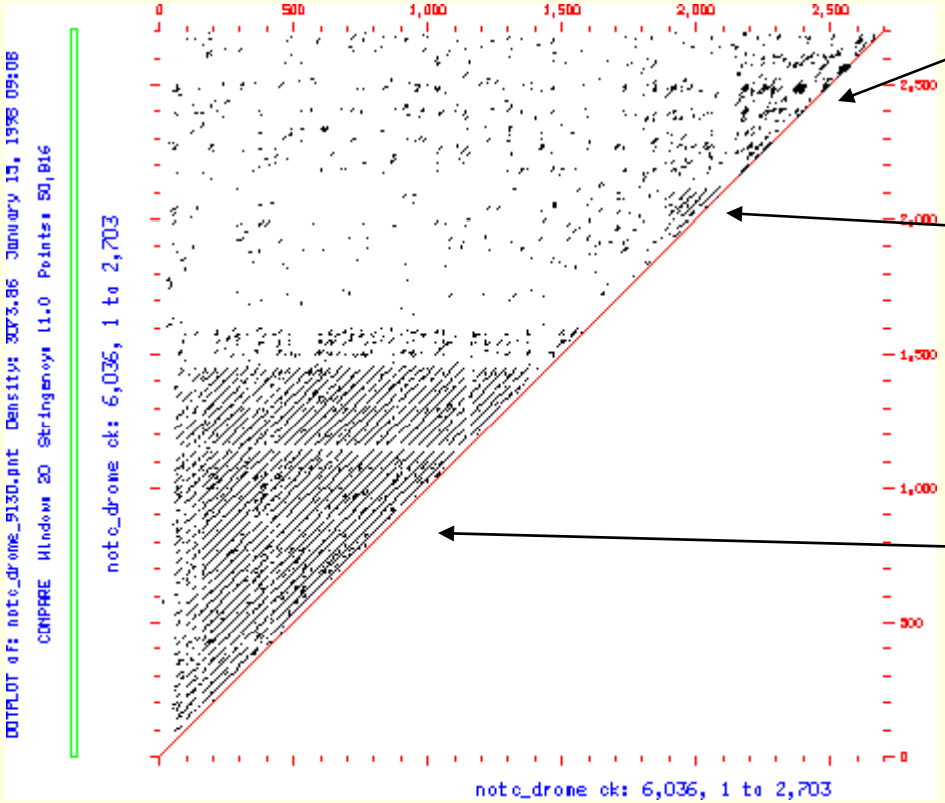
Genomics

Dot Matrix Plots

- ***What can you see in dotplots?***
  - Similar regions
  - Repeated sequences
  - Rearrangements
  - RNA structures



Genomics



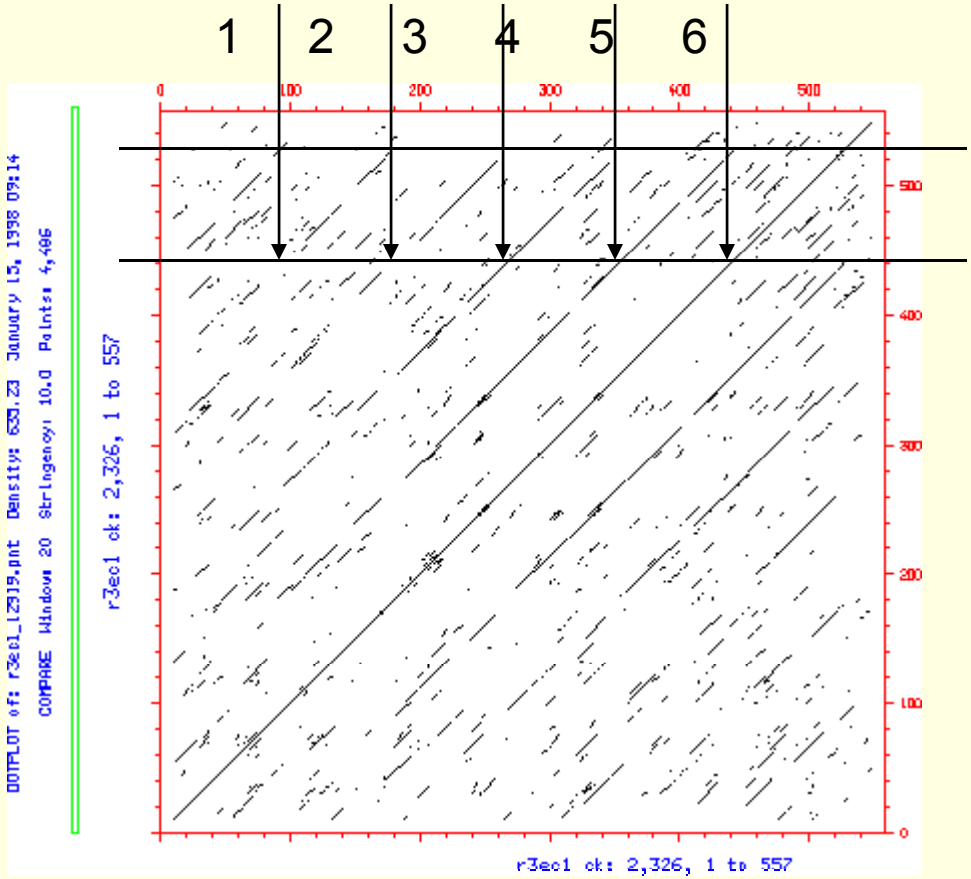
"low entropy" sequences

Lin12 repeats

EGF repeats

Drosophila Notch protein

Genomics

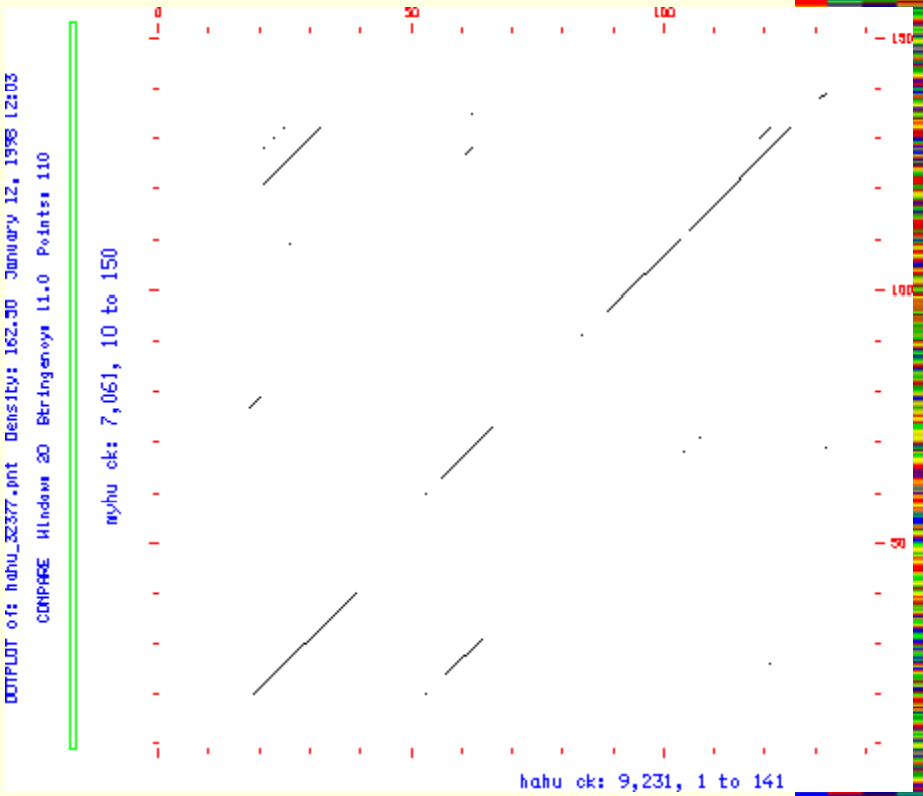


Repeated sequence in *E. coli* ribosomal protein S1

Genomics

Homology

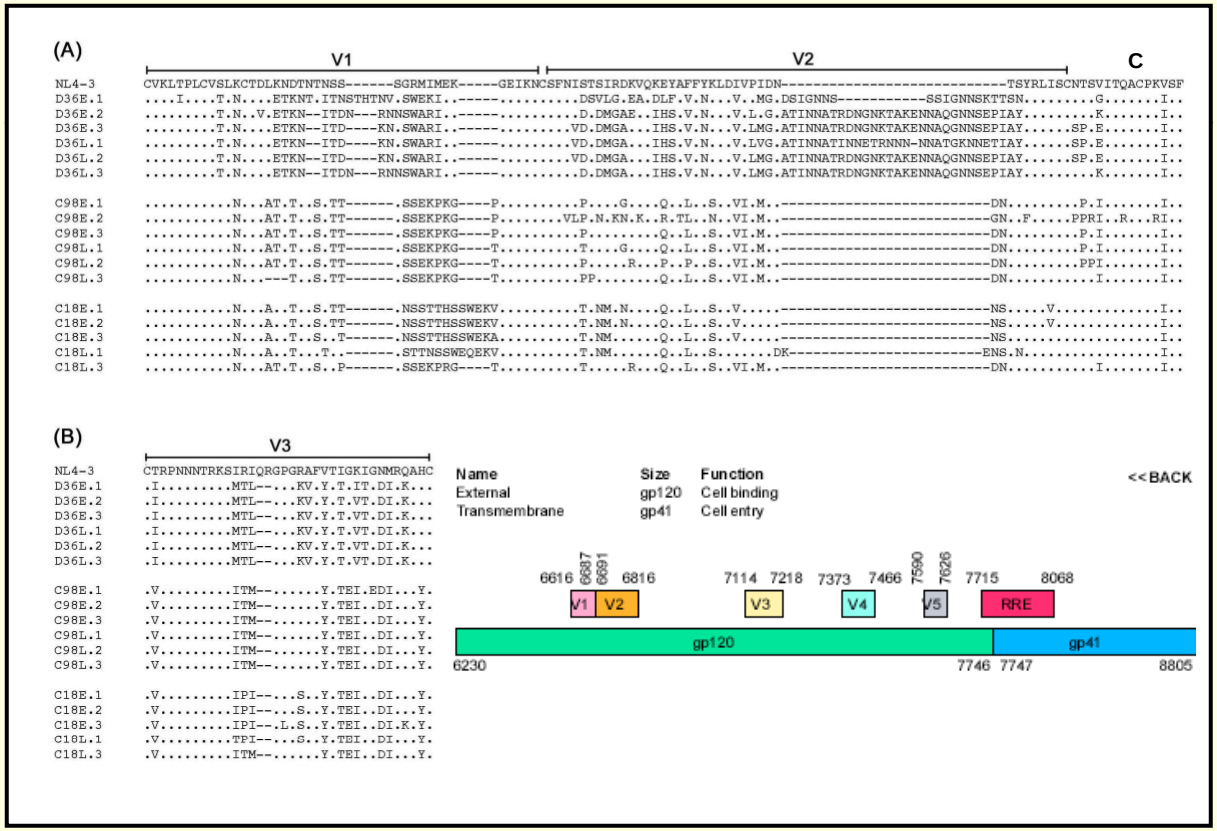
- *Homologous sequences show up as diagonal lines in dotplots*
- *Basis for methods to find homologous sequences*

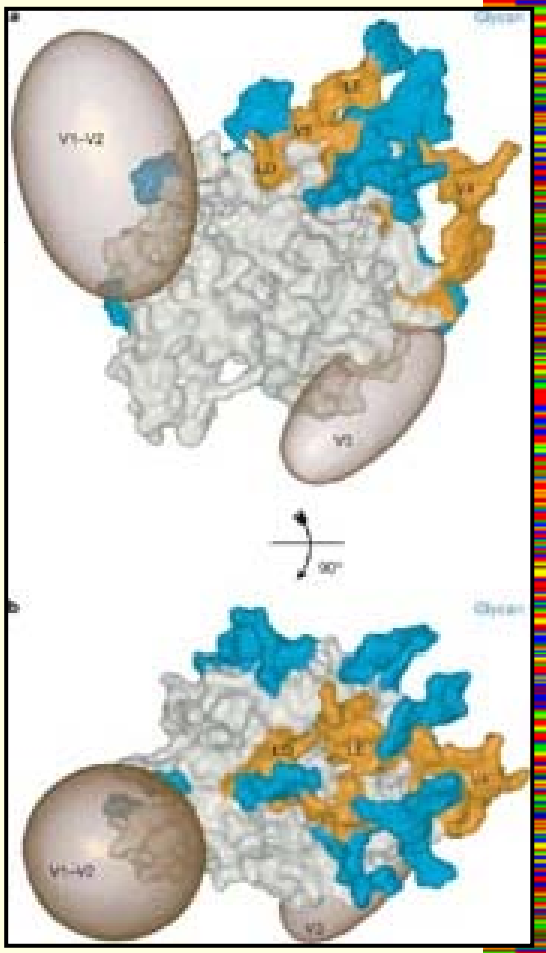


Genomics

HIV Env Gene

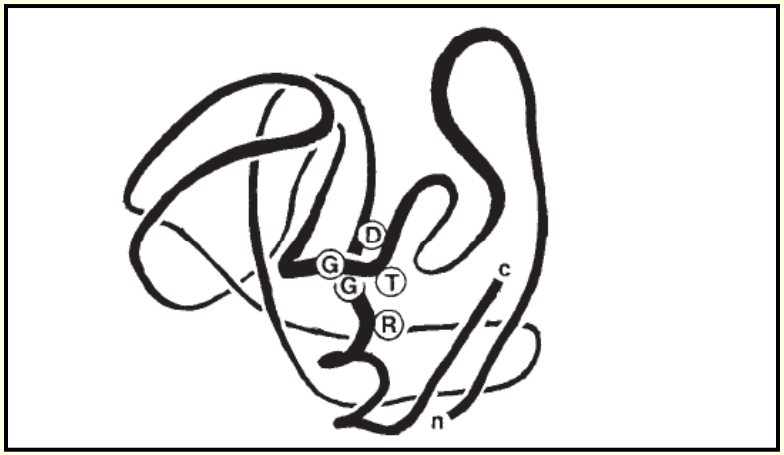
- **Conserved (C) and Variable (V) regions**

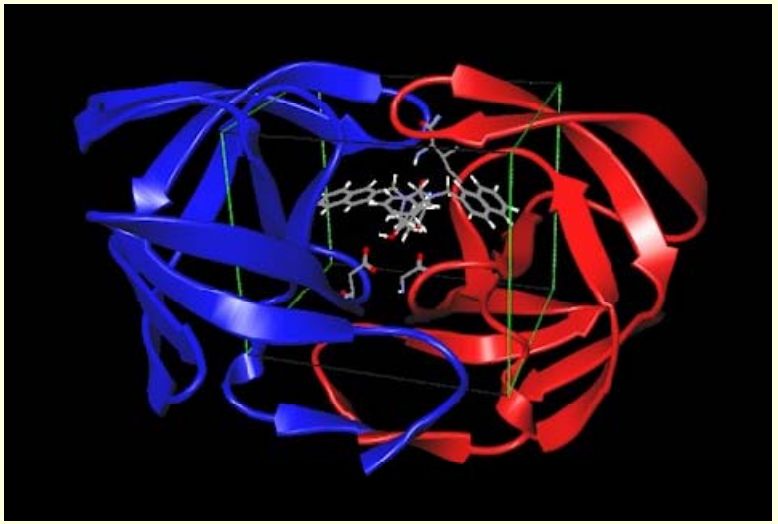




Genomics

HIV Protease





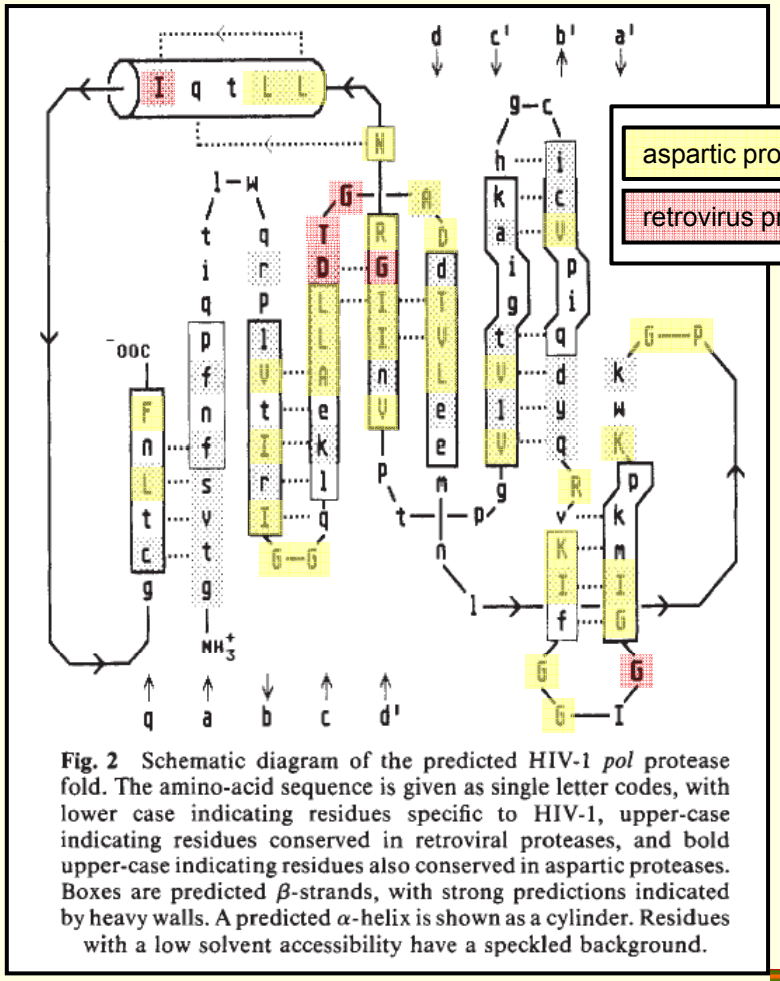


Fig. 2 Schematic diagram of the predicted HIV-1 *pol* protease fold. The amino-acid sequence is given as single letter codes, with lower case indicating residues specific to HIV-1, upper-case indicating residues conserved in retroviral proteases, and bold upper-case indicating residues also conserved in aspartic proteases. Boxes are predicted β-strands, with strong predictions indicated by heavy walls. A predicted α-helix is shown as a cylinder. Residues with a low solvent accessibility have a speckled background.

Pearl & Taylor, Nature 329, 351-354 ,1987

Genomics

Database Searching

- *Big problem is database size*
- *Bigger database means longer search. Alignments are $O(n^2)$*
- *Bigger database means wo*
- *Sequence data doubles every 12-14 months*