Sequence Comparison

DotPlots & Alignments

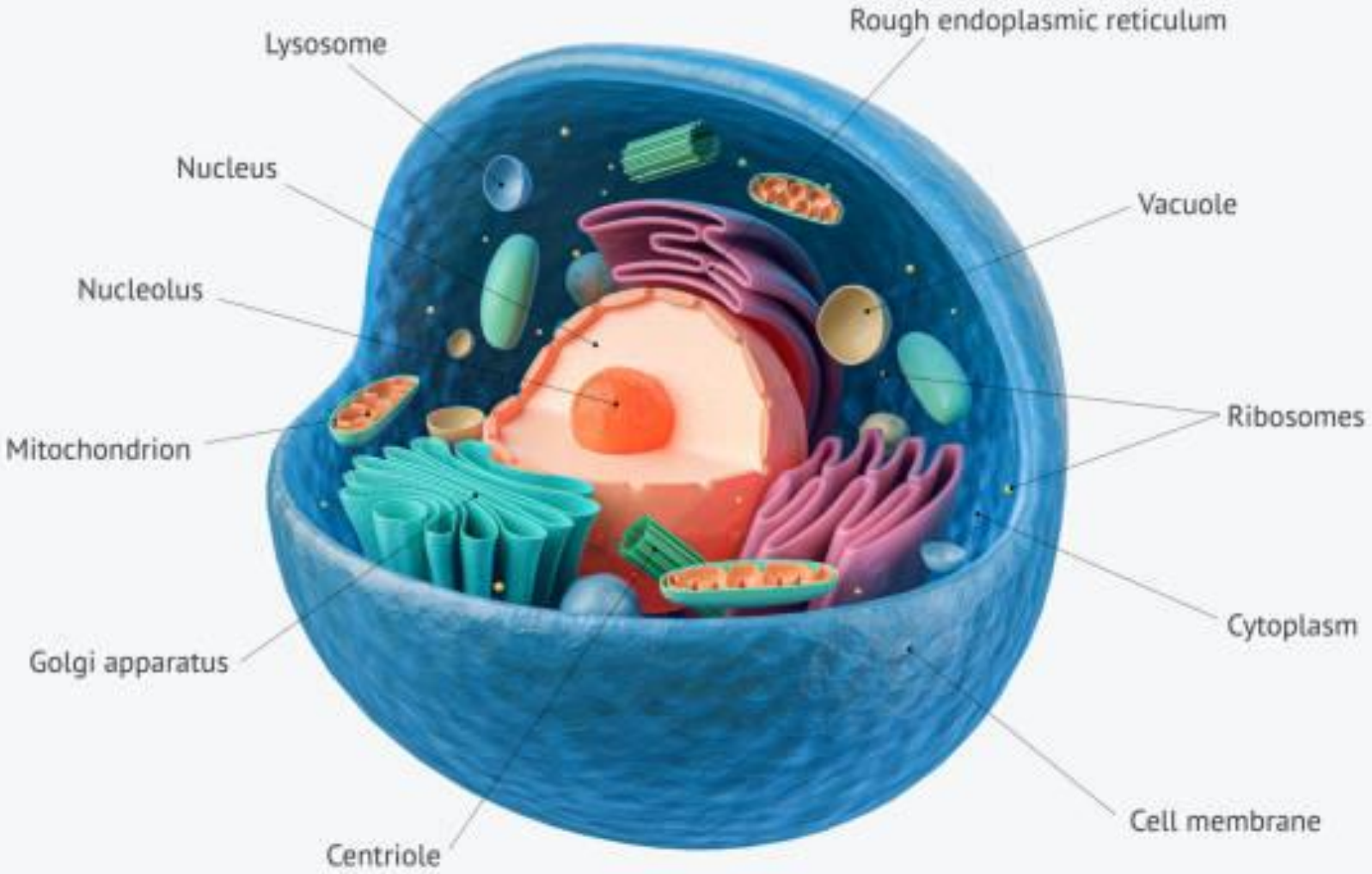
Mystery of the Chilean Blob

- A 13-tonne blob containing no skin, bone or cells was washed ashore on a Chilean beach in July 2003
- Hypotheses ranged from remains of a giant squid, octopus, whale blubber, to some sea monster, alien
- DNA samples were extracted from the blob and sequenced (NADH2, control region)
- Search against the database unambiguously established identity of the blob: sperm whale blubber

Pierce, S.K *et al.* 2004, *Biological Bulletin*v206, 125-133

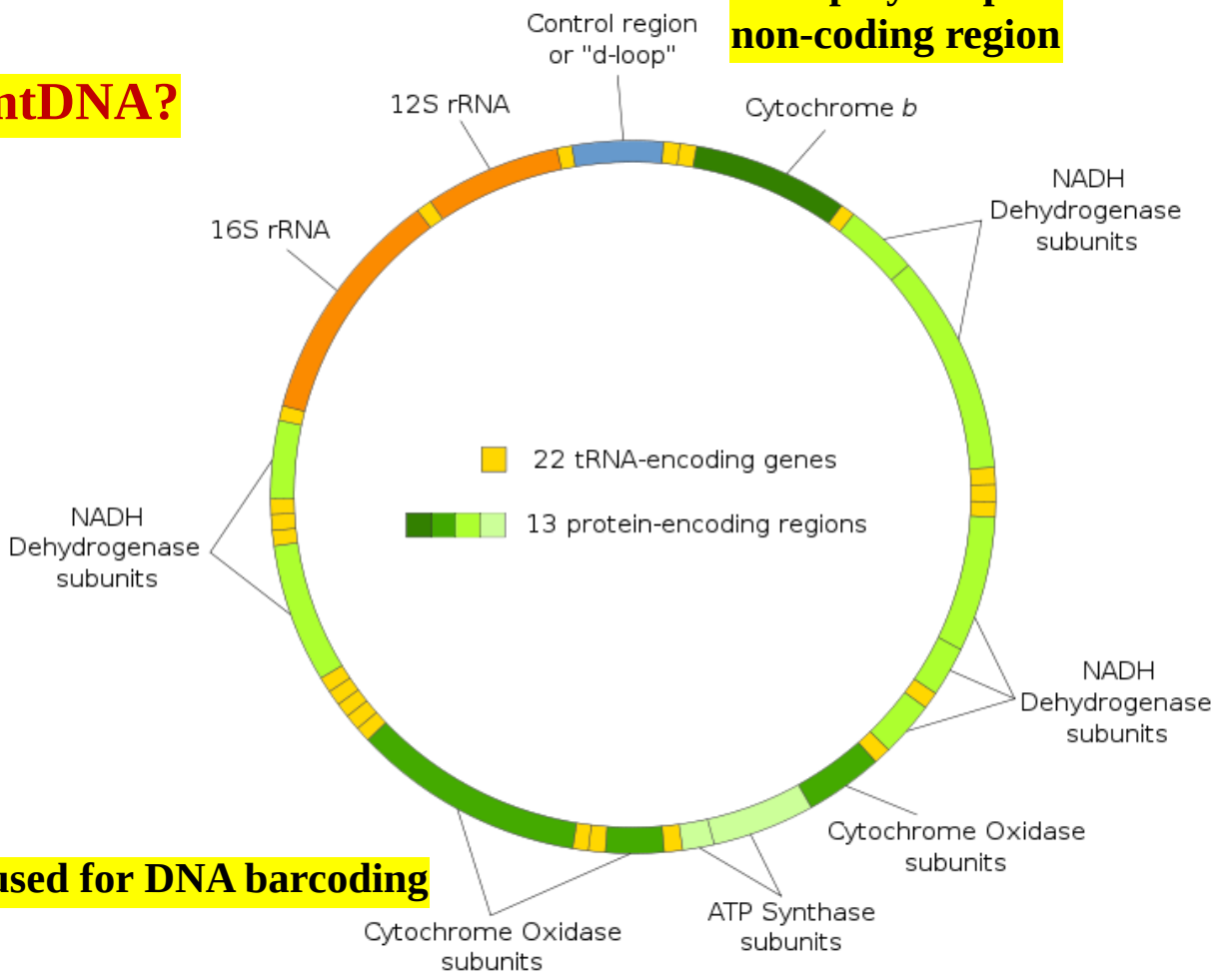
Is that all, or can we say more about this blubber?

ANIMAL CELL STRUCTURE



Why mtDNA?

most polymorphic non-coding region



used for DNA barcoding

Rapid mutation rate (in animals) makes mtDNA useful for assessing genetic relationships of individuals / groups within a species, for identifying & quantifying the phylogeny (evolutionary relationships) among different species.

Mystery of the Chilean Blob

Apart from the species it came from, can we say:

- **Which individual whale it belonged to?**
- **Where was that individual born?**
- **Can we understand the relationship it had during its lifetime, both with its immediate family and members of social group?**
- **Can this sample of whale DNA allow us to track whale movements across the globe, both in space and time?**

Mystery of the Chilean Blob

- **Can we use it to trace this whale's family history back through the ages, exploring how the whales responded to a changing world as ice ages came and went?**
- **Can it help us reconstruct the evolution of important whale adaptation such as echolocation.**
- **Can it help to identify the nearest mammalian relatives of the whales?**

DNA gives us the opportunity to look at evolution beyond fossil records

Echolocation – whales emit bursts of ultrasonic sound and use the sound reflected back from their surroundings to locate prey.

Mystery of the Chilean Blob

Answer to all these Qs is YES!

The reason the DNA sequences can provide information at all of these different levels of evolutionary history

- **from individuals to populations to species, families, phyla and Kingdoms of life**
- **is that different parts of the genome evolve at different rates**

⇒ different parts of the genome can be selected to tell different stories

Individual identification

- **DNA mutations accumulate over time**
- **Differences are minimal between parent & offspring (~100), and keep increasing through generations**
- **Given enough DNA sequences from a group of whales it is possible to establish relationships between individuals by comparing DNA sequences**
- **More generally one can infer how a population of whales interacts & interbreeds.**

Same rational applies to forensic DNA analysis to identify biological samples left at crime scenes

Observation of Sperm Whale Behaviour

- **Social groups of 10-30 animals - includes mature females & immature whales of both sexes**
- **Duration of social bonds within groups highly variable**
- **Some associations persist for several years, others last few days**
- **Mature males are never seen to be long term members of social groups, implying they must at some point disperse from their natal groups**

Observation of Sperm Whale Behaviour

- **Naturally sloughed skin of free-living whales were collected from 3 different regions/groups**
- **Microsatellite markers, mtDNA & male-specific SRY gene were used to construct genetic profiles of individuals, compare relatedness within & between groups**
- **Social groups were defined through photographic identification of individuals.**
- **Each group contained ~26 members, mostly female (79%)**

- consistency of matrilineal model was confirmed

Richard et al. (1996), *PNAS* 93, 8792

Where the Individual was Born

- **Sperm whales are found in every ocean**
- **Males move around different oceans of the world, while females live in the oceans they were born.**
- **In sperm whales also the mtDNA is passed from mother to offspring**
- **As a result, each whale will tend to have the mtDNA sequence typical of the ocean it was born in.**

$\Rightarrow$ **sequencing the mtDNA one can identify in which region of the world that whale was born.**

Estimation of Population Ssize

- Measures of genetic similarity over the whole population gives an indication of population size & mating structure.
- Low diversity in the population implies either in-breeding or small population; Large populations have more diversity
- Samples collected from 37 whales from different oceanic areas - mtDNA control region extracted & sequenced
- Unusually low diversity was found for which two explanations were investigated: 1) low evolution rate, 2) recent common ancestry
- Current estimates suggest that size of global sperm whale population is ~ 3,60,000 (based on average no. of genetic differences between individuals)

Lyrholm et al. (1996) *Mol. Bio. Evol.* 13, 1318

Estimation of Population Size

- **Genetic diversity of this population is surprisingly low**
- **Examination of other mtDNA reveals normal evolution rate, suggesting the low diversity is likely to be related to an evolutionarily recent ancestry of the lineages.**

  ⇒ all have descended from a small no. of founding mothers who survived the last ice age.

Most recent glacial period peaked ~21,000 yrs ago and ended ~11,500 yrs ago

Estimation of Population Size

- **Authors suggest a time estimate since common mtDNA ancestry in the sperm whale to be roughly 6000-25000 yrs, i.e. if a bottleneck occurred in the sperm whale population, it could have coincided with the last ice age.**

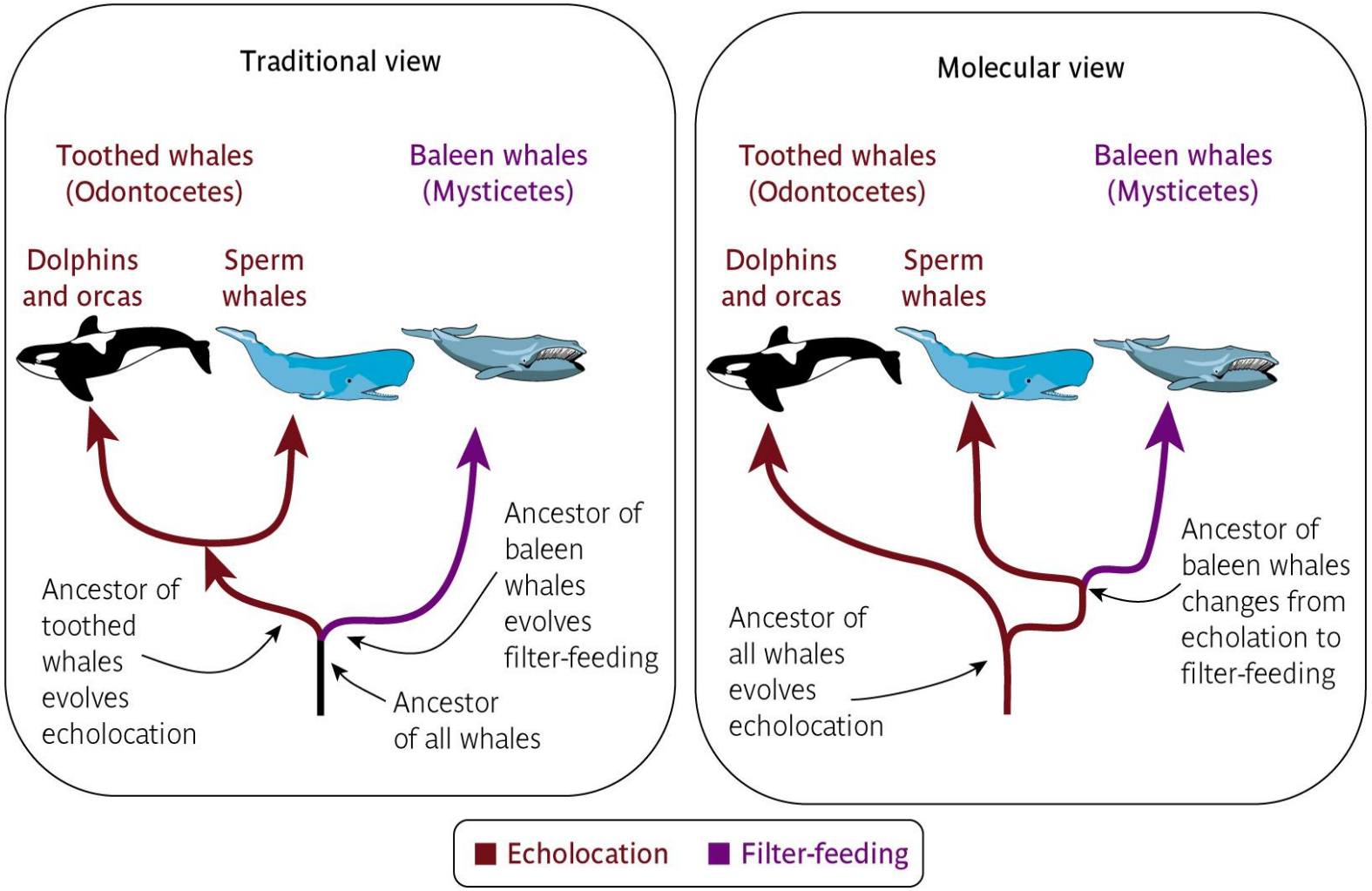
DNA sequences can be used to study how the population has changed over time

Using transition/transversion ratio R and No. of substitutions inferred from parsimony analysis, age of sperm whale mtDNA diversity is estimated to be ~6,000-25,000 years

Echolocation in Sperm Whale: Ancient or Derived Characteristic

- **Sperm whales use echolocation to hunt, a characteristic shared with other predatory toothed marine mammals, such as dolphins and killer whale**
- **Baleen whales (e.g., Blue whales) have filter plates in their mouth that sieve planktons out of the water**
- **The theory is that the two groups evolved from a common ancestor ~30MYA, and each adapted to their different lifestyle by developing echolocation as in sperm whale and baleen as in blue whale**

Genome of sperm whales tell a different story



Common ancestral whale must have had echolocation; it was lost in the baleen whales as they adapted to a new way of life

- evidence of vestigial "melons" (echolocation sounding chambers) in baleen whales provides support to this hypothesis

- **DNA sequences are valuable sources of information**
- **Evolution both adds and takes away complex characteristics, thus sometimes obscuring the history of species.**
- **But while the appearance of a species changes, the genome continues to record evolutionary history, and hence molecular data can sometimes give a clearer view of a species evolutionary past then can its highly modified morphology.**

Computational Molecular Biology

Genome Analysis/ Sequence Analysis

- involves identifying characteristic features in a genome

Some important analytical approaches involve:

- **Sequence Alignment - to identify regions of similarity (Pairwise & Multiple)**
- **Pattern search - identifying repeats, motifs, domains, etc.**
- **Database search - sequence/pattern-based search to identify similar sequences in the database**
- **Statistical measures – *ab initio* methods based on certain characteristic features of sequence (e.g. gene prediction), evaluating significance of alignment/motifs in Db search.**

Types of Mutations

- **Mutations - are local changes in DNA content, caused by inexact replication. There are various kinds of mutations:**
- **Substitution - a wrong base is incorporated instead of a true copy. A substitution may or may not alter the protein sequence depending on the place it occurs, e.g., GUU, GUC, GUA, GUG all code for – Valine, GGU – Glycine, CUU – Leucine**
- **Insertion / Deletion - addition / removal of one or more bases - leads to frame-shift in coding regions.**
- **Rearrangement - a change in the order of complete segments along a chromosome, e.g., human and mouse genome are very similar – major difference being the internal order of DNA segments.**

Mutations are important for several reasons:

- **are the source of phenotypic variation on which natural selection acts, creating species & changing them.**
- **are responsible for inherited disorders and diseases such as cancer, which involve alterations in gene.**

To understand evolution we need to know the various types of mutations that occur, frequency/distribution of their occurrence, and their effect.

For disease diagnosis, we need to understand the types of mutations, their inheritance pattern, their phenotype, etc.

Sequence Comparison

Why compare sequences?

Why Compare Sequences?

Sequencing of the genomes – has outputted an enormous amount of sequence data on new proteins

Fundamental problem – determination of the function of a new protein

If there is significant sequence similarity between a pair of sequences, we can extrapolate the functional annotation of one sequence to the other.

Computational Methods in Sequence Comparison:

- **Graphical methods - visual /qualitative comparison - dotplots**
- **Sequence Alignment: Determine residue-residue comparison to identify patterns of conservation and variability.**
  - pairwise alignment

  e.g., identify genes/proteins belonging to the same family.
- **Database Search: Look for homologs of query genes/proteins in the database**

Computational Methods in Sequence Comparison:

- **Knowledge-based prediction: known examples are extracted into empirical rules (stored in motif libraries) representing sequence-structure or sequence-function relationships. The query sequence is checked with a motif dictionary and if any motif is present it is an indication of a functional site.**

  – multiple alignment

  Useful for motif identification, remote homologs of genes/proteins

Dot Plots - Graphical Comparison of Sequences

One of the simplest method for comparing two sequences, described by Gibbs & McIntyre (1970)

A dot plot is a visual representation of the regions of similarity within a sequence/between two sequences.

A dot plot can identify

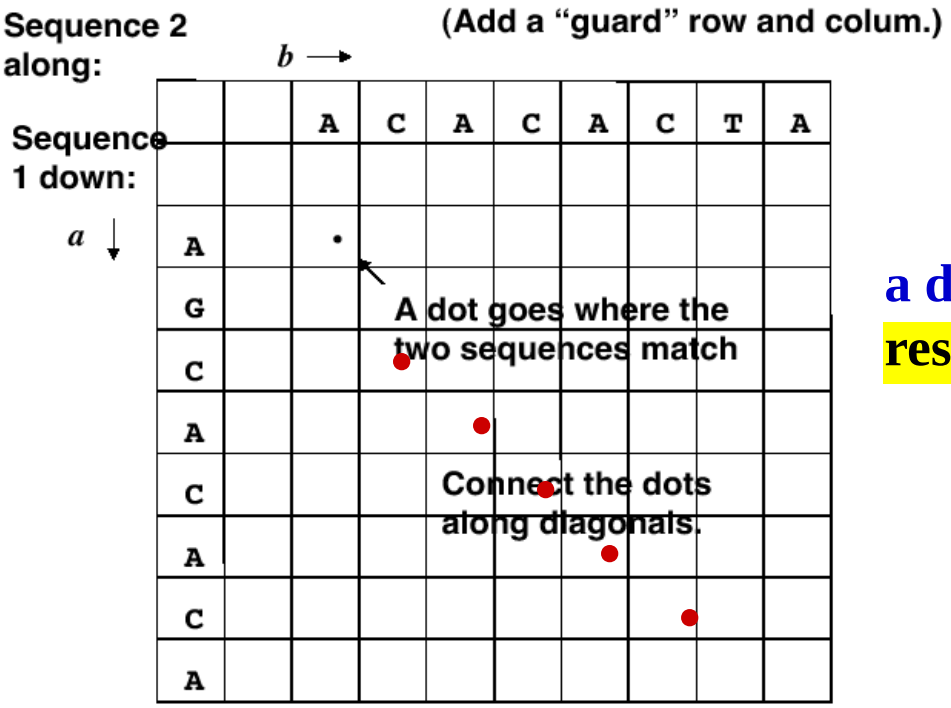
- **regions of similarity**
- **overlap regions**
- **rearrangement events**

Comparing two/more sequences

- **internal repeats, multiple copies of domains**
- **self-complementary regions in RNA sequences**

Self-comparison

Dot Plots



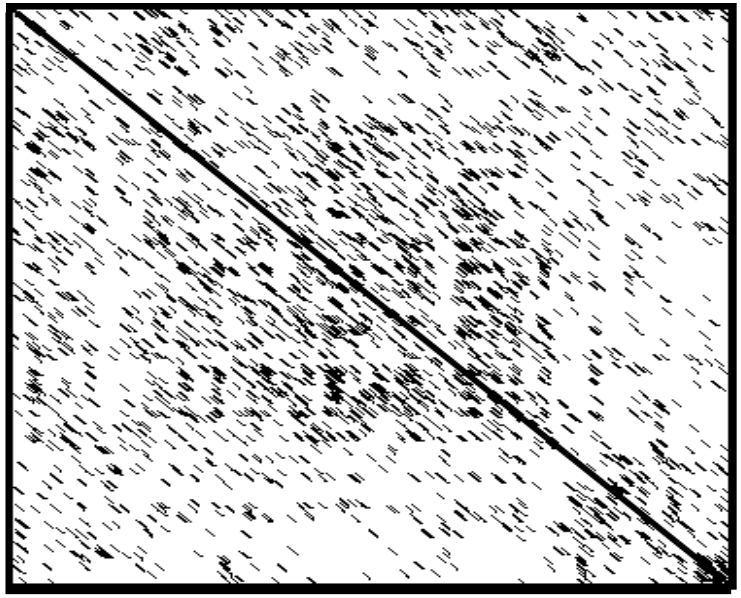
a dot is drawn for residue-residue match

Where the two sequences have substantial regions of similarity, many dots align to form diagonal lines

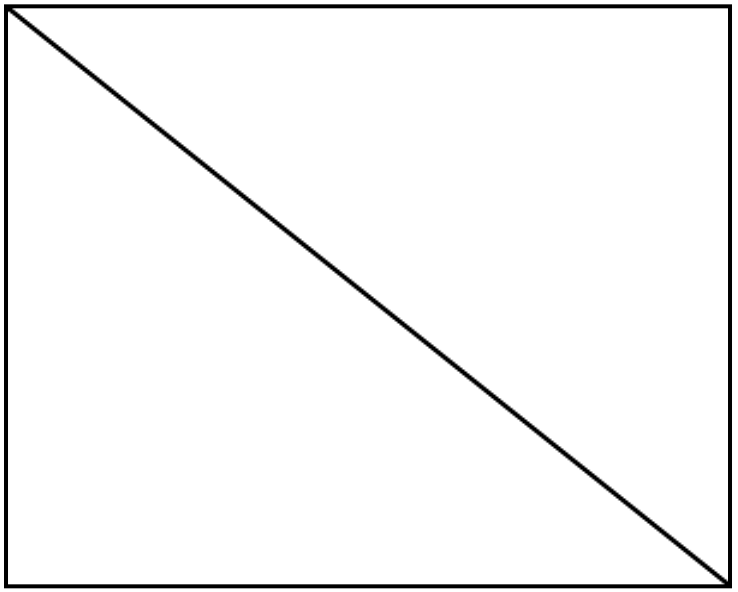
Dot Plots

If two sequences share similarity over their entire length:

Non-stringent, self-dot plot



Very stringent, self-dot plot



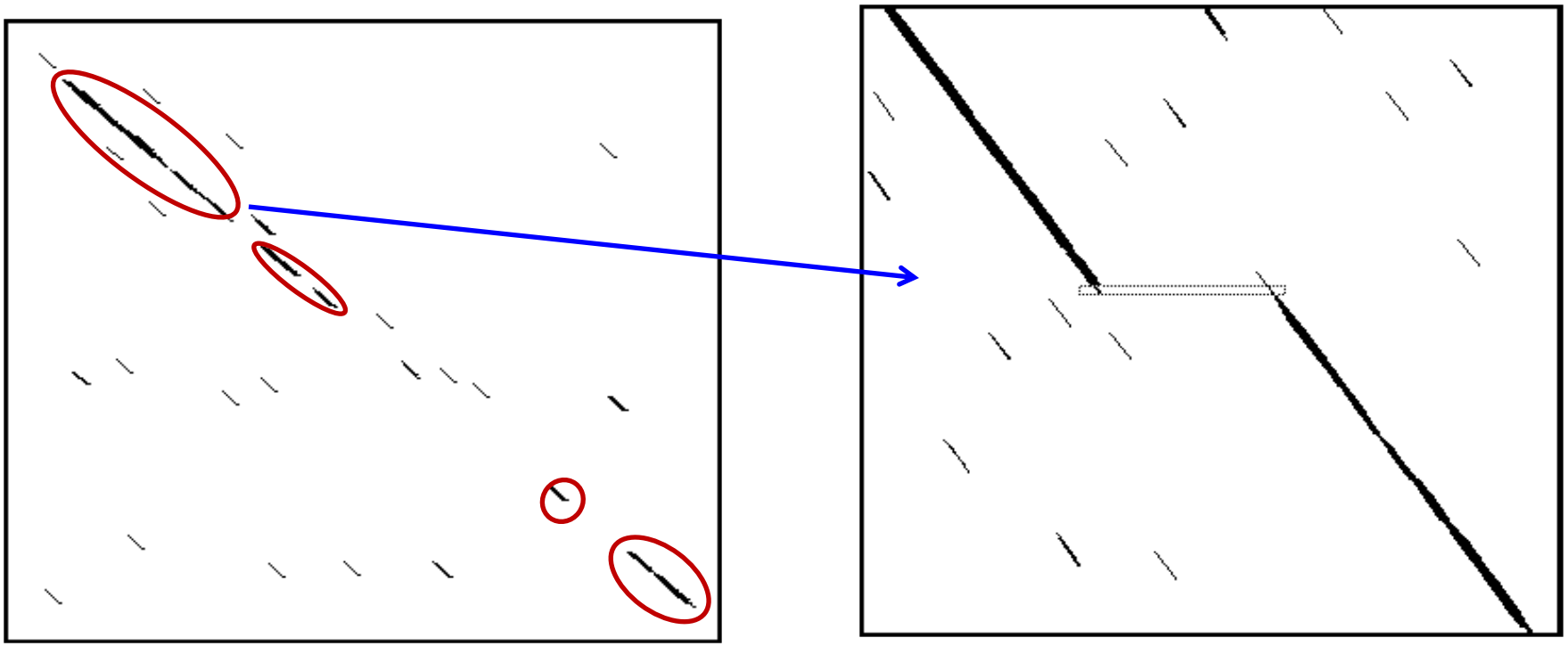
Every residue in one sequence is compared to every residue in the other sequence - nothing is missed

Dot Plots

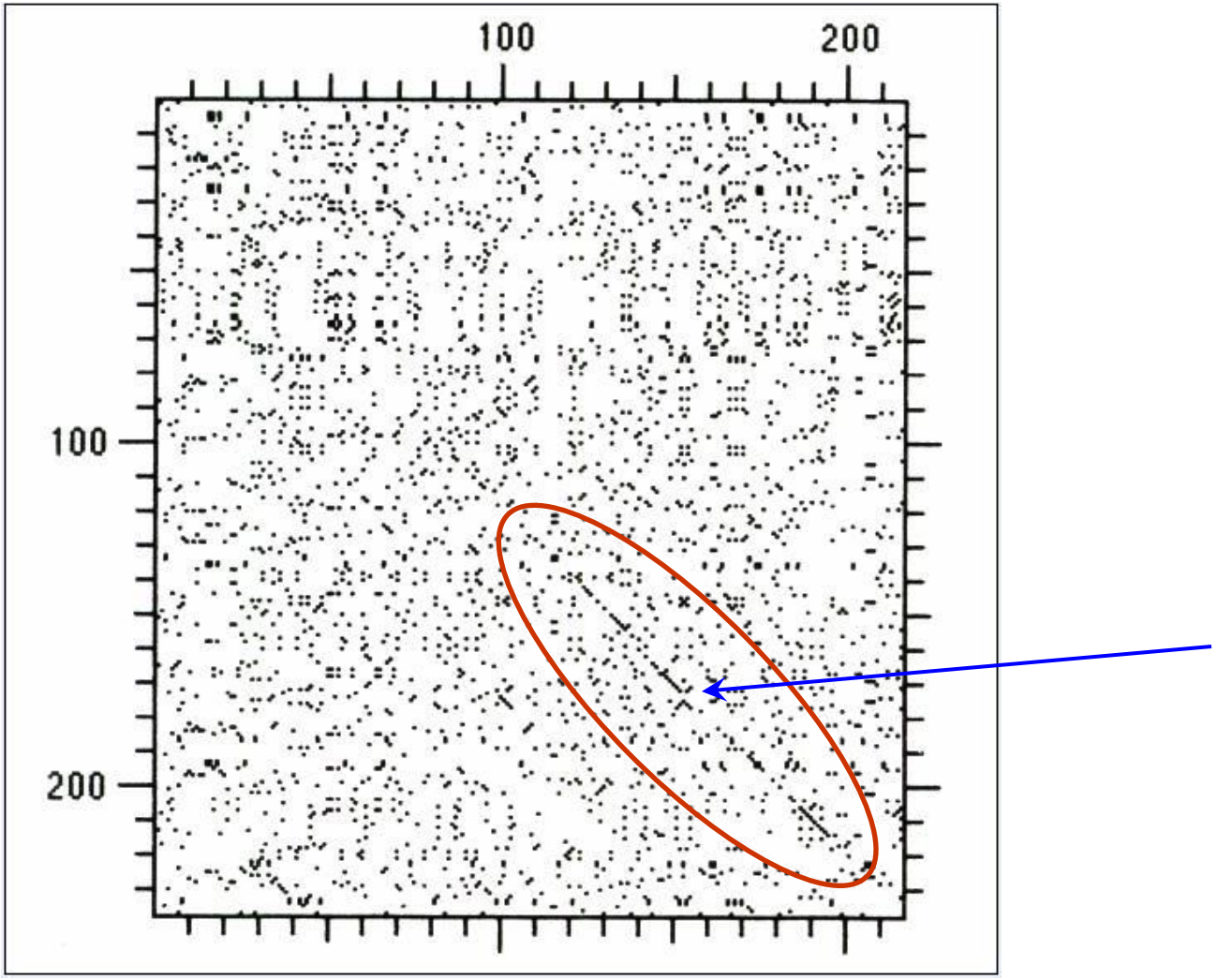
If two sequences are evolutionarily related and share patches of similarity:

Two similar, but not identical sequences

Insertion or Deletion



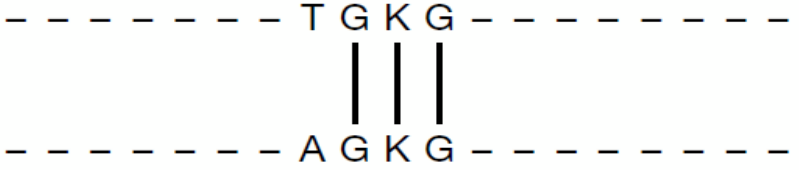
Dot matrix analysis of amino acid sequences of the phage λ cI and phage P22 c2 repressors



Dot Plots

- **Major advantage of dot matrix method for finding sequence alignment - all possible matches of residues between two sequences are found, leaving investigator choice of identifying the most significant ones**
- **Based on the dot plot, user can decide whether he deals with a case of global, i.e. end-to-end similarity, local similarity, or overlapping (similarity at the ends)**

Global alignment

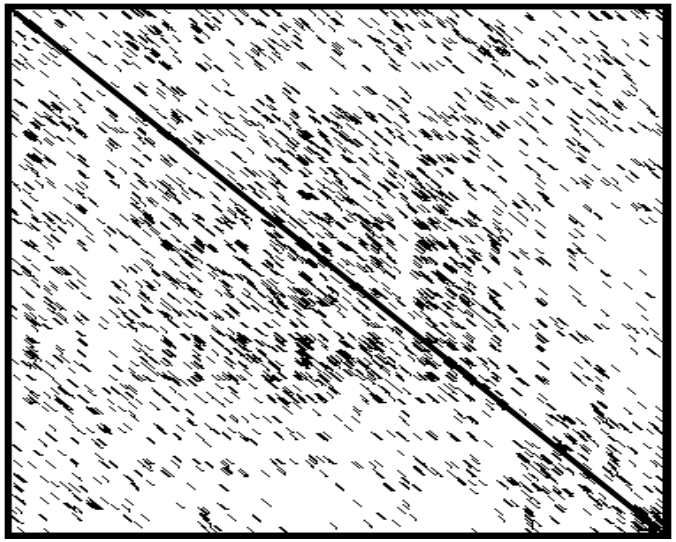


Local alignment

Dot Plots

Detection of matching region is improved by filtering out random matches in a dot matrix - by using a sliding window to compare the two sequences.

Instead of comparing every base, a window of adjacent positions in the two sequences is compared and a dot is printed only if a certain minimal number of matches occur.

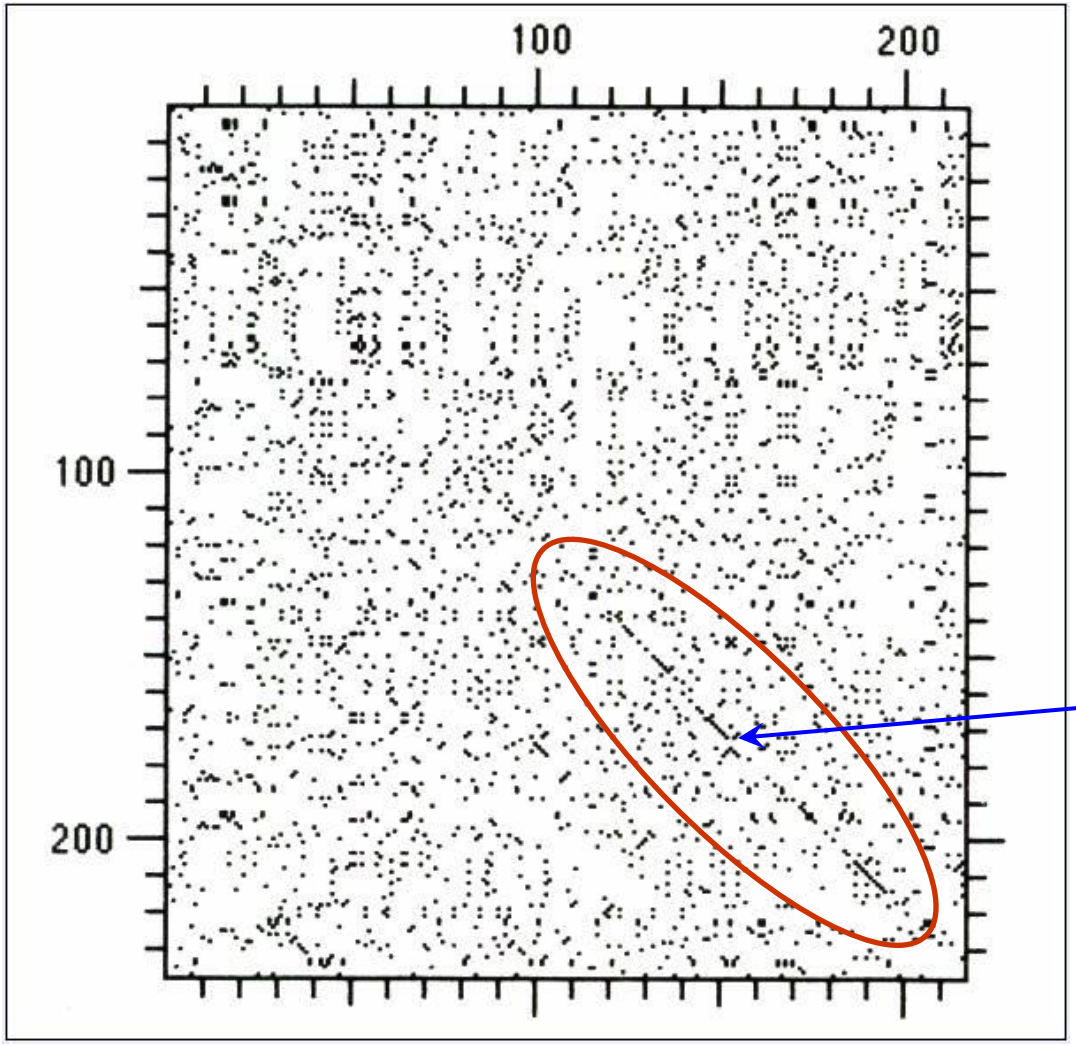


Extensions of Dot Plots

Detection of matching region is improved by filtering out random matches in a dot matrix - by using a sliding window to compare the two sequences

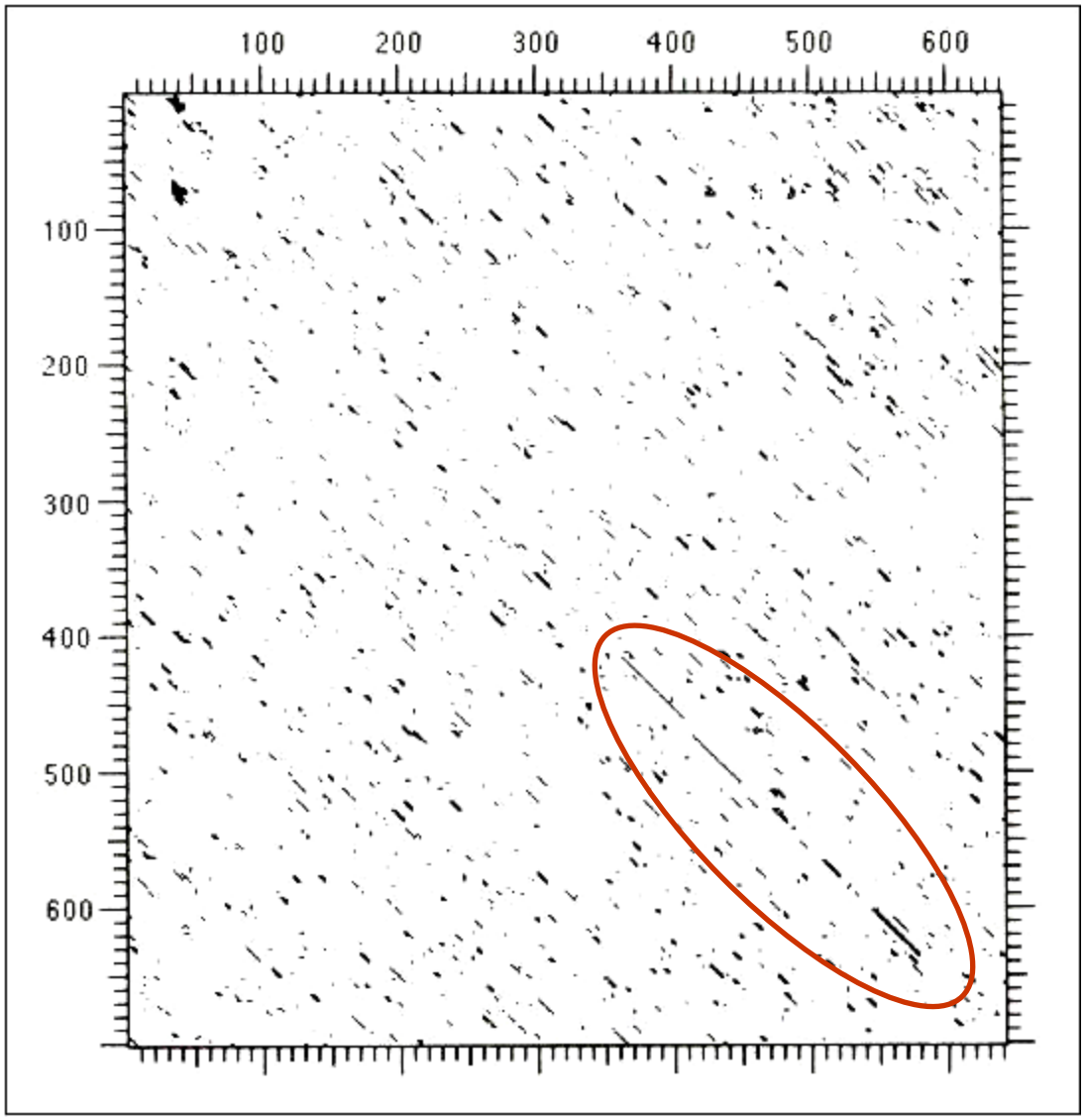
- **Window:** **size of diagonal strip centered on an entry, over which matching is accumulated, and**
- **Stringency:** **the extent of agreement required over the window, before a dot is placed at the central entry.**

Dot matrix analysis of amino acid sequences of the phage λ cI and phage P22 c2 repressors



Window size: 1
Stringency: 1

Dot matrix analysis of DNA sequences encoding the E. coli phage λ cI (horizontal) & phage P22 c2 (vertical) repressors



Window size: 11
Stringency: 7

Suggesting similarity in the C-terminal domains of the encoded proteins

Dot Plots

A large window size is generally used for DNA sequences.

- typically a window size of 15 and a suitable match requirement of 10.

For protein sequences, the matrix is often not filtered, but a window size of 2 or 3 and a match requirement of 1or 2 will highlight matching regions.

Why?

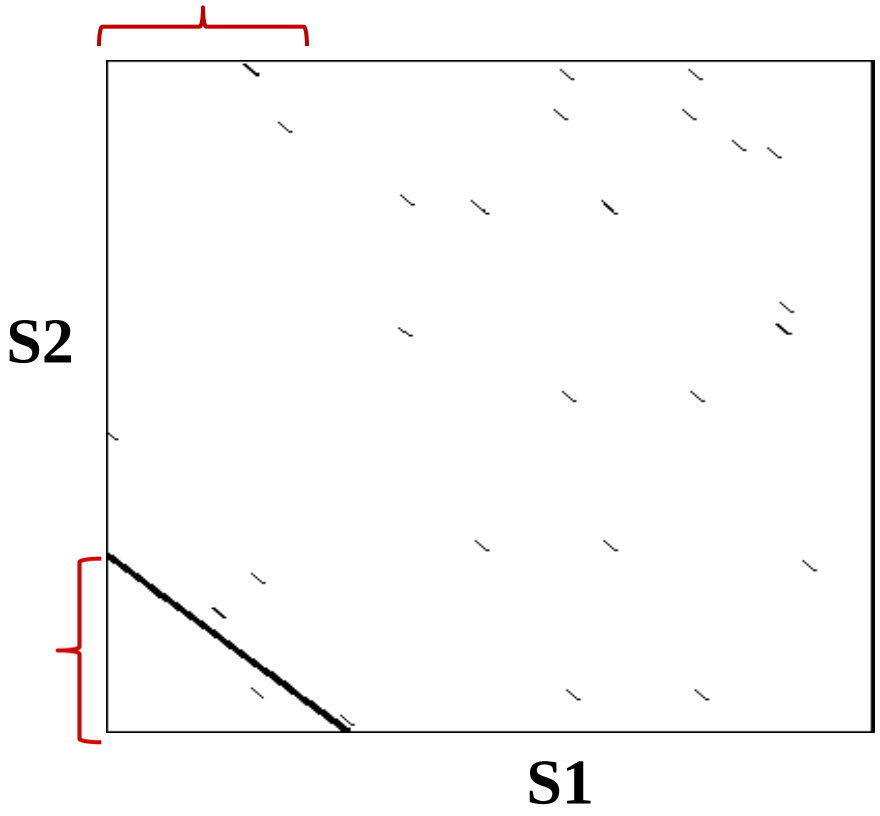
Dot Plots

If two proteins are expected to be related but have long regions of dissimilar sequence with only a small proportion of identities, such as similar active sites,

- a large window, e.g., 20, and a small stringency, e.g., 5, should be useful for seeing any similarity.

- the reason being, residues in an active site are not necessarily contiguous in the sequence, and only the positions involved in interaction are conserved.

Identifying Overlapping Sequences Dot Plots



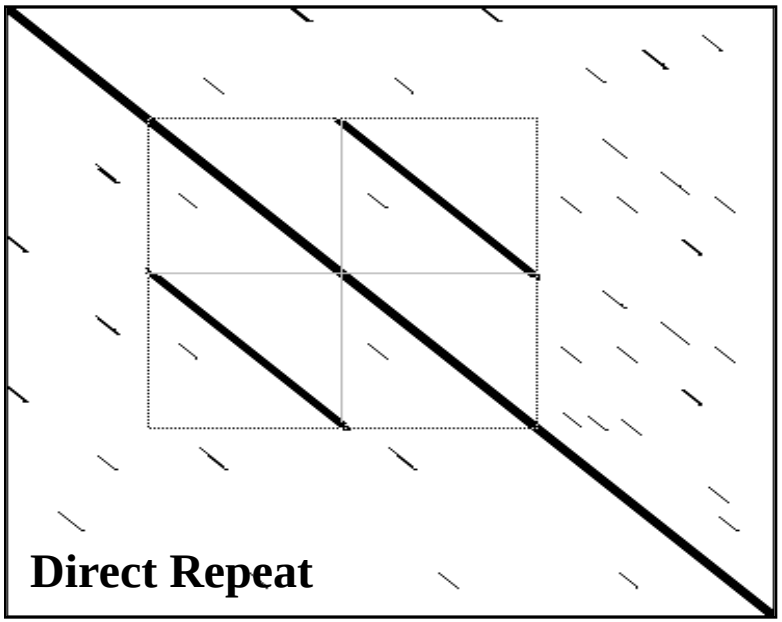
S2

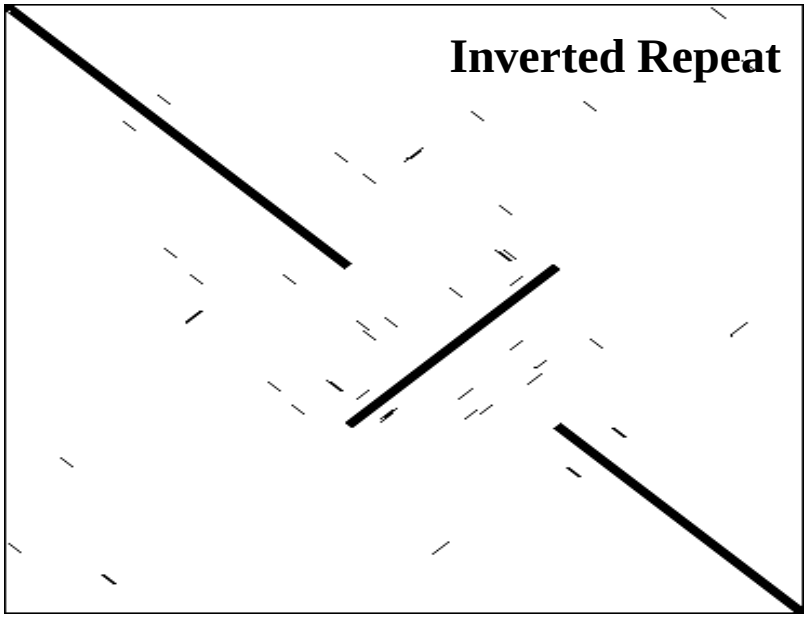
S1

When do we expect to find overlapping sequences?

Dot Plots

Self-dot plot of a tandem duplication

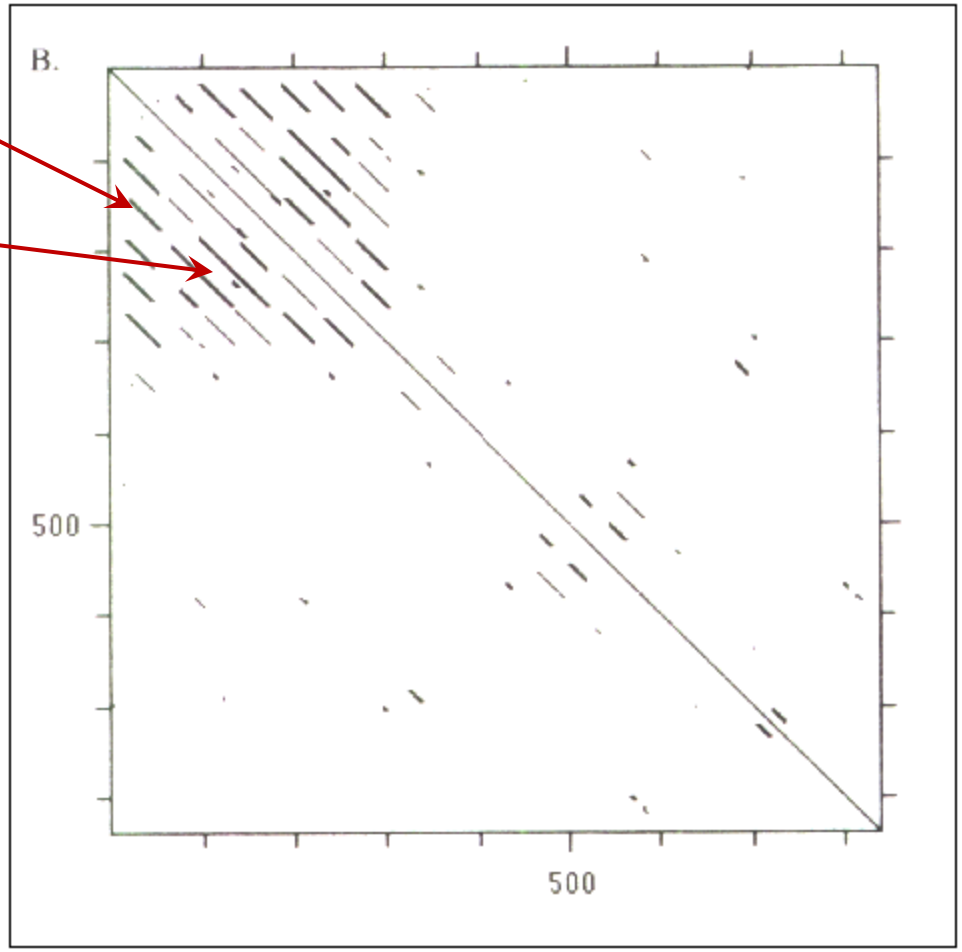




We can compare a sequence to itself - it reveals repeat regions in the sequence

Dot matrix analysis of the human LDL receptor against itself

Repeats of different lengths



Window size: 23
Stringency: 7

At Stringency: 11/23 – few remain
At Stringency: 15/23 – all disappear

Proteins composed of multiple copies of a single domain can be identified by dot plots

Self-Complementary Regions in RNA Sequences

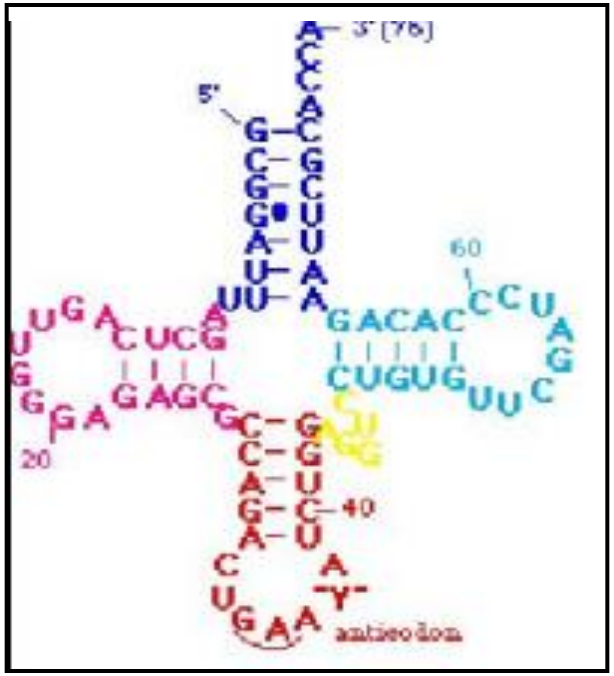
RNA secondary structure analysis begin with the identification of self-complementary regions

- these represent regions that can potentially self-hybridize to form RNA double strands

- once identified, the compatible regions may be used to predict a minimum free-energy structure.

- simplest way of identifying stretches of self-complementary regions in RNA sequence is a dot plot analysis

– there are two approaches.



Self-Complementary Regions in RNA Sequences

Method-1: Sequence is listed in 5′ to 3′ direction along the horizontal axis and its complementary sequence is listed along the vertical axis, also in the 5′ to 3′ direction.

Matrix is then scored for identities

Self-complementary regions appear as rows of dots going from upper left to lower right.

For RNA, these regions represent sequences that can potentially form A/U and G/C base pairs

	G	A	U	C	G	G
C				•		
C				•		
G	•				•	•
A		•				
U			•			
C				•		

Self-Complementary Regions in RNA Sequences

Method-2: **Alternative approach - list the RNA sequence along the horizontal axis and also along the vertical axis,**

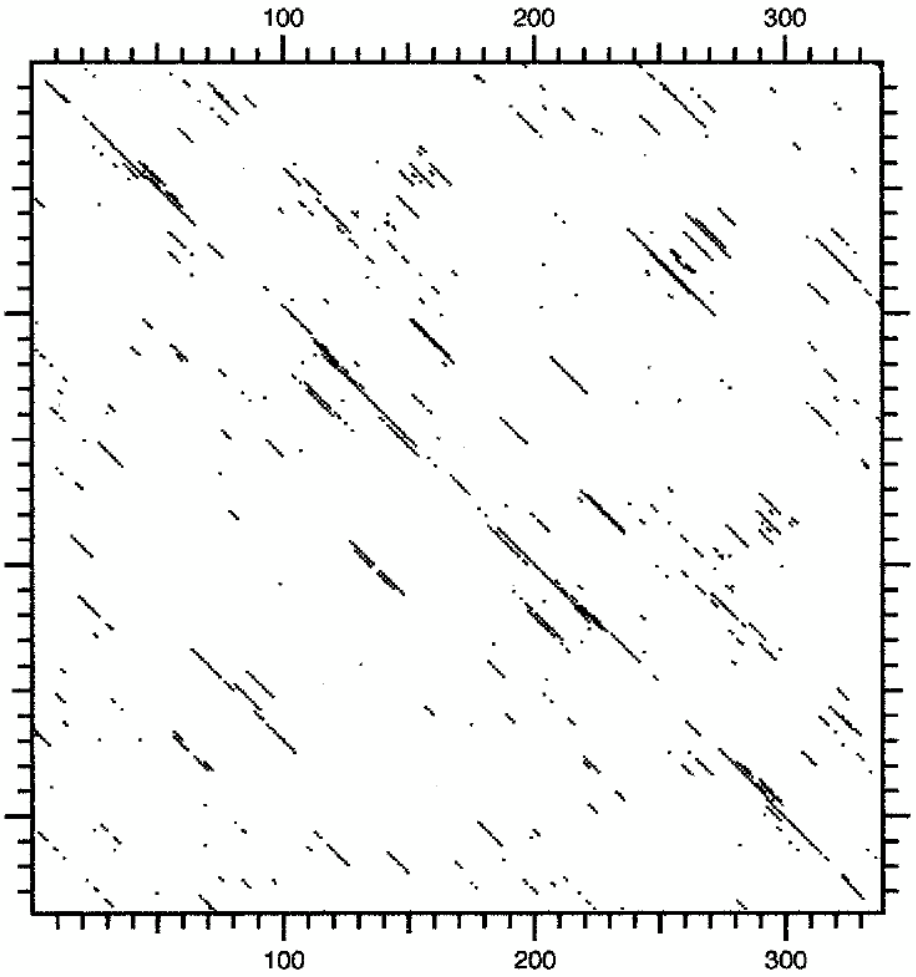
- Score matches of complementary bases G/C, A/U, and G/U instead of identities (as in the earlier method)

Diagonals indicating complementary regions will go from upper right to lower left in this matrix.

This type of matrix is used to produce an energy matrix for RNA secondary structure prediction.

	G	A	U	C	G	G
G				•		
A			•			
U		•				
C	•				•	•
G				•		
G				•		

Dot matrix Analysis of Potato Spindle Tuber Viroid for RNA Secondary Structure Analysis



Window: 15
Stringency: 11

Note: mirror image of diagonal from center to upper left and from center to lower right

Tools for Dot Plots

- **Dotter**
- **Dottup - EMBOSS (dotmatcher, dotpath, polydot)**
- **Diagon**
- **Compare & dotplot - GCG package**

EMBOSS - European Molecular Biology Open Software Suite:

http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/alignment_dot_plots_group.html

Assignment: Find out the functionalities of the various dotplot programs in EMBOSS: Dottup, dotmatcher. Dotpath, polydot?

Summarize

By analyzing the diagonal segments, dot plots can be used:

- **to find local regions of similarity, i.e., conserved and less conserved parts of homologous proteins**
  - as long diagonal lines
- **to identify domain homologies between proteins not homologous overall**
- **to identify overlapping sequences, e.g., in sequence assembly**
  - as a diagonal on a corner of the plot
- **to identify internal repeats and duplications**
  - as lines parallel to the diagonal
- **to identify insertions and deletions**
  - as breaks or discontinuities in the diagonal lines
- **to identify self-complementary regions**
  - in RNA secondary structure analysis