Sequence Comparison

DotPlots & Alignments

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, CDS, 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; Val & Leu – non-polar, Gly - polar**
- **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, allowing them to adapt to changes in the environment, etc.**
- **are responsible for inherited disorders and diseases including 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 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.

Any other reasons for Sequence Comparison?

Comparison of Sequences

Any other reasons for Sequence Comparison?

- **Identifying species – as in the case of DNA barcoding**
- **Phylogenetic analysis – to find evolutionary relatedness between species**
- **Genome comparison between individuals in a population – for structural variation analysis**
- **Genome comparison between diseased (e.g., cancer) and normal cells – for identifying variations responsible for the disease**
- **Genome comparison between species – for understanding genome evolution**

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
- **Knowledge-based prediction:** extract **empirical rules** from known examples representing **sequence-structure or sequence-function relationships**.

  – multiple alignment

  e.g., motif identification, identifying remote homologs

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

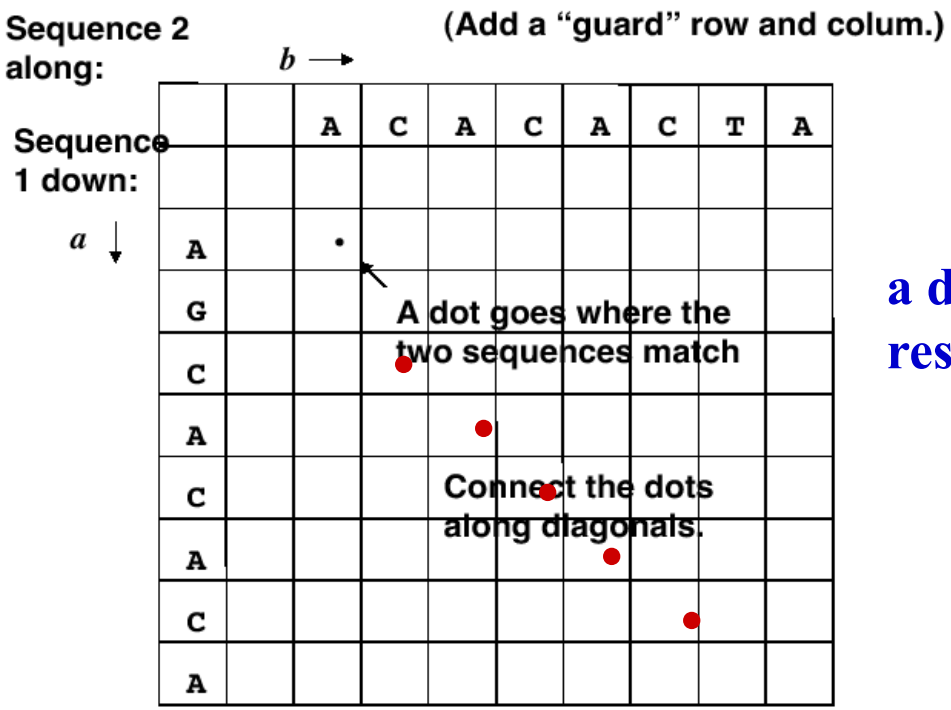
- **regions of similarity**
- **overlap regions**
- **rearrangement events**

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

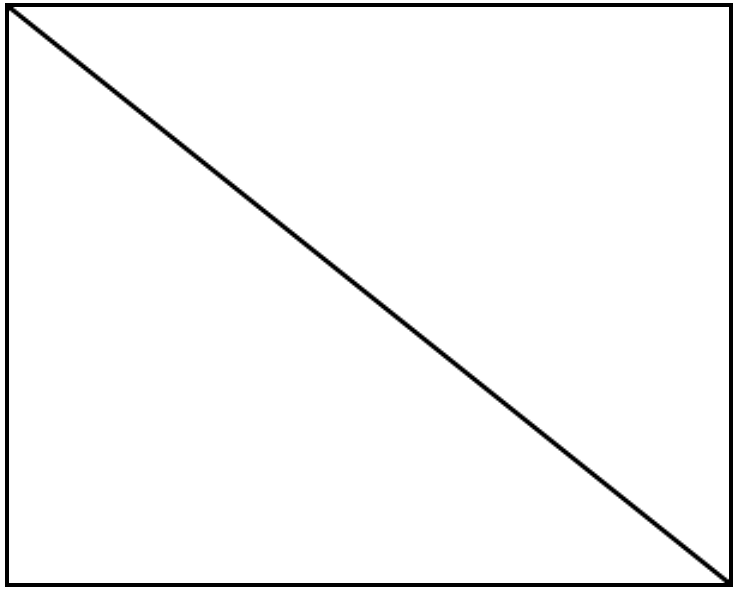
Where the two sequences have substantial regions of similarity, many dots align to form diagonal lines

Dot Plots

When two sequences share similarity over their entire length, a diagonal line will extend from one corner of the dot plot to the diagonally opposite corner.

Non-stringent, self-dot plot

Very stringent, self-dot plot

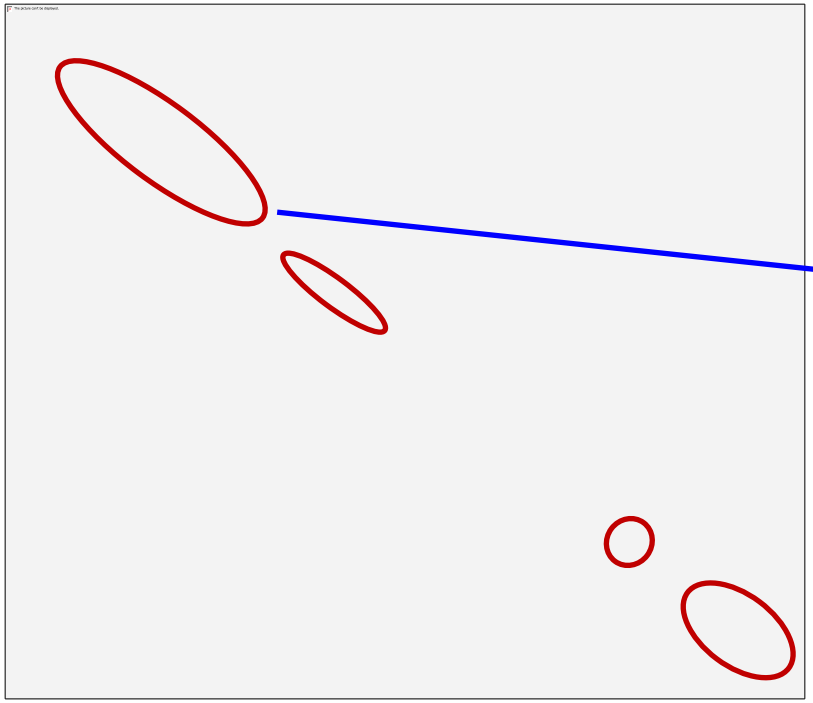


Every residue in one sequence is compared to every residue in the other sequence - nothing is missed

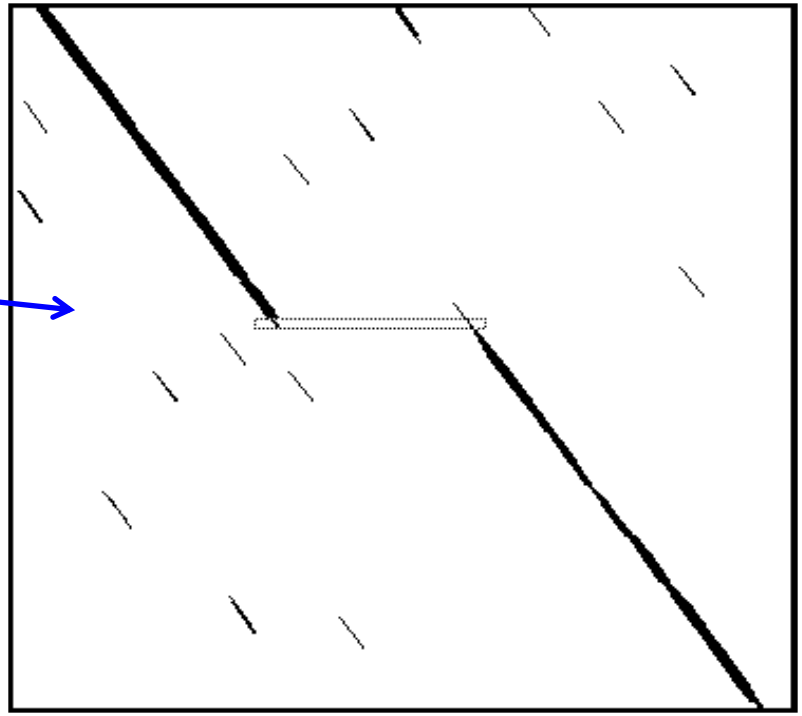
Dot Plots

If two sequences only share patches of similarity this will be revealed by short diagonal stretches.

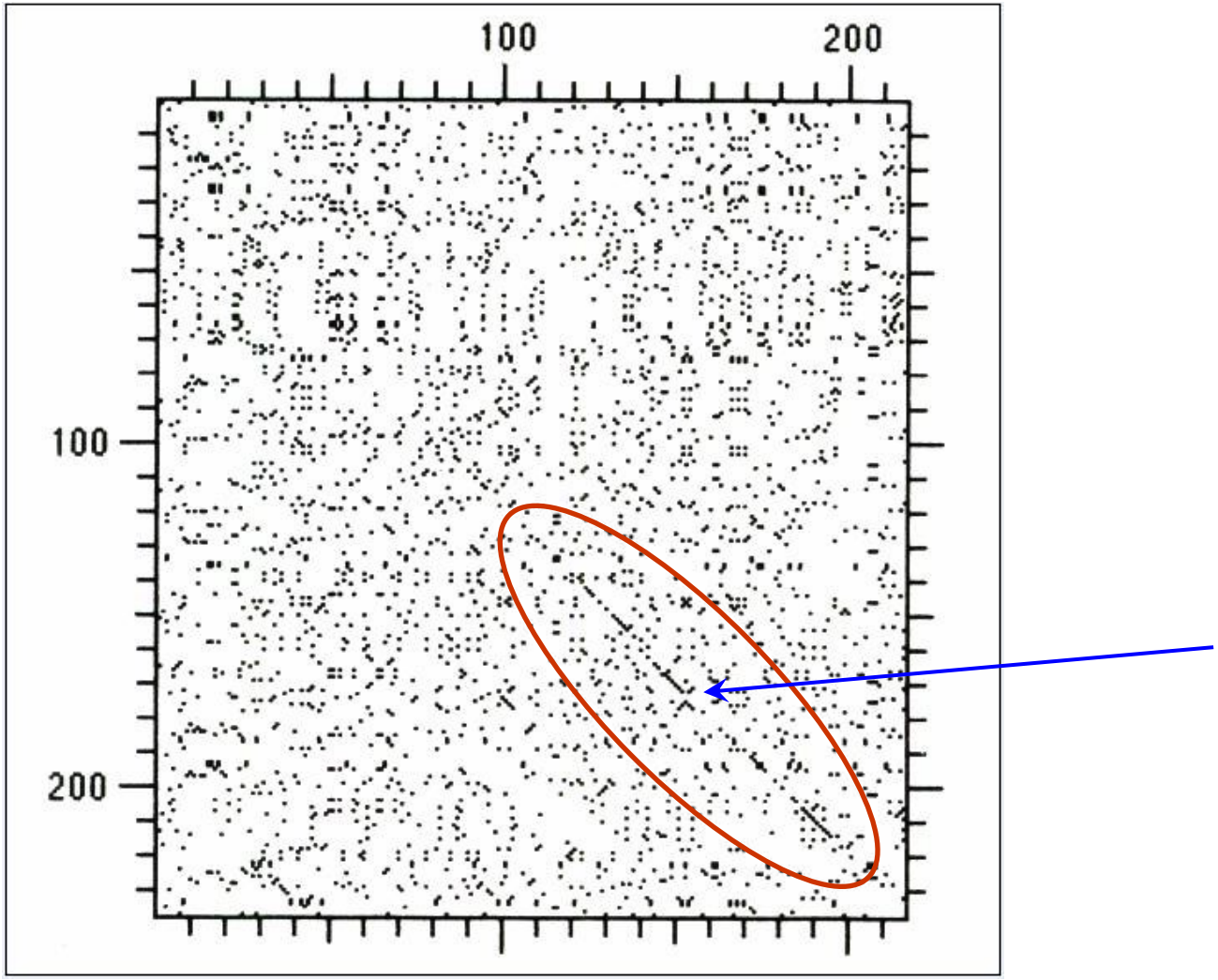
Two similar, but not identical sequences



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)**

```
L G P S S K Q T G K G S – S R I W D N
|       |     | | |       |     |
L N – I T K S A G K G A I M R L G D A
```

Global alignment

```
– – – – – – – T G K G – – – – – – – –
              | | |
– – – – – – – A G K G – – – – – – – –
```

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

Thus, for window analysis of dot plots we define:

- **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 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

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?

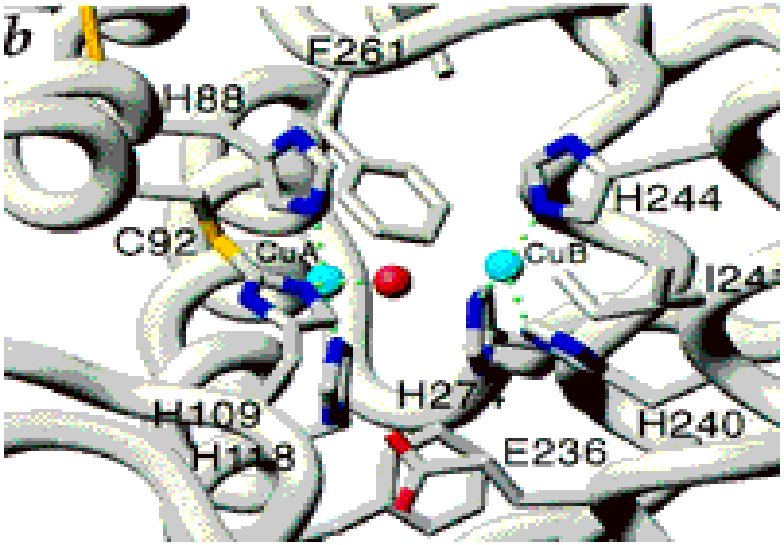
- the no. of random matches is more in case of DNA due to the use of 4 nucleotides symbols as compared to 20 amino acid symbols for proteins.

Dot Plots

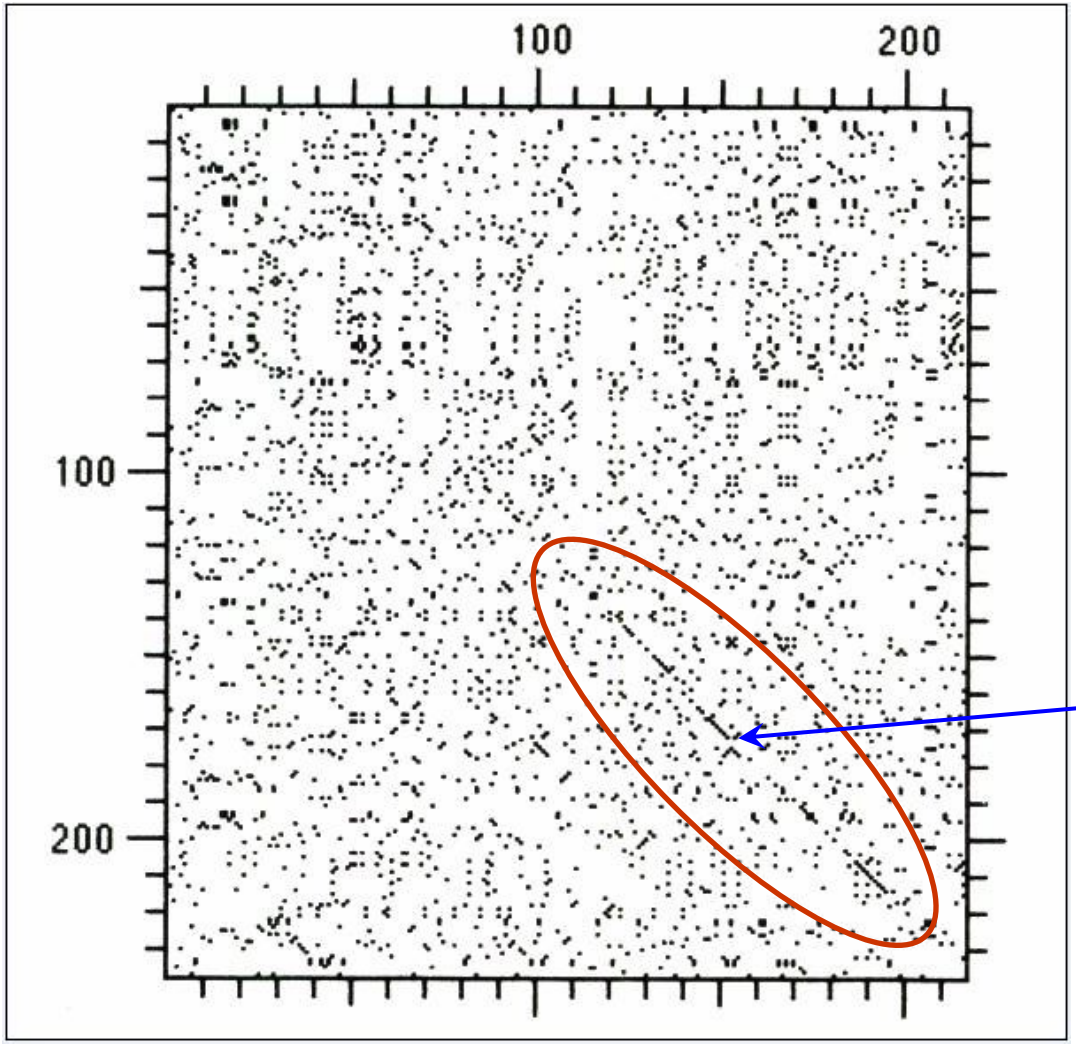
If two proteins are expected to be related but have long regions of dissimilar sequence with only a small proportion of identities, such as similar active sites,

- 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.

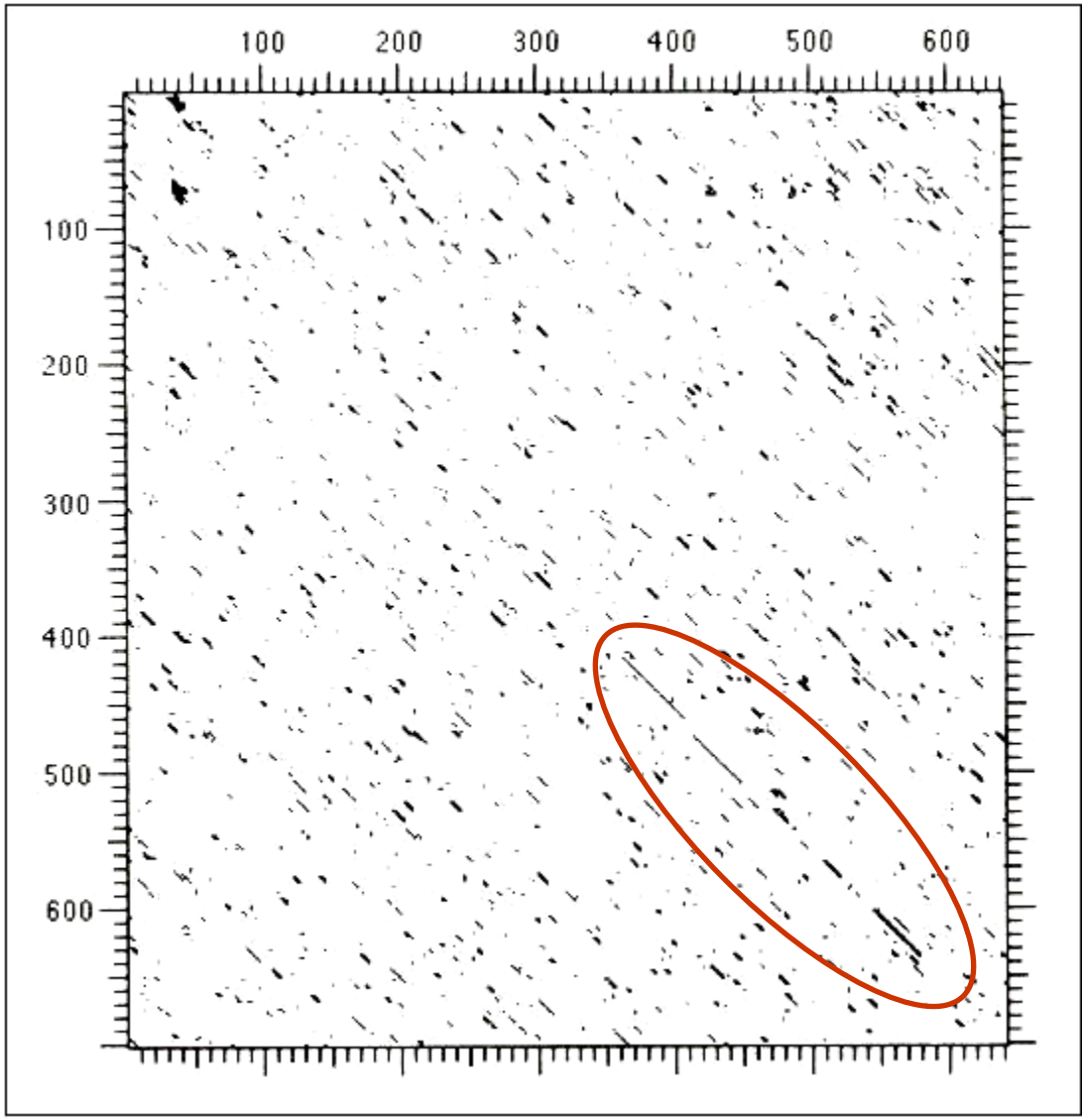


Dot matrix analysis of amino acid sequences of the phage λ cI and phage P22 c2 repressors



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

There are three types of variations in the analysis of protein sequences by the dot matrix method.

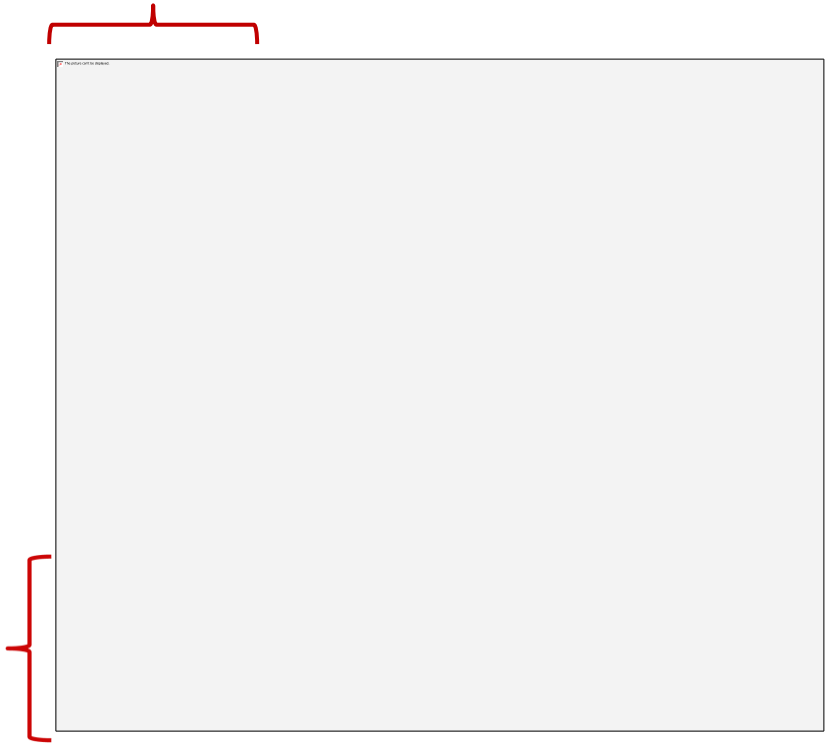
- **First, chemical similarity of amino acid R group or some other feature for distinguishing amino acids may be used to score similarity.**
- **Second, scoring matrices may be used to provide scores for matches based on their occurrence in aligned protein families.**

When these tables are used, a dot is placed in the matrix only if a minimum similarity score is found.

These table values may also be used in a sliding window option, which averages the score within the window, and prints a dot only above a certain average score.

- improves the sensitivity of a dotplot while comparing protein sequences

Identifying Overlapping Sequences Dot Plots



When do we expect to find overlapping sequences?

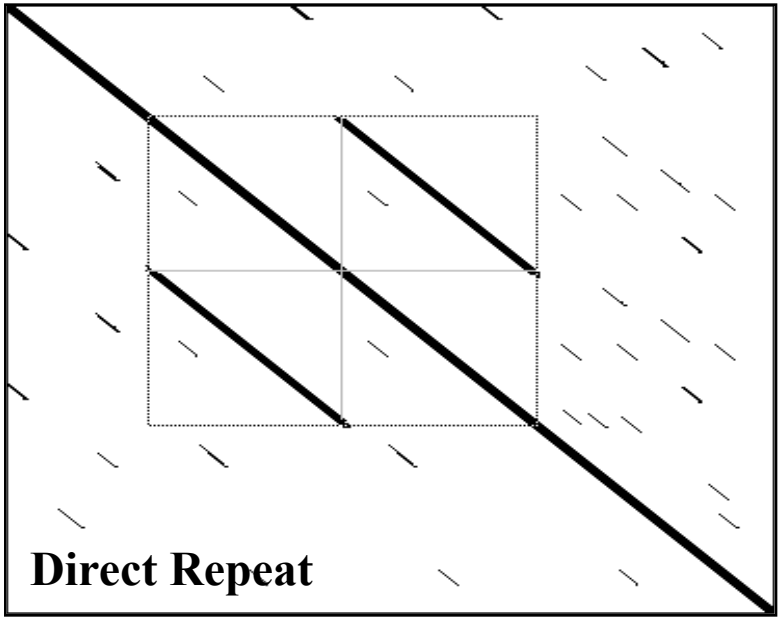
- during sequence assembly, aligning ESTs to gene / genomic sequences

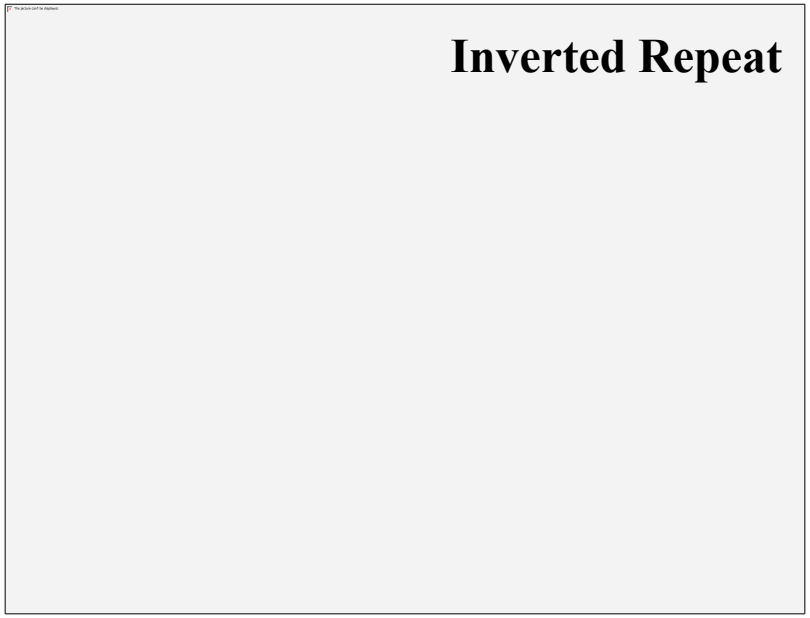
Dot Plots

- **Sequences may contain regions of self-similarity termed internal repeats. A dot plot comparison of the sequence with itself will reveal internal repeats by displaying several parallel diagonals.**
- **Presence of repeats of the same character many times (low-complexity regions) appear as - horizontal or vertical rows of dots that sometimes merge into rectangular or square patterns**

Dot Plots

Self-dot plot of a tandem duplication





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

Sequence Repeats

Identifying direct and inverted repeats within sequences using Dot matrix analysis.

Sequence is aligned against itself and the presence of repeats is revealed by rows of dots parallel to the diagonal

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

	G	A	T	T	A	G
G	•					•
A		•			•	
T			•	•		
T			•	•		
A		•			•	
G	•					•

Repeats of a Single Sequence Symbol

A dot matrix analysis can also reveal the presence of repeats of the same sequence character many times.

- these repeats become apparent on the dot matrix as horizontal or vertical rows of dots, merging into rectangular or square patterns.

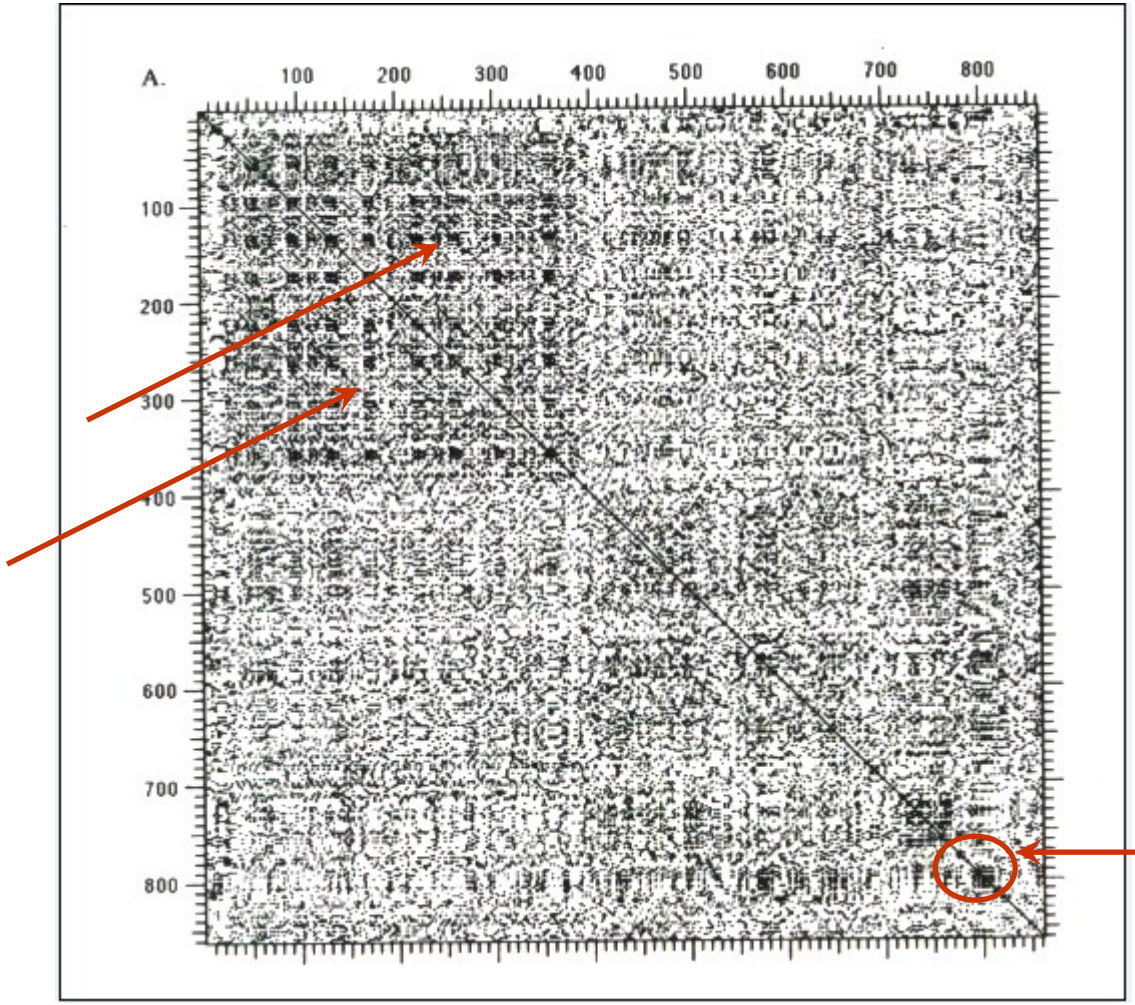
- as seen in the lower-right regions of the dot matrix of the human LDL receptor

Occurrence of such repeats of the same character increases the difficulty of aligning sequences as they create alignments with artificially high scores

- Mask these repeats during database searches

Programs: DUST (DNA), SEG (Protein)

Dot matrix analysis of the human LDL receptor against itself

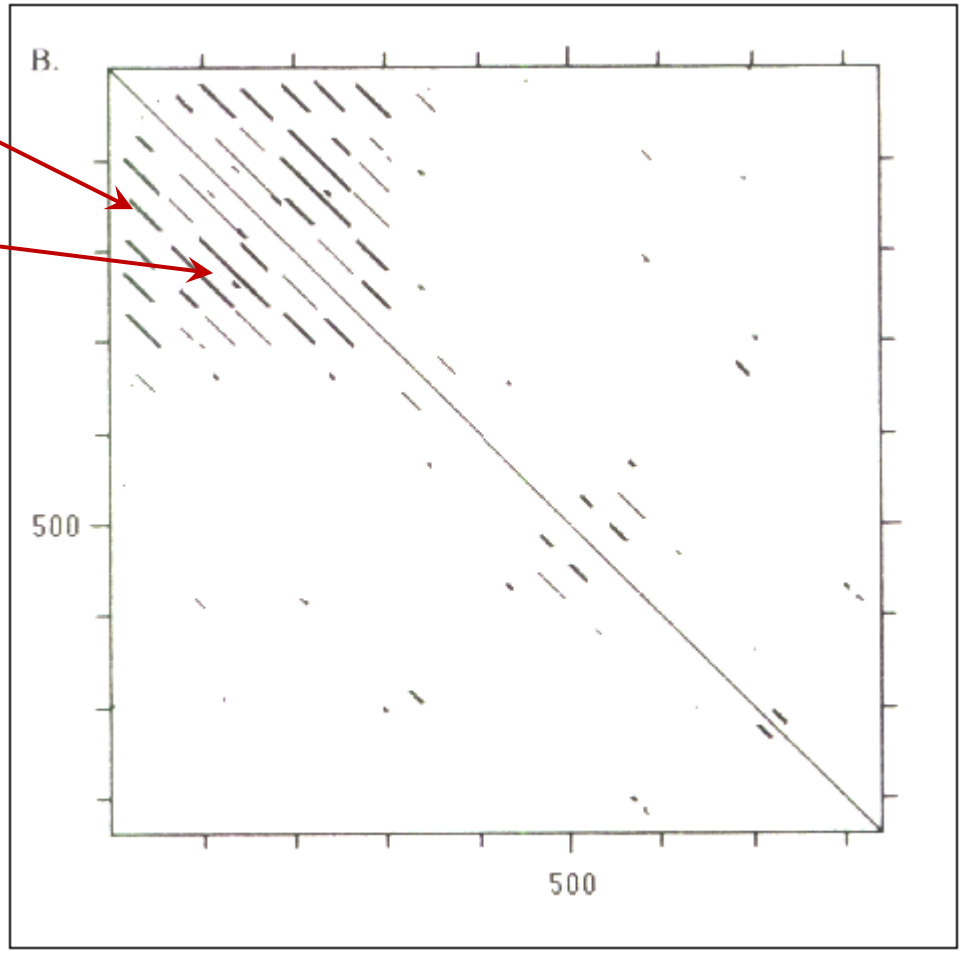


Window size: 1
Stringency: 1

Low-complexity region

Dot matrix analysis of the human LDL receptor against itself

Repeats of different lengths



Window size: 23
Stringency: 7

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

Self-Complementary Regions in RNA Sequences

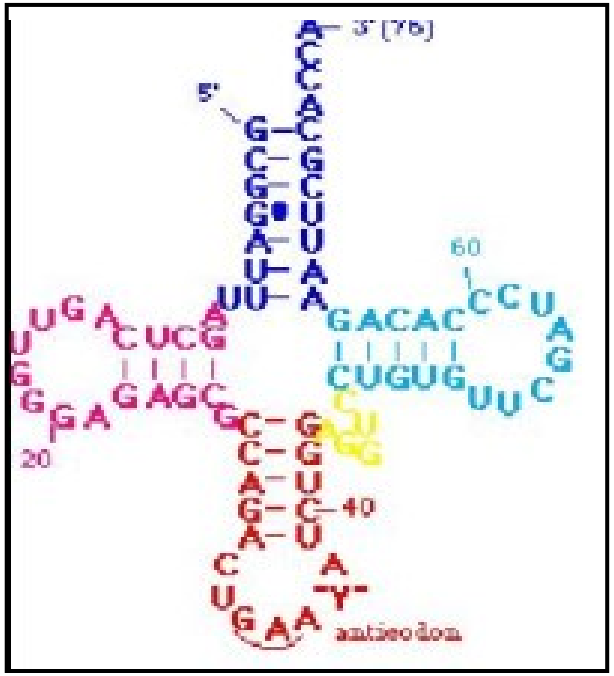
RNA secondary structure analysis begin with the identification of self-complementary regions

- 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

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

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/U base pairs <u>not</u> included in this simple analysis because they play a less significant role in base-pairing.

Self-Complementary Regions in RNA Sequences

As with matching DNA sequences, there are many random matches between the four bases in RNA, and the diagonals are difficult to visualize.

A long nucleotide window and a requirement for a large number of matches within this window are used to filter out the random matches.

Self-Complementary Regions in RNA Sequences

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

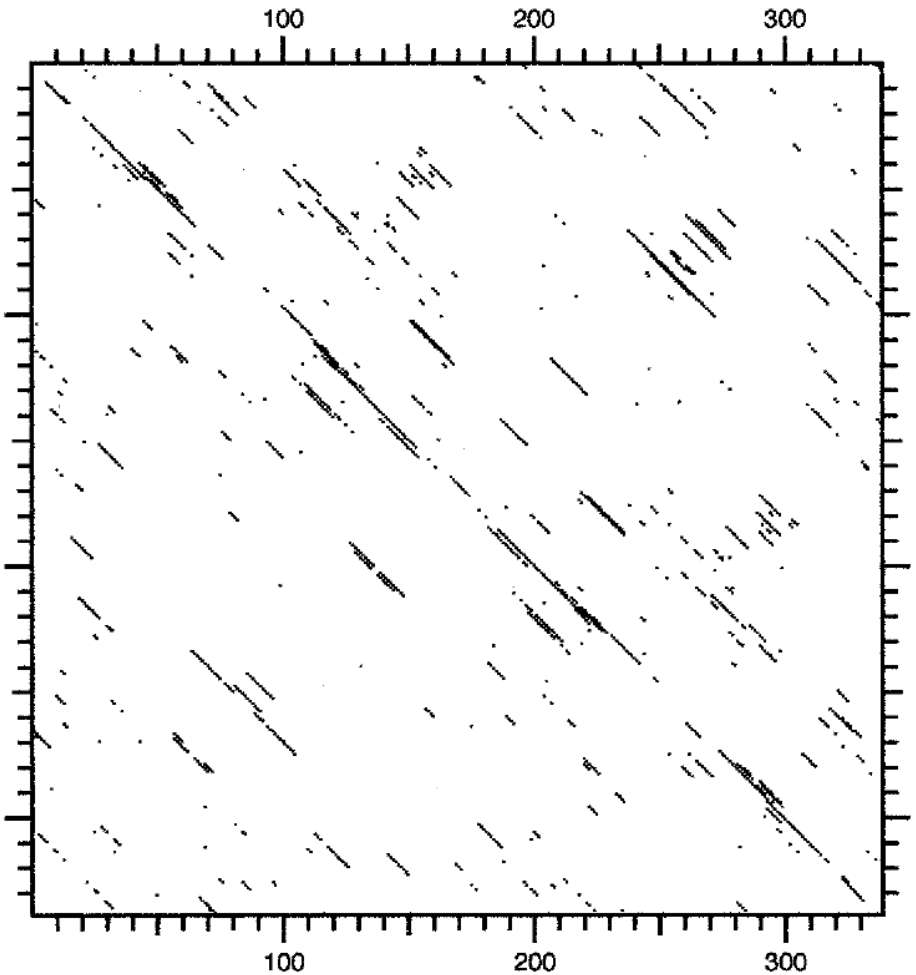
- Score matches of complementary bases G/C, A/U, and G/U instead of identities (as in the earlier method)

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

EMBOSS - European Molecular Biology Open Software Suite

- is a suite of free software tools for sequence analysis. It consists of a wide variety of programs ranging in application from database search to presentation of sequence data.

https://www.ebi.ac.uk/Tools/emboss/

dottup

EMBOSS dottup - displays a wordmatch dotplot of two sequences

It looks for places where words (tuples) of a specified length have an exact match in both sequences and draws a diagonal line over the position of these words.

Using a longer tuple size displays less random noise, runs extremely quickly, but is less sensitive.

Shorter word sizes are more sensitive to shorter or fragmentary regions of similarity, but also display more random points of similarity (noise) and runs slower

For what tasks is this program suitable?

dottup

For what tasks is dottup program suitable?

- **When comparing a cDNA sequence (mRNA sequence converted to double stranded DNA sequence) to the genomic sequence, we expect an exact match, and dottup is suitable in such situations.**
- **Comparing very closely related sequences, when we expect a large no. of exact matches.**

Other Dot Plot programs in EMBOSS:

- **dotmatcher – displays a threshold dotplot of two sequences**

  - a sliding window analysis along the diagonal; displays a line over the window if the sum of the comparisons (using a substitution matrix) exceeds a threshold. It is slower but much more sensitive.

- **dotpath - Displays a non-overlapping wordmatch dotplot of two sequences**

- **polydot - Displays all-against-all dotplots of a set of sequences**

Difference between dottup and dotpath programs in EMBOSS? Explain with application

Assignment:

Find out the functionalities of the various dotplot programs in EMBOSS.

emboss

About • Applications • GUIs • Servers • Downloads • Licence • User docs • Developer docs • Administrator docs • Get involved • Support • Meetings • News • Credits

About EMBOSS

Overview • Uses • FAQ Citing EMBOSS

A high-quality package of free, Open Source software for molecular biology ... *more >*

Applications

EMBOSS • EMBASSY • Groups Proposed

Hundreds of useful, well documented applications for molecular sequence and other analyses ... *more >*

GUIs

Jemboss • GUIs • Web • Others

We support the Jemboss GUI but many others are available... *more >*

Servers

Portals • Servers • Mirrors • Misc

Many EMBOSS portals, servers and mirrors are available ... *more >*

Downloads

Stable release • Developers (CVS) version • Getting started

EMBOSS is open source software and is freely available to all ... *more >*

Licence

Licensing terms

EMBOSS uses the General Public Licence (GPL) and Library GPL ... *more >*

EMBOSS

vers

[sort alphabetically]

ALIGNMENT CONSENSUS

cons
megamerger
merger

ALIGNMENT DIFFERENCES

diffseq

ALIGNMENT DOT PLOTS

dotmatcher
dotpath
dottup
polydot

ALIGNMENT GLOBAL

alignwrap
est2genome
needle
stretcher

ALIGNMENT LOCAL

DOTTUP

(Displays a wordmatch dotplot of two sequences)

Help ?

Fields with a coloured background are optional and can safely be ignored...

[Hide optional fields]

1. SET THE PARAMETERS FOR THE RUN (OR ACCEPT THE DEFAULTS...)

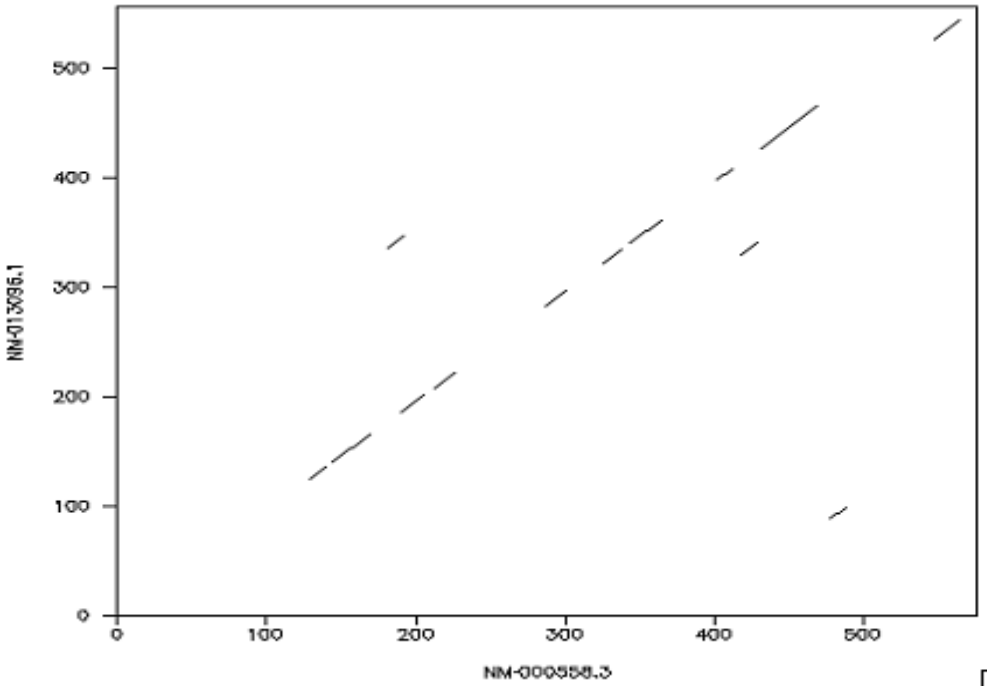
input section

Select an input sequence.

Use one of the following three fields:

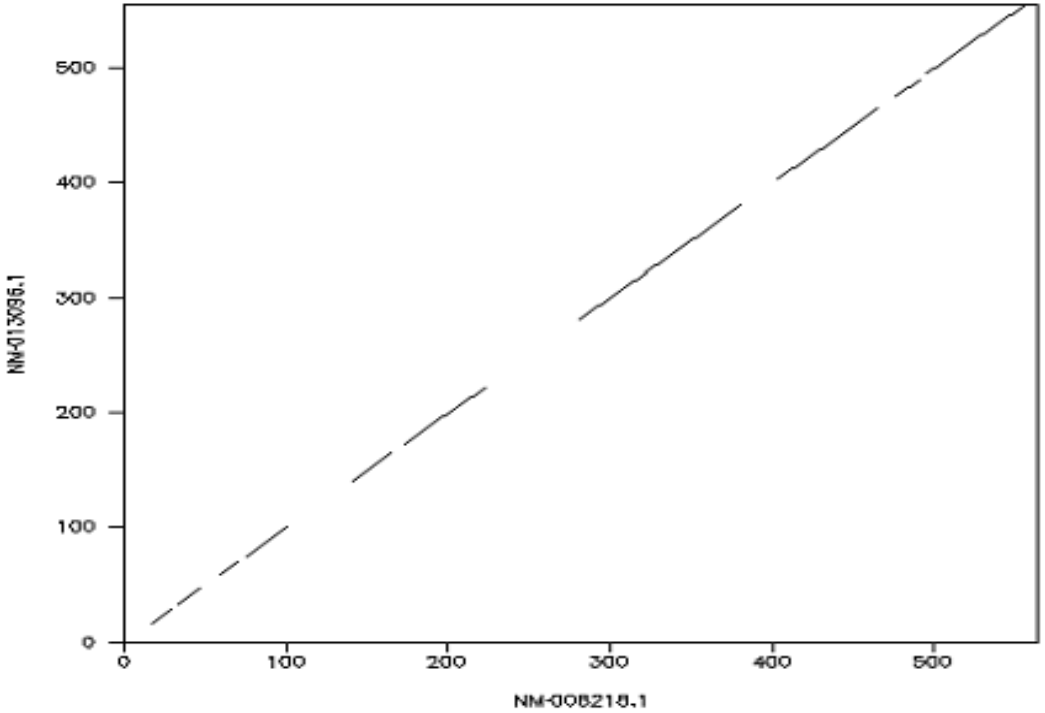
1. To access a sequence from a database, enter the USA path here: *(dbname:entry)*
2. Or, upload a sequence file from your local computer here: Browse...
3. Or enter the sequence data manually here:

1. **>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA**
2. **>gi|6981009|ref|NM_013096.1| Rattus norvegicus hemoglobin alpha, adult chain 1 (Hba-a1), mRNA**
3. **>gi|6680174|ref|NM_008218.1| Mus musculus hemoglobin alpha, adult chain 1 (Hba-a1), mRNA**



Homo sapiens
vs
Mus Musculus

Mus Musculus
vs
Rattus norvegicus



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

Summarize

- **For DNA sequence dot matrix comparisons, use <u>long</u> windows and <u>high</u> stringencies, e.g., 11 & 7, 15 & 11.**
- **For protein sequences, use <u>short</u> windows, e.g., 2 & 1 for window and stringency, respectively.**
- **When looking for a short domain of partial similarity in otherwise not-similar protein sequences, e.g. sharing similar active sites**

  - use a <u>longer</u> window and a small stringency, e.g., 15 & 5, for window and stringency, respectively.

Sequence Alignment

Sequence alignment - a scheme of writing one sequence on top of another where the residues in one position are deemed to have a common evolutionary origin

If the same letter occurs in both sequences then this position has been conserved in evolution.

If the letters differ it is assumed that the two derive from an ancestral letter (could be one of the two or neither)

Comparison of Sequences

Sequence alignment of two sequences basically involves

- identifying regions of similarity, i.e., *conserved regions*, between them

- to find out if the two sequences are related or not

- enable us to extrapolate knowledge of the known sequence, or family, to the unknown query sequence

Any other reasons for Sequence Comparison?

Comparison of Sequences

Sequence alignment of two sequences basically involves

- identifying regions of similarity, i.e., *conserved regions*, between them

- to find out if the two sequences are related or not

- enable us to extrapolate knowledge of the known sequence, or family, to the unknown query sequence

- identifying species, evolutionary analysis

Statistical measures have been proposed to evaluate the significance of alignment, i.e.,

- decide whether the alignment is more likely to have occurred because they are related, or just by chance

Sequence Alignment

A letter or a stretch of letters may be paired up with dashes in the other sequence to signify an insertion or deletion event.

Since an insertion in one sequence can always be seen as a deletion in the other, one frequently uses the term "*indel*"

I	**BANANA-** **-ANANAS**	**BANANA** **PANAMA**	**II**
	Score: 10	**Score: 2**	

Sequence Alignment

Using a simple evolutionarily motivated scoring scheme, an alignment mediates the definition of a distance for two sequences:

Assign 0 to a match, some positive number (say, +1) to a mismatch and a larger positive number (say, +5) to an *indel*.

By adding these values along an alignment one obtains a score for this alignment:

BANANA-	**BANANA**
- ANANAS	**PANAMA**
Score: 10	**Score: 2**

Sequence Alignment

A distance function for two sequences can be defined by looking for the alignment which yields the *minimum score*

Using dynamic programming this minimization can be effected without explicitly enumerating all possible alignment of two sequences.

The idea of assigning a score to an alignment and then minimizing over all alignments is at the heart of all biological sequence alignments.

Sequence Alignment

Note: one may either define a distance or a similarity function to an alignment.

- difference lies mainly in the interpretation of the values

A distance function defines 0 for a match and positive values for mismatches or gaps, and then aims at minimizing this distance

A similarity function assigns high positive values to matches and negative values to mismatches and gaps, and then maximize the resulting score.

Basic structure of the algorithm is the same for both cases.

When would you use a distance function and a similarity function for scoring an alignment?

Sequence Alignment

Thus, an alignment is:

- **a mutual arrangement of two sequences**
- **It exhibits where the two sequences are similar, and where they differ**
- **An 'optimal' alignment is one that exhibits the most correspondences, and the least differences**
- **'Optimal' alignment need not reflect the true evolutionary relationship between two sequences, though it usually does**

Similarity ⇒ Homology

Why is this not true?

Sequence Alignment

Differences between similarity and homology:

- **Similarity is simply a measure of expression how alike two sequences are**
- **Homology means there is an evolutionary relationship between two sequences - there are no degrees of homology.**
- **Extending this to individual residues they are 'identical' or 'similar' residues - similar implies that they share certain physicochemical properties**
- **Homology cannot be observed, it is only an inference**

Differences between similarity and homology

Identical protein sequences result in identical 3-D structures - similar sequences may result in similar structures, and this is usually the case.

The converse is not true: identical 3-D structures do not necessarily indicate identical sequences. It is because of this that there is a distinction between "homology" and "similarity".

There are examples of proteins in the databases that have nearly identical 3-D structures, and are therefore homologous, but do not exhibit significant (or detectable) sequence similarity

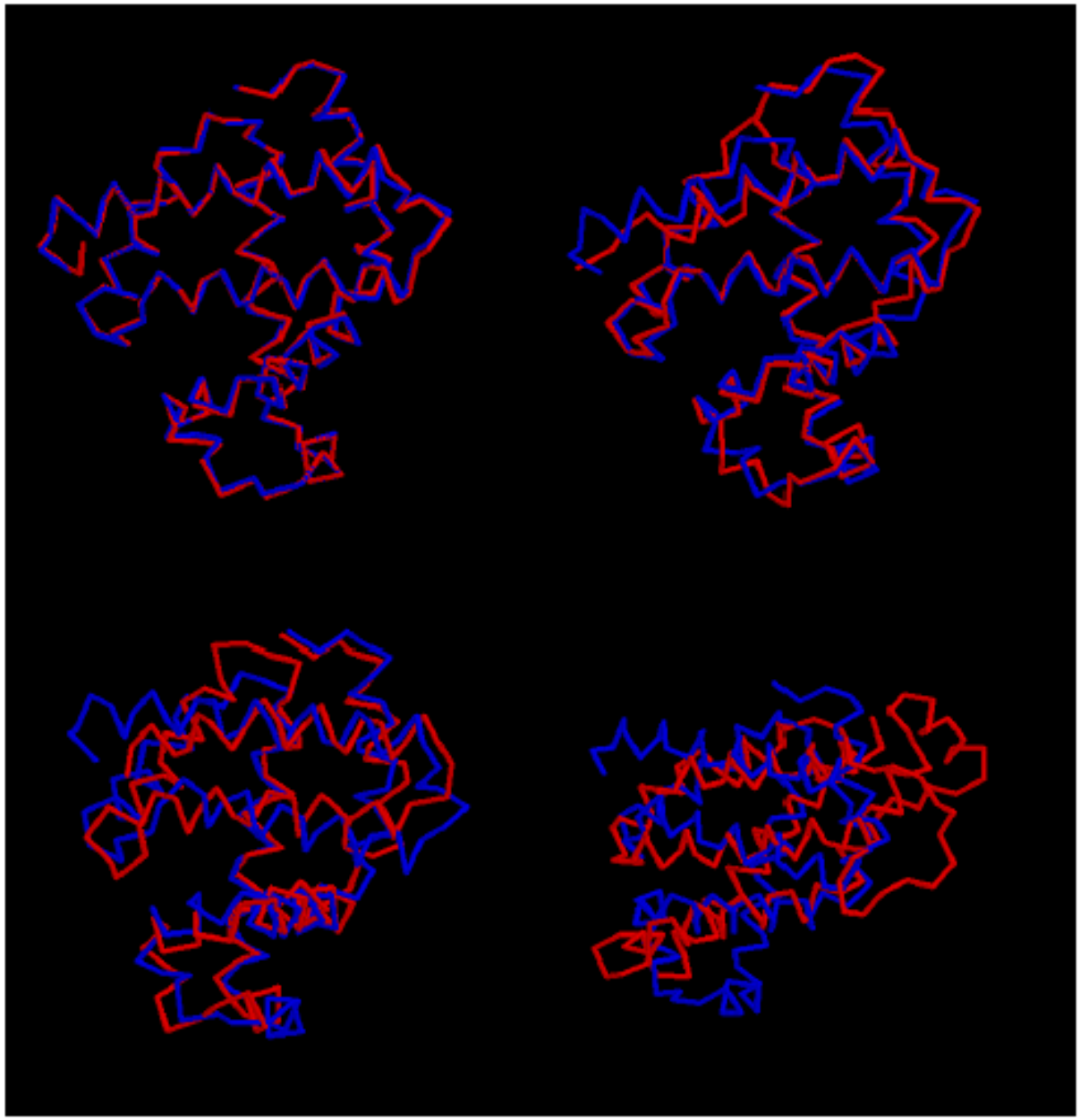
Sequence identity and rmsd of Sperm Whale myoglobin

myoglobin
pig

rmsd = 0.5 Å
id = 86%

haemoglobin
pig

rmsd = 1.5 Å
id = 28%



globin-3
P. piclitum

rmsd = 2.2 Å
id = 18%

phycocyanin
F. diplosiphon

rmsd = 3.3 Å
id = 8%

Summarize

- **Comparison of an unknown sequence to an annotated sequence permits us to infer structural, functional & evolutionary relationships**
- **Wherever possible use the protein sequence since this confers more information**
- **Substitutions, deletions and insertions all occur as part of the natural evolutionary process**
- **Homology implies an evolutionary relationship between two sequences**