Sequence Alignment: I

Dr Safdar Ali
Department of Biological Sciences
Aliah University, Kolkata
Date: 04/04/2020

Why Align Sequences?

- Alignment can be done for DNA/RNA/Protein sequences.
- Each alignment gives different information
- Some of the applications include
  - Phylogenetic analysis
  - Functional domain search
  - Comparative genetics and genomics
  - Sequence polymorphisms and variations

Topics to be covered

- Sequence formats
- Dot Plot
- Global and Local alignment
- Scoring Matrices
- Needleman Wunsch Algorithm
- Smith Waterman Algorithm
- BLAST and FASTA
- HMM
- Multiple sequence alignment

Sequence Formats

- The nucleotide sequence itself can be written in multiple formats like
  - Plain text
  - FASTQ
  - EMBL
  - FASTA
  - Rich sequence
  - GenBank
  - IUPAC
  - Genomatix

https://www.genomatix.de/online_help/help/sequence_formats.html

FASTA Format

- Begins with “>” symbol followed by a description of the sequence
- Sequence begins from the next line
- Can be used for both nucleotide and protein sequence
- Most databases have options for retrieval in FASTA format
- Most softwares accept FASTA format for analysis
- Example

>ABCC gene Human, partial sequence

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCAC
GGCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGA
GCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC

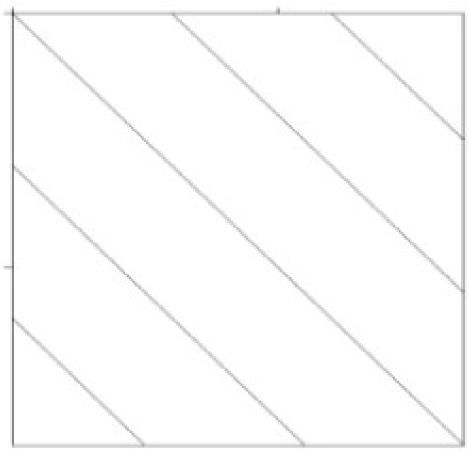
DOT Plot

- Simplest method of alignment
- Two sequences placed on X and Y axis and matched for similarity/difference
- Example

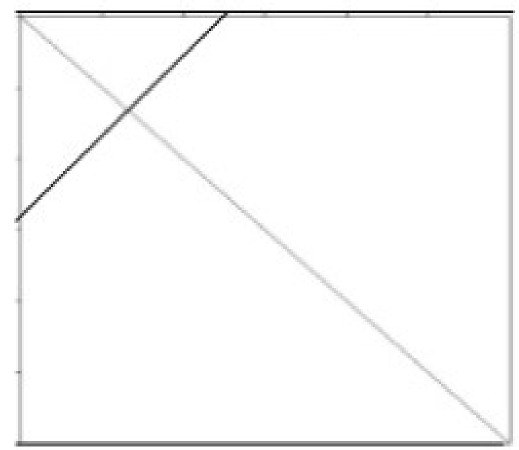
	A	C	A	T	A	G	G
A	X		X		X		
C		X					
C		X					
C		X					
A	X		X		X		
G						X	X
G						X	X

- If the same sequence is written on both the axes or a sequence is aligned with itself then the resultant graph identifies
  - Repeat regions
  - Low complexity regions

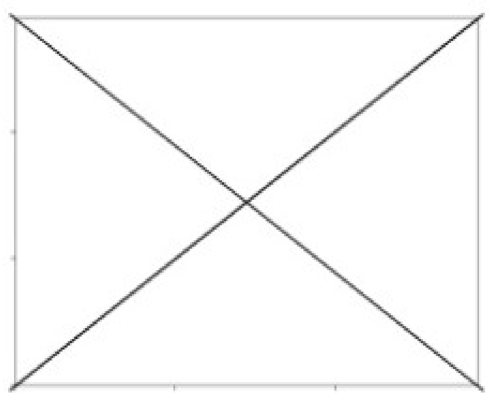
Different Types of DOT Plots



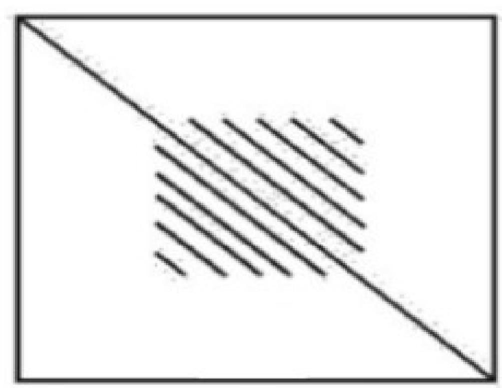
Direct Repeat



Inverted Repeat



Palindromic Sequences



Low Complexity Regions

Common tools for DOT Plot

- GEPARD (https://omictools.com/gepard-tool)
- EMBOSS (https://omictools.com/emboss-tool)
- DNAMAN (https://omictools.com/dnaman-tool)

Global and Local Alignment

Global Alignment	Local Alignment
Matching residues of two sequences along their entire length	Matching two sequences from regions similar to each other
Useful for similar and equal length sequences only	Useful for dissimilar sequences suspected to contain regions of similarity
Based on dynamic programming using Needleman Wunsch algorithm	Based on dynamic programming using Smith Waterman algorithm
Application Comparing homologous genes	Application Finding conserved regions

Scoring Matrices

PAM	BLOSUM
Percent Accepted Mutations	Block Substitution Matrix
Based on global alignment of closely related proteins	Based on local alignments of conserved sequence blocks
PAM 1 matrix is calculated from comparison of sequences with no more than 1% divergence	BLOSUM62 matrix is calculated from comparison of sequences with at most 62% identity in the blocks
Other PAM matrices are extrapolated from PAM1 matrix	All BLOSUM matrices are based on observed alignments, no extrapolation
Higher the number after PAM means more divergent sequences	Higher the number after BLOSUM means more conserved sequences
Used often for phylogenetic analysis	Suitable for local sequence alignments

PAM Matrix	BLOSUM Matrix	Use
PAM100	BLOSUM90	Less divergent
PAM120	BLOSUM80	
PAM160	BLOSUM62	
PAM200	BLOSUM52	
PAM250	BLOSUM45	More divergent

Supratim Choudhuri, in Bioinformatics for Beginners, 2014

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Orzechowski, Patryk & Boryczko, Krzysztof. (2015). Parallel approach for visual clustering of protein databases. Computing and Informatics.

Thank You