Overview

Comparison and alignment of sequences

- Alphabets: DNA, RNA, Protein
- Motivation: Why compare biological sequences?
- Dot plots
- Definition of alignment
- Edit distance
- Alignments and evolution

Symbol alphabets

A *symbol alphabet* Z is a finite set of symbols:

The protein alphabet:

A = {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}

The DNA alphabet:

A = {A, T, C, G}

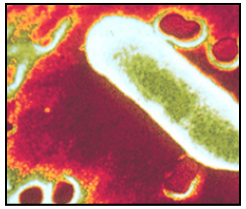
The RNA alphabet:

A = {A, U, C, G}

When we consider alignments with gaps, we will also include the gap or indel symbol indicated by ”-” in the alphabet.

Why compare and align sequences?

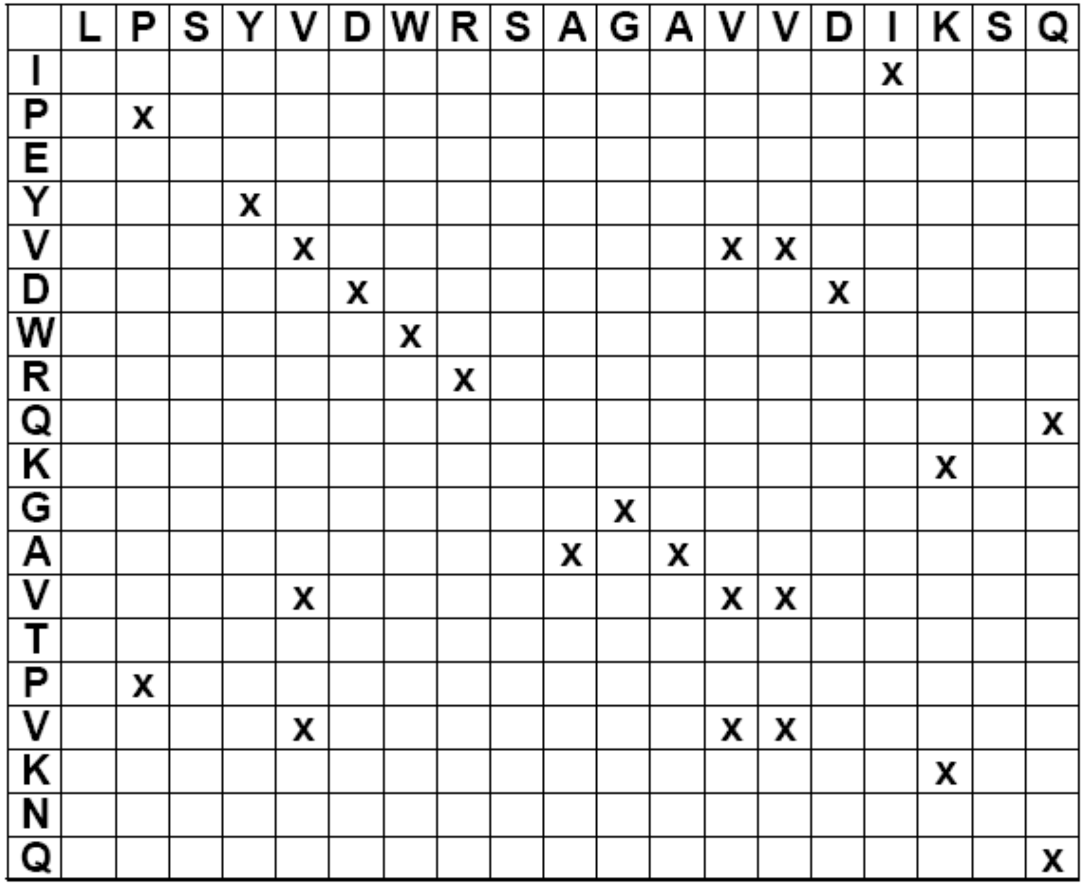
- Why align sequences?
  - An alignment is usually the best way of comparing sequences
  - It may indicate whether the sequences are homologous or not (i.e. they have a common evolutionary ancestor)
  - Alignments will give us a evolutionary perspective of the sequences
  - Information may be derived from genes of known function to genes of unknown function based on sequence homology
  - Biological hypotheses may be created



- Comparing sequences is fundamental in sequence analysis and bioinformatics

Dot plots

- A dot plot is a simple way to compare sequences.
- It gives a visual impression of the similarity between two sequences and may reveal interesting information.



	L	P	S	Y	V	D	W	R	S	A	G	A	V	V	D	I	K	S	Q
I																x			
P		x																	
E																			
Y				x															
V					x								x	x					
D						x									x				
W							x												
R								x											
Q																			x
K																	x		
G											x								
A										x		x							
V					x								x	x					
T																			
P		x																	
V					x								x	x					
K																	x		
N																			
Q																			x

Dot plot basics

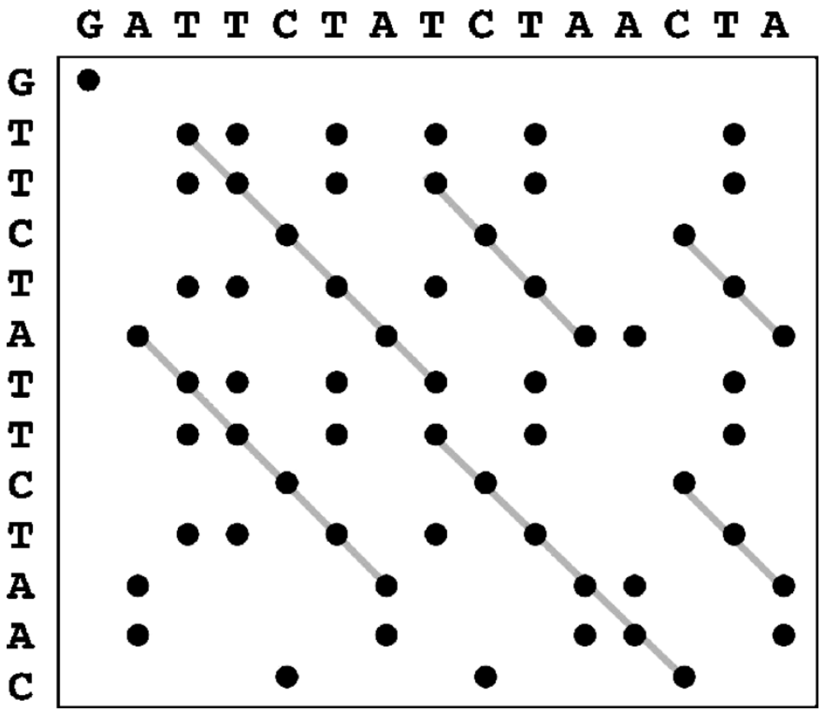


Figure 3.3: Example of comparing two sequences using dot plots. Lines linking the dots in diagonals indicate sequence alignment. Diagonal lines above or below the main diagonal represent internal repeats of either sequence.

- The gray lines indicate windows of at least 3 consecutive matches on the same diagonal

Dot plot example: identical sequences

```
IPEYVDWRQKGAVTPVKNQG
IPEYVDWRQKGAVTPVKNQG
```

	I	P	E	Y	V	D	W	R	Q	K	G	A	V	T	P	V	K	N	Q	G
I	X																			
P		X													X					
E			X																	
Y				X																
V					X								X			X				
D						X														
W							X													
R								X												
Q									X										X	
K										X							X			
G											X									X
A												X								
V					X								X			X				
T														X						
P		X													X					
V					X								X			X				
K										X							X			
N																		X		
Q									X										X	
G											X									X

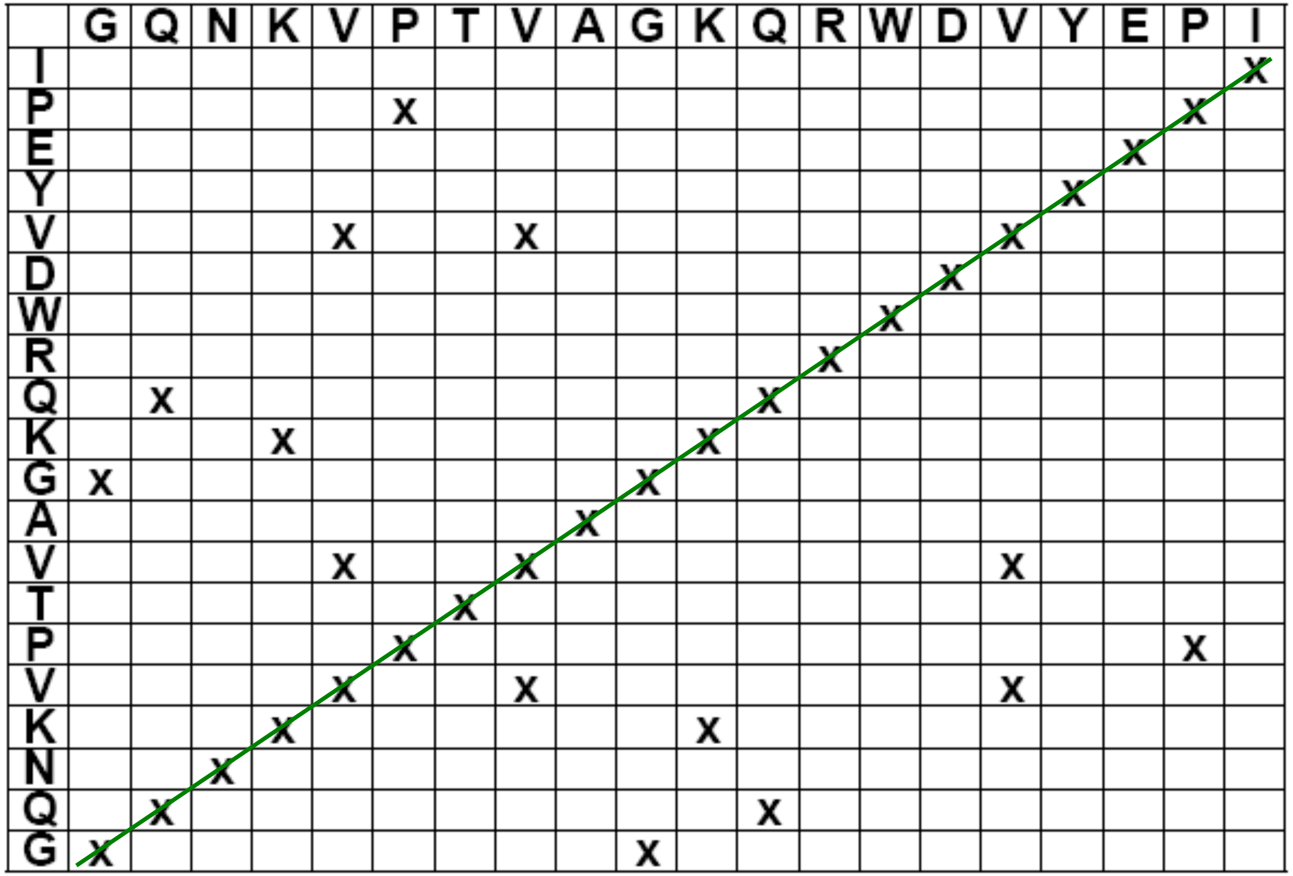
Dot plot example: several gaps

```
IPE---WRQKGAVTPV--QG
IPEYVDWRQK---TPVKNQG
```

	I	P	E	W	R	Q	K	G	A	V	T	P	V	Q	G
I	X														
P		X										X			
E			X												
Y															
V										X			X		
D															
W				X											
R					X										
Q						X								X	
K							X								
T											X				
P		X										X			
V										X			X		
K							X								
N															
Q						X								X	
G								X							X

Dot plot example: inverted segments

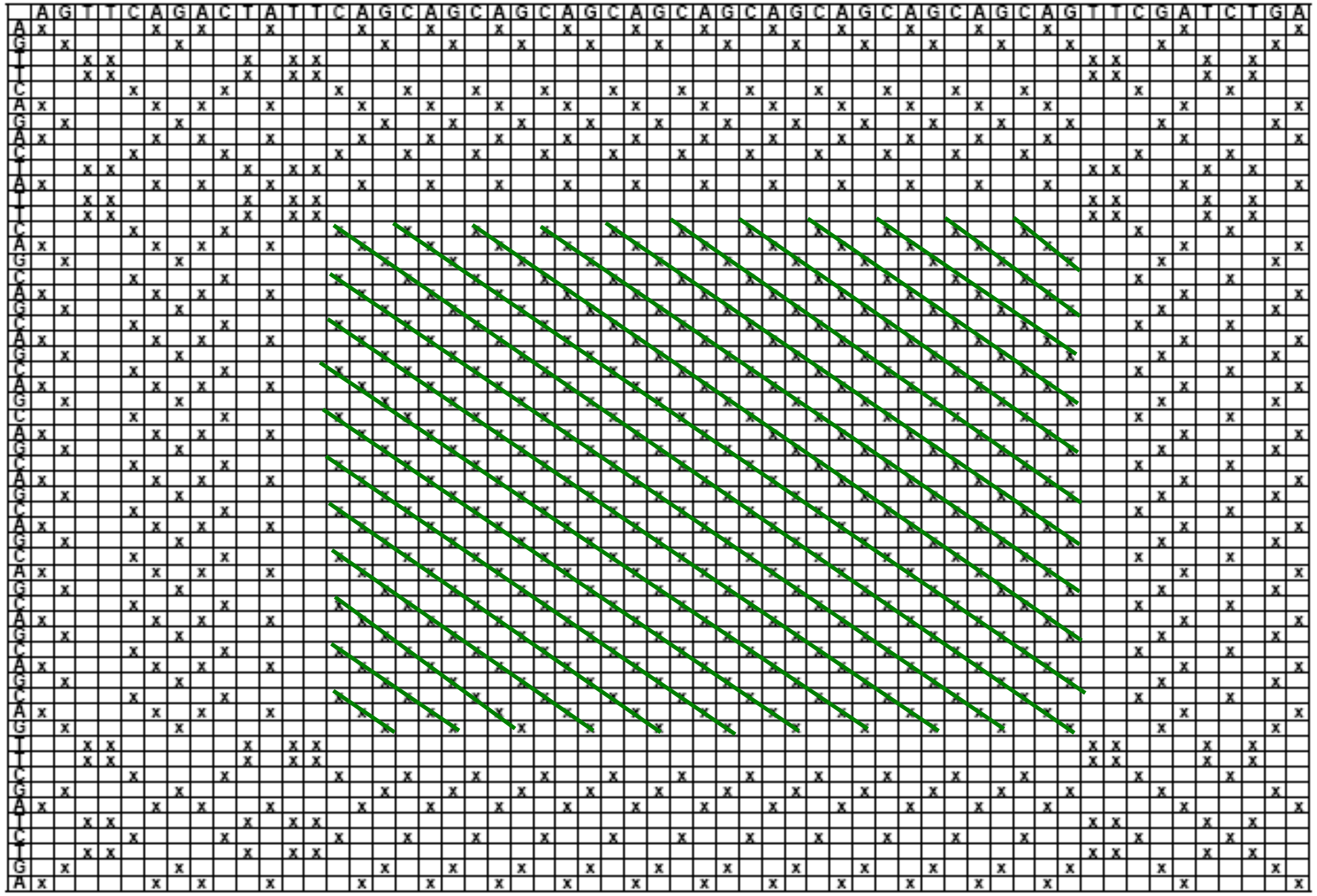
```
GQNKVPTVAGKQRWDVYEPI
IPEYVDWRQKGAVTPVKNQG
```



	G	Q	N	K	V	P	T	V	A	G	K	Q	R	W	D	V	Y	E	P	I
I																				X
P						X													X	
E																		X		
Y																	X			
V					X			X								X				
D															X					
W														X						
R													X							
Q		X										X								
K				X							X									
G	X									X										
A									X											
V					X			X								X				
T							X													
P						X													X	
V					X			X								X				
K				X							X									
N			X																	
Q		X										X								
G	X									X										

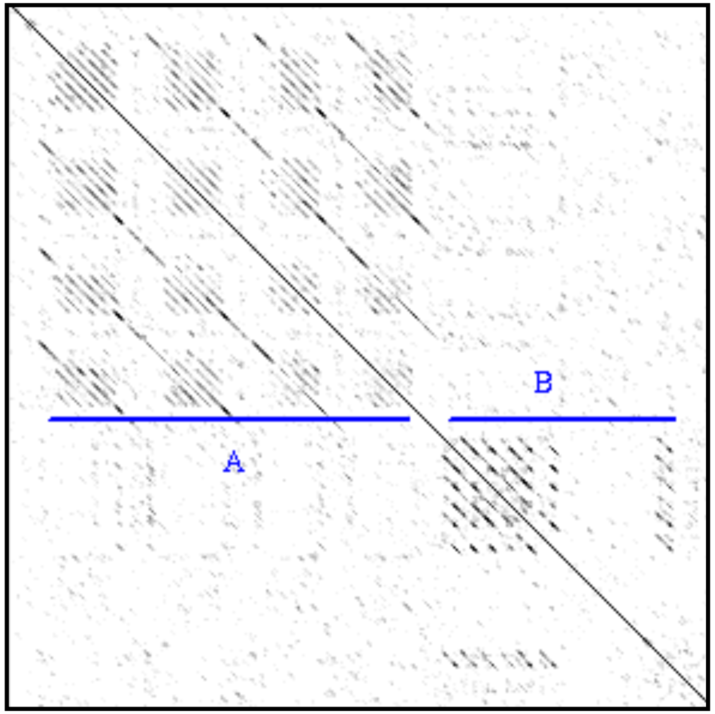
Dot plot example: repetitive sequences

```
AGTTCAGACTATTCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTCGATCTGA
AGTTCAGACTATTCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTCGATCTGA
```

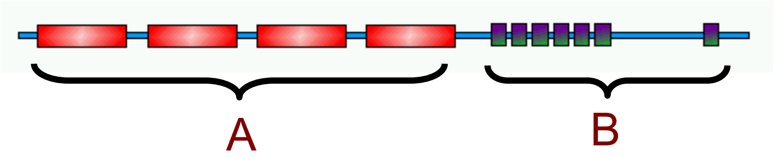


Example using Dotlet

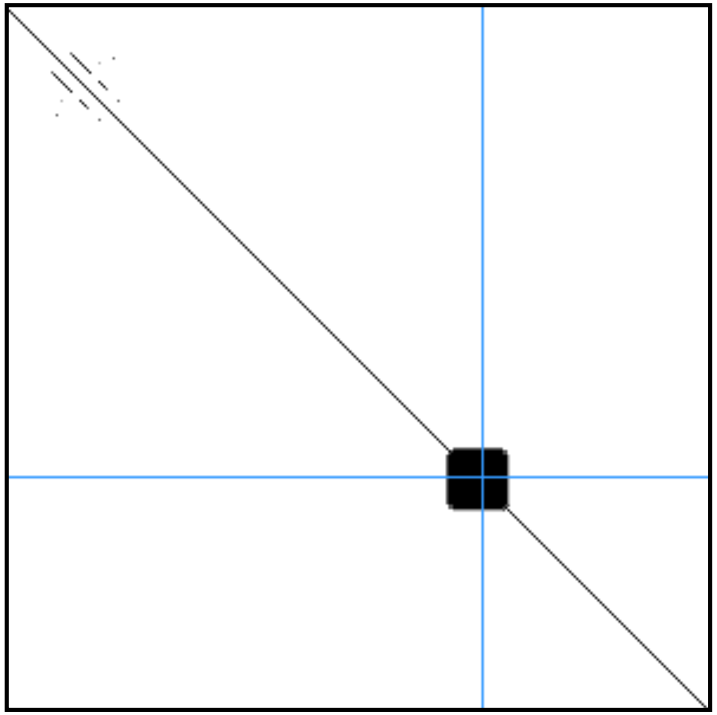
Drosophila melanogaster SLIT protein plotted against itself:



A series of repeated regions is shown.



Plasmodium falciparum protein plotted against itself:



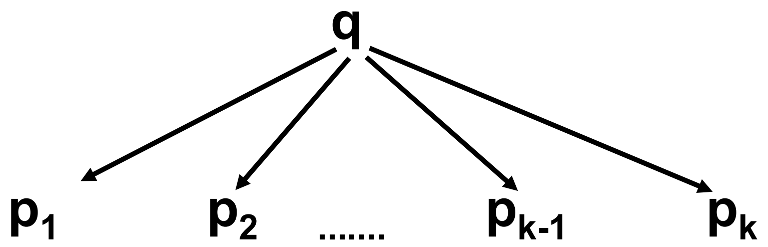
An area with many repeats is shown – a ”low complexity region”.

Dotlet: http://myhits.isb-sib.ch/cgi-bin/dotlet or https://dotlet.vital-it.ch

Sequence alignment basics

Starting point:

- Sequences p_1, p_2, ..., p_k (DNA, RNA, or protein)
- We have a hypothesis that the sequences (or parts of them) have developed during generations from a common unknown seqence q through a series of mutations, insertions and deletions in DNA



Goal:

- Identify the most likely residue by residue correspondance between the sequences. Residues may be amino acids or nucleotides.
- Determine whether the similarity between the sequences is significantly better than expected by chance

Edit distance

- The edit distance indicates the difference or dissimilarity between two strings
- The edit distance between two strings or sequences A and B is the total number of operations needed to transform the first string A into the second string B using the following operations:
  - Substitute one symbol with another
  - Delete a symbol
  - Insert a symbol
- The edit distance is also called the Levenshtein distance
- A low edit distance indicates that the two strings are similar
- Example (kitten -> sitting): 3 operations
  - kitten -> sitten (substitute k with s)
  - sitten -> sittin (substitute e with i)
  - sittin -> sitting (insert g at end)

Alignments and evolution

- We would like to identify the evolutionary relation between two sequences. What has happened during evolution?
- The sequences may be nucleotide sequences (DNA, RNA) or amino acid sequences (protein).
- We consider four different possible fates for a symbol during evolution:
  - No change of the symbol
  - Substitution of symbol a to symbol b (a → b)
  - Deletion of symbol (a →)
  - Insertion of symbol (→ a)

- Indel = insertion or deletion (when the direction of evolution is unknown)

Alignments and evolution: Example 1

Example with known history:
q= GLISVT, d=GIVT, h=GLVST

History of events:

`GLISVT`

`GLIS-T` (deletion of V)

`GLVS-T` (substitution of I to V)

`GLV--T` (deletion of S)

`GIV--T` (substitution of L to I)

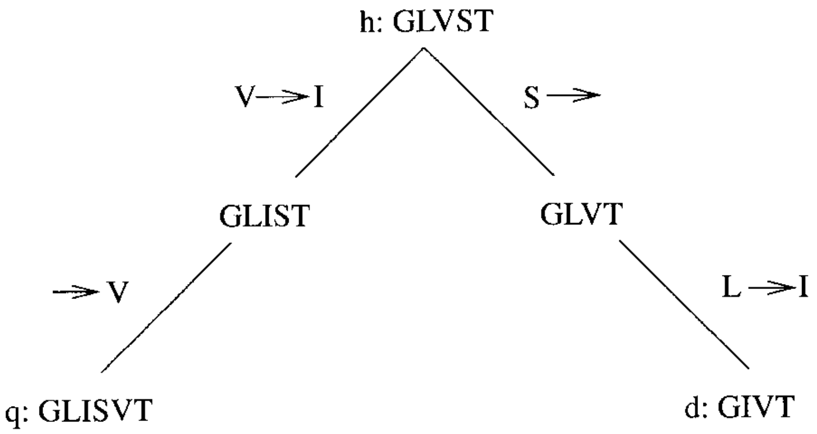


Figure 1.1 An evolution from h to q and d.

Alignments and evolution: Example 2

Example with unknown history
q= GLISVT, d=GIVT

Operations to consider (examples):
I ↔ V ; L ↔ I ; V ↔ - ; S ↔ -

Possible history of events:
`GLISVT`
`GLI-VT` (deletion av S)
`G-I-VT` (deletion av L)

- The alignment may be different depending on what is known about the evolution of the sequences.
- In lack of additional information we choose the simplest explanation of the evolution (Occam's razor).