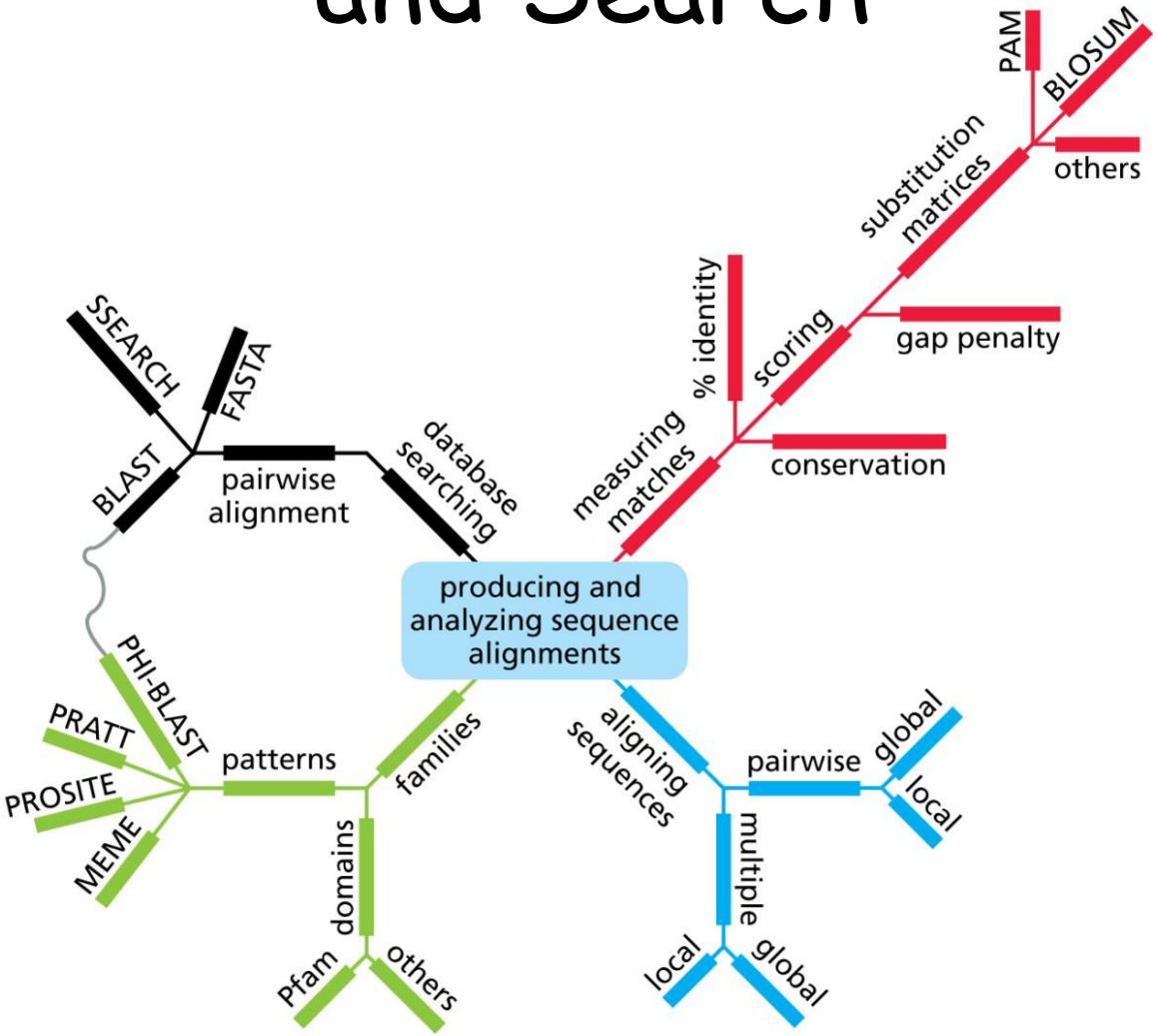
Sequence Similarity/Alignment

by Ahmet Sacan
http://sacan.biomed.drexel.edu

Images are from "Understanding Bioinformatics" and Wikipedia.

Sequence Similarity/Alignment, and Search

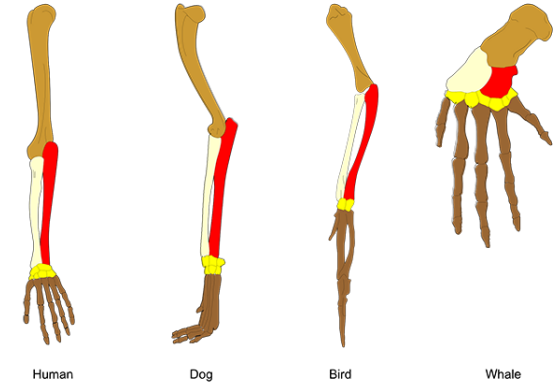


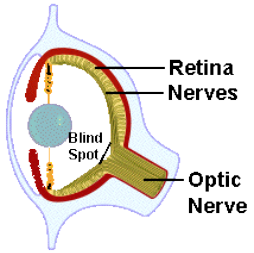
Why align sequences?

- Are two sequences related?
- What are the corresponding residues?
- Find unique microarray probes.
- Shotgun assembly
- Find related sequences in other species
- Find motifs important for function

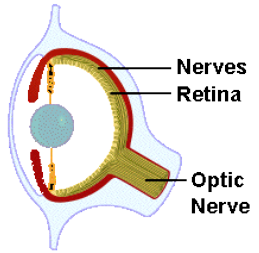
Homology

- Divergent Evolution
  - Shared ancestry
  - Homologous ~~ Similar
- Convergent Evolution
  - acquire same structure or function independently (Analogous)





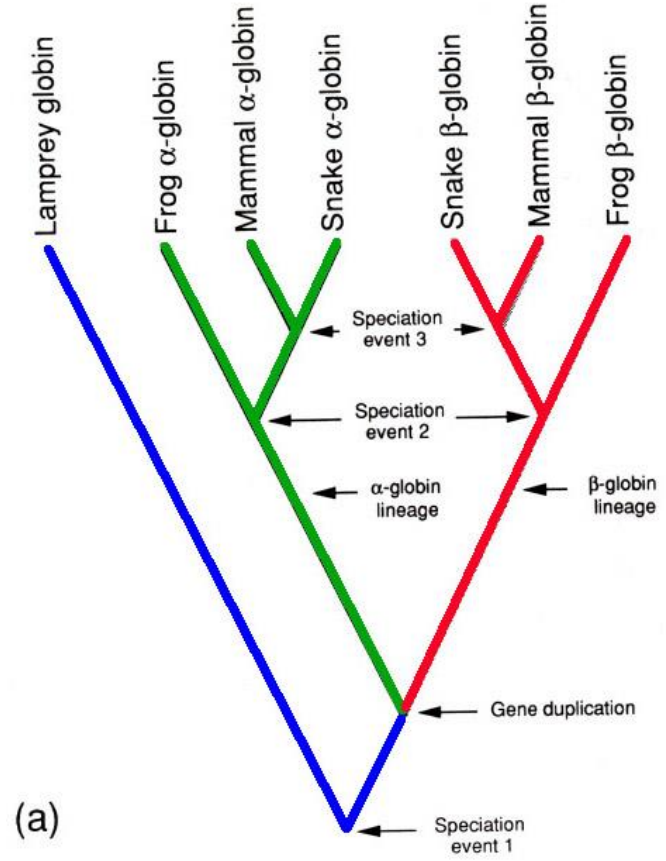
Vertebrate



Cephalopod

Homology

- Orthologs
  - speciation
- Paralogs
  - duplication
- Xenologs
  - Horizontal gene transfer



Sources of variation

- Substitution
  - Replication error
  - Chemical reaction
- Insertions or deletions (indels)
- Duplication
  - Entire gene
  - Part of a gene: Domain duplication, Exon shuffling
  - Entire chromosome: polysomy
  - Part of a chromosome: partial polysomy
  - Entire genome: polyploidy

Sequence similarity and alignment

```
THATSEQUENCE      THISISASEQUENCE   THEFREQUENCE

A:1,T:2,E:3,...   A:1,T:1,E:3,...   A:0,T:1,E:4,..
```

RAIL SAFETY

```
TH----ATSEQUENCE
||    | ||||||||
THISISA-SEQUENCE
```

FAIRY TALES

```
THATSEQUENCE
||   |||||||
THEFREQUENCE
```

Sequence alignment

- Ancestor

```
WHATSEQUENCE
```

- Mutations

```
THATSEQUENCE
||  ||||||||
THISSEQUENCE
```

- Insertions

```
THATSEQUENCE
||      |  |
THISISASEQUENCE

TH----ATSEQUENCE
||    | ||||||||
THISISA-SEQUENCE
```

- Alternative alignments

```
AGGCTAGTT-
AGCGAAGTTT
match,mismatch,gap: 6,3,1

AGGCTA-GTT-
AG-CGAAGTTT
match,mismatch,gap: 7,1,3

AGGC-TA-GTT-
AG-CG-AAGTTT
match,mismatch,gap: 7,0,5
```

Evaluating alignments

- Percent identity

```
TH----ATSEQUENCE
||    | ||||||||
THISISA-SEQUENCE
```

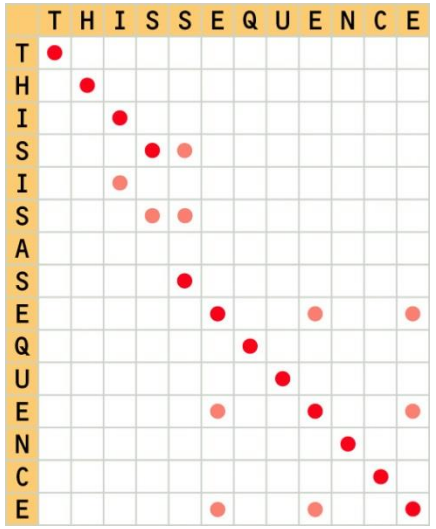
11/16≈69%

```
THAT---SEQUENCE
||     ||||||||
THISISASEQUENCE
```

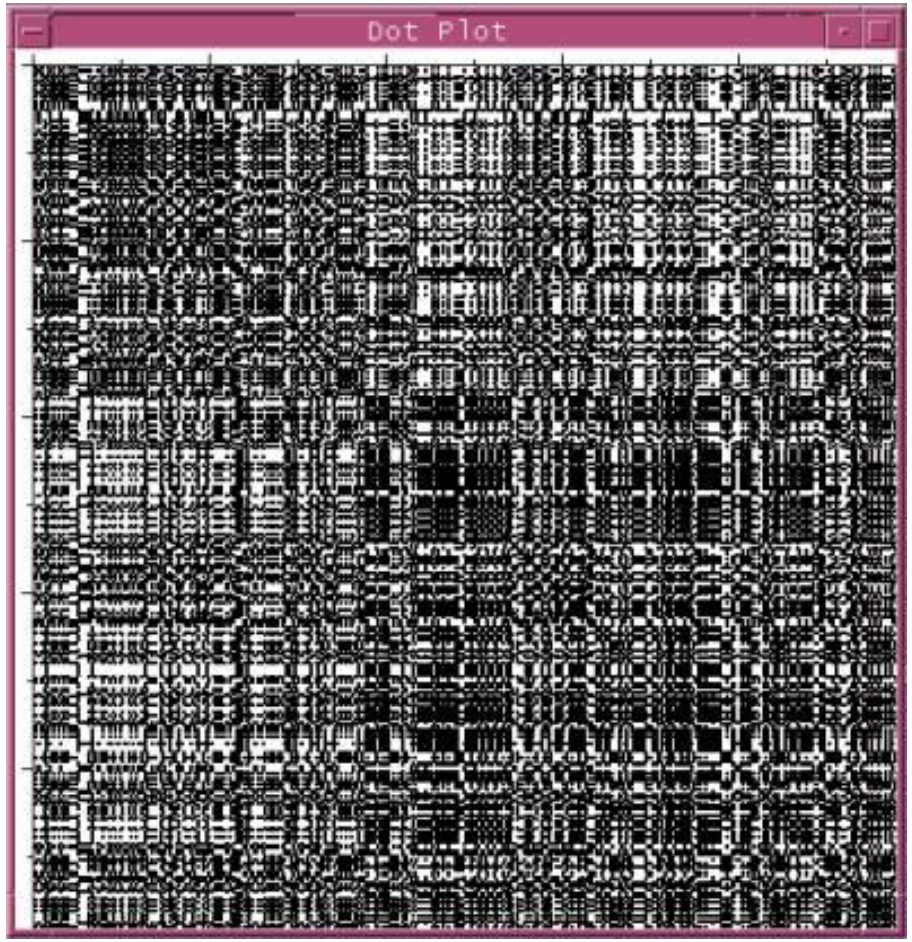
10/15 ≈ 67%

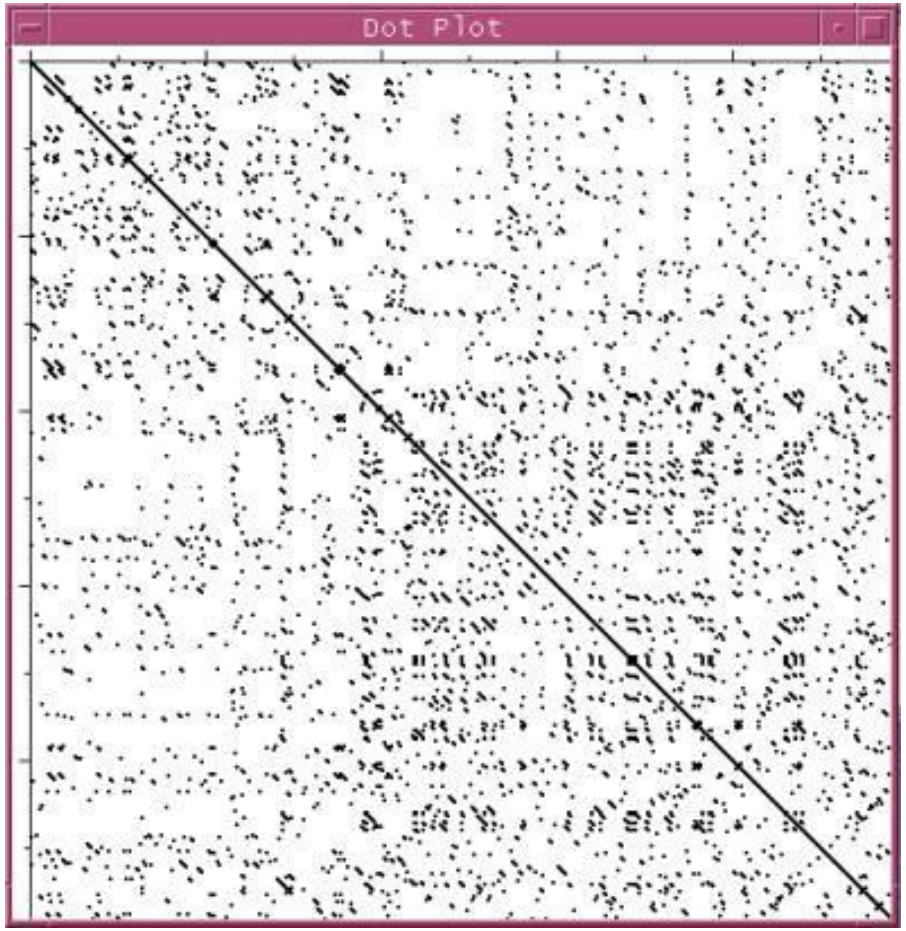
- Scoring matrix

- Dot-plot (visual)



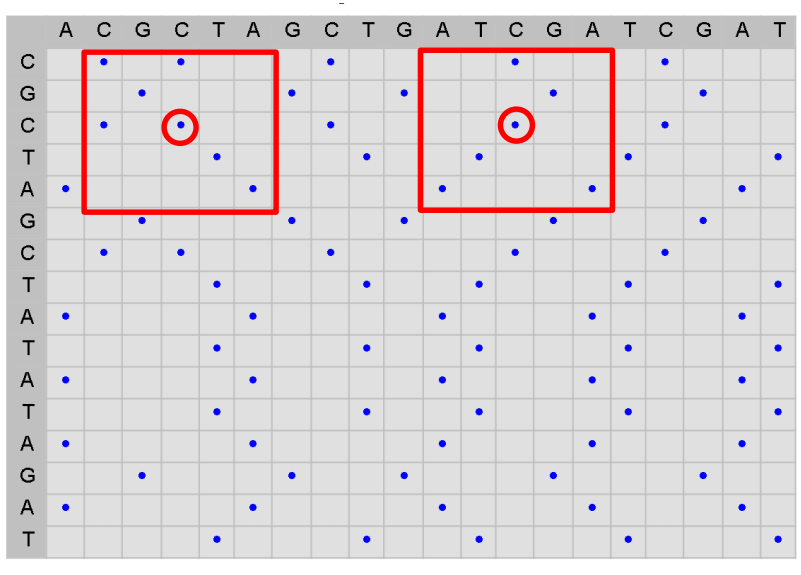
Apply filter to remove noise

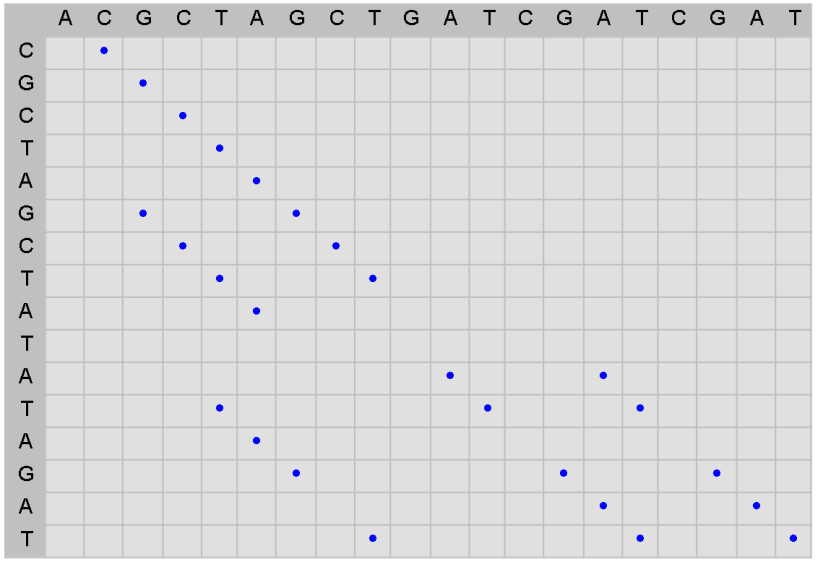




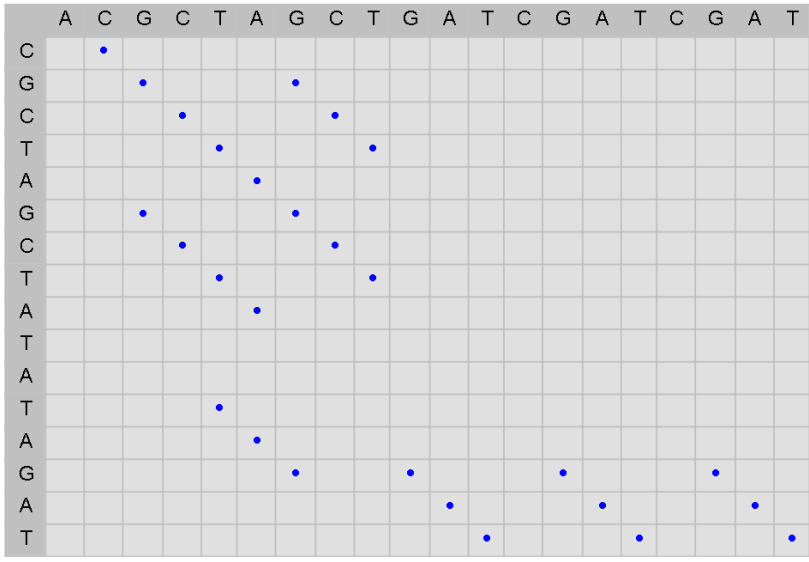
http://www.quretec.com/u/vilo/edu/2002-03/Tekstialgoritmid_I/Loengud/Loeng3_Edit_Distance/bcorum_copy/seq_align2.htm

Apply filter to remove noise





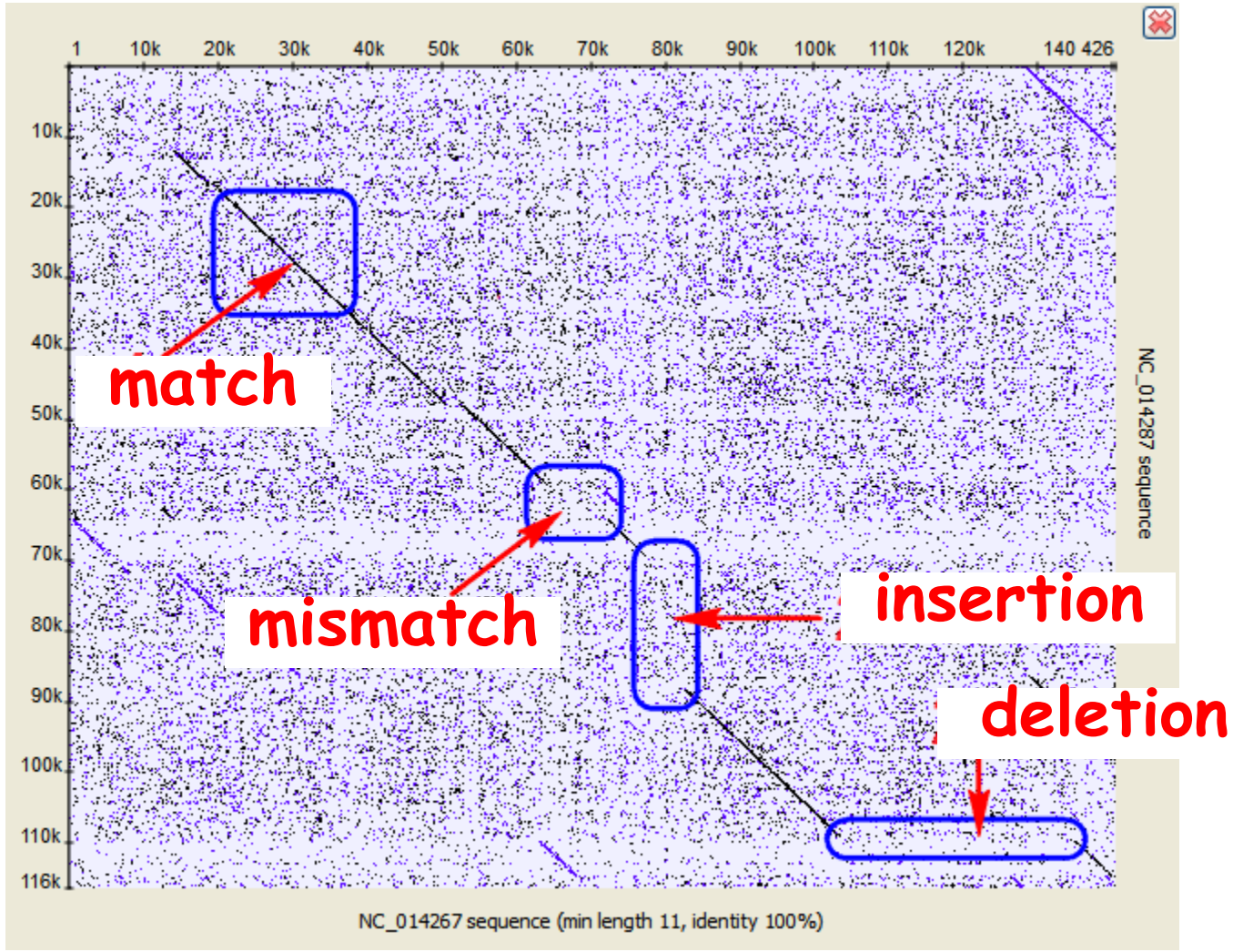
Window Size=5, match>=60%

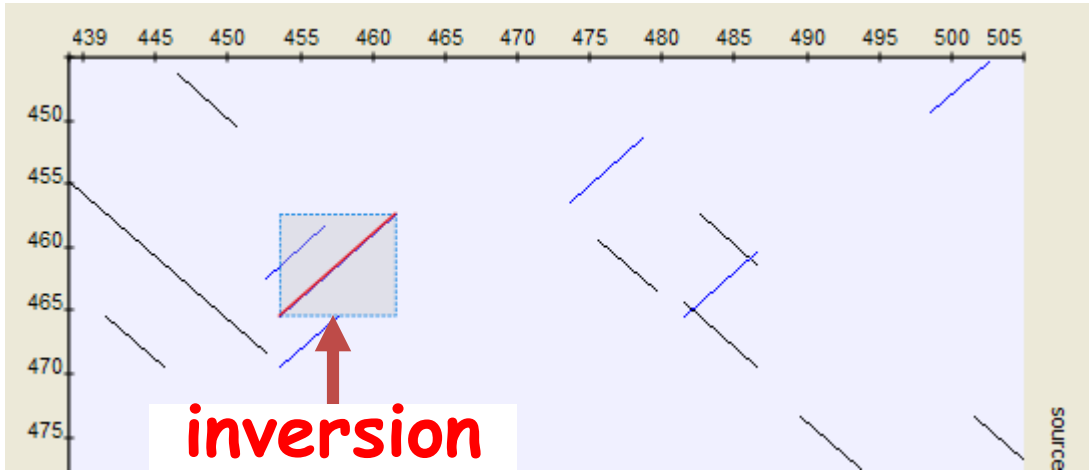


Window Size=3, match>=99%
"3-letter word filter"

http://www.quretec.com/u/vilo/edu/2002-03/Tekstialgoritmid_I/Loengud/Loeng3_Edit_Distance/bcorum_copy/seq_align2.htm

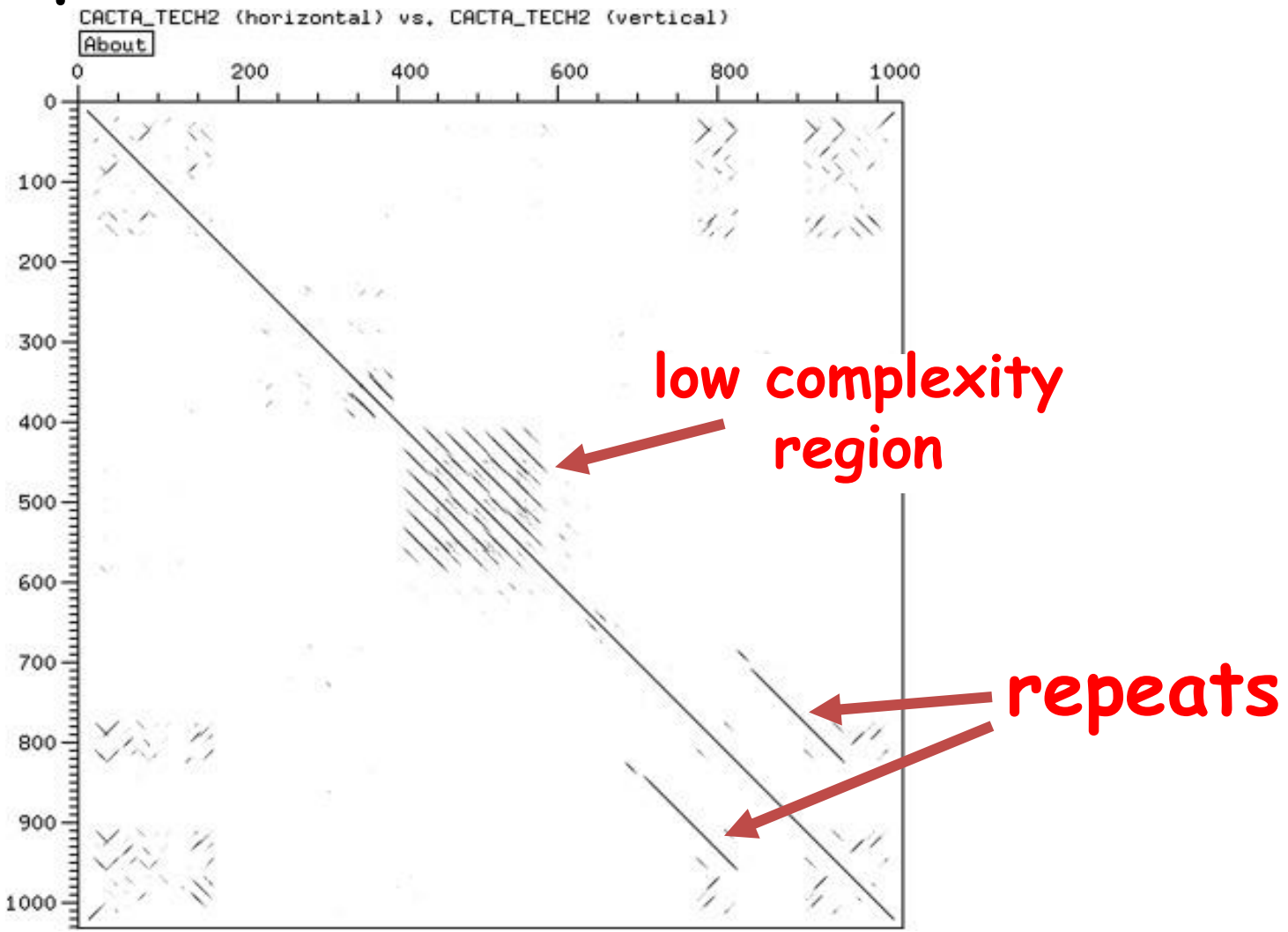
Interpreting Dot Plots





https://ugene.net/wiki/pages/viewpage.action?pageId=4227426

Self-comparison Dot-plots can identify repeats and inversions



https://www.researchgate.net/figure/Dot-plot-analysis-of-the-CACTA-transposon-Tech2-identified-in-Chrysanthemoides2-Chry2_fig5_259809378

Genuine matches do not have to be identical

- Some amino acids are more similar in hydrophobicity, charge, size than others.

```
THAT---SEQUENCE
||      ||||||||
THISISASEQUENCE
```

  - Alanine & Isoleucine are both hydrophobic
  - Threonine & Serine both have an -OH group on their side chain and are polar

Substitution matrices assign scores to aligned residues

- BLOSUM-62 matrix

```
T H  A T S E Q U E N C E
| |  | | | | | | | | | |
T H  I S S E Q U E N C E
5 8 -1 1 4 5 5 0 5 6 9 5
total: 52
```

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

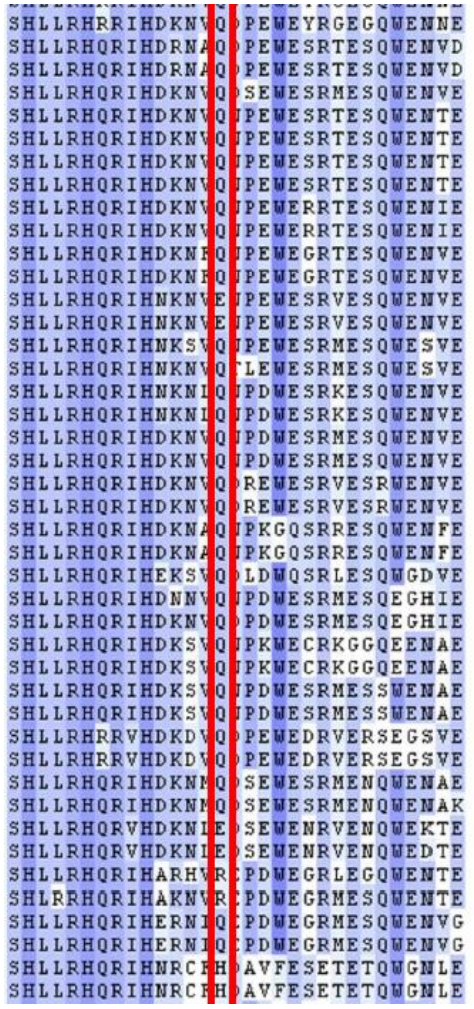
BLOSUM (BLOck Substitution Matrix)

- Odds Ratio:

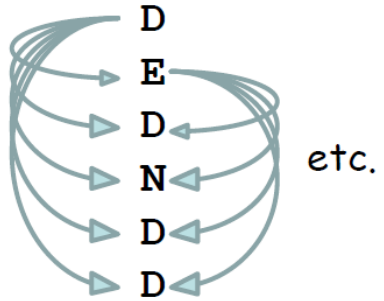
$$\frac{p_{ij}}{q_i * q_j}$$

- BLOSUM score:

$$S_{ij} = 2\log_2 \frac{p_{ij}}{q_i * q_j}$$



http://elbo.gs.washington.edu/courses/GS_559_11_wi/slides/4A-Sequence_comparison-ScoreMatrices.pdf



6 D-D pairs
4 D-E pairs
4 D-N pairs
1 E-N pair

BLOSUM example

- Aligned Sequences:

VVAD
AVAD
DVAD
DAAA

- $q_A = \frac{7}{16} = 0.44$, $q_D = \frac{5}{16} = 0.31$, …
- $N_{AA} = 12$, $N_{AD} = 5$, …
- $p_{AA} = \frac{12}{48} = 0.25$, $p_{AD} = \frac{5}{48} = 0.10$ …
- $S_{AA} = 2\log_2 \frac{0.25}{0.44*0.44} = 0.74$
- $S_{AD} = 2\log_2 \frac{0.10}{0.44*0.31} = -0.9$

Interpretation of BLOSUM alignment score

- Alignment score has a probabilistic interpretation.

```
T H  A T S E Q U E N C E
| |  | | | | | | | | | |
T H  I S S E Q U E N C E
5 8 -1 1 4 5 5 0 5 6 9 5
total: 52
```

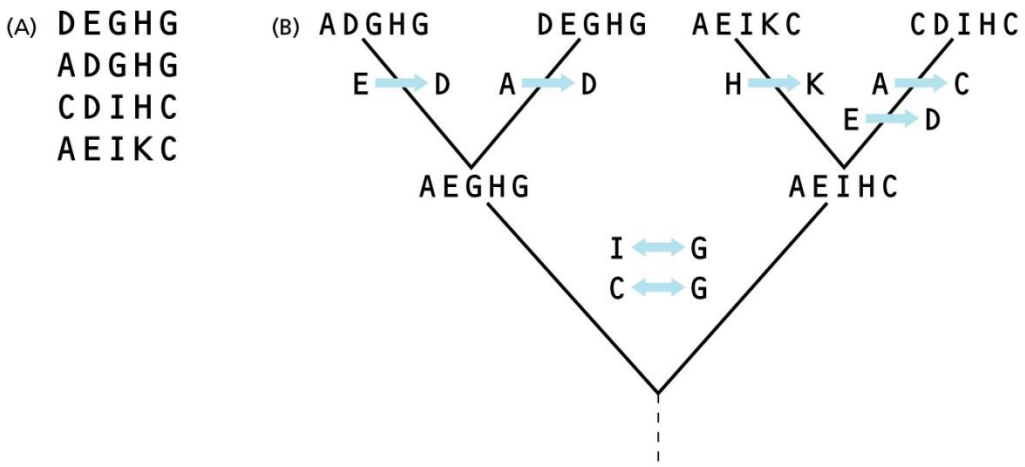
- An alignment with score of 52 is 2^{26} times more likely to be seen in a real alignment than expected from a random alignment.

BLOSUM-62

- BLOSUM sequences are clustered at different % identity levels. The clustering affects how the pairs are counted.

- BLOSUM-80
  - highly similar sequences
- BLOSUM-45
  - highly divergent sequences
- BLOSUM-62
  - most common

PAM (Point Accepted Mutation) matrices trace evolutionary origins



(C)

	A	C	D	E	G	H	I	K
A		1	1					
C	1				1			
D	1			2				
E			2					
G		1					1	
H								1
I					1			
K						1		

$$S_{i,j} = \log \frac{p_i \cdot M_{i,j}}{p_i \cdot p_j} = \log \frac{M_{i,j}}{p_j} = \log \frac{observed\ frequency}{expected\ frequency}$$

Gap Penalty

```
THAT---SEQUENCE
||     ||||||||
THISISASEQUENCE
```

```
TH-A-T-SEQUENCE
||     ||||||||
THISISASEQUENCE
```

- Linear gap penalty

$$gappenalty = -n_{gaps} * E$$

- Affine gap penalty
  - Insertions/deletions tend to be several residues long rather than just a single residue long

$$gappenalty = -n_{gapopen} * I - n_{gapextend} * E$$

Global vs. Local alignment

- Global alignment aligns sequences in their entirety
- Local alignment finds parts of the sequences that are most similar

Global Alignment:

```
HEAGAWGHEEAHGEGAE
--|-||-|-||--|-||
--P-AW-H-EA--E-HE
```

Local Alignment:

```
AWGHEEAH
||-|||||
AW-HEAEH
```

Local alignment is useful especially for multi-domain proteins