Sequence comparison and alignment

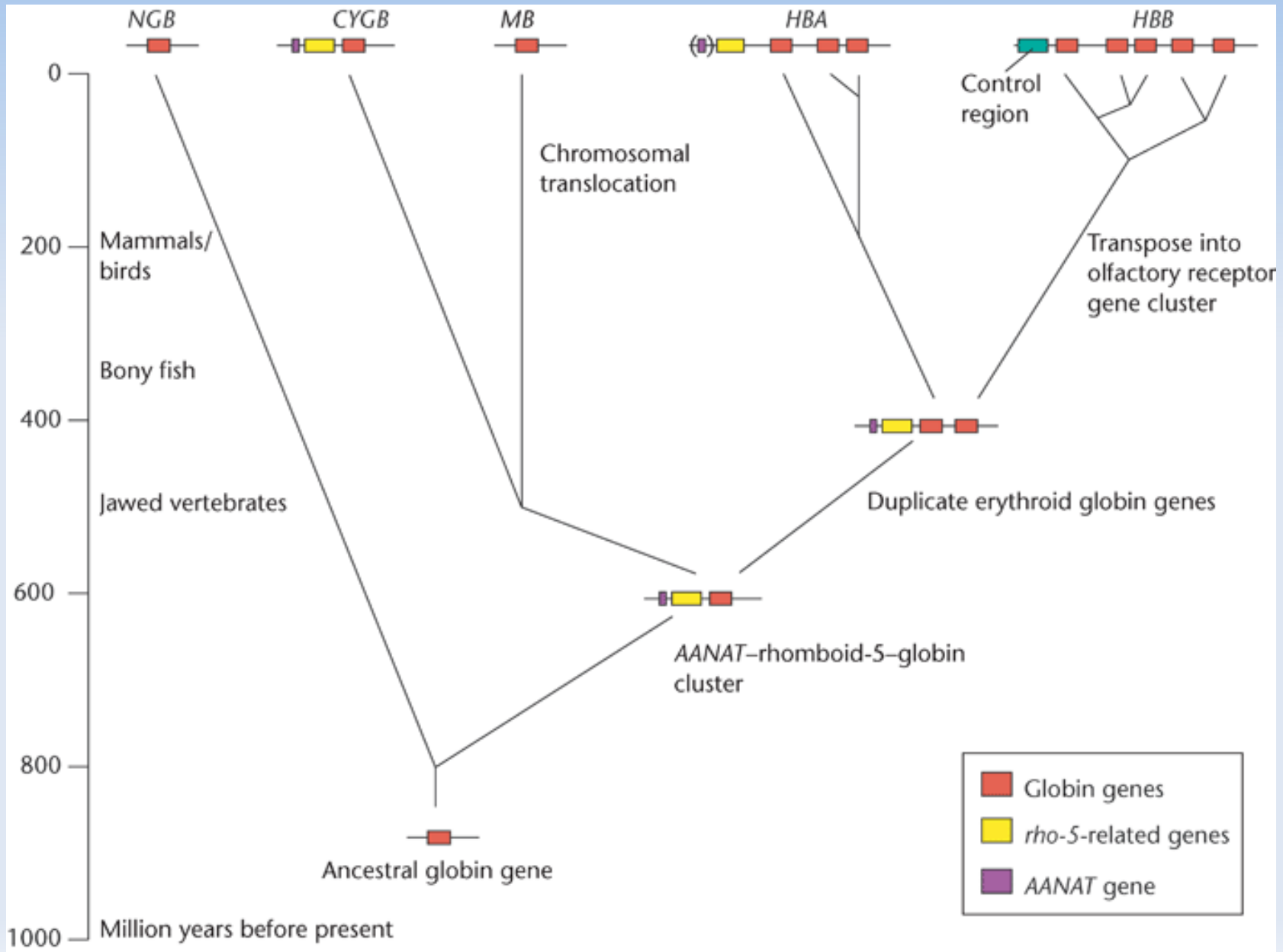
Eszter Ari
Dept. of Genetics, ELTE
arieszter@gmail.com

genetics.elte.hu
username: ***genetika2017***
password: ***genetika2017***

What we will talk about?

- Sequence similarity, alignment
- Number of possible alignments
- Pairwise sequence alignments
  - Pairwise comparisons: „Dot-plot”
  - Scoring systems, substitution matrices: PAM, BLOSUM
  - Optimal alignment
  - Global and local alignments
    - Dynamic programming algorithms:
      Needleman - Wunsch, Smith - Waterman
- Multiple sequence alignment

The evolution of vertebrate globin genes



Hardison, R. C. 2008. Globin Genes: Evolution. eLS

Differences between homolog sequences

10 million years ago

```
Seq:   ATCTCGTTTA
```

5 million years ago: gene duplication

```
SeqA:  ATCTCGTTTA
SeqB:  ATCTCGTTTA
```

5 million years - today: independent mutations

```
SeqA:  AATCTCGT(T/C)TA
SeqB:  ATCGTCGTTT(A/T)
```

today:

```
SeqA:  AATCTCTCTA
SeqB:  ATCGTCGTTTT
```

Alignment that reflect the evolution:

```
SeqA:  AATC-TC-TCTA
SeqB:  -ATCGTCGTTTT
```

Human Alpha and Beta hemoglobin

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
>sp|P68871|HBB_HUMAN Hemoglobin subunit beta
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

The goals of sequence alignment

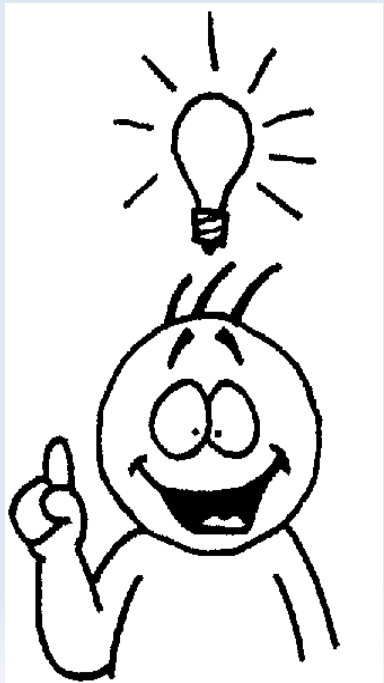
- Direct goal:
  - Insert gaps between the residues so that identical or similar characters are aligned in successive sites.

```
seq1   T A A G C T T C
seq2   T G A T G C G T C
```

```
seq1   T - A A G C T T C
seq2   T G A T G C G T C
```

The goals of sequence alignment

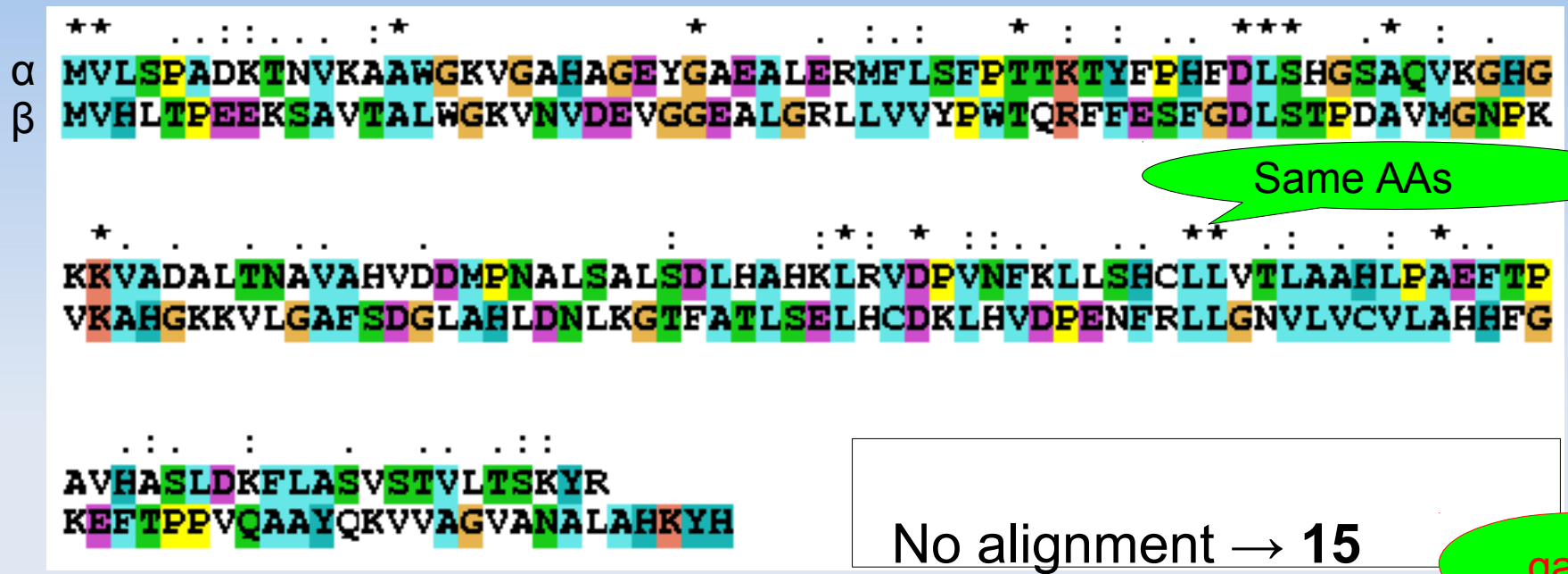
- A lot of bioinformatic tasks contains sequence alignmet → the result is dependent on the quality of alignmnet
- Ideal alignment: reflect to evolution (substitutions, in/dels)
- Indirect goals:
  - Similarity searches in sequence databases
  - Phylogenetic and populationgenetic analyses
  - Prediction of structure and function
  - Gene prediction and annotation
  - Comparison of whole genome sequences
  - Pattern search: i.e promoter regions
  - Genome sequencing: assembly, mapping

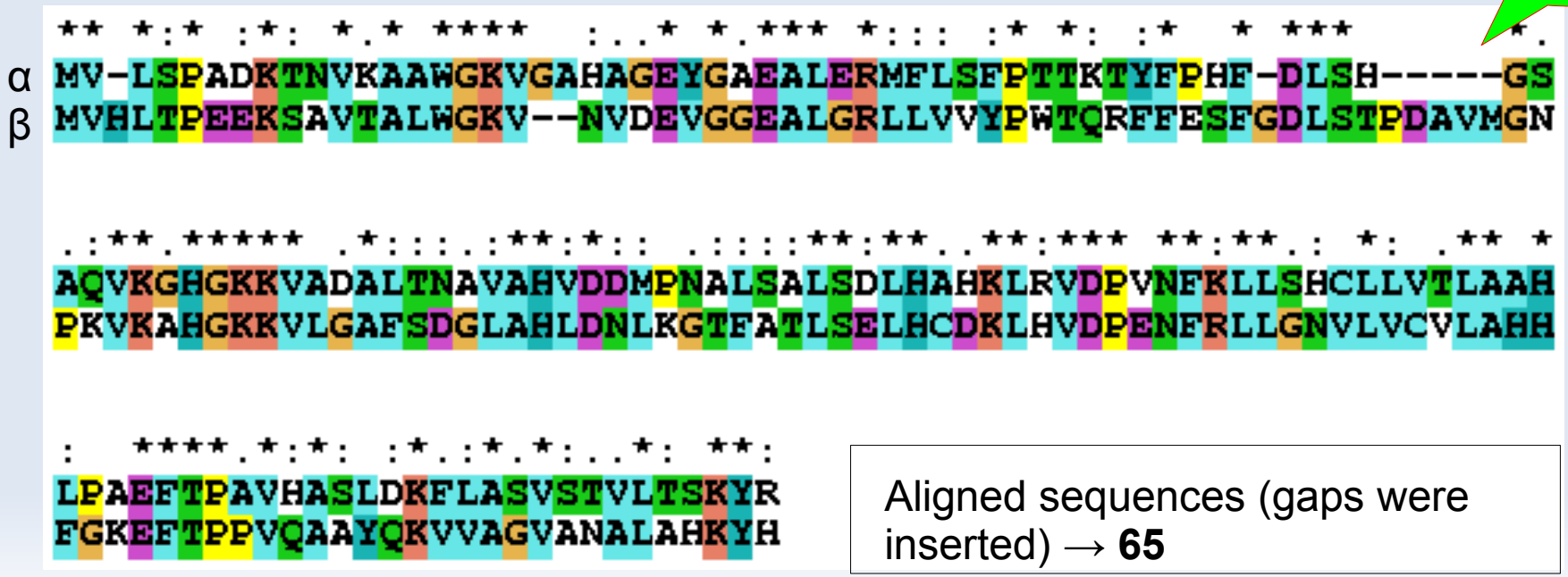


The two approaches of sequence analysis

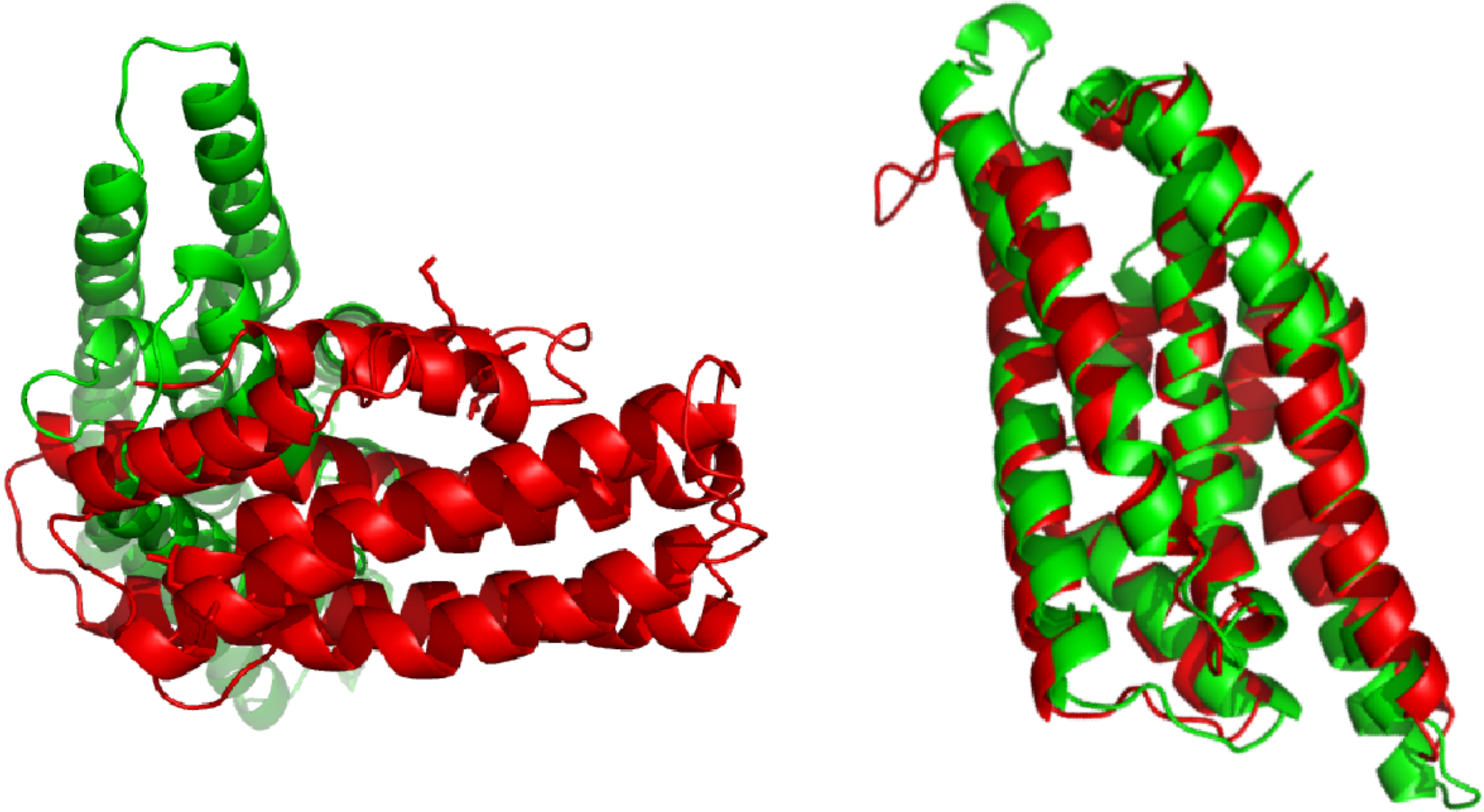
- Based on sequence similarity:
  - Structure and function prediction
  - Similar sequences → similar structures → similar functions
    - In most cases this is valid, except when not
- *Ab initio* prediction:
  - DNA sequence → protein sequence prediction → protein structure prediction → protein function prediction
  - So far very limited application...

Why do we need alignment softwares?





Structural alignmnet

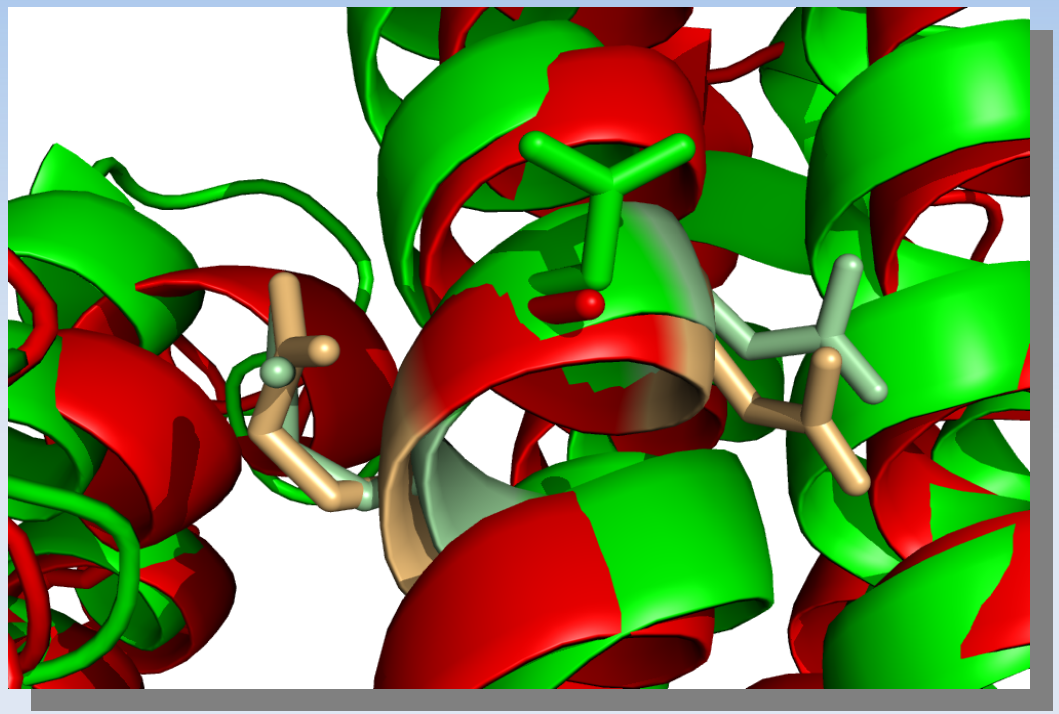


Before structural alignmnet

After structural alignment

Bacteria toxin: 1ji6 and insecticide protein: 1i5p

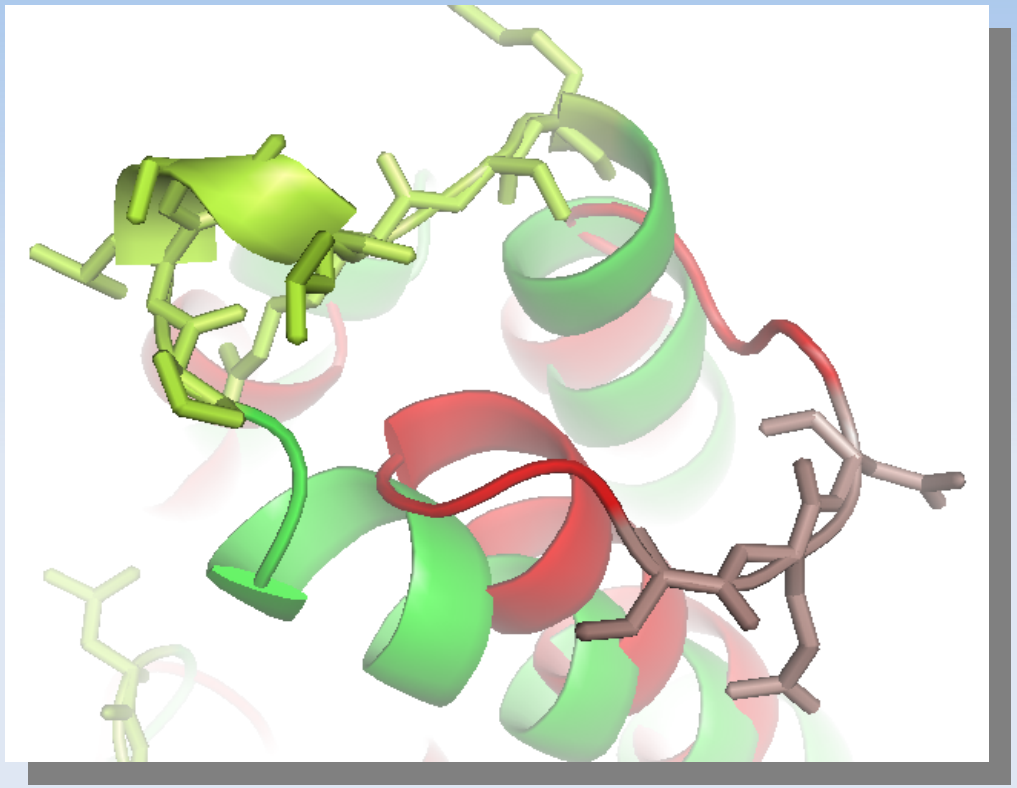
Structural alignmnet



- Establish homology between two or more polymer structures based on their shape and three-dimensional conformation.

```
1i5p:  ...ELIGLQANIREFNQQVDNF...
          ||||||||||||||||||||
ji6:   ...ELQGLQNNFEDYVNALNSW...
```

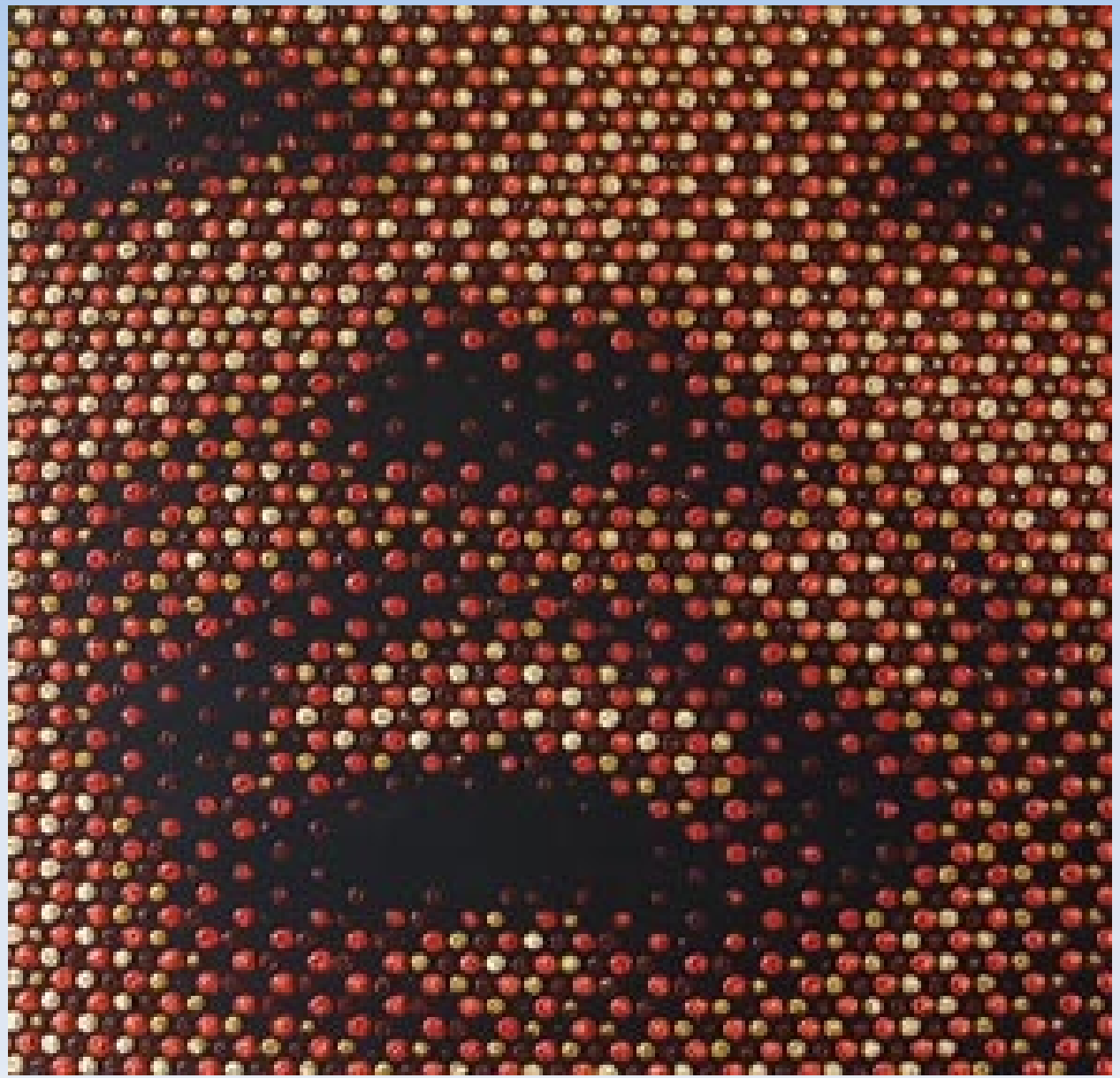
Szerkezeti egyezés



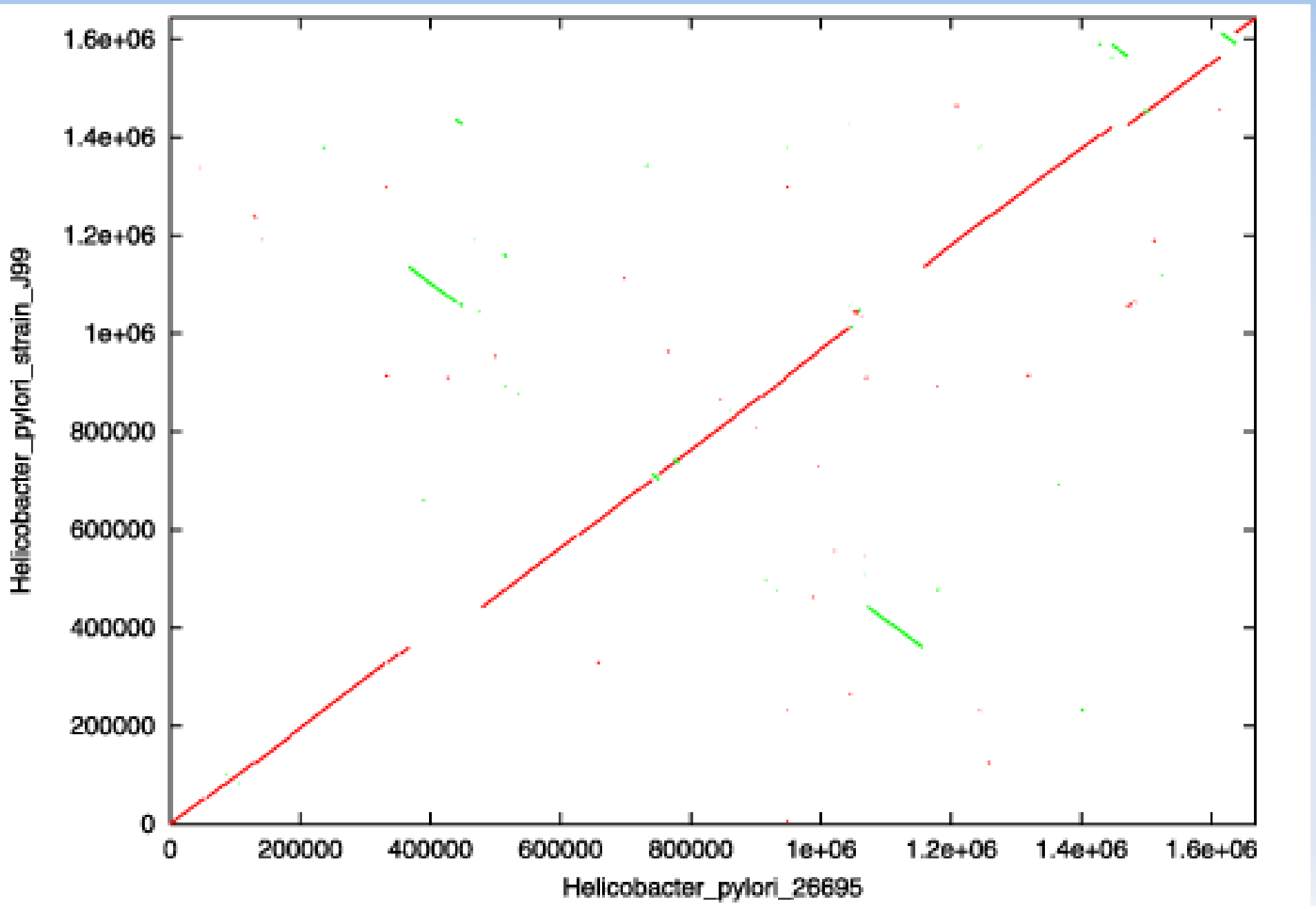
- Unalignable regions
  → gaps are inserted

```
1i5p:  ...DNFLNPTQN----PVPLSITSSVN...
          ||||||        ||||||||||||
ji6:   ...NSWKKTPLSLRSKRSQDRIRELFS...
```

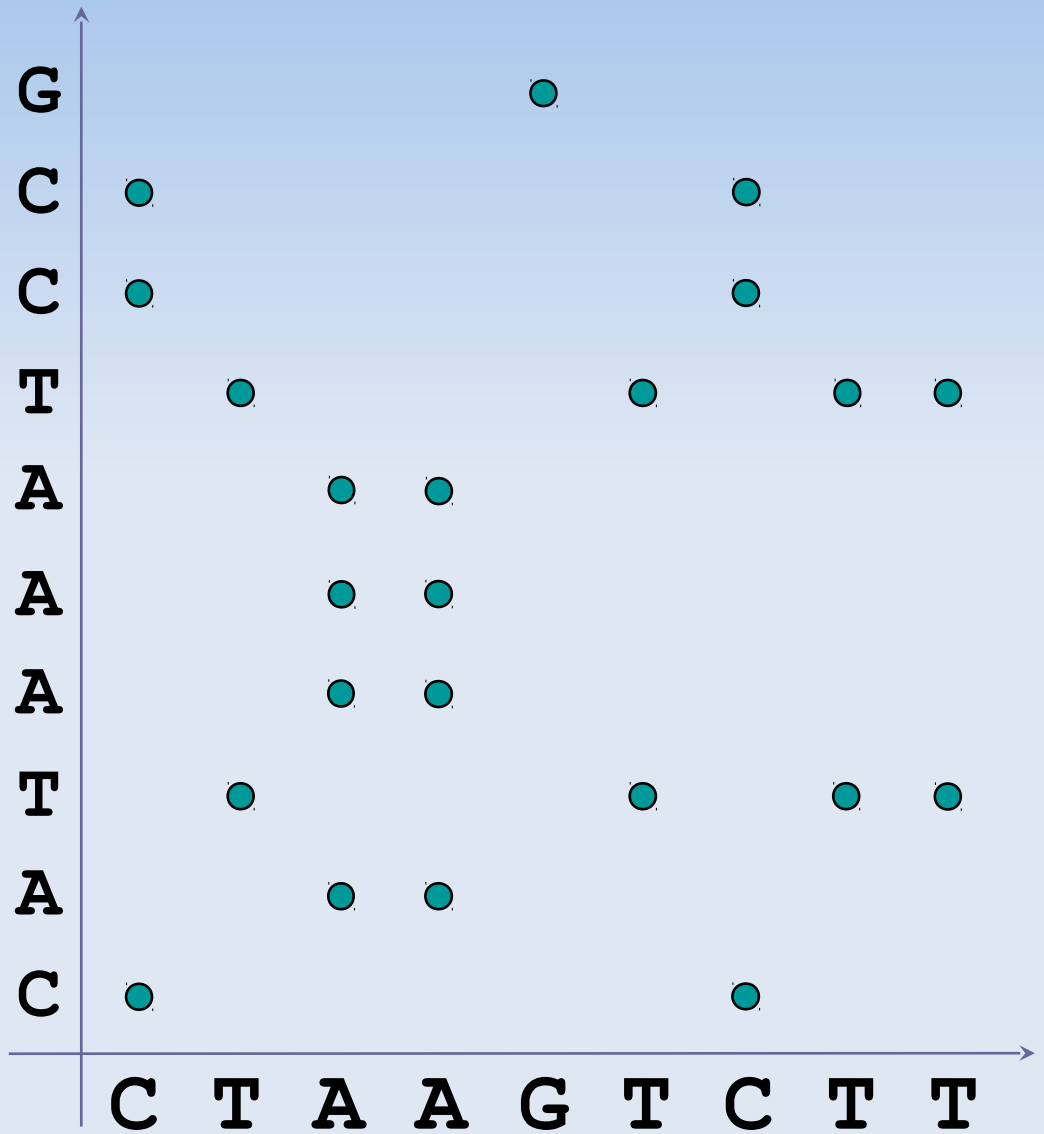
Pairwise sequence alignments: „Dot-plot”



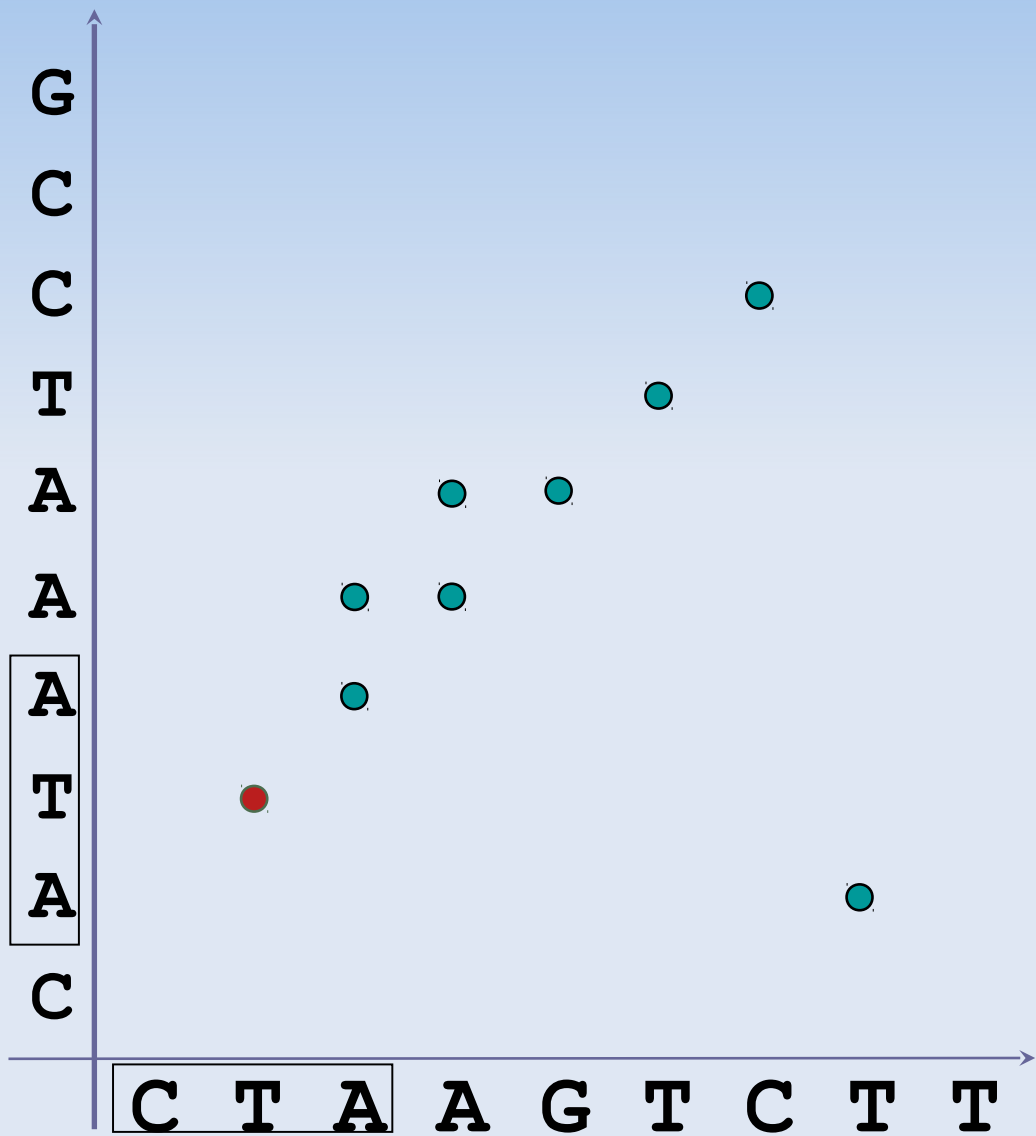
The „dot-plot”


1.6e+06
1.4e+06
1.2e+06
1e+06
800000
600000
400000
200000
0
Helicobacter_pylori_strain_J99
0
200000
400000
600000
800000
1e+06
1.2e+06
1.4e+06
1.6e+06
Helicobacter_pylori_26695

The „dot-plot”



window, word = 1

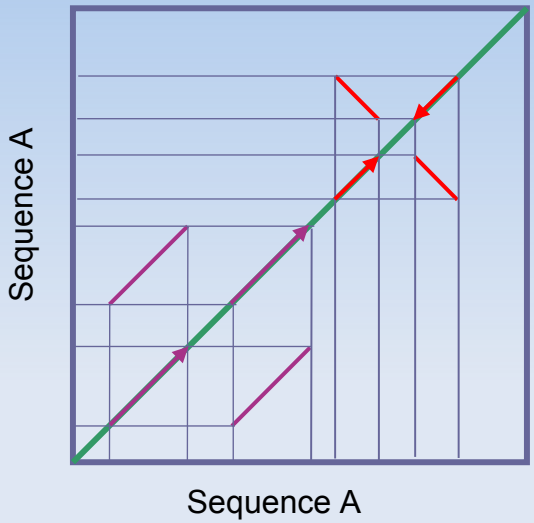


window = 3 stringency = 2

How to interpred the Dot-plot?

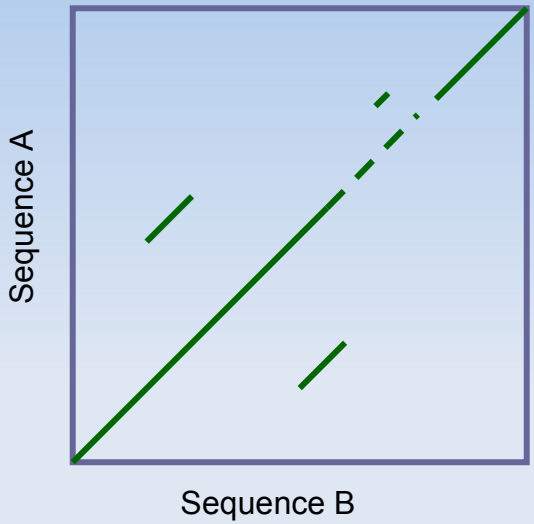
2 same sequences

A



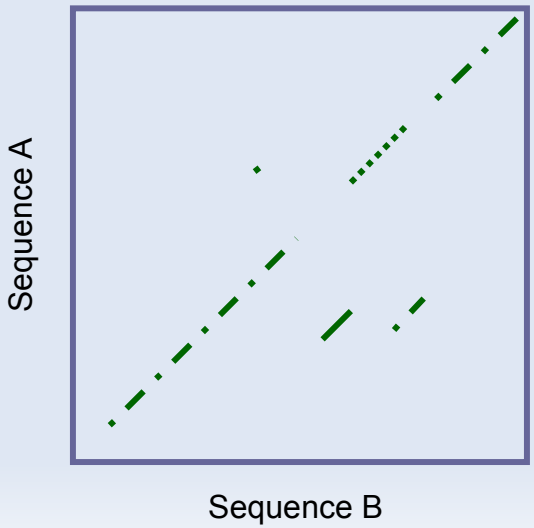
2 very similar sequences

B



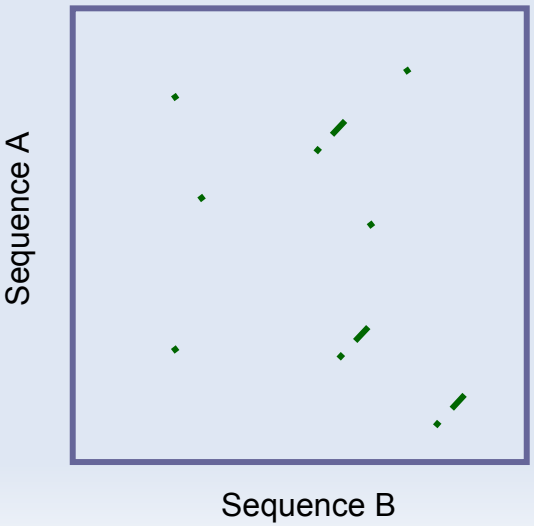
2 moderately similar sequences

C

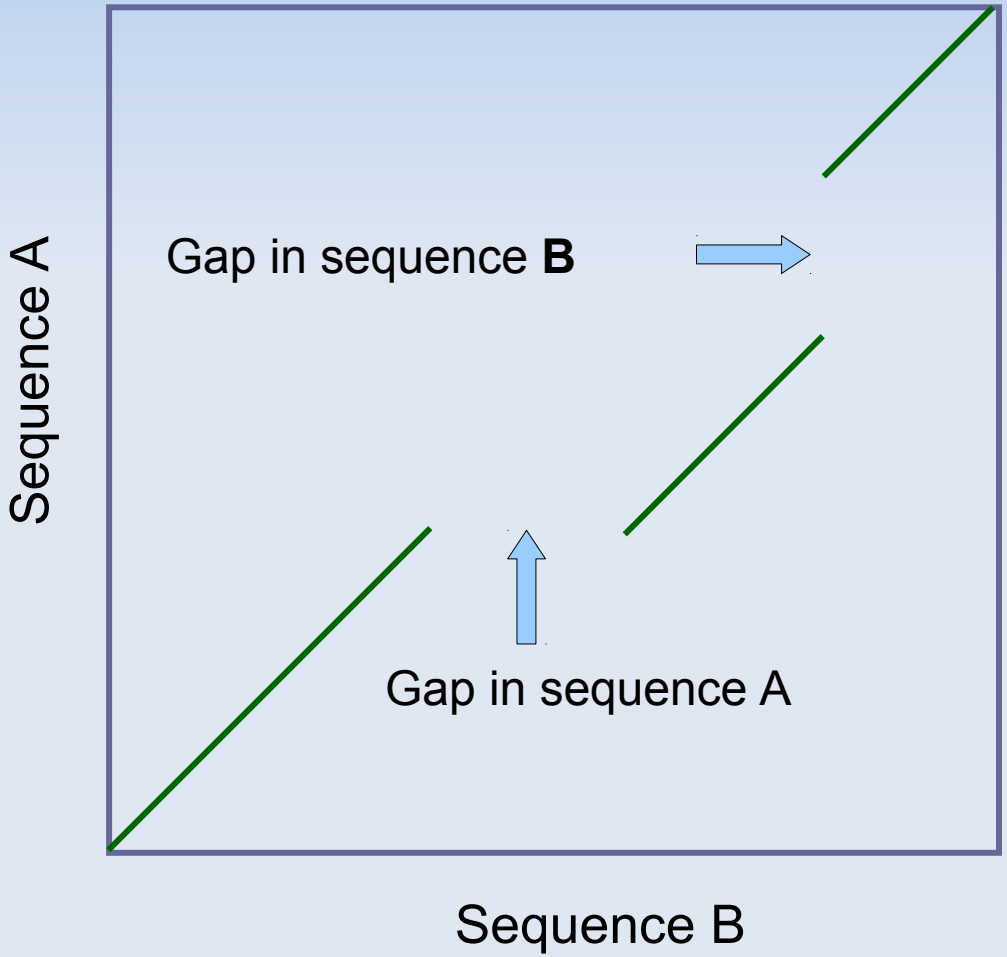


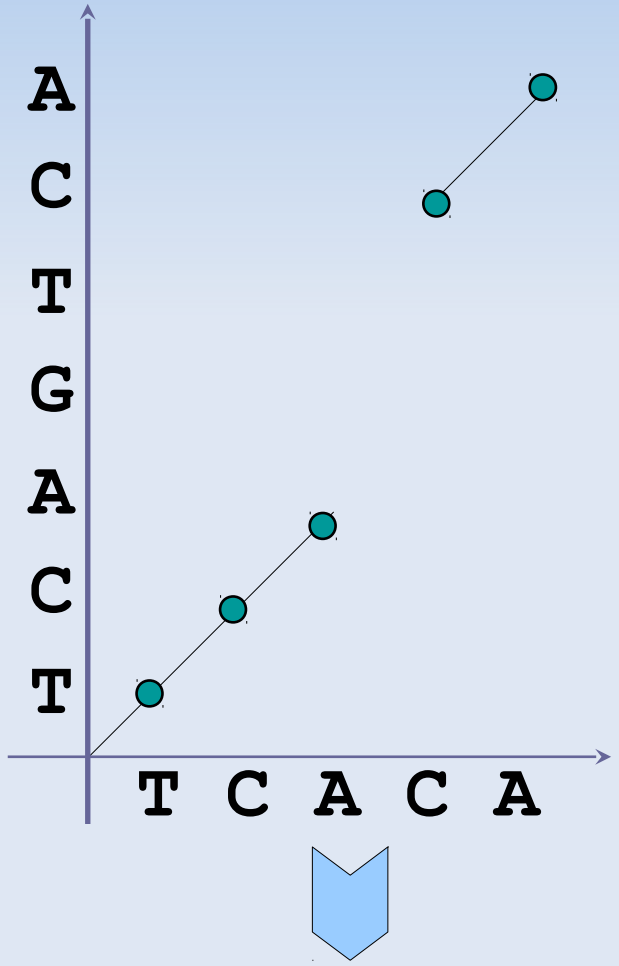
No similarty between sequences

D



Places of gaps





T C A G T C A
T C A – – C A

Dot-plot sofwares

- EMBOSS
  - dottup
  - dotpath
  - polydot
  - dotmatcher
- WWW, Java:
  - http://myhits.isb-sib.ch/cgi-bin/dotlet
  - http://pgrc.ipk-gatersleben.de/jdotter/

Number of possible alignments

How many possible sequence alignments are there?

```
SeqA:M
SeqB:T
SeqC:A
```

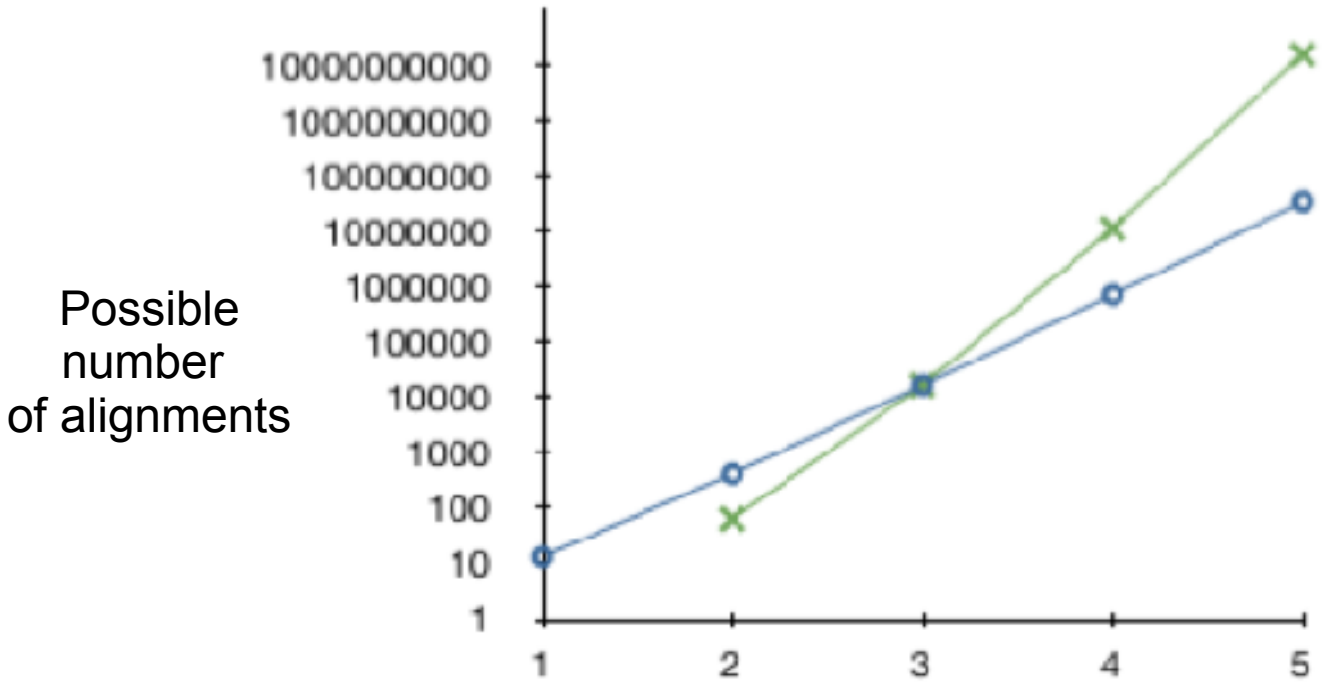
```
SeqA:M-    SeqA:M-    SeqA:M-    SeqA:-M    SeqA:-M    SeqA:M-
SeqB:T-    SeqB:-T    SeqB:-T    SeqB:-T    SeqB:T-    SeqB:-T
SeqC:-A    SeqC:-A    SeqC:A-    SeqC:A-    SeqC:A-    SeqC:A-
```

```
SeqA:M--   SeqA:M--   SeqA:-M-   SeqA:--M   SeqA:--M   SeqA:-M-
SeqB:-T-   SeqB:--T   SeqB:T--   SeqB:T--   SeqB:-T-   SeqB:--T
SeqC:--A   SeqC:-A-   SeqC:--A   SeqC:-A-   SeqC:A--   SeqC:A--
```

13 possible alignments

Number of possible alignments

- Growing (faster than) exponentially!



Seq. lengths (in case of 3 sequences)

Nr. of sequences (in case of 3 as/nt long sequences)

Alignment scores: Scoring matrices

Scoring of an alignment

- **Scores based on a substitution matrix**
  - PAM250:

A	Ala
C	Cys
D	Asp
E	Glu
F	Phe
G	Gly
H	His
I	Ile
K	Lys
L	Leu
M	Met
N	Asn
P	Pro
Q	Gln
R	Arg
S	Ser
T	Thr
Y	Tyr
V	Val
W	Trp

```
Seq 1:   M  N  A  L  S  D  R  T
Seq 2:   M  S  D  R  T  T  E  T
score    6 +1 +0 -3 +1 +0 -1 +3  =  7
```

Scoring systems, substitution matrices

- To score the level of similarity in between two sequences.
- Main types:
  - Identity matrix
    - Identity: 1, difference: 0 (or -4 and 5).
    - Mostly applied for nucleotide sequences
  - Chemical property based matrices
    - Polar, apolar, size, shape, charge, etc.
  - Substitution matrices
    - Describes the rate at which one character in a sequence changes to other character states over time.
    - The similarity between sequences depends on their divergence time and the substitution rates as represented in the matrix.

Identity matrices

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Nucleotides

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Amino acids

Substitution matrices: PAM

- Invented by Margaret Dayhoff (1970s)
  - PAM = APM = Accepted Point Mutation
  - is the replacement of a single amino acid in the primary structure of a protein with another single amino acid, which is accepted by the processes of natural selection.
  - Each entry in a PAM matrix indicates the likelihood of the amino acid of that row being replaced with the amino acid of that column
  - The calculation of these matrices were based on 1572 observed mutations in the phylogenetic trees of 71 families of closely related proteins. The proteins to be studied were selected on the basis of having high similarity (at least 85% identity).

PAM 120

Positive score – frequency of substitutions is greater than would have occurred by random chance.

Zero score – frequency is equal to that expected by chance.

Negative score – frequency is less than would have occurred by random chance.

	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

small, polar

small, nonpolar

polar or acidic

basic

large, hydrophobic

aromatic

Different PAM matrices

Difference %	PAM (evol. distance)
1	1
10	11
20	23
30	38
40	56
50	80
60	112
70	159
80	246

- Disadvantages of PAM matrices:
  - Below 85% if identity all other matrices were just extrapolated from the original PAM matrices
  - Based on a limited number of sequences.

PAM 10 matrix (~ 90% seq. identity)

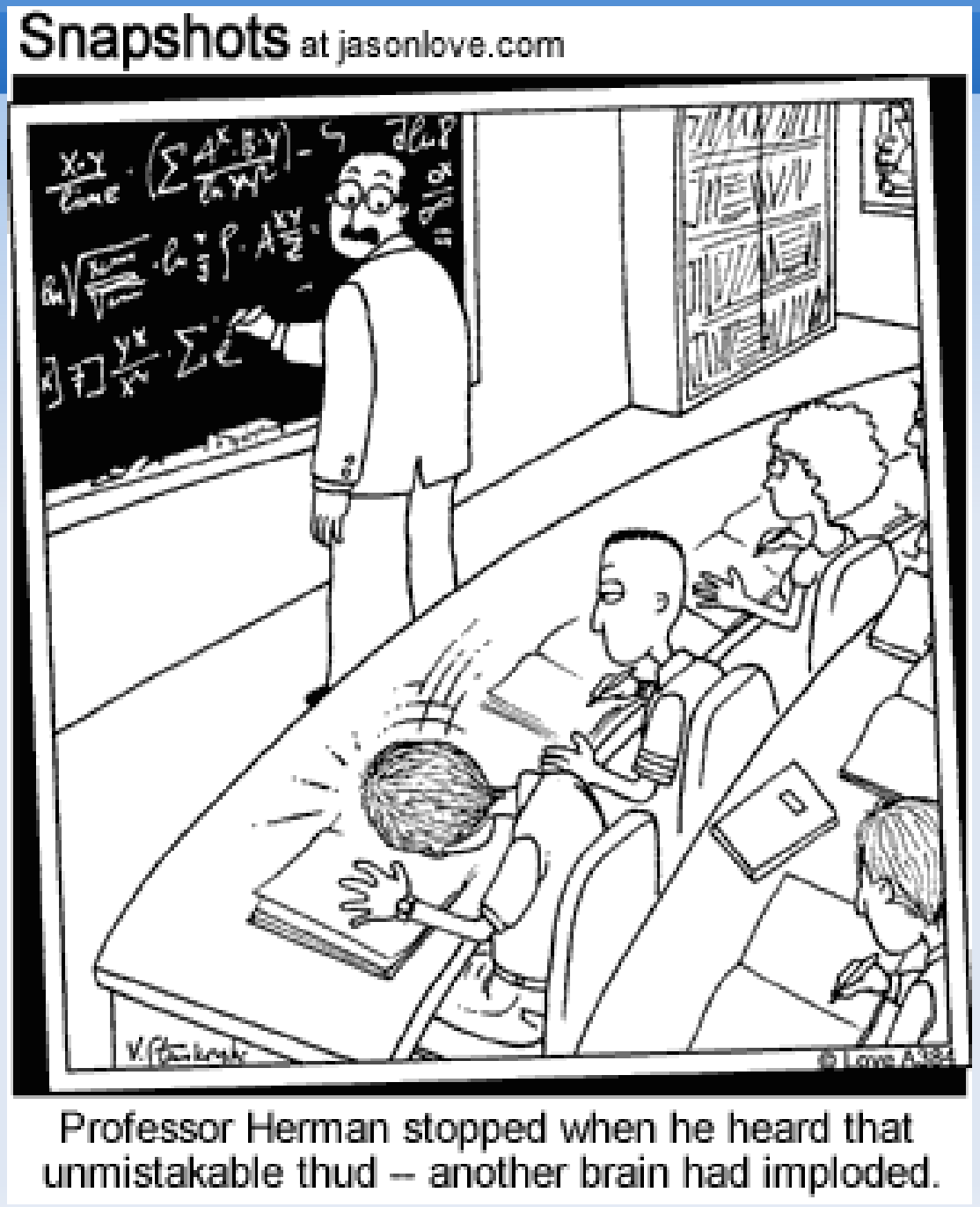
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-20	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8

A	Ala
C	Cys
D	Asp
E	Glu
F	Phe
G	Gly
H	His
I	Ile
K	Lys
L	Leu
M	Met
N	Asn
P	Pro
Q	Gln
R	Arg
S	Ser
T	Thr
Y	Tyr
V	Val
W	Trp

PAM 250 matrix (~ 20% seq. identity)

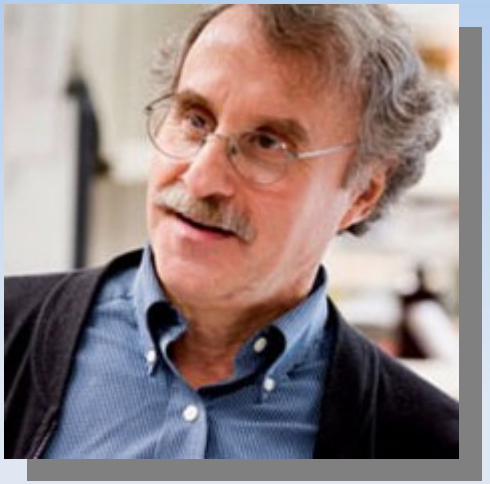
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I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

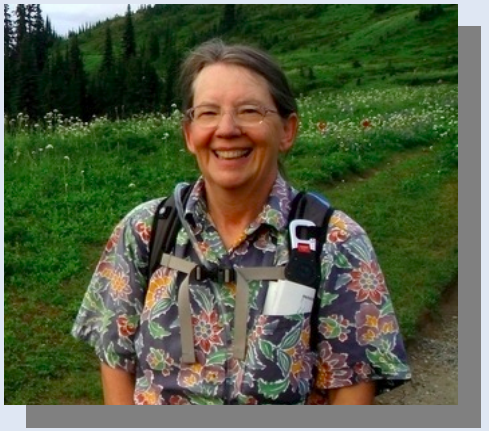
A	Ala
C	Cys
D	Asp
E	Glu
F	Phe
G	Gly
H	His
I	Ile
K	Lys
L	Leu
M	Met
N	Asn
P	Pro
Q	Gln
R	Arg
S	Ser
T	Thr
Y	Tyr
V	Val
W	Trp


Snapshots at jasonlove.com
Professor Herman stopped when he heard that unmistakable thud – another brain had imploded.

Az észlelt helyettesítéseken alapuló mátrixok II.

- BLOSUM mátrixok:
  BLOcks SUbstitution Matrix
  - Steven Henikoff & Jorja G. Henikoff 1992
  - Based on BLOCKS database: (http://blocks.fhcrc.org/)
  - Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.





BLOCKS databes → BLOSUM

- They scanned the BLOCKS database for very conserved regions of protein families
  - that do not have gaps in the sequence alignment
  - and then counted the relative frequencies of amino acids and their substitution probabilities.
- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins like the PAM Matrices.
- BLOSUM62: midrange
- BLOSUM80: more related proteins
- BLOSUM45: distantly related proteins
- In most of the cases gives better results than using PAM matrices.

BLOSUM62 scoring matrix

	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	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-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

A	Ala
C	Cys
D	Asp
E	Glu
F	Phe
G	Gly
H	His
I	Ile
K	Lys
L	Leu
M	Met
N	Asn
P	Pro
Q	Gln
R	Arg
S	Ser
T	Thr
Y	Tyr
V	Val
W	Trp

BLOSUM – PAM correspondences

BLOSUM90	BLOSUM80	**BLOSUM62**	BLOSUM45
PAM30	**PAM120**	PAM180	PAM240

More similar sequences → More different sequences

Scoring with PAM250:

```
Identities = 36/52 (69%), Positives = 47/52 (90%)

seq A: KMGPGFTKALGHGVDLGHIYGDNLERQYQLRLFKDGKLKYQVLDGEMYPPSV
         GP+FTK+  HGVDL+HIYG++LERQ +LRLFKDGK+KYQ+++GEMYPP+V
seq B: ERGPAFTKGKNHGVDLSHIYGESLERQHKLRLFKDGKMKYQMINGEMYPPTV
```

Scoring with BOLSUM62:

```
Identities = 36/52 (69%), Positives = 46/52 (88%)

seq A: KMGPGFTKALGHGVDLGHIYGDNLERQYQLRLFKDGKLKYQVLDGEMYPPSV
       + GP FTK   HGVDL HIYG++LERQ++LRLFKDGK+KYQ+++GEMYPP+V
seq B: ERGPAFTKGKNHGVDLSHIYGESLERQHKLRLFKDGKMKYQMINGEMYPPTV
```

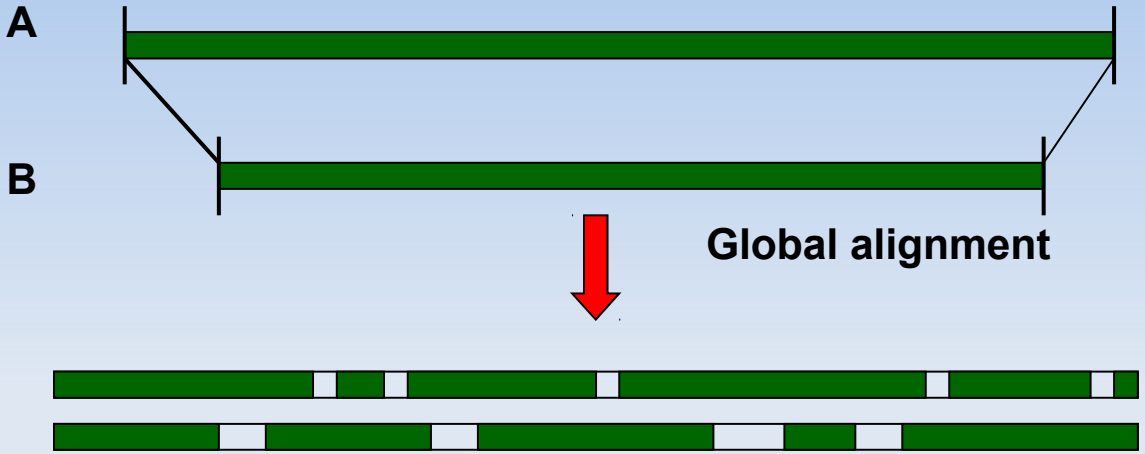
A	Ala
B	Asp,Asn
C	Cys
D	Asp
E	Glu
F	Phe
G	Gly
H	His
I	Ile
K	Lys
L	Leu
M	Met
N	Asn
P	Pro
Q	Gln
R	Arg
S	Ser
T	Thr
V	Val
W	Trp
X	Xxx
Y	Tyr
Z	Glu,Gln
*	End

Pairwise sequence alignment algorithms

„Optimal” alignments

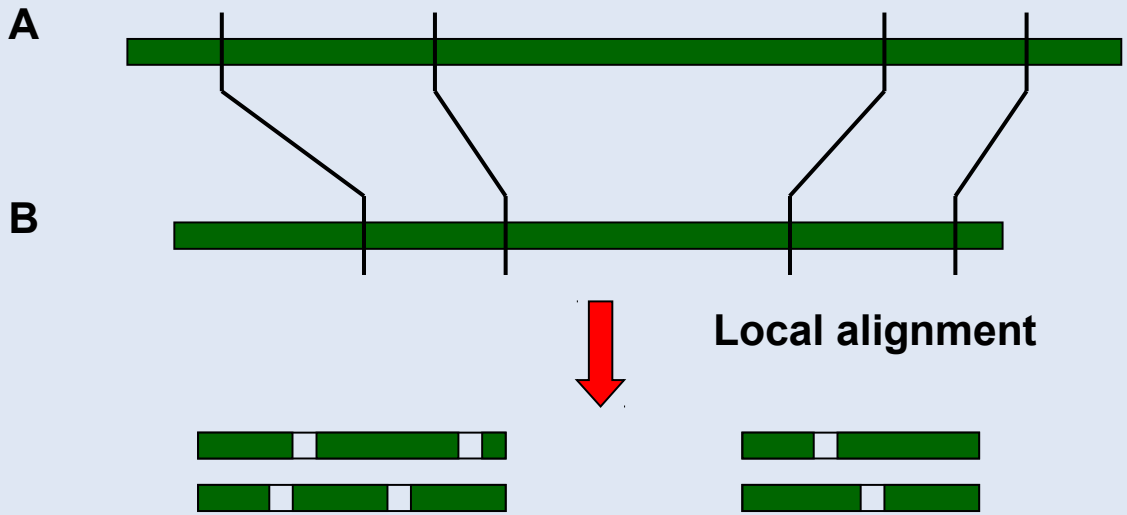
- Sequence alignmnet: Shows where 2 sequences are similar or different from each other
- Mathematically optimal alignment: which maximize the similarity measure of sequences (alignment score).
- The result is based on the scoring matrix and the applied alignment method.
  - We can overrule the alignment by hand if necessary

Global and local pairwise alignments



Needleman − Wunsch algorithm

attempt to align every residue in every sequence



Smith − Waterman algorithm

Align the similar regions only

Gap penalties

- Fixed
- or based on the length of the inserted gap
  - $\rightarrow$ affine gap penalty: w_x
  - Gap opening penalty (bigger): g
  - Gap extension penalty (smaller): r

$$w_x = g + r \times x$$

  - x: length of the gap
- The scores are related to the scoring matrix

Gap penalties

- Using affine gap penalty:

```
 A T G T A G T G T A T A G T A C A T G C A
 A T G T A G - - - - - - - T A C A T G C A
+5+5+5+5+5+5-5-1-1-1-1-1-1+5+5+5+5+5+5+5+5
```

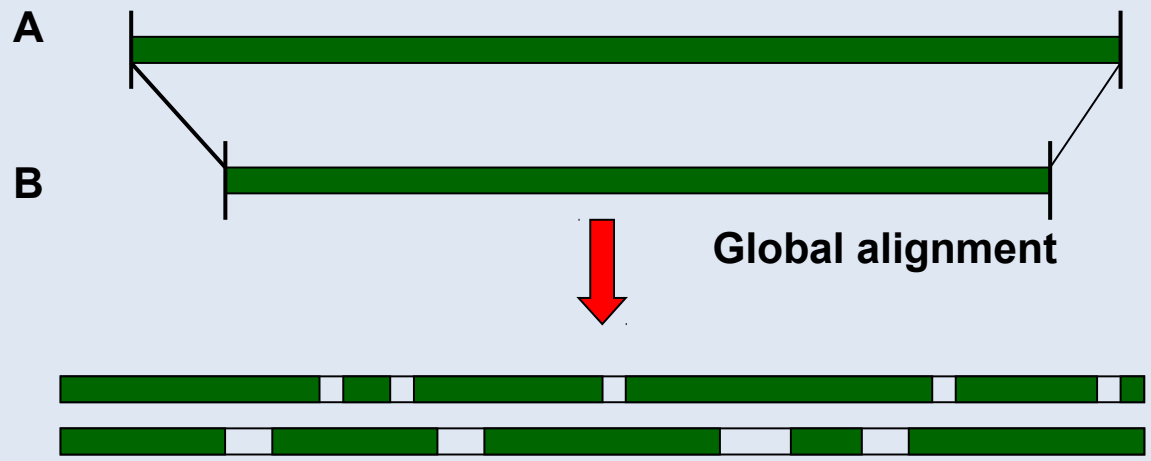
- Without affine gap penalty:

```
 A T G T A G T G T A T A G T A C A T G C A
 A T G T A - - G - - T A - - - C A T G C A
+5+5+5+5+5-5-5+5-5-5+5+5-5-5-5+5+5+5+5+5+5
```

- Which has more relevance in biological sense?

Global alignment

- Needleman - Wunsch algorithm
  - Needleman and Wunsch 1970
  - divides a large problem (e.g. the full sequence) into a series of smaller problems and uses the solutions to the smaller problems to reconstruct a solution to the larger problem.
  - Using a Dynamic Programming algorithm



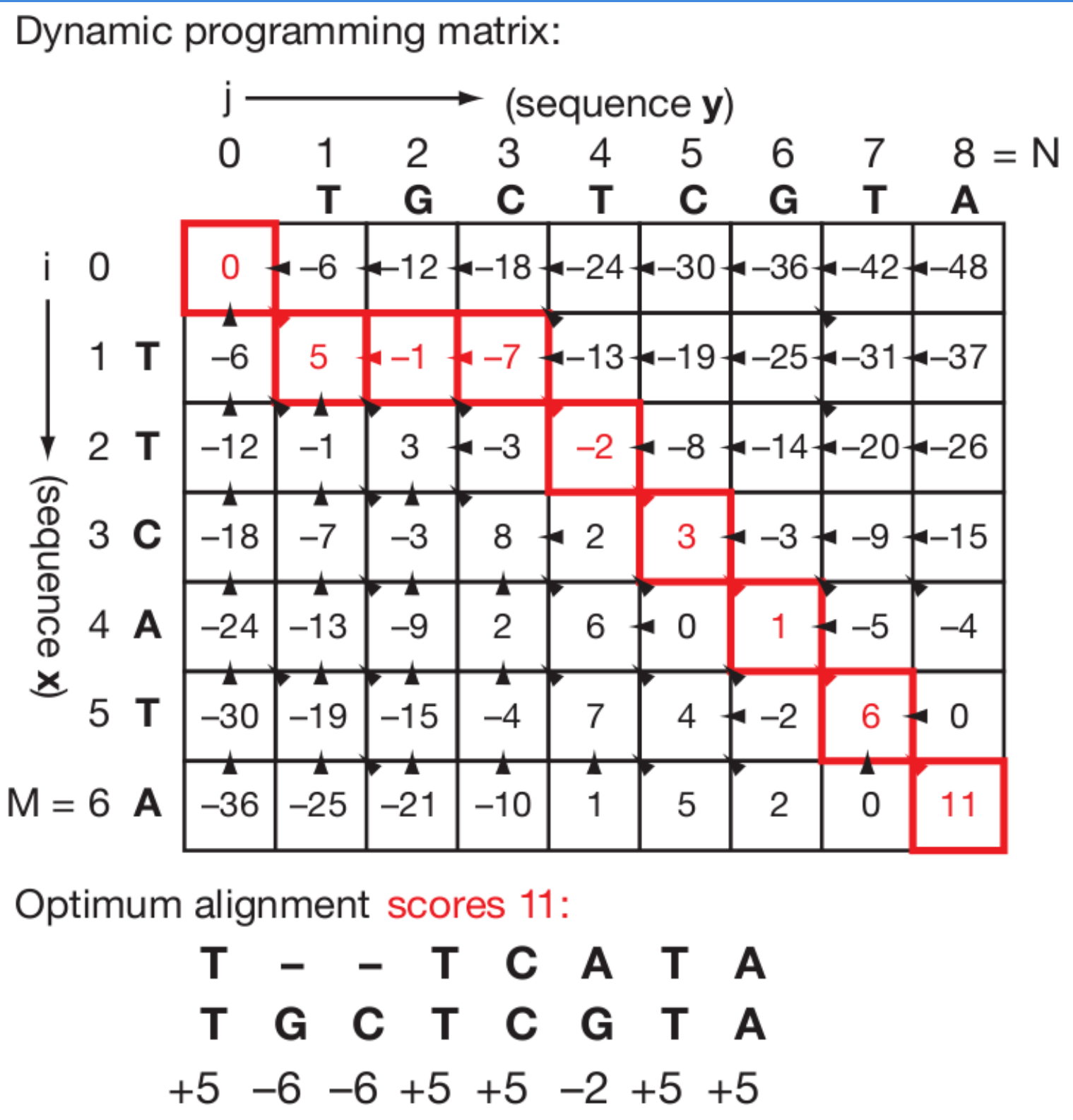
Example of Dynamic programming

scoring:

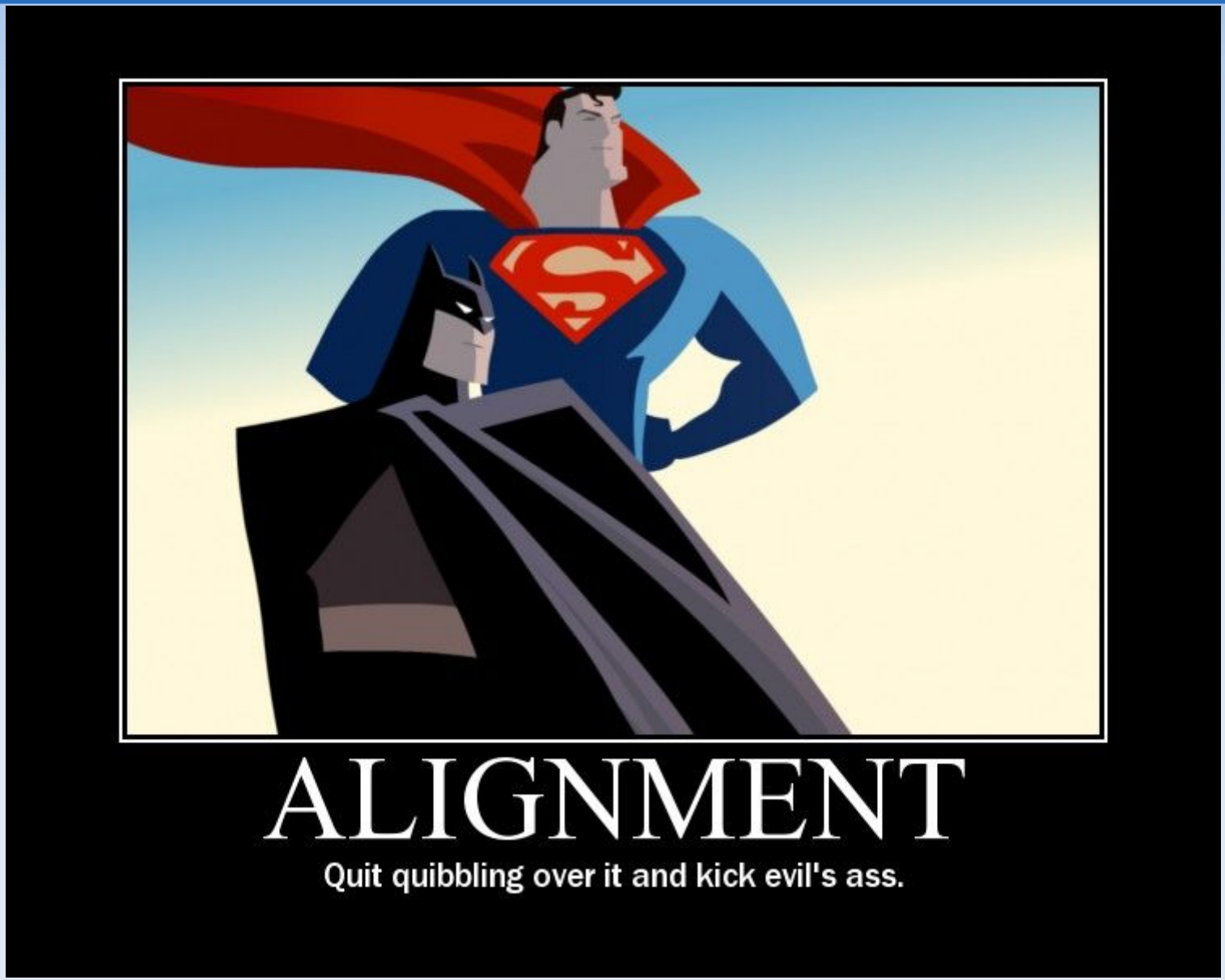
match: +5

mismatch: -2

in/del (gap): -6

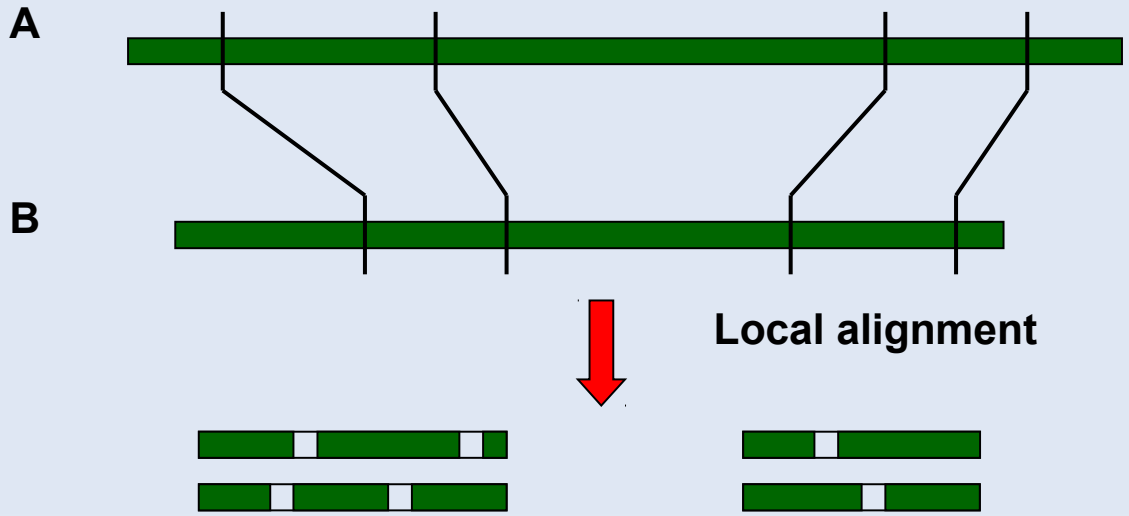


Eddy SR. (2004)
What is dynamic programming?
Nat Biotechnol, 22(7):909-10.


ALIGNMENT
Quit quibbling over it and kick evil's ass.

Local alignment

- Smith - Waterman algorithm
  - Smith and Waterman 1981
  - determining similar regions between two sequences.
  - Instead of looking at the entire sequence, compares segments of all possible lengths and optimizes the similarity measure.
  - Uses Dynamic programming algorithm.



Pairwise alignments

- Global alignment

```
seq1   M  -  N  A  L  S  D  R  T
seq2   M  G  S  D  R  T  T  E  T
score  6 -12 1  0 -3  1  0 -1  3  =  -5
```

- Global alignment with no gap penalties at the sequence ends

```
seq1   M  N  A  L  S  D  R  T  -  -  -
seq2   -  -  M  G  S  D  R  T  T  E  T
score  0  0 -1 -4  2  4  6  3  0  0  0  =  10
```

- Local alignment

```
seq1                S  D  R  T
seq2                S  D  R  T
score               2  4  6  3  =  15
```

What software should I use?

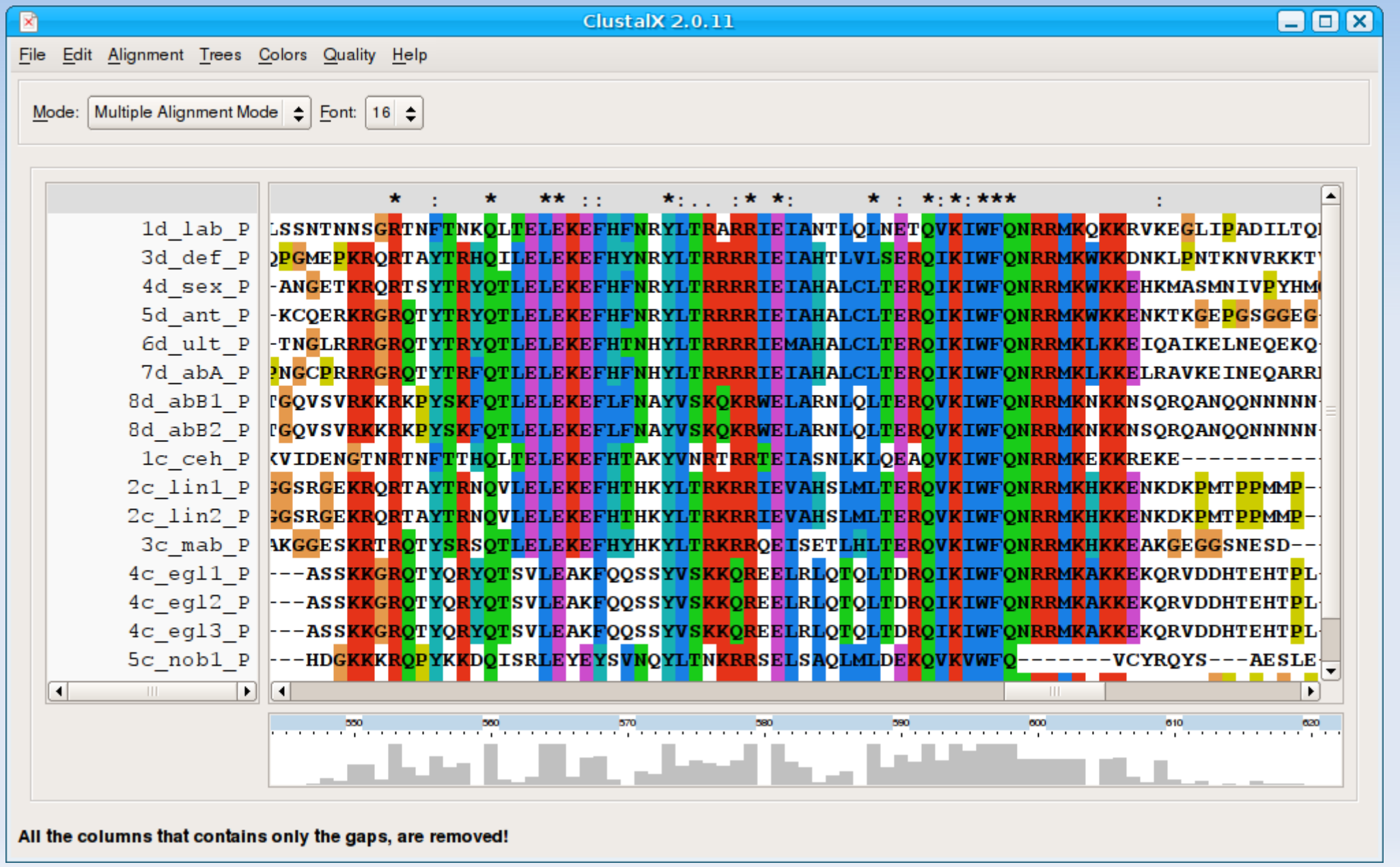
- Global alignment (Needleman - Wunsch)
  - EMBOSS: needle; stretcher (for long sequences)
- Local alignment (Smith - Waterman)
  - EMBOSS: water; matcher (for long sequences)
  - sim (optimal and sub-optimal alignments)
  - FASTA3 package: ssearch3; lalign (even suboptimal alignmnets)
- Web:
  - http://bioweb2.pasteur.fr/
  - Similarity searches in databases:
    - EBI: SSEARCH (http://www.ebi.ac.uk/services/other-software) based on the Smith - Waterman algorithm
- Warning: Softwares will give you a result even where there is no biological sense!

Differences in between alignments

Clustal

Muscle

Multiple sequence alignment



Is it simple or complicated?

```
GCGGCCCA TCAGGTAGTT GGTGG
GCGGCCCA TCAGGTAGTT GGTGG
GCGTTCCA TCAGCTGGTT GGTGG
GCGTCCCA TCAGCTAGTT GGTGG
GCGGCGCA TTAGCTAGTT GGTGA
******** ********** *****
```

- Simple

```
TTGACATG CCGGGG---A AACCG
TTGACATG CCGGTG--GT AAGCC
TTGACATG -CTAGG---A ACGCG
TTGACATG -CTAGGGAAC ACGCG
TTGACATC -CTCTG---A ACGCG
******** ?????????? *****
```

- Complicated
  - Because of in/dels

3 base methods

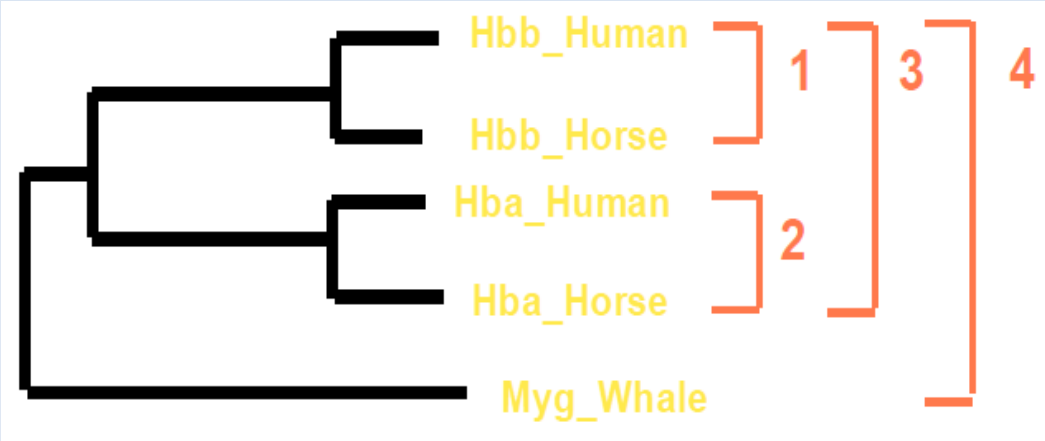
- Manually
  - Applicable when sequences are very similar, almost identical
- Automatic
  - Using an alignment software (e.g. ClustalW, -Ω, T-Coffee, MUSCLE)
- Combining the above two
  - Correcting the software provided alignment by hand
  - Based on other information (e.g. structure, phylogeny)

Automatic methods for multiple alignment

- With dynamic programming algorithms
  - Huge computational demand
  - → maximum 10 average length protein sequences
    - *MSA* (Lipman et al., 1989) Global Optimal **M**ultiple **S**equence **A**lignment Program
    - *DCA* (Stoye et al., 1997) **D**ivide-and-**C**onquer **A**lignment
- Stochastic methods, iterative strategies, progressive alignment
  - Robust, less sensitive to the number of sequences
  - It is not guaranteed that it will find the optimal alignment

Progressive multiple alignment

```
Hbb_Human 1 -
Hbb_Horse 2 .17 -
Hba_Human 3 .59 .60 -
Hba_Horse 4 .59 .59 .13 -
Myg_Whale 5 .77 .77 .75 .75 -
```

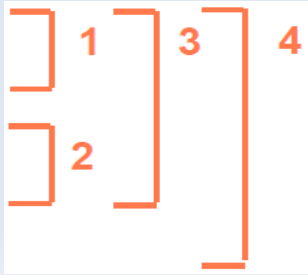


1. Fast pairwise alignments:
   → distance matrix

2. Neighbor-joining guide tree

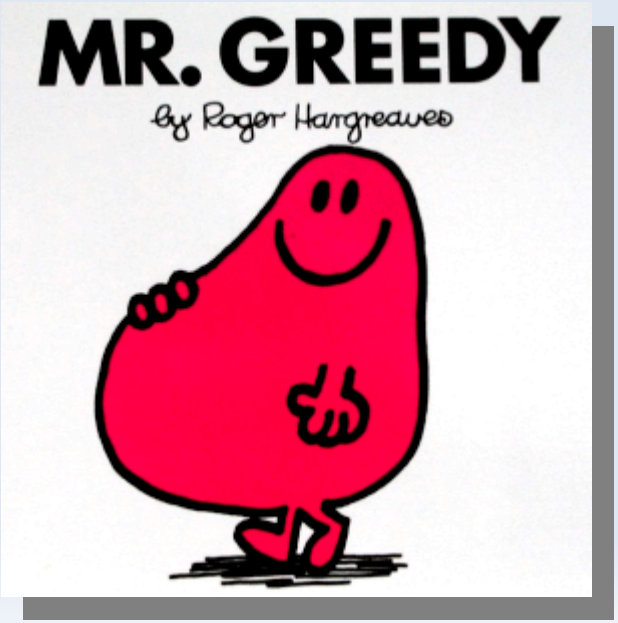
3. Progressive alignmnet using the guide tree

```
Hbb_Human PEEKSAVTALWGKVN--VDEVGG
Hbb_Horse GEEKAAVLALWDKVN--EEEVGG
Hba_Human PADKTNVKAAWGKVGAHAGEYGA
Hba_Horse AADKTNVKAAWSKVGGHAGEYGA
Myg_Whale EHEWQLVLHVWAKVEADVAGHGQ
```



Drawbacks of progressive alignment methods

- Local minimum problem:
  - Greedy algorithm
  - The algorithm doesn't correct misaligned regions
  - Once a gap is inserted it will stay in the alignment
  - Causes of problems in most of the times: Incorrect guide tree



Softwares for multiple alignment

- *ClustalW* (*ClustalW2, ClustalX, ClustalX2, ClustalΩ*)
  - Most cited paper, mostly applied sequence aligner
  - Pro: relatively fast and uses not too much memory
  - Contra: for global alignment only.
  - Windows, Linux, Mac installers: http://www.clustal.org/
  - On-line: e.g. http://www.ebi.ac.uk/Tools/msa/clustalo/
- *Multalin*
  - Iterative: it regenerate the guide tree after alignment and restart to align
  - http://bioinfo.genotoul.fr/multalin/multalin.html
- *TCoffee*
  - Pro: Better than ClustalW when aligning less similar sequences
  - Contra: 2X slower than ClustalW
  - http://www.tcoffee.org/
- http://en.wikipedia.org/wiki/List_of_sequence_alignment_software

Sources

- I used some slides of Dr. Aidan Budd and Dr. Gábor Tóth (with their approval).
- Thanks to the original authors.

Thank you for the attention!