In the Name of God, the Merciful, the Compassionate

Introduction to Bioinformatics
06 : Multiple Sequence Alignment

Instructor: Hossein Zeinali

Amirkabir University of Technology

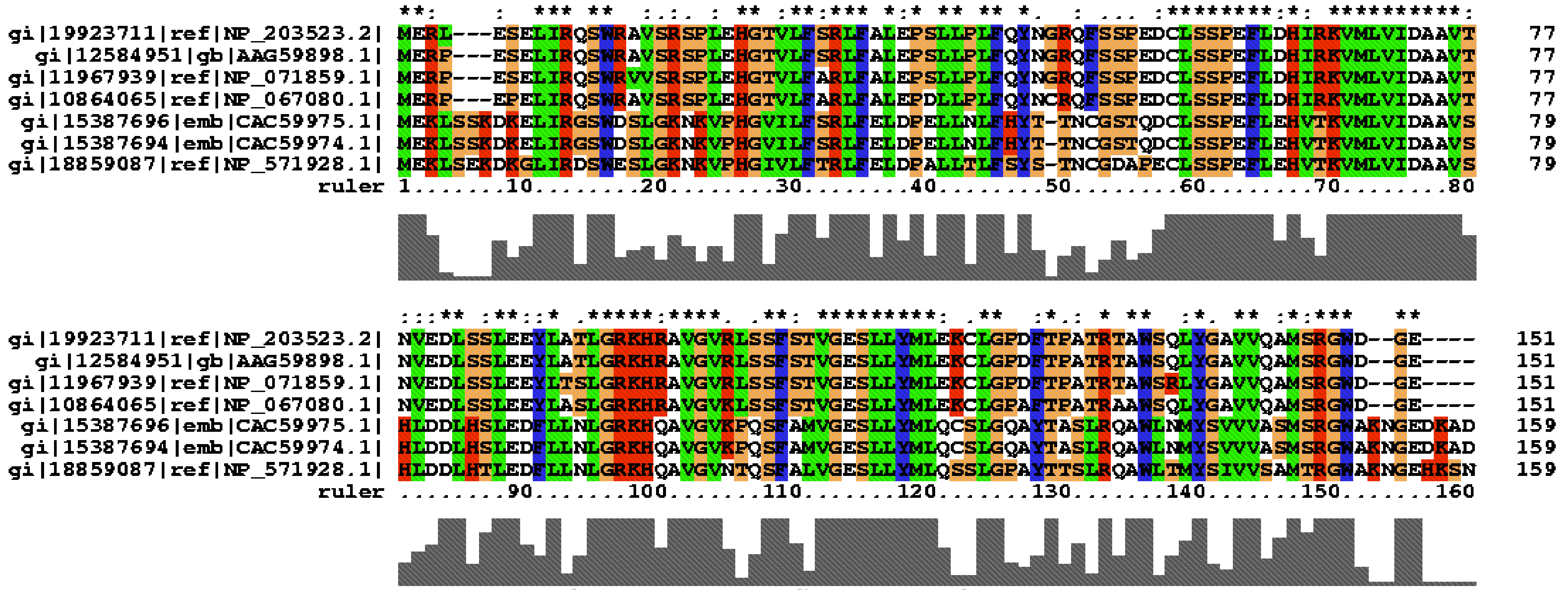
Overview

1. What is a multiple sequence alignment (MSA)?
2. Where/why do we need MSA?
3. What is a *good* MSA?
4. Algorithms to compute a MSA

Multiple Sequence Alignment

- Generalize pairwise alignment of sequences to include **> 2** *homologous (related)* sequences
- Analyzing more than 2 sequences gives us much more information:
  - Which amino acids are required? Correlated?
  - Evolutionary/phylogenetic relationships
- Similar to PSI-BLAST idea (not yet covered):
  - Use a set of homologous sequences to provide more "sensitivity"

Multiple Sequence Alignments



```
                                **:    :  *** **  :.:. : ** :**:*** *:* ** ** *.  : ... :********:*: **********:
gi|19923711|ref|NP_203523.2|    MERL---ESELIRQSWRAVSRSPLEHGTVLFSRLFALEPSLLPLFQYNGRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVT     77
  gi|12584951|gb|AAG59898.1|    MERP---ESELIRQSWRAVSRSPLEHGTVLFSRLFALEPSLLPLFQYNGRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVT     77
gi|11967939|ref|NP_071859.1|    MERP---ESELIRQSWRVVSRSPLEHGTVLFARLFALEPSLLPLFQYNGRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVT     77
gi|10864065|ref|NP_067080.1|    MERP---EPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCRQFSSPEDCLSSPEFLDHIRKVMLVIDAAVT     77
 gi|15387696|emb|CAC59975.1|    MEKLSSKDKELIRGSWDSLGKNKVPHGVILFSRLFELDPELLNLFHYT-TNCGSTQDCLSSPEFLEHVTKVMLVIDAAVS     79
 gi|15387694|emb|CAC59974.1|    MEKLSSKDKELIRGSWDSLGKNKVPHGVILFSRLFELDPELLNLFHYT-TNCGSTQDCLSSPEFLEHVTKVMLVIDAAVS     79
gi|18859087|ref|NP_571928.1|    MEKLSEKDKGLIRDSWESLGKNKVPHGIVLFTRLFELDPALLTLFSYS-TNCGDAPECLSSPEFLEHVTKVMLVIDAAVS     79
                       ruler    1.......10........20........30........40........50........60........70........80

                                :::** :**::* .*****:****. .**: *********: .**  :*.: * **  :*. ** :*:***   **
gi|19923711|ref|NP_203523.2|    NVEDLSSLEEYLATLGRKHRAVGVRLSSFSTVGESLLYMLEKCLGPDFTPATRTAWSQLYGAVVQAMSRGWD--GE----    151
  gi|12584951|gb|AAG59898.1|    NVEDLSSLEEYLATLGRKHRAVGVRLSSFSTVGESLLYMLEKCLGPDFTPATRTAWSQLYGAVVQAMSRGWD--GE----    151
gi|11967939|ref|NP_071859.1|    NVEDLSSLEEYLTSLGRKHRAVGVRLSSFSTVGESLLYMLEKCLGPDFTPATRTAWSRLYGAVVQAMSRGWD--GE----    151
gi|10864065|ref|NP_067080.1|    NVEDLSSLEEYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKCLGPAFTPATRAAWSQLYGAVVQAMSRGWD--GE----    151
 gi|15387696|emb|CAC59975.1|    HLDDLHSLEDFLLNLGRKHQAVGVKPQSFAMVGESLLYMLQCSLGQAYTASLRQAWLNMYSVVVASMSRGWAKNGEDKAD    159
 gi|15387694|emb|CAC59974.1|    HLDDLHSLEDFLLNLGRKHQAVGVKPQSFAMVGESLLYMLQCSLGQAYTASLRQAWLNMYSVVVASMSRGWAKNGEDKAD    159
gi|18859087|ref|NP_571928.1|    HLDDLHTLEDFLLNLGRKHQAVGVNTQSFALVGESLLYMLQSSLGPAYTTSLRQAWLTMYSIVVSAMTRGWAKNGEHKSN    159
                       ruler    ........90.......100.......110.......120.......130.......140.......150.......160
```

What is a MSA?

```
ATT-GC
ATTTGC
TTTTG
```

Not a MSA

```
AT-TGC
ATTTGC
ATTTG-
```

MSA

```
AT-T-GC
ATTT-GC
ATTT-G-
```

Not a MSA

Why?

Definition of MSA

- Given a set of sequences, a multiple sequence alignment is an assignment of gap characters, such that
  - Resulting sequences have same length
  - No column contains only gaps

```
ATT-GC
ATTTGC
ATTTG
```

No

```
AT-TGC
ATTTGC
ATTTG-
```

Yes

```
AT-T-GC
ATTT-GC
ATTT-G-
```

No

Applications of MSA

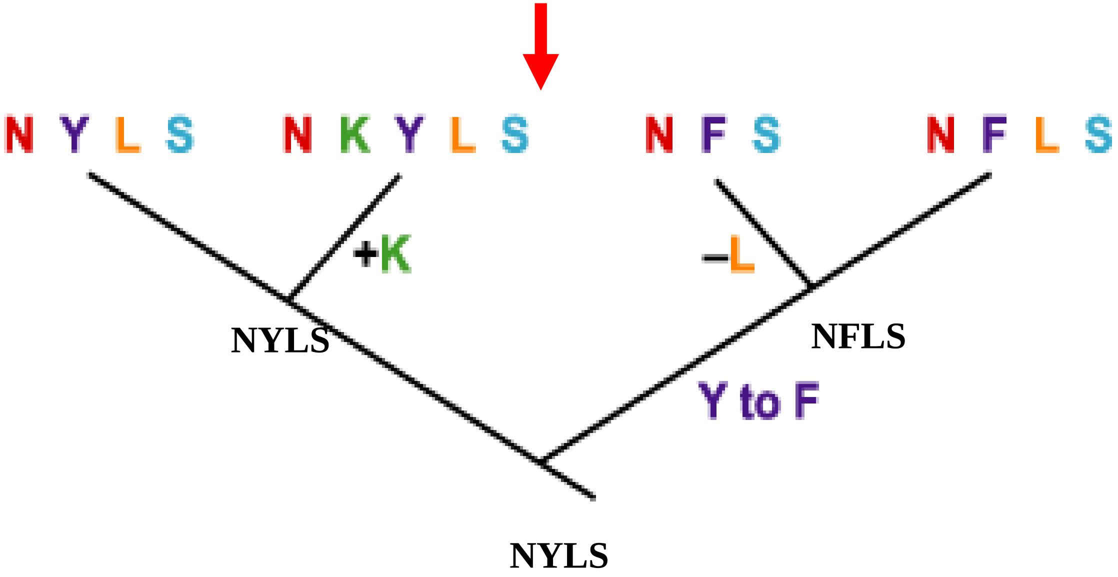
- Building phylogenetic trees and doing phylogenetic analysis of sequence families.
- Finding conserved patterns, e.g.:
  - Regulatory motifs
  - Protein domains
- Identifying and characterizing protein families
  - Find out which protein domains have same function
  - Prediction of protein secondary and tertiary structures
- DNA fragment assembly (in genomic sequencing)

Scoring an Alignment

Goal: Align homologous positions.

But: Without knowledge of phylogenetic tree is this very hard (sometimes impossible) to achieve!

seqA	N	•	F	L	S
seqB	N	•	F	–	S
seqC	N	K	Y	L	S
seqD	N	•	Y	L	S



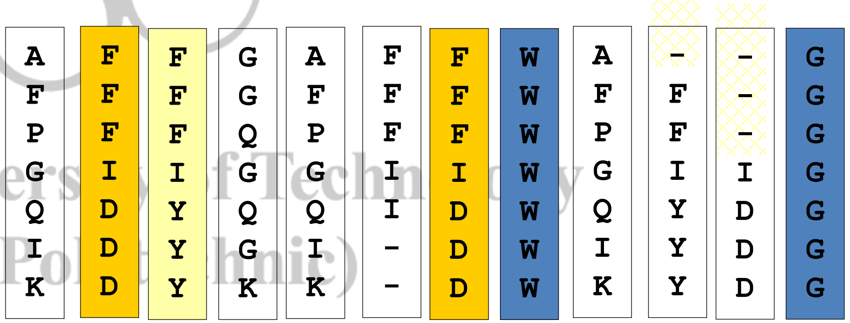
Scoring an Alignment

- In practice, simple scoring functions are used: usually, columns are scored independently, i.e.

$$S(m)=\sum_i S(m_i)+G$$

gap penalty

*i*th column of alignment *m*



A	F	F	G	A	F	F	W	A	-	-	G
F	F	F	G	F	F	F	W	F	F	-	G
P	F	F	Q	P	F	F	W	P	F	-	G
G	I	I	G	G	I	I	W	G	I	I	G
Q	D	Y	Q	Q	I	D	W	Q	Y	D	G
I	D	Y	G	I	-	D	W	I	Y	D	G
K	D	Y	K	K	-	D	W	K	Y	D	G

Scoring Function

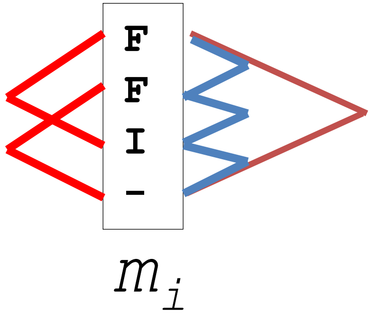
- Sum of Pairs (SP) Score = sum of scores of all possible pairs of sequences in an MSA based on a particular scoring matrix

- Compute for each column *c*

$$S(m_i) = \Sigma_{k<l}\, s(m_i^k, m_i^l)$$

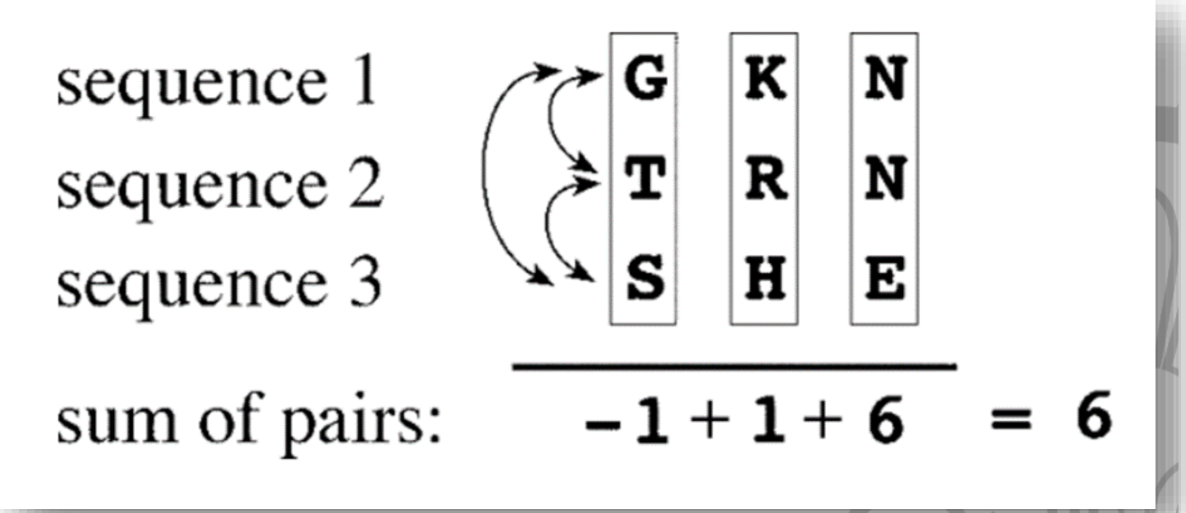
PAM or BLOSUM score

residue *l*



A	F	F	G	A	F	F	W	A	-	-	G
F	F	F	G	F	F	F	W	F	F	-	G
P	F	F	Q	P	I	F	W	P	F	D	G
G	I	I	G	G	-	I	W	G	Y	D	G

Example

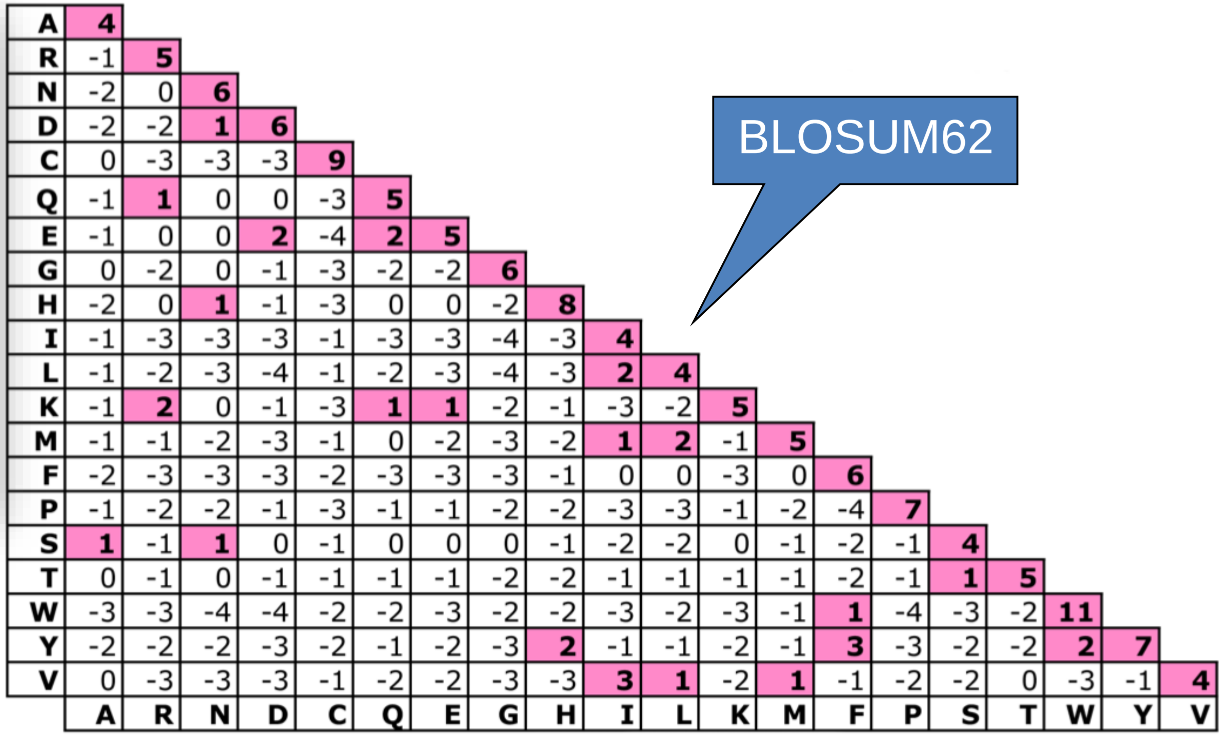


sequence 1 G K N

sequence 2 T R N

sequence 3 S H E

sum of pairs: −1 + 1 + 6 = 6



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

BLOSUM62

How Score Gaps in MSAs?

- Want to align gaps with each other over all sequences.
- A gap in a pairwise alignment that “matches” a gap in another pairwise alignment should cost less than introducing a totally new gap.
  - Possible that a new gap could be made to “match” an older one by adjusting older pairwise alignment
  - Change gap penalty near conserved domains of various kinds (e.g. secondary structure elements, hydrophobic regions)

Example of SP Score with Gap

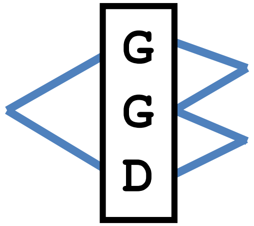
	F	Y	G	D
F	5	-2	-2	-1
Y		7	1	-5
G			4	-3
D				5

BLOSUM 60

$$m=\begin{array}{|c|\,|c|\,|c|} F & - & G \\ F & - & G \\ F & Y & D \end{array}$$

Gap penalty = -8

$s(-,-) = 0$



$$\begin{aligned} S(m) &= S(m_1) + S(m_2) + S(m_3) \\ &= 3s(F,F) + 2s(-,Y) + s(-,-) + s(G,G) + 2s(G,D) \\ &= 15 - 16 + 0 + 4 - 6 = -3 \end{aligned}$$

Algorithms for MSA

Exhaustive Methods

- Multidimensional dynamic programming (DP)
  - Divide-and-Conquer Alignment (DCA) - "semi-exhaustive"
  - Full DP Optimal Global Alignment?
    - *Prohibitive in both time & space requirements for more than 10 sequences!!*

Heuristic Methods

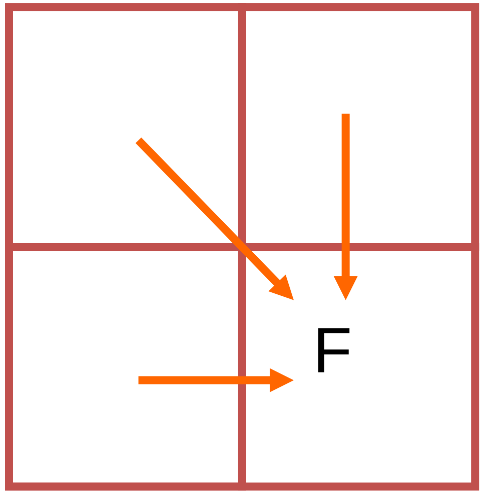
- Progressive alignments
  - We will cover Clustal, Star Alignment, T-Coffee, POA
  - Others: DbClustal and PRALINE -see text-book

Algorithms for MSA (Cont.)

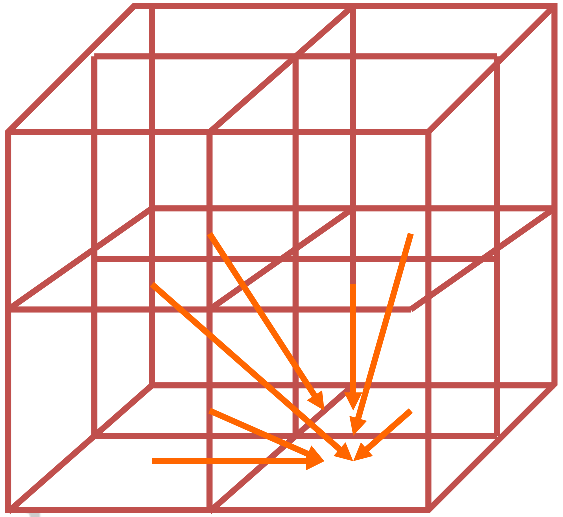
- Iterative methods
  - – Idea: optimal solution can be found by repeatedly modifying existing suboptimal solutions (eg: PRRN)
- Block-based Alignment
  - – Multiple re-building attempts to find best alignment (eg: DIALIGN2 & Match-Box)
- Local alignments
  - – Profiles, Blocks, Patterns - more on these soon!

Dynamic Programming for MSA

- As with pairwise alignments, multiple sequence alignments can be computed by dynamic programming



2D

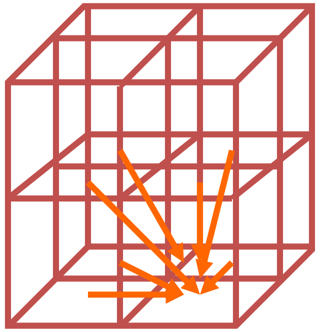


3D

Generalized Needleman-Wunsch Algorithm

- Given 3 sequences x, y, and z:
- Main iteration loop:

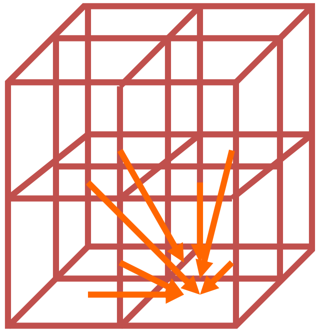
$$F(i,j,k) = \max \left(\begin{array}{l} F(i-1, j-1, k-1) + S(x_i, y_j, z_k), \\ F(i-1, j-1, k) + S(x_i, y_j, -), \\ F(i-1, j, k-1) + S(x_i, -, z_k), \\ F(i-1, j, k) + S(x_i, -, -), \\ F(i, j-1, k-1) + S(-, y_j, z_k), \\ F(i, j-1, k) + S(-, y_j, -), \\ F(i, j, k-1) + S(-, -, z_k) \end{array} \right)$$



3D

What Happens to Computational Complexity?

- Given k sequences of length n:
  - Space for matrix: **$O(n^k)$**
  - Neighbors/cell: **2^k-1**
  - Time to compute SP score: **$O(k^2)$**
  - Overall runtime: **$O(k^2 2^k n^k)$**



3D

- So, full dynamic programming is limited to small datasets of less than ten short sequences.

An Example of DP's Running Time

- Overall runtime: $O(k^2 2^k n^k)$

# sequences	running time
2	1 second
3	2 minutes
4	5 hours
5	3 weeks
6	9 years

Sequences: globins (≈ 150 aa)

Don't worry, there are fast heuristics

- Implementation example:
  - Divide-and-Conquer Alignment (DCA): semi-exhaustive
    - Breaking each of the sequences into two smaller sections.
  - http://bibiserv.techfak.uni-bielefeld.de/dca

Progressive Alignment

Heuristic procedure:

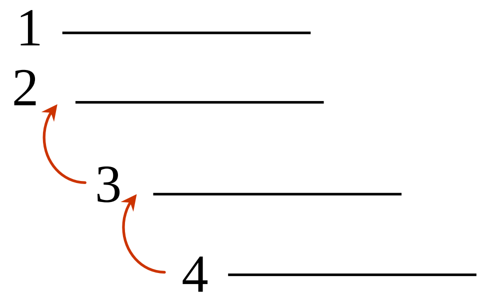
1. Align *most* similar sequences first
2. Add sequences progressively

- Often ***guide trees*** is used to determine order of alignments.

Examples: Star alignment

ClustalW

Multiple Alignment by adding sequences



What is a Consensus Sequence?

- A single sequence that represents *most common* residue of each column in a MSA
- **Example**:

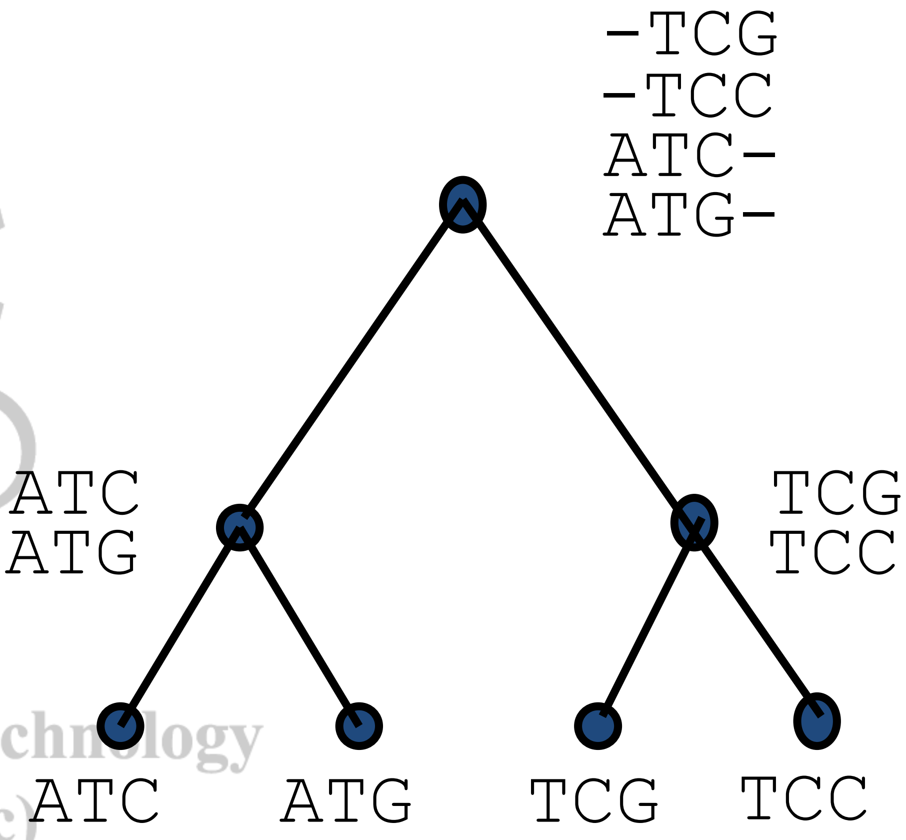
```
FGGHL-GF
F-GHLPGF
FGGHP-FG
FGGHL-GF
```

- **Steiner consensus sequence:** given sequences $s_1,\ldots, s_k$, find a sequence s^* that maximizes $\Sigma_i S(s^*,s_i)$

Guide Tree

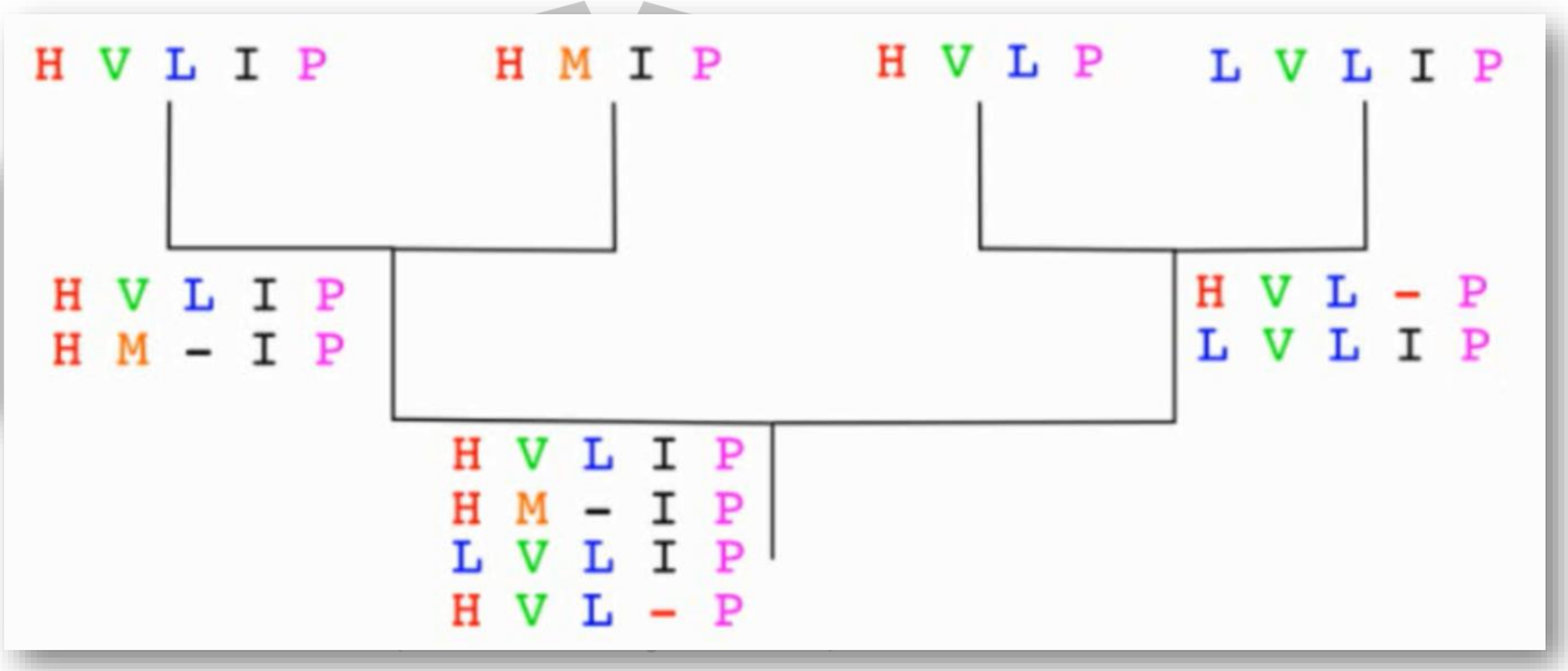
Binary tree

- Leaves correspond to sequences
- Internal nodes represent alignments
- Root corresponds to final MSA
- Is created using neighbor-joining method
- Is a simple phylogenetic tree



Example

S1: HVLIP
S2: HMIP
S3: HVLP
S4: LVLIP



H V L I P

H M I P

H V L P

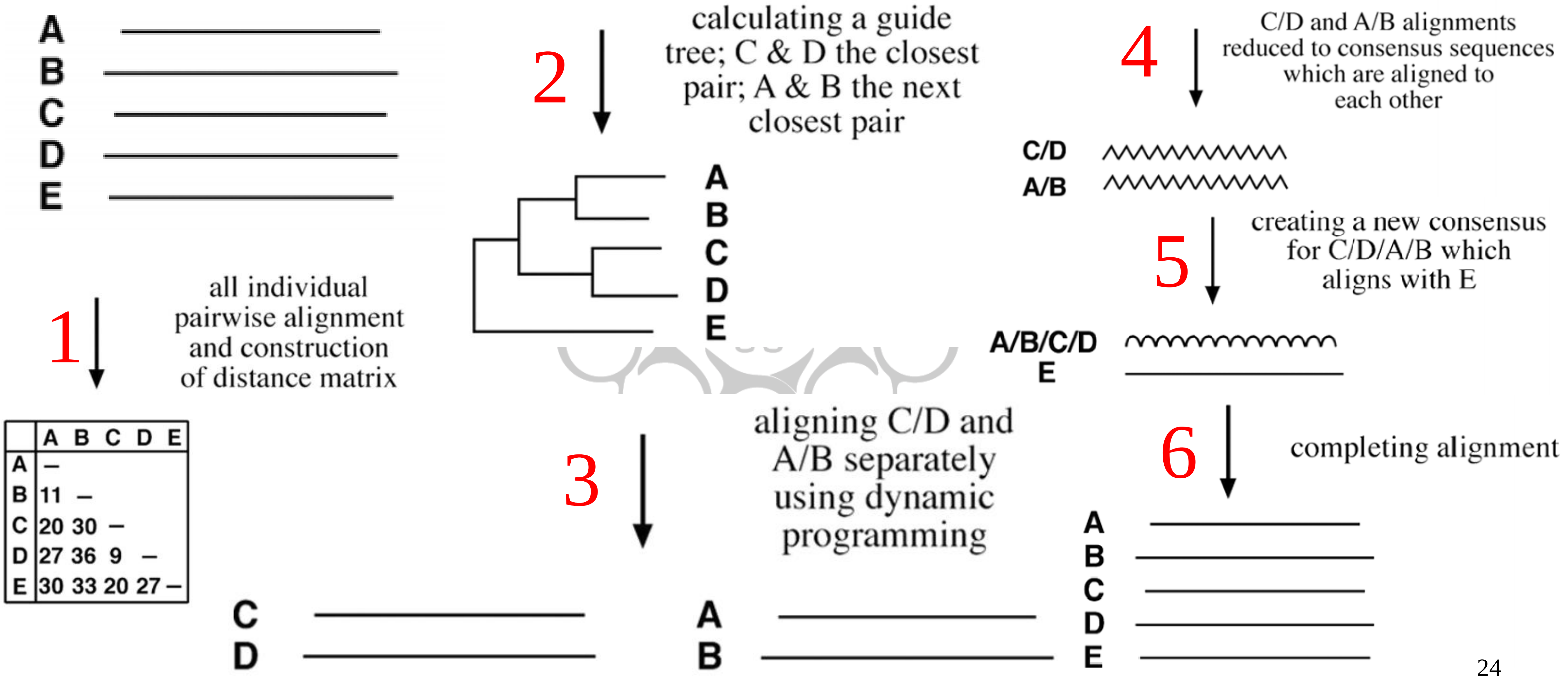
L V L I P

```
H V L I P
H M - I P
```

```
H V L - P
L V L I P
```

```
H V L I P
H M - I P
L V L I P
H V L - P
```

Progressive Alignment Steps



1 all individual pairwise alignment and construction of distance matrix

	A	B	C	D	E
A	–				
B	11	–			
C	20	30	–		
D	27	36	9	–	
E	30	33	20	27	–

2 calculating a guide tree; C & D the closest pair; A & B the next closest pair

3 aligning C/D and A/B separately using dynamic programming

4 C/D and A/B alignments reduced to consensus sequences which are aligned to each other

5 creating a new consensus for C/D/A/B which aligns with E

6 completing alignment

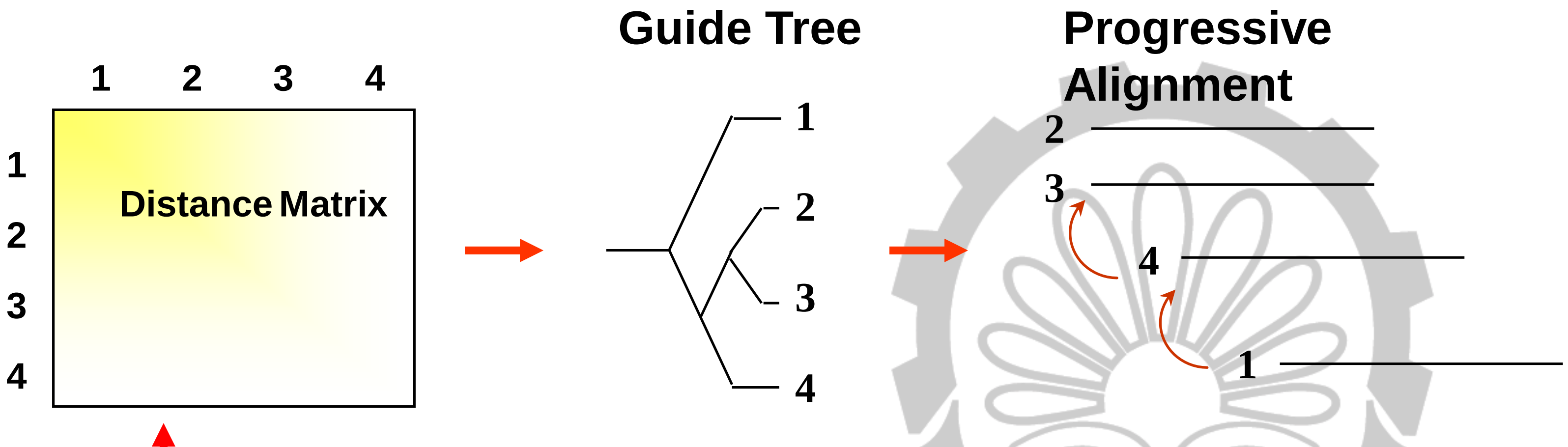
Clustal Program

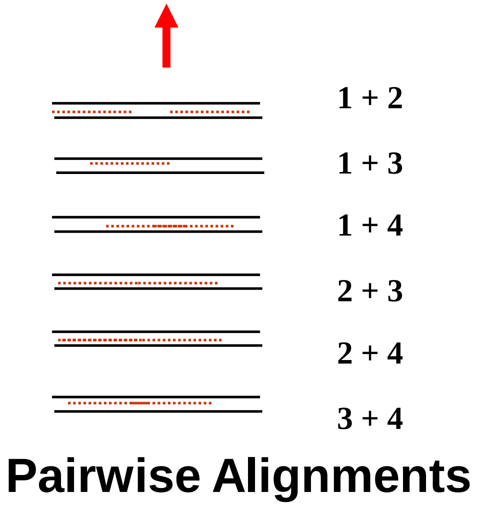
- The most well-known progressive alignment program
  - www.ebi.ac.uk/clustalw
- ClustalW and ClustalX: stand-alone programs which run on UNIX and Macintosh respectively.
- Does not rely on a single substitution matrix
  - Applies different scoring matrices depending on degrees of similarity.
- Uses of adjustable gap penalties
  - allow more insertions and deletions in regions that are outside the conserved domains, but fewer in conserved regions.
- Applies a weighting scheme to increase the reliability of aligning divergent sequences.

Clustal

1. Perform pair-wise alignments between all pairs of sequences (n * (n-1)/2 possibilities)
2. Generate distance matrix
   - Distance between a pair = number of mismatched positions in alignment divided by total number of matched positions
3. Generate a Neighbor-Joining ‘**guide tree**’ from distance table
4. Use guide tree to progressively align sequences in pairs from tips to root of tree

CLUSTAL: Overview



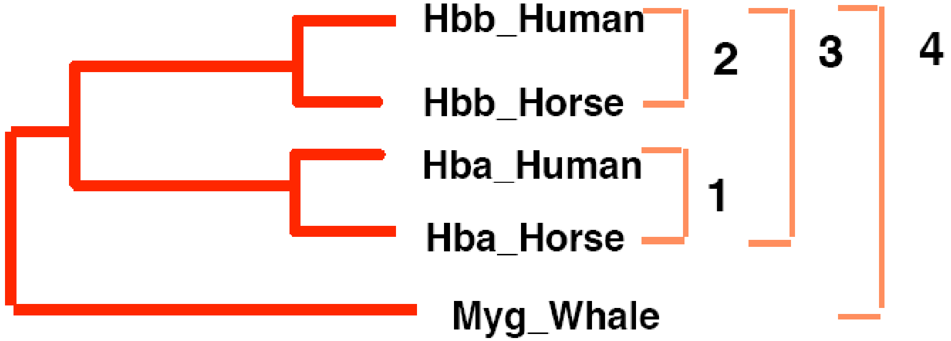


1. Compute pairwise alignments (DP)
2. Convert similarities into distances

   Distance between a pair = # of mismatched positions in alignment (divided by total # of matches)
3. Build <u>guide tree</u> from distances by *Neighbor Joining*
4. Align with respect to guide tree

ClustalW

```
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   -
```



alpha-helices

```
1   PEEKSAVTALWGKVN--VDEVGG   2   3   4
2   GEEKAAVLALWDKVN--EEEVGG
3   PADKTNVKAAWGKVGAHAGEYGA
4   AADKTNVKAAWSKVGGHAGEYGA   1
5   EHEWQLVLHVWAKVEADVAGHGQ
```

CLUSTAL W

Quick pair wise alignment
calculate distance matrix

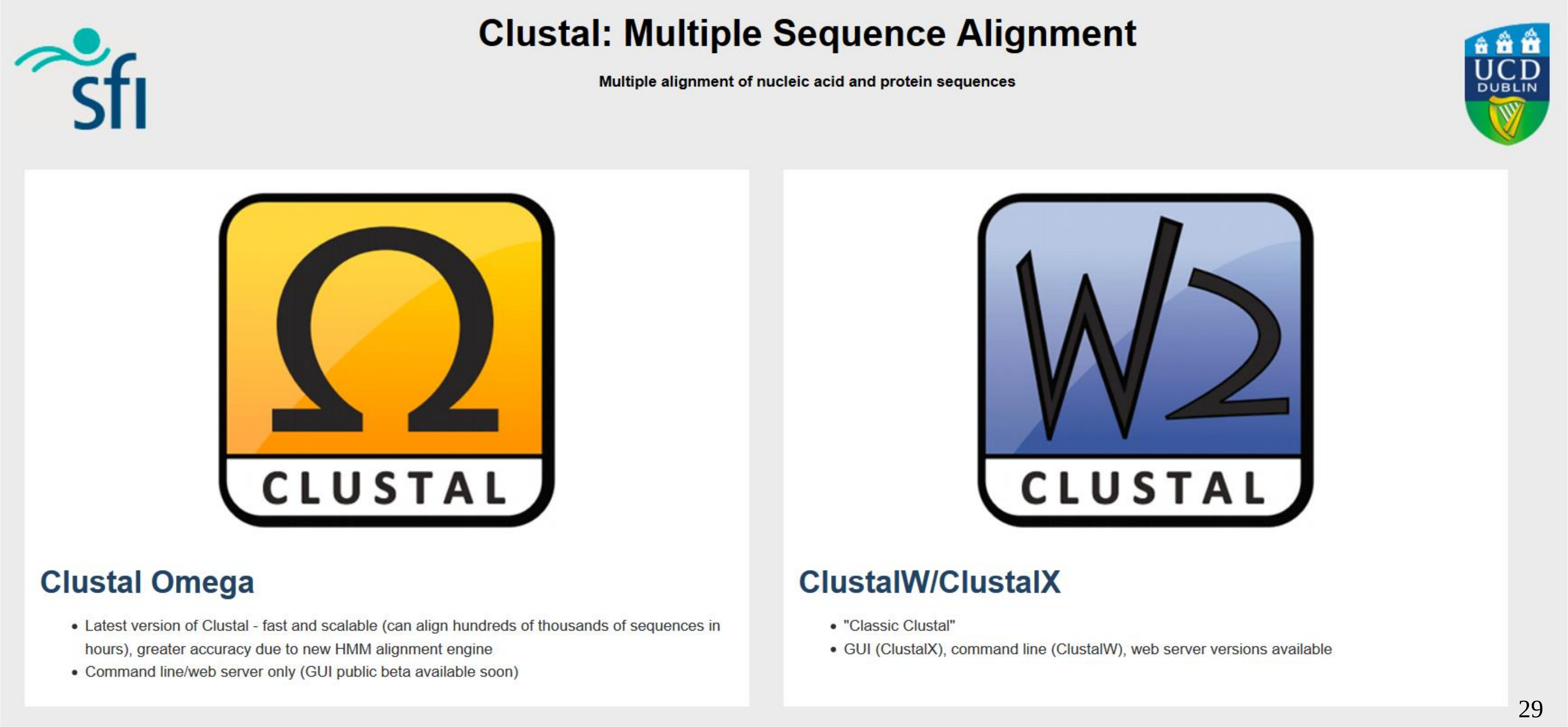
↓

Neighbour-joining tree
(guide tree)

↓

Progressive alignment
following guide tree

Clustal Versions (http://www.clustal.org/)



Clustal: Multiple Sequence Alignment

Multiple alignment of nucleic acid and protein sequences

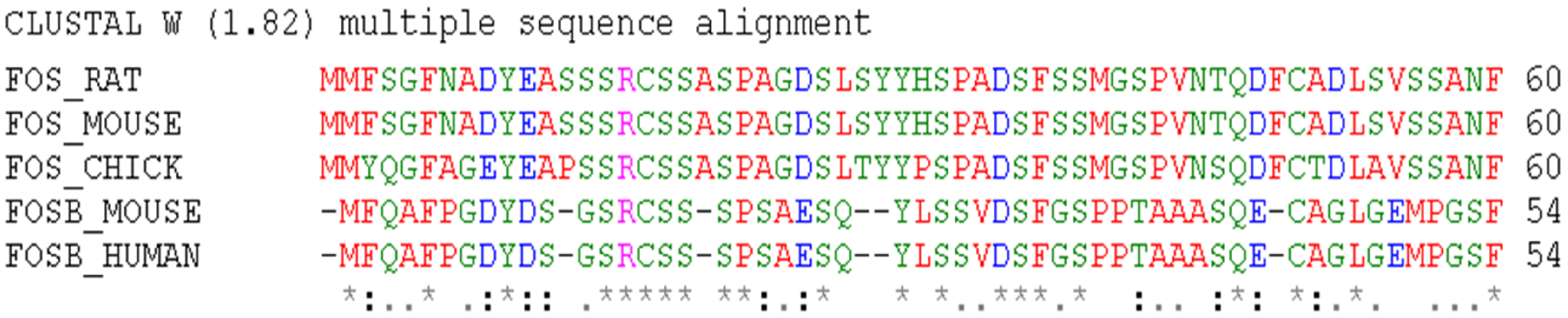
Clustal Omega

- Latest version of Clustal - fast and scalable (can align hundreds of thousands of sequences in hours), greater accuracy due to new HMM alignment engine
- Command line/web server only (GUI public beta available soon)

ClustalW/ClustalX

- "Classic Clustal"
- GUI (ClustalX), command line (ClustalW), web server versions available

Displaying MSAs using ClustalW



```
CLUSTAL W (1.82) multiple sequence alignment

FOS_RAT         MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_MOUSE       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_CHICK       MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF 60
FOSB_MOUSE      -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
FOSB_HUMAN      -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
                 *:..* .:*:: .***** **:.:*   * *..***.*  :.. :*: *:.*.  ...*
```

RED: AVFPMILW (small)

BLUE: DE (acidic, negative chg)

MAGENTA: RHK (basic, positive chg)

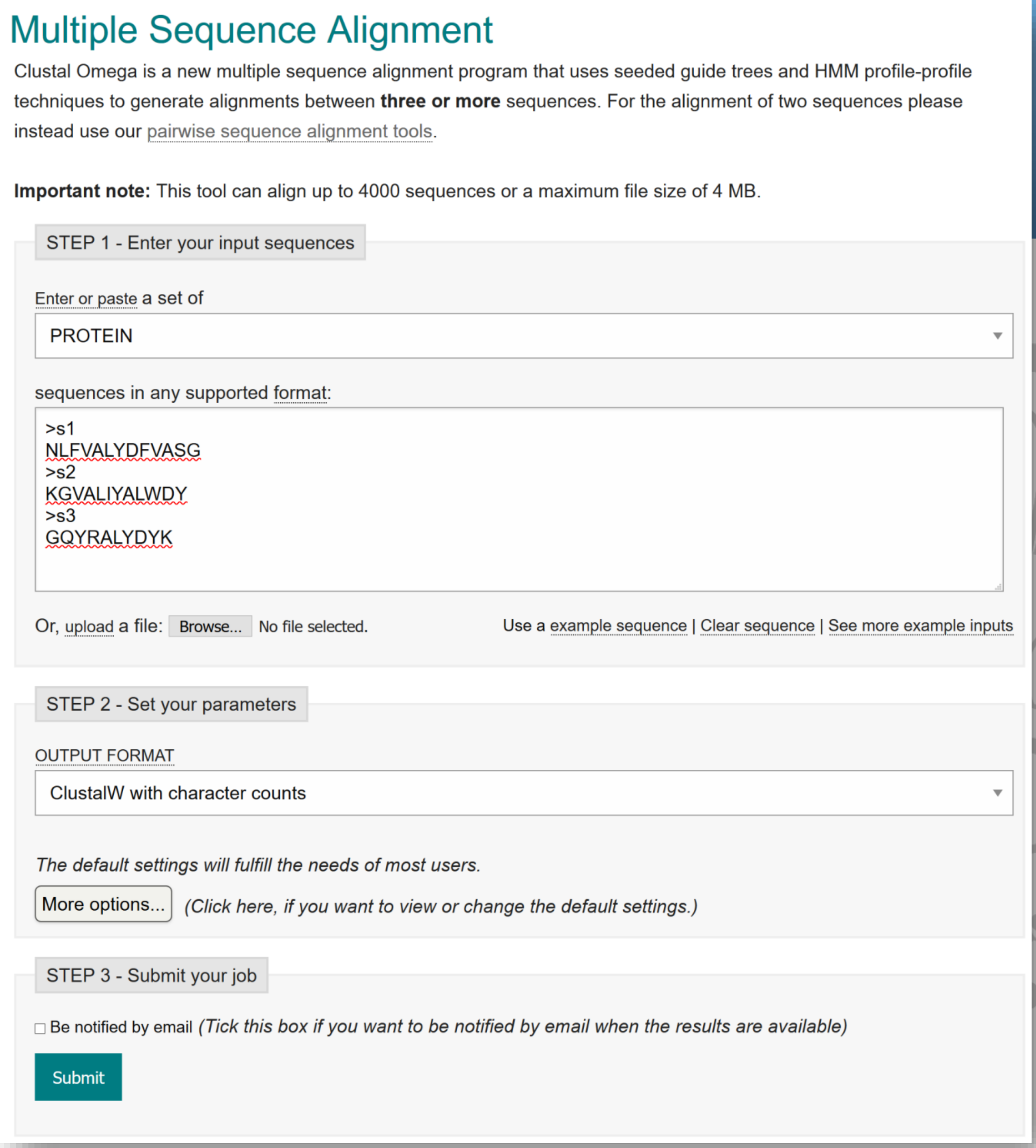
GREEN: STYHCNGQ (hydroxyl + amine + basic)

* entirely conserved column

: all residues have ~ same size *AND* hydropathy

. all residues have ~ same size *OR* hydropathy

Clustal Omega



Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

```
>s1
NLFVALYDFVASG
>s2
KGVALIYALWDY
>s3
GQYRALYDYK
```

Or, upload a file: Browse... No file selected.

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

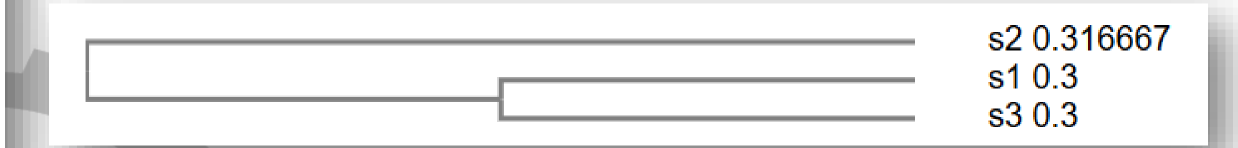
The default settings will fulfill the needs of most users.

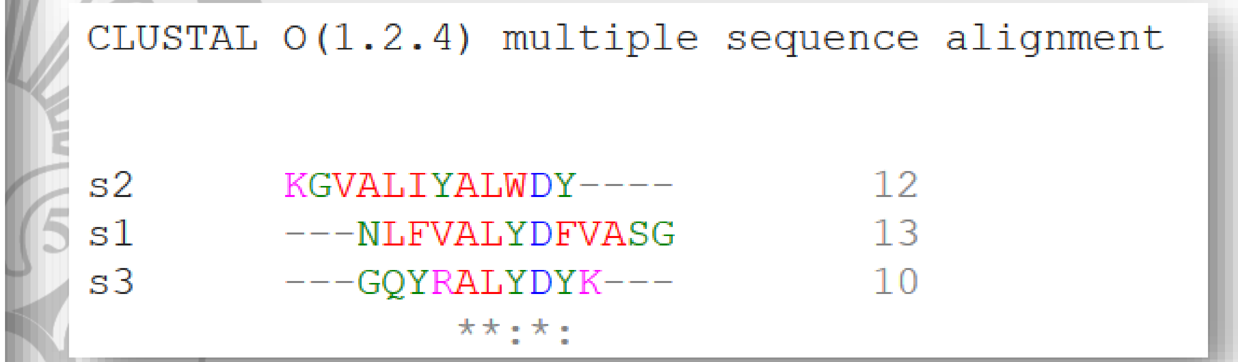
More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit





```
CLUSTAL O(1.2.4) multiple sequence alignment


s2      KGVALIYALWDY----      12
s1      ---NLFVALYDFVASG      13
s3      ---GQYRALYDYK---      10
               **:*:
```

Clustal Drawbacks

- Is not suitable for comparing sequences of different lengths because it is a global alignment–based method.
- The final alignment result is influenced by the order of sequence addition.
- The “greedy” nature of the algorithm: it depends on initial pairwise alignment.
- Any errors made in first steps cannot be corrected.
- To alleviate some of the limitations, a new generation of algorithms have been developed.

Star Alignment

- Fast heuristic to compute MSA
- Good approximation of *optimal* MSA, if scoring scheme satisfies triangle inequality

Algorithm:

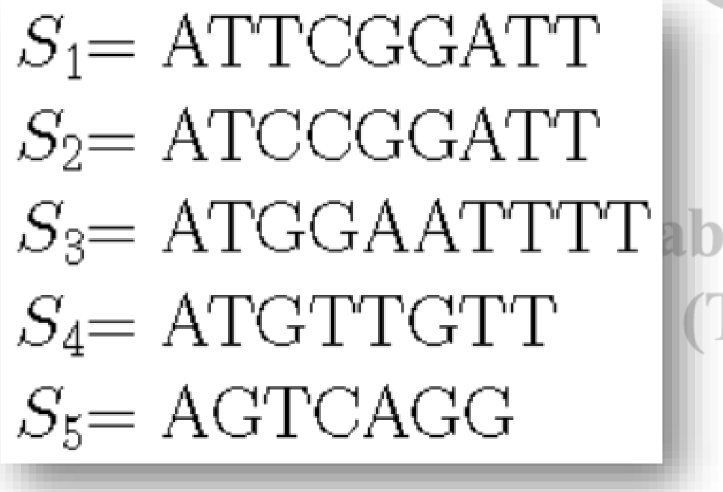
1. ***Compute pairwise similarities***
2. ***Select center s_c*** that maximizes $\Sigma_{i \neq c}$ S(s_c,s_i)
3. ***Add sequences in decreasing order of similarity*** to *center* s_c
   – Rule: “once a gap, always a gap”

Step 2 - Select Center

Does that function look familiar?

- ***Recall: Consensus sequence:*** single sequence (*more accurately*; "model") that represents *most common* residue of each column in MSA.
- ***Recall: Steiner consensus sequence:*** Given sequences $s_1, \ldots, s_k$, find a sequence s* that maximizes $\Sigma_i\ S(s^*, s_i)$

```
FGGHL-GF
F-GHLPGF
FGGHP-FG
________
FGGHL-GF
```



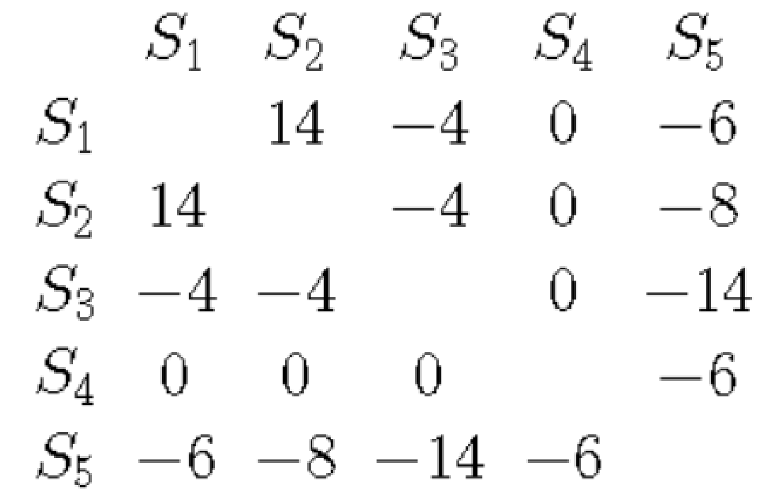
S_1= ATTCGGATT

S_2= ATCCGGATT

S_3= ATGGAATTTT

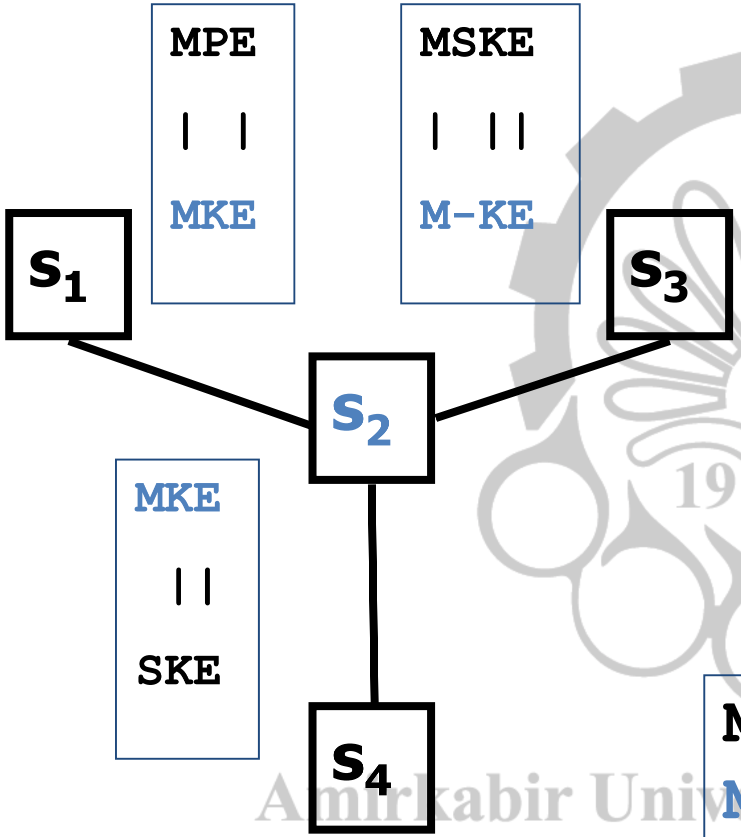
S_4= ATGTTGTT

S_5= AGTCAGG

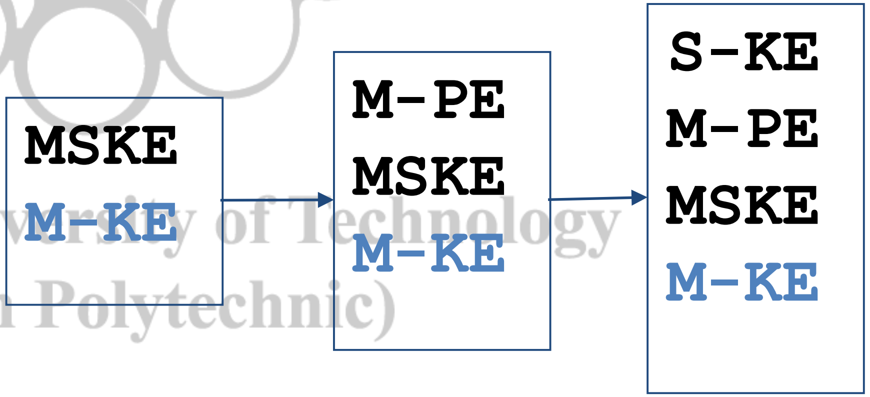


	S_1	S_2	S_3	S_4	S_5
S_1		14	−4	0	−6
S_2	14		−4	0	−8
S_3	−4	−4		0	−14
S_4	0	0	0		−6
S_5	−6	−8	−14	−6	

Step 3 - Add sequences in decreasing order

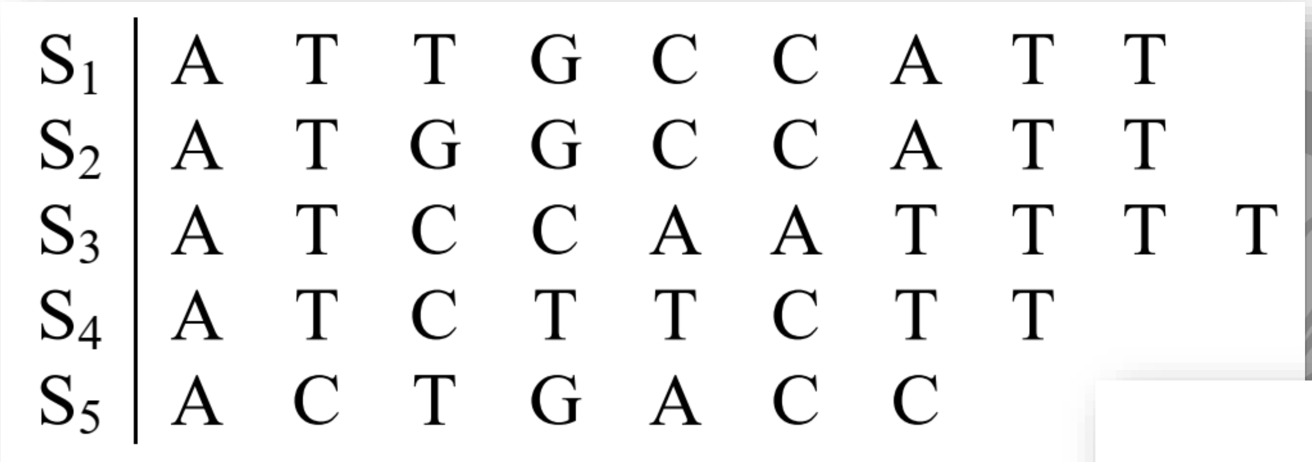


s_1:	MPE
s_2:	MKE
s_3:	MSKE
s_4:	SKE

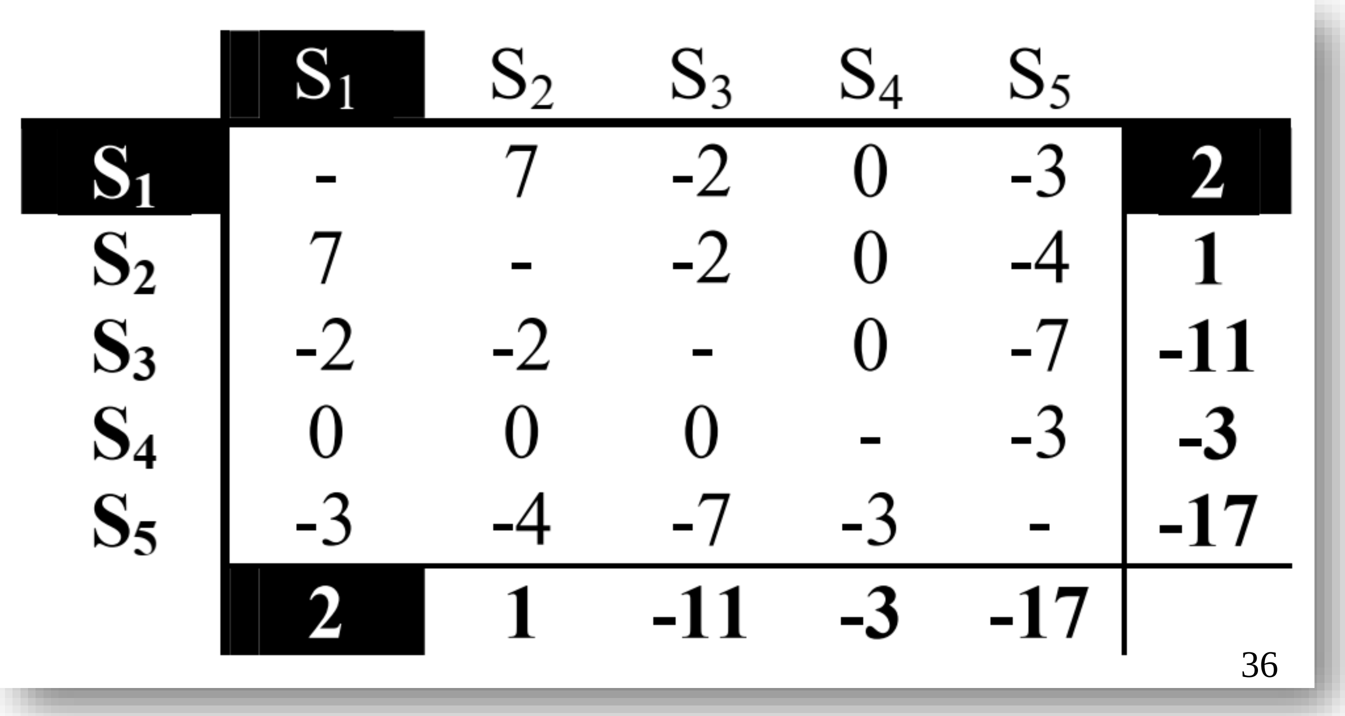


S_2+S_3 +S_1 +S_4

Star-Alignment Example

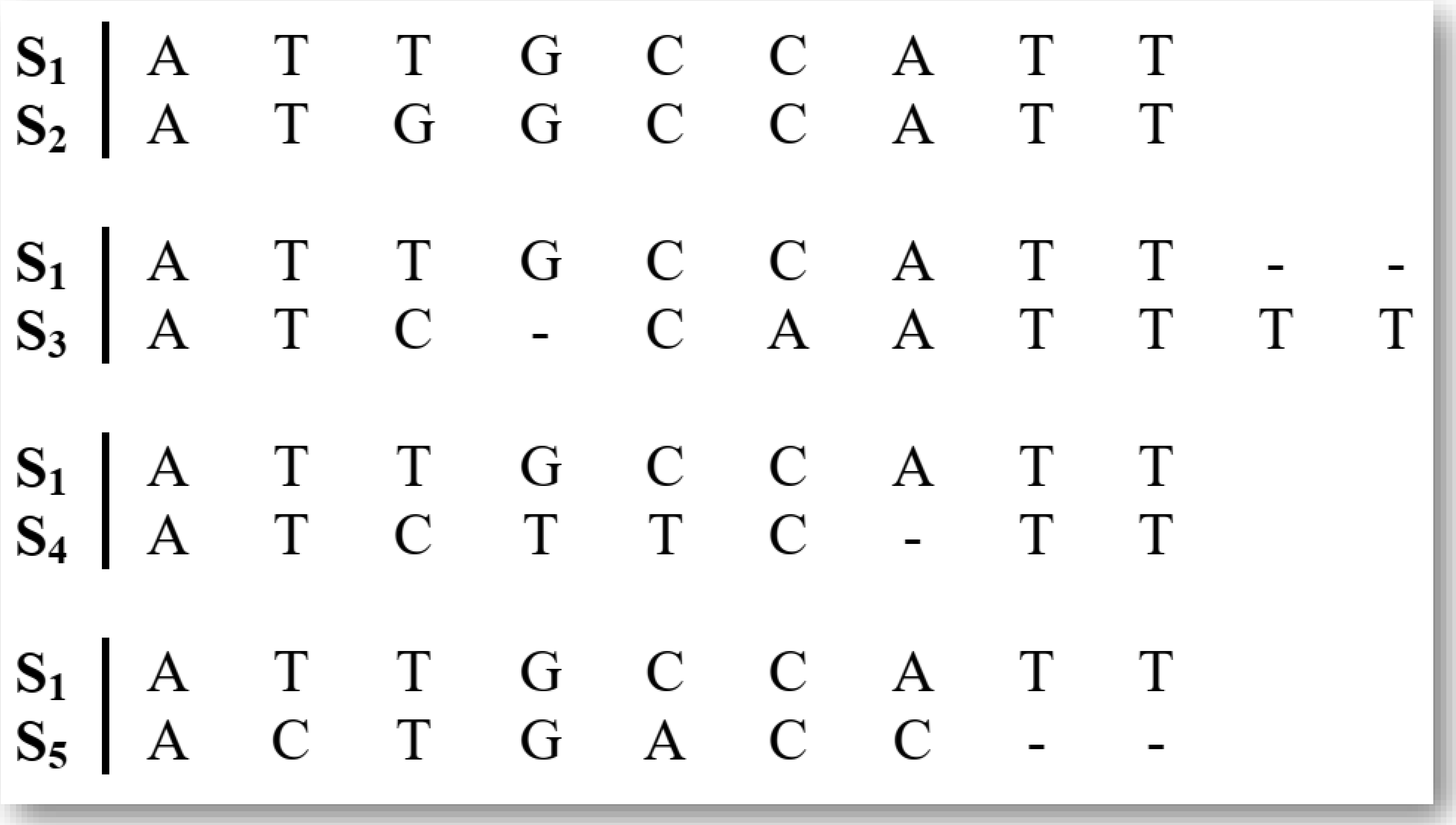


S_1	A	T	T	G	C	C	A	T	T	
S_2	A	T	G	G	C	C	A	T	T	
S_3	A	T	C	C	A	A	T	T	T	T
S_4	A	T	C	T	T	C	T	T		
S_5	A	C	T	G	A	C	C			



	S_1	S_2	S_3	S_4	S_5	
S_1	-	7	-2	0	-3	**2**
S_2	7	-	-2	0	-4	**1**
S_3	-2	-2	-	0	-7	**-11**
S_4	0	0	0	-	-3	**-3**
S_5	-3	-4	-7	-3	-	**-17**
	2	**1**	**-11**	**-3**	**-17**	

Star-Alignment Example (Cont.)



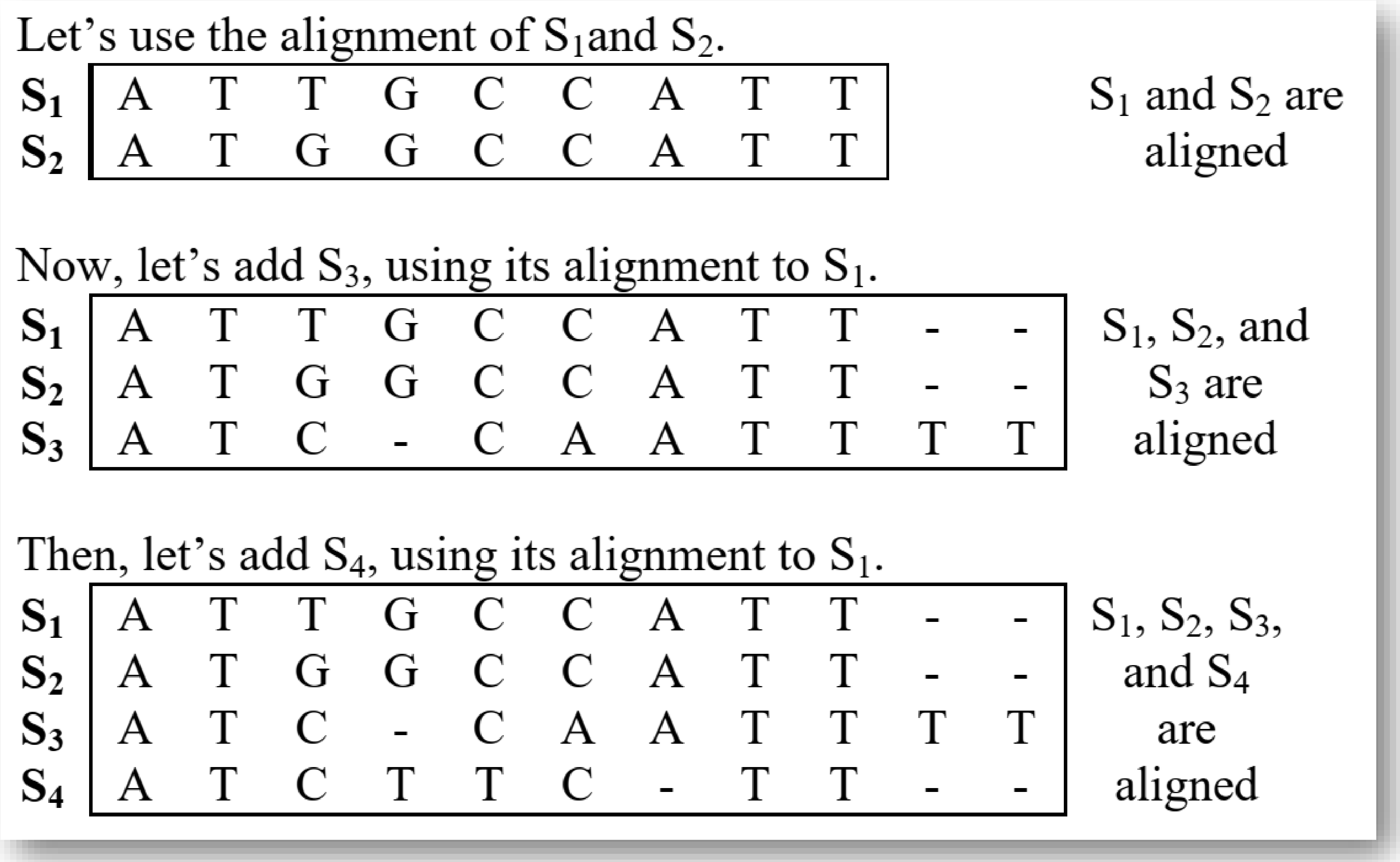
S_1	A	T	T	G	C	C	A	T	T
S_2	A	T	G	G	C	C	A	T	T

S_1	A	T	T	G	C	C	A	T	T	-	-
S_3	A	T	C	-	C	A	A	T	T	T	T

S_1	A	T	T	G	C	C	A	T	T
S_4	A	T	C	T	T	C	-	T	T

S_1	A	T	T	G	C	C	A	T	T
S_5	A	C	T	G	A	C	C	-	-

Star-Alignment Example (Cont.)



Let’s use the alignment of $S_1$and S_2.

S_1	A	T	T	G	C	C	A	T	T	S_1 and S_2 are
S_2	A	T	G	G	C	C	A	T	T	aligned

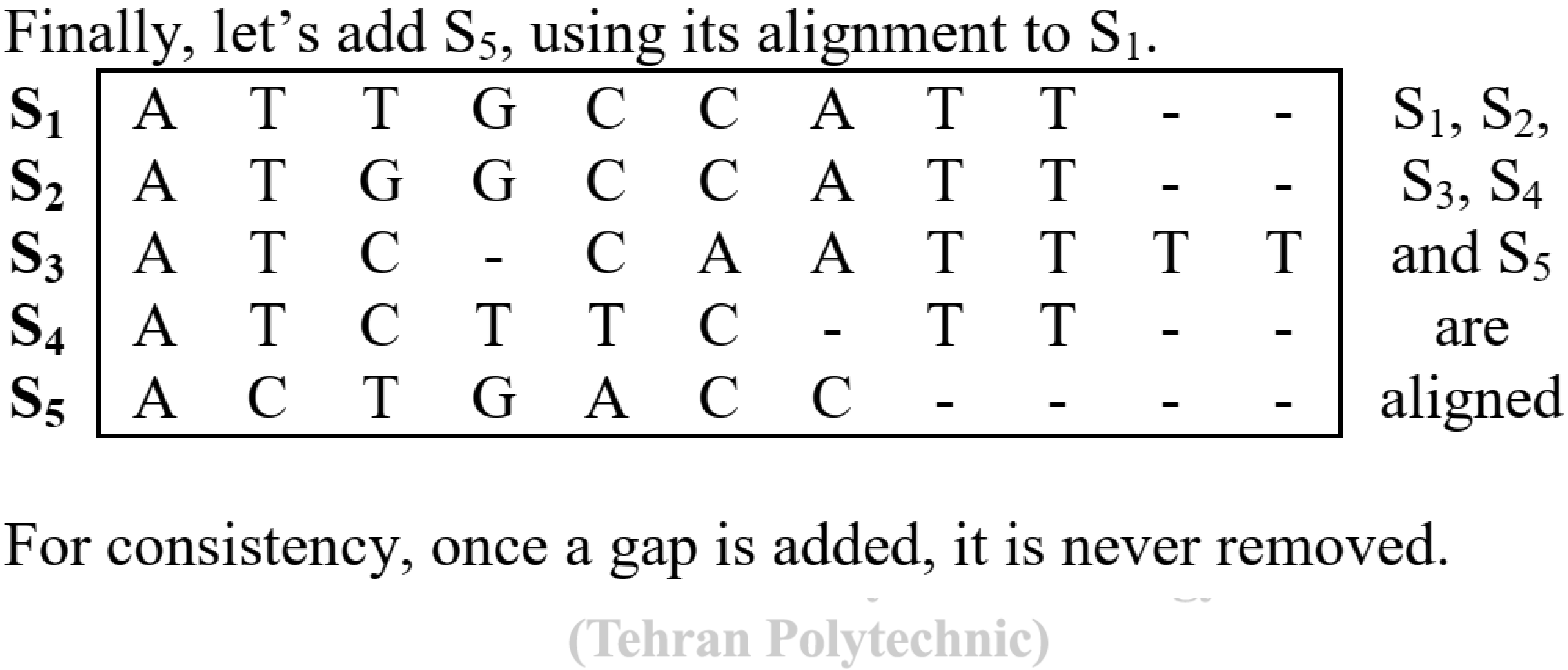
Now, let’s add S_3, using its alignment to S_1.

S_1	A	T	T	G	C	C	A	T	T	-	-	S_1, S_2, and
S_2	A	T	G	G	C	C	A	T	T	-	-	S_3 are
S_3	A	T	C	-	C	A	A	T	T	T	T	aligned

Then, let’s add S_4, using its alignment to S_1.

S_1	A	T	T	G	C	C	A	T	T	-	-	S_1, S_2, S_3,
S_2	A	T	G	G	C	C	A	T	T	-	-	and S_4
S_3	A	T	C	-	C	A	A	T	T	T	T	are
S_4	A	T	C	T	T	C	-	T	T	-	-	aligned

Star-Alignment Example (Cont.)



Finally, let's add S_5, using its alignment to S_1.

$\mathbf{S_1}$	A	T	T	G	C	C	A	T	T	-	-
$\mathbf{S_2}$	A	T	G	G	C	C	A	T	T	-	-
$\mathbf{S_3}$	A	T	C	-	C	A	A	T	T	T	T
$\mathbf{S_4}$	A	T	C	T	T	C	-	T	T	-	-
$\mathbf{S_5}$	A	C	T	G	A	C	C	-	-	-	-

S_1, S_2, S_3, S_4 and S_5 are aligned

For consistency, once a gap is added, it is never removed.

(Tehran Polytechnic)

Complexity of Star Alignment?

Given *k* sequences of length *n*, and an upper bound *l* for alignment length. We need:

- $O(k^2n^2)$ to compute the alignments
- $O(k^2)$ to compute the center
- $O(k^2l)$ to build multiple alignment

Overall: **$O(k^2n^2)$**

Is this really much better than **$O(k^22^kn^k)$?**

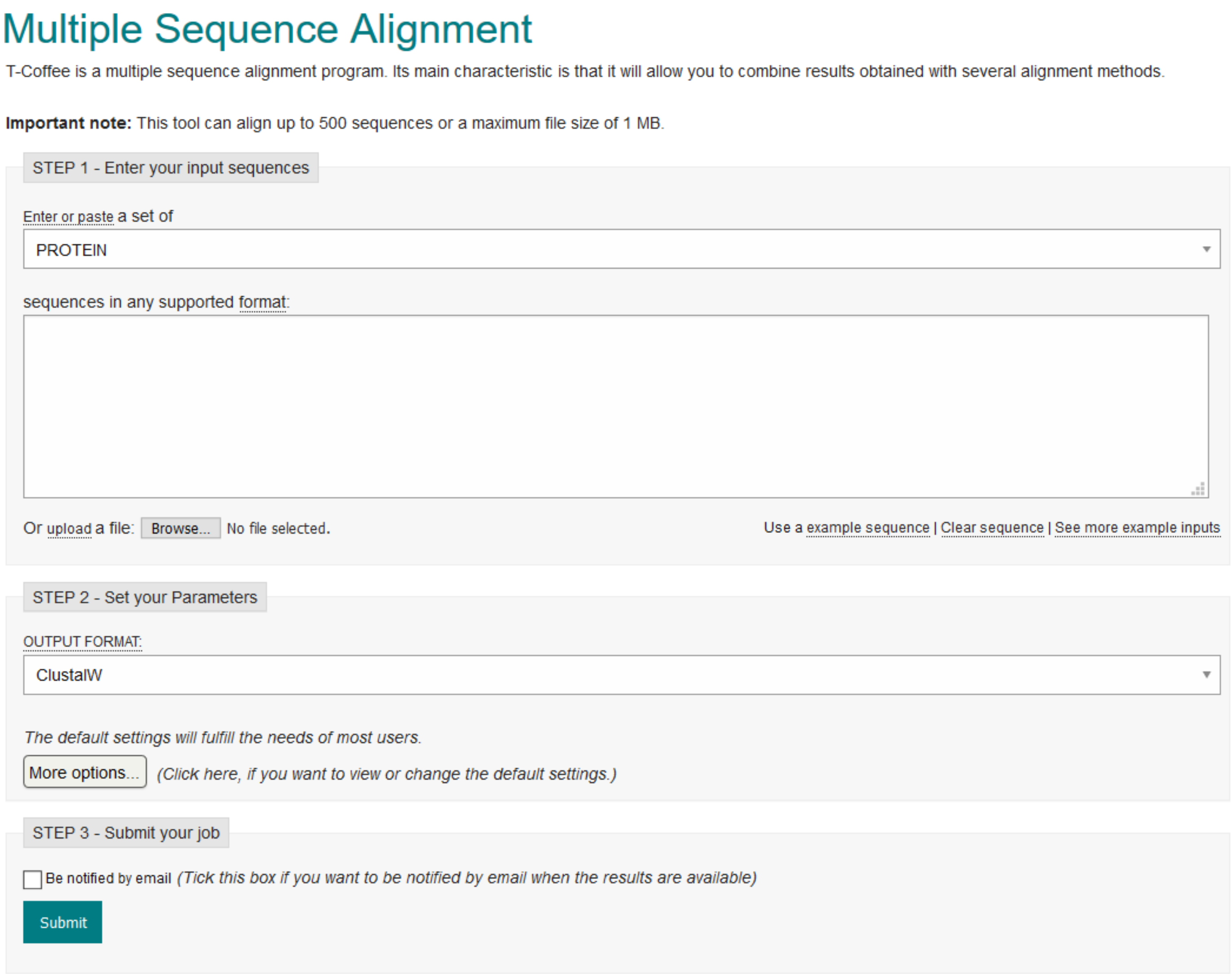
YES! *Remember: k = # of sequences*

n = length of sequences

T-Coffee Program

- T-Coffee (Tree-based Consistency Objective Function for alignment Evaluation) performs progressive sequence alignments as in Clustal.
  - www.ch.embnet.org/software/TCoffee.html
- The main difference: T-Coffee performs both global and local pairwise alignment for all possible pairs involved.
- The global pairwise alignment is performed using the Clustal program.
- The local pairwise alignment is generated by the ***Lalign*** program
  - The top ten scored alignments are selected.
- T-Coffee outperforms Clustal when aligning moderately divergent sequences. However, it is also slower than Clustal.

https://www.ebi.ac.uk/Tools/msa/tcoffee/



Multiple Sequence Alignment

T-Coffee is a multiple sequence alignment program. Its main characteristic is that it will allow you to combine results obtained with several alignment methods.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or upload a file: Browse... No file selected.

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your Parameters

OUTPUT FORMAT:

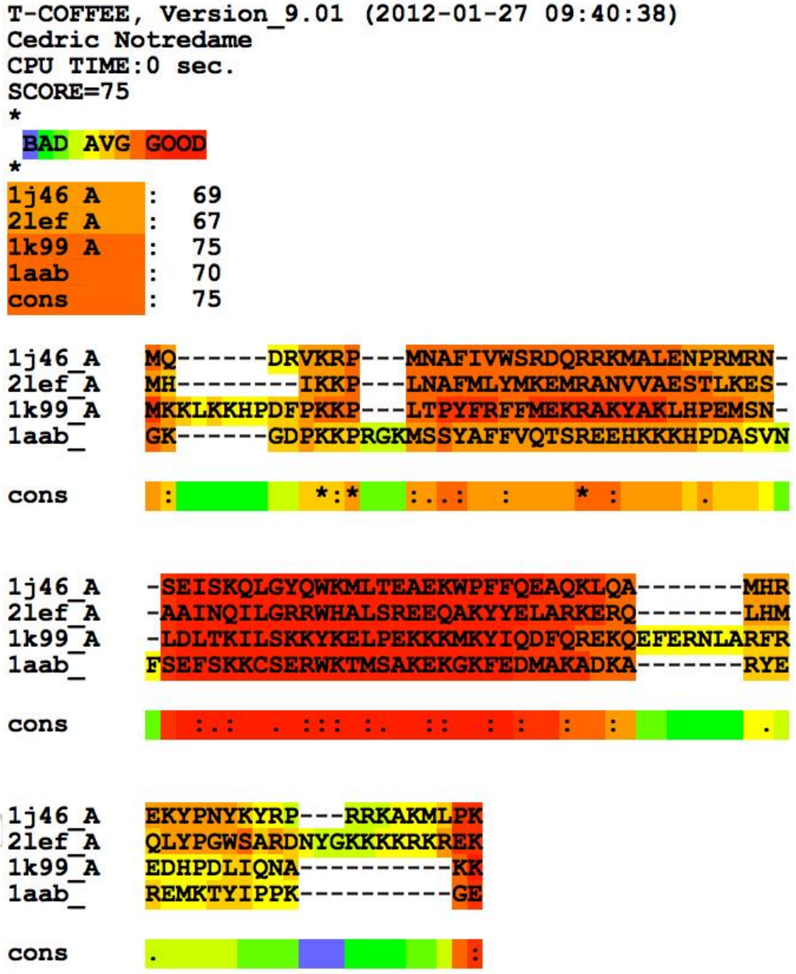
ClustalW

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

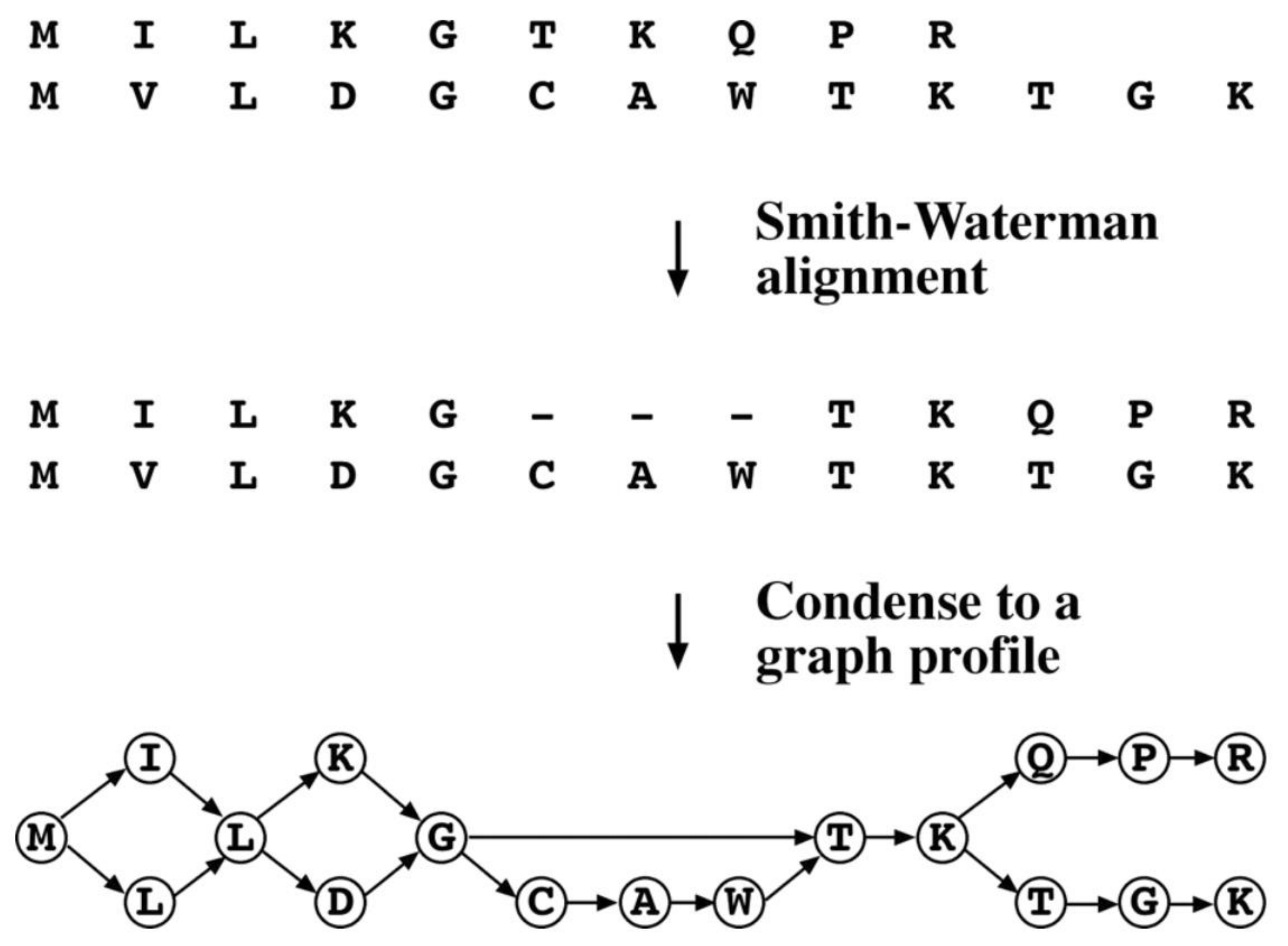


```
T-COFFEE, Version_9.01 (2012-01-27 09:40:38)
Cedric Notredame
CPU TIME:0 sec.
SCORE=75
*
 BAD AVG GOOD
*
1j46 A     :   69
2lef A     :   67
1k99 A     :   75
1aab       :   70
cons       :   75

1j46_A     MQ------DRVKRP---MNAFIVWSRDQRRKMALENPRMRN-
2lef_A     MH--------IKKP---LNAFMLYMKEMRANVVAESTLKES-
1k99_A     MKKLKKHPDFPKKP---LTPYFRFFMEKRAKYAKLHPEMSN-
1aab_      GK------GDPKKPRGKMSSYAFFVQTSREEHKKKHPDASVN

cons        :         *:*   :..:  :    * :     .


1j46_A     -SEISKQLGYQWKMLTEAEKWPFFQEAQKLQA-------MHR
2lef_A     -AAINQILGRRWHALSREEQAKYYELARKERQ-------LHM
1k99_A     -LDLTKILSKKYKELPEKKKMKYIQDFQREKQEFERNLARFR
1aab_      FSEFSKKCSERWKTMSAKEKGKFEDMAKADKA-------RYE

cons          :.:  . ::: :.  ::  : :  :  :         .


1j46_A     EKYPNYKYRP---RRKAKMLPK
2lef_A     QLYPGWSARDNYGKKKKRKREK
1k99_A     EDHPDLIQNA----------KK
1aab_      REMKTYIPPK----------GE

cons       .                    :
```

POA (Partial Order Alignments)

- POA is a progressive alignment program that does not rely on guide trees.
  - The multiple alignment is assembled by adding sequences in the order they are given.
  - A partial order graph is used to represent a growing multiple alignment.
  - www.bioinformatics.ucla.edu/poa/
- Each time a new sequence is added, it is aligned with every sequence within the partial order graph individually using the Smith–Waterman algorithm.
- POA is local alignment-based and has been shown to produce **more accurate** alignments than Clustal. It is also **faster** than Clustal.

POA



```
M   I   L   K   G   T   K   Q   P   R
M   V   L   D   G   C   A   W   T   K   T   G   K
```

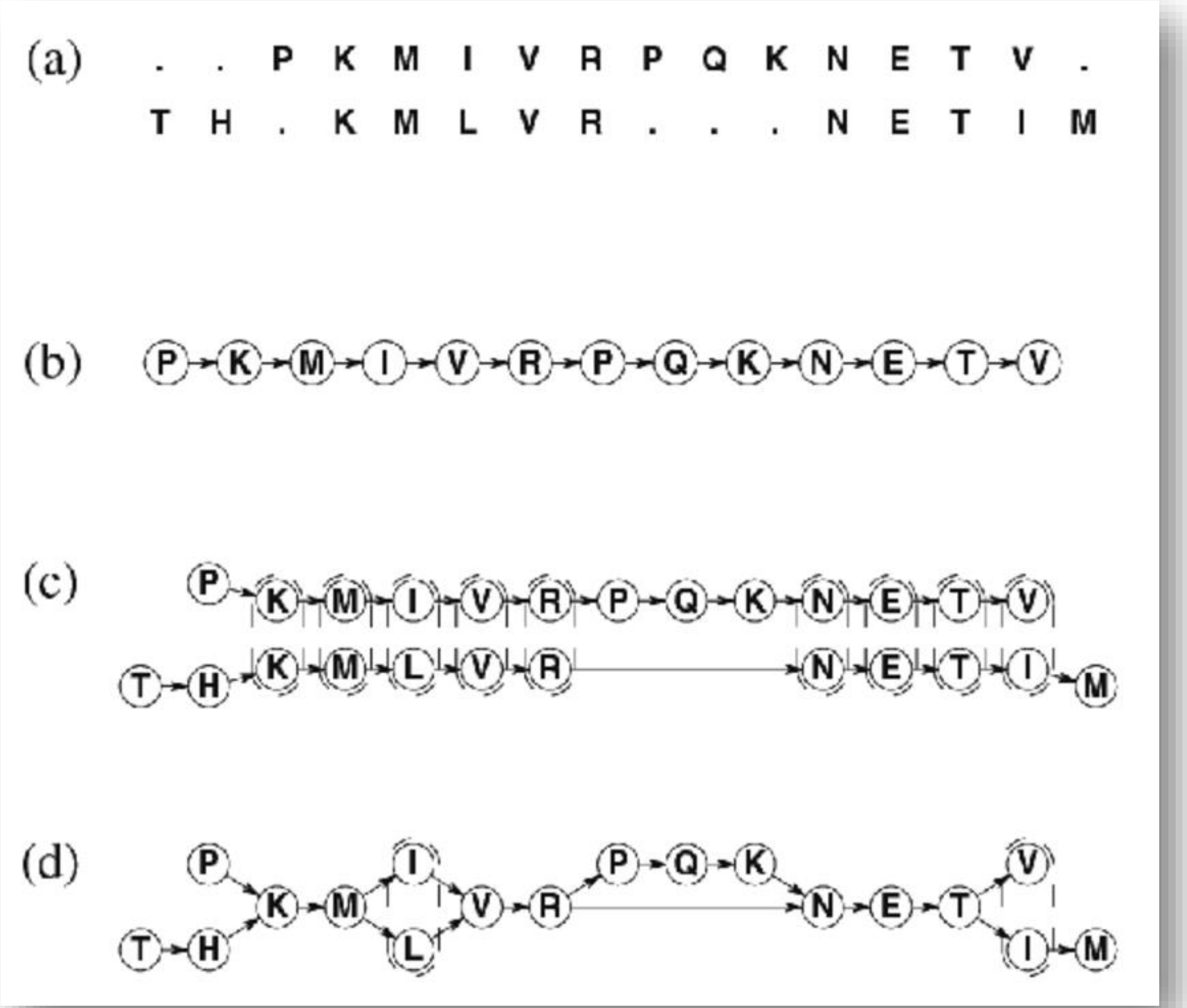
Smith-Waterman alignment

```
M   I   L   K   G   -   -   -   T   K   Q   P   R
M   V   L   D   G   C   A   W   T   K   T   G   K
```

Condense to a graph profile

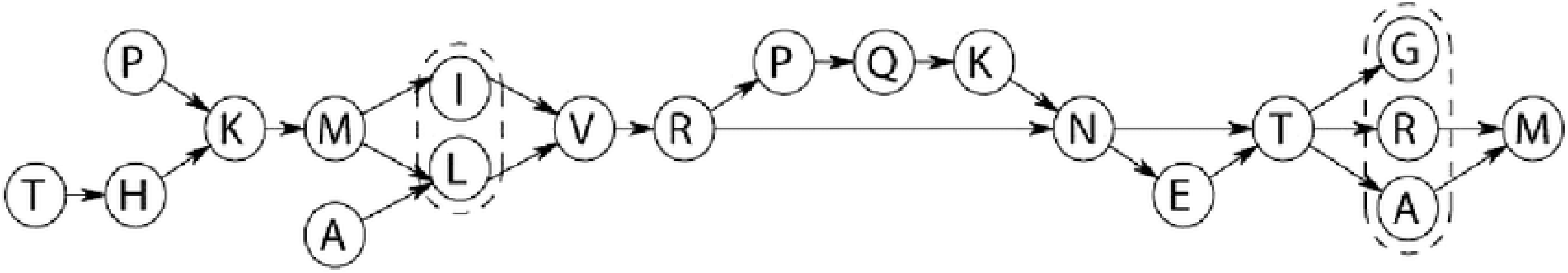
- Conversion of a sequence alignment into a graphical profile in the POA algorithm. Identical residues in the alignment are condensed as nodes in the partial order graph.

POA Example


(a)
. . P K M I V R P Q K N E T V .
T H . K M L V R . . . N E T I M
(b)
P K M I V R P Q K N E T V
(c)
P K M I V R P Q K N E T V
T H K M L V R N E T I M
(d)
P K M I V R P Q K N E T V
T H L I M

POA Example 2

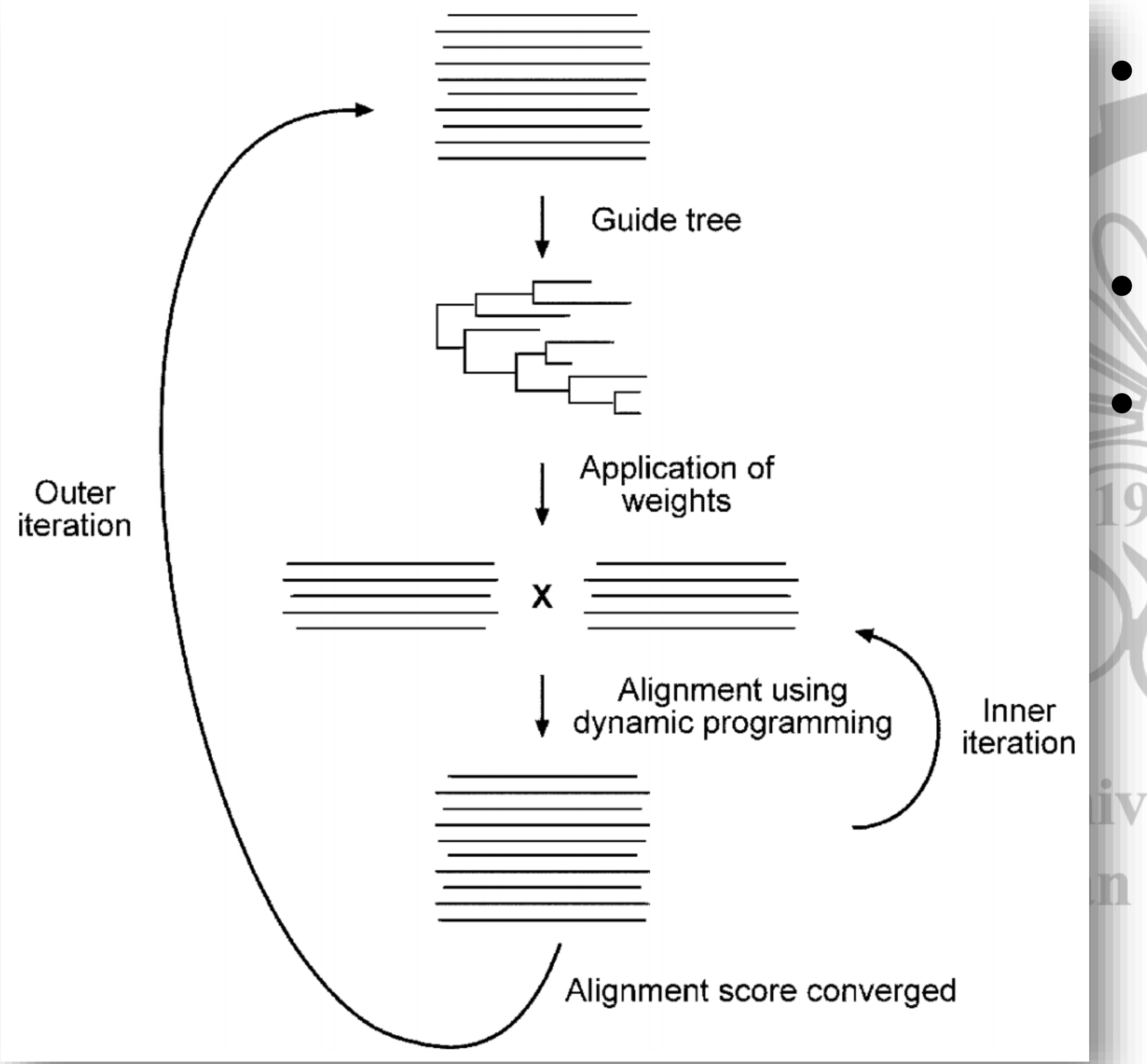
.	.	P	K	M	.	I	V	R	P	Q	K	N	E	T	G	.
.	.	.	.	.	A	L	V	R	P	Q	K	N	.	T	R	M
T	H	.	K	M	.	L	V	R	.	.	.	N	E	T	A	M



Iterative Alignment

- Idea: an optimal solution can be found by repeatedly modifying existing suboptimal solutions.
- The procedure starts by producing a low-quality alignment and gradually improves it by iterative realignment.
- The method may reduce the “greedy” problem of the progressive strategy:
  - Because the order of the sequences used for alignment is different in each iteration.
- The method is also heuristic in nature and does not have guarantees for finding the optimal alignment.

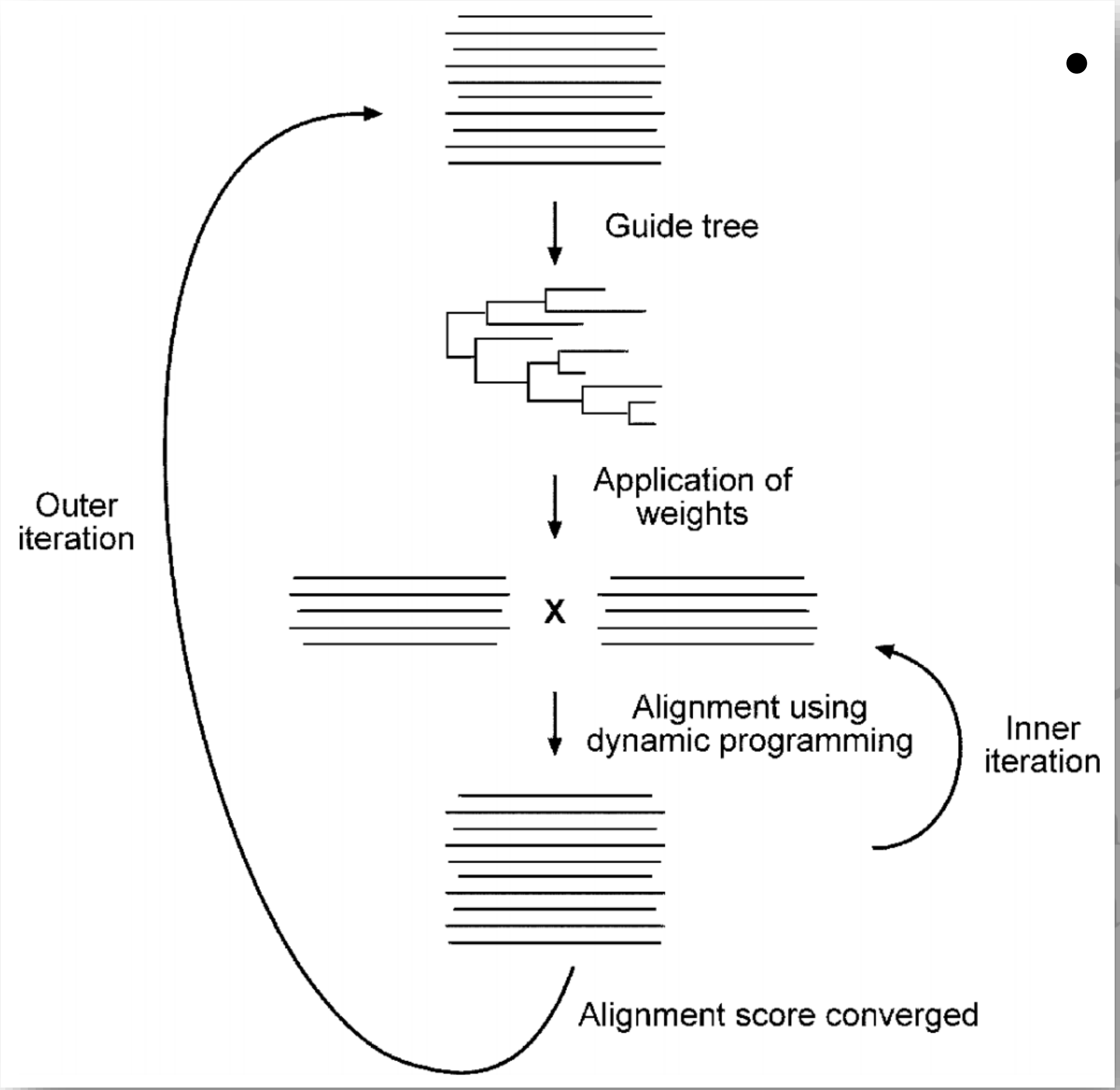
PRRN



- Uses a double nested iterative strategy.
- Two sets of iterations:
- *Outer iteration:*
  - An initial random alignment is generated that is used to derive a guide tree.
  - Weights are subsequently applied to optimize the alignment.

http://prrn.ims.u-tokyo.ac.jp/

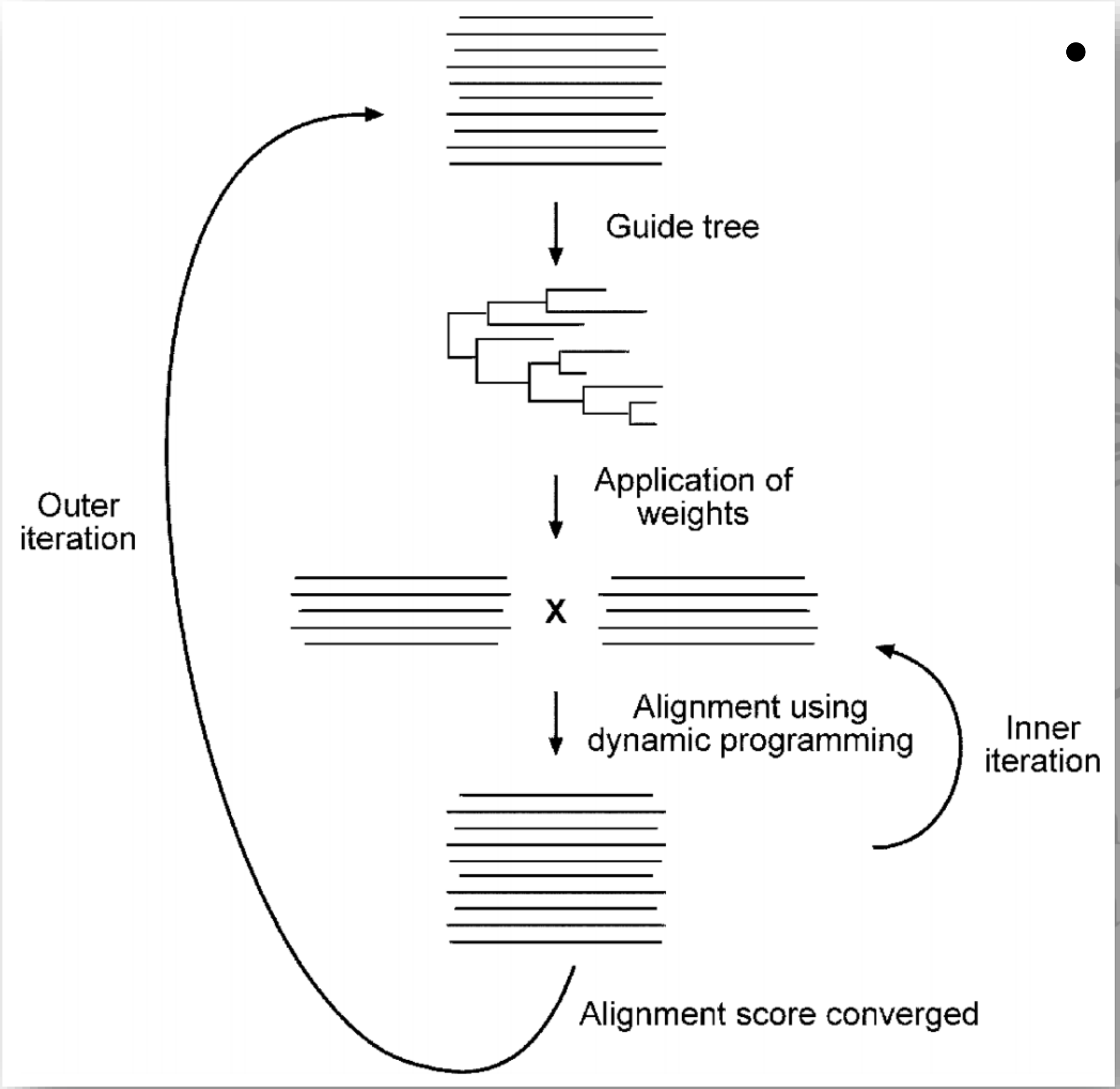
PRRN (Cont.)



- *Inner iteration*:
  - The sequences are randomly divided into two groups.
  - Randomized alignment is used for each group in the initial cycle, after which the alignment positions in each group are fixed.
  - The two groups, each treated as a single sequence, are then aligned to each other using global dynamic programming.

http://prrn.ims.u-tokyo.ac.jp/

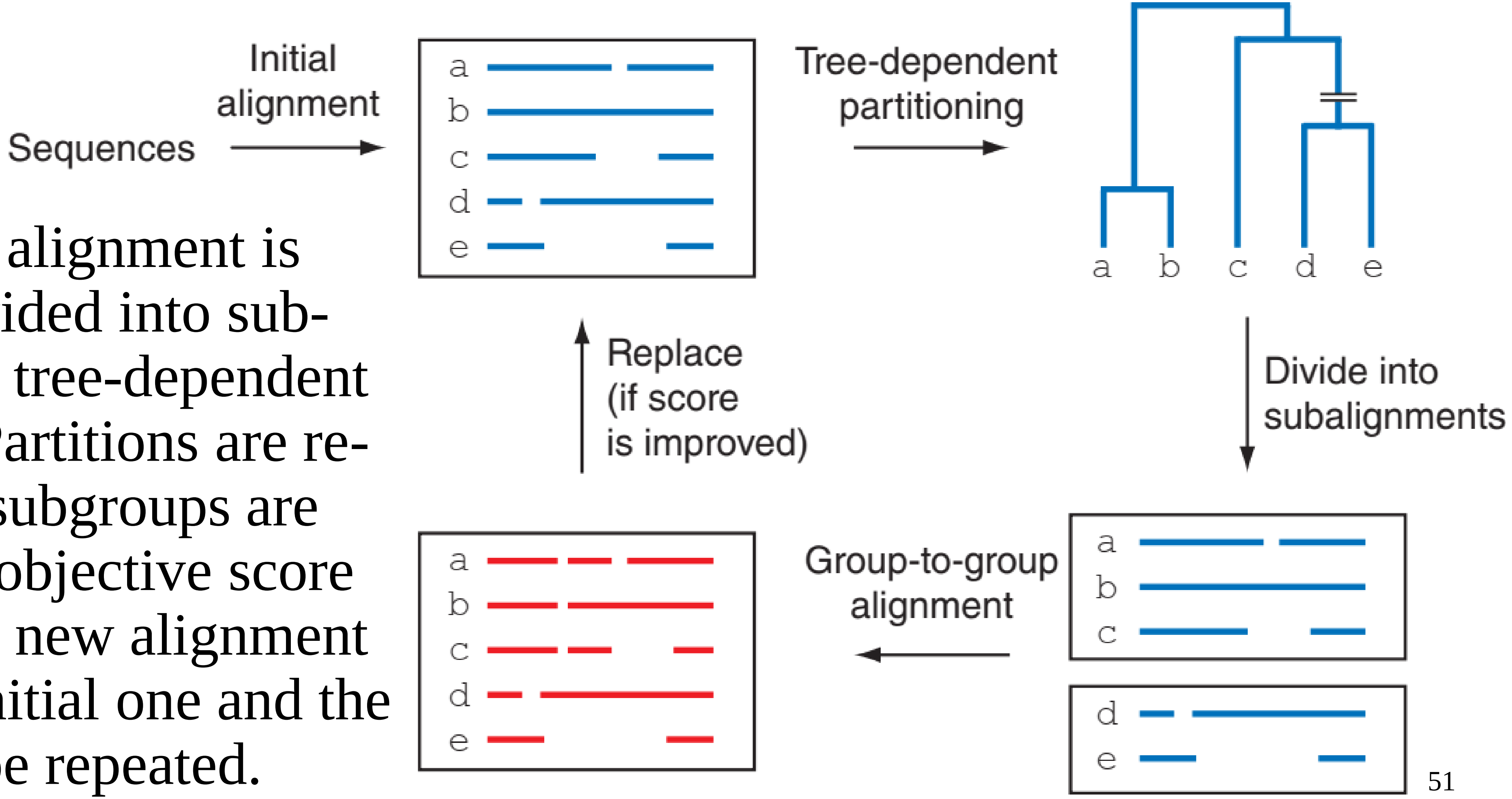
PRRN (Cont.)



- *Inner iteration (Cont.)*:
  - The process is repeated through many cycles until the total SP score no longer increases.
  - At this point, the resulting alignment is used to construct a new guide tree. New weights are applied to optimize alignment scores.
  - The newly optimized alignment is subject to further realignment in the inner iteration.
  - This process is repeated over many cycles until there is no further improvement in the overall alignment scores.

http://prrn.ims.u-tokyo.ac.jp/

MAFFT (Multiple Alignment using FFT)



A progressive alignment is made then divided into sub-alignments by tree-dependent partitioning. Partitions are re-aligned, then subgroups are aligned. If an objective score improves, this new alignment replaces the initial one and the process may be repeated.

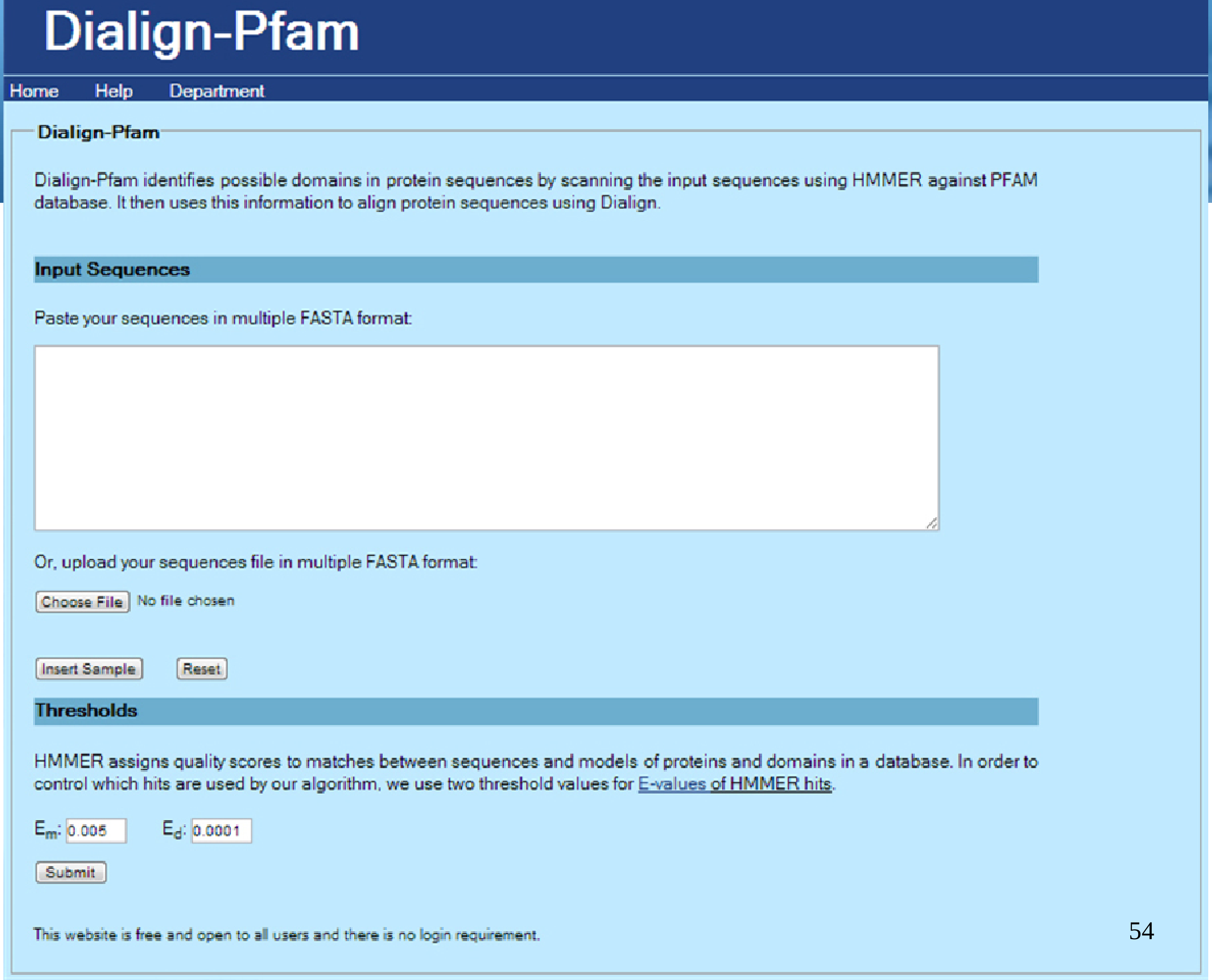
Block-Based Alignment

- The progressive and iterative alignment strategies are largely global alignment based:
  - May fail to recognize conserved domains and motifs among highly divergent sequences of varying lengths.
- For such divergent sequences that share only regional similarities, a local alignment based approach has to be used.
- The strategy identifies a **block of ungapped alignment** shared by all the sequences, hence, the block-based local alignment strategy.

DIALIGN

- It does not apply gap penalties and thus is not sensitive to long gaps.
- The method breaks each of the sequences down to smaller segments and performs all possible pairwise alignments between the segments.
- High-scoring segments, called *blocks*, among different sequences are then compiled in a progressive manner to assemble a full multiple alignment.
- The program has been shown to be especially suitable for aligning *divergent sequences* with only local similarity.

Example



Dialign-Pfam

Dialign-Pfam identifies possible domains in protein sequences by scanning the input sequences using HMMER against PFAM database. It then uses this information to align protein sequences using Dialign.

Input Sequences

Paste your sequences in multiple FASTA format:

Or, upload your sequences file in multiple FASTA format:

Thresholds

HMMER assigns quality scores to matches between sequences and models of proteins and domains in a database. In order to control which hits are used by our algorithm, we use two threshold values for E-values of HMMER hits.

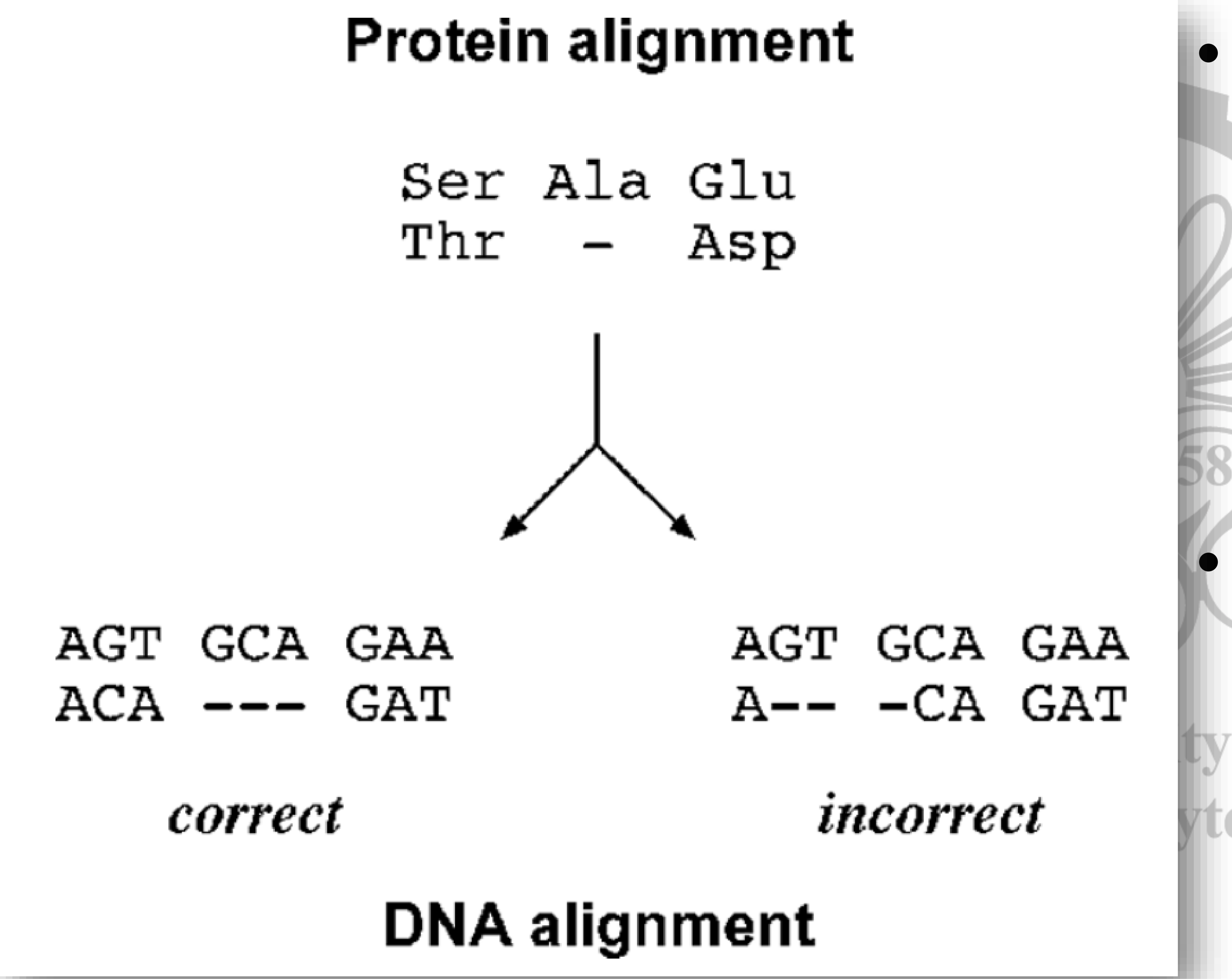
E_m: 0.005 E_d: 0.0001

This website is free and open to all users and there is no login requirement.

Protein-Coding DNA Sequences

- Alignment at the protein level is more sensitive than at the DNA level.
- Sequence alignment directly at the DNA level can often result in **frameshift** errors
  - in DNA alignment gaps are introduced irrespective of codon boundaries.
- In the process of achieving maximum sequence similarity at the DNA level, mismatches of genetic codons occur that violate the *accepted evolutionary scenario*
  - insertions or deletions occur in units of codons
- There are occasions when sequence alignment at the DNA level is often necessary, for example, in constructing DNA-based molecular phylogenetic trees.

Protein-Coding DNA Sequences (Cont.)



Protein alignment

```
Ser Ala Glu
Thr  -  Asp
```

```
AGT GCA GAA        AGT GCA GAA
ACA --- GAT        A-- -CA GAT
```

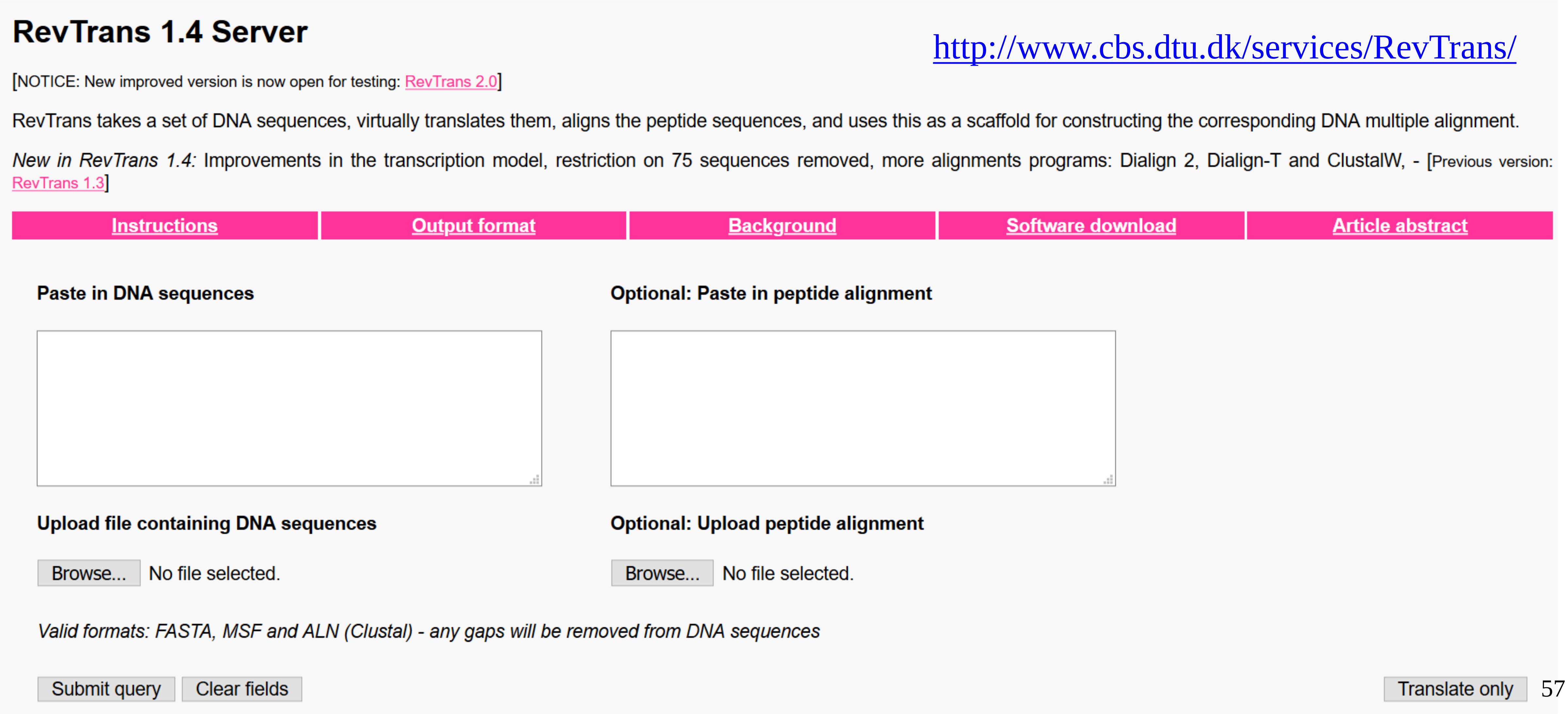
correct *incorrect*

DNA alignment

- DNA can be translated into an amino acid sequence before carrying out alignment to avoid the errors of inserting gaps within codon boundaries.
- After alignment of the protein sequences, the alignment can be converted back to DNA alignment.

Example: RevTrans Program

http://www.cbs.dtu.dk/services/RevTrans/



RevTrans 1.4 Server

[NOTICE: New improved version is now open for testing: RevTrans 2.0]

RevTrans takes a set of DNA sequences, virtually translates them, aligns the peptide sequences, and uses this as a scaffold for constructing the corresponding DNA multiple alignment.

New in RevTrans 1.4: Improvements in the transcription model, restriction on 75 sequences removed, more alignments programs: Dialign 2, Dialign-T and ClustalW, - [Previous version: RevTrans 1.3]

Paste in DNA sequences

Optional: Paste in peptide alignment

Upload file containing DNA sequences

Browse... No file selected.

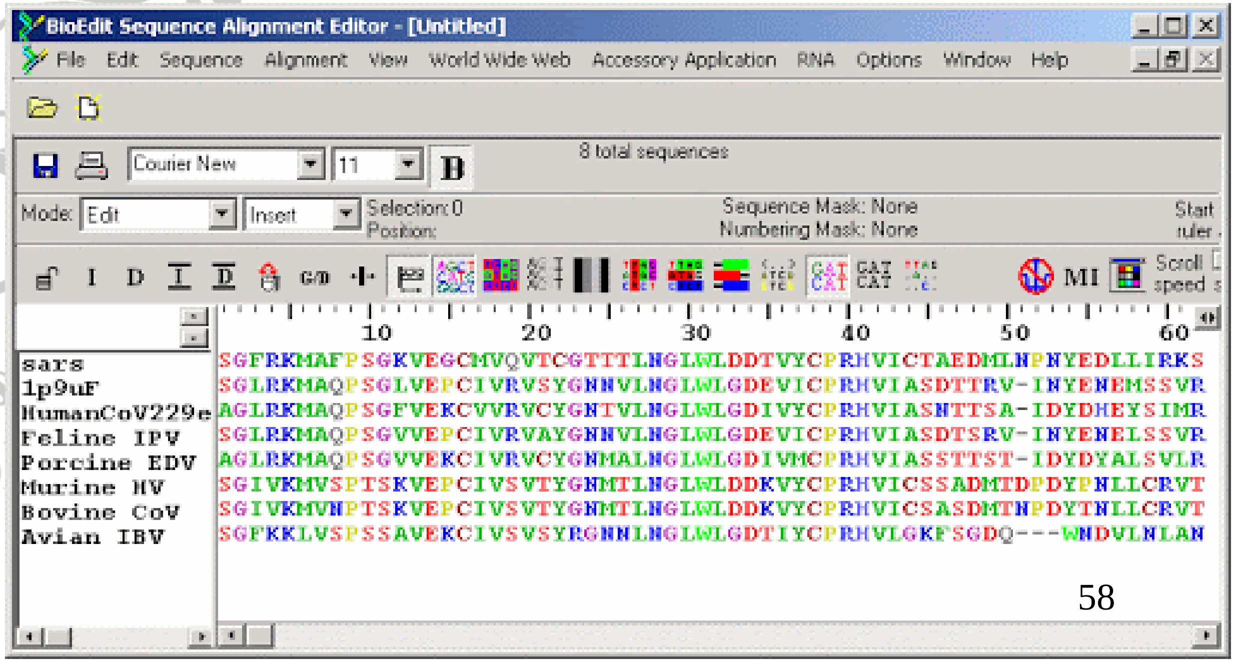
Optional: Upload peptide alignment

Browse... No file selected.

Valid formats: FASTA, MSF and ALN (Clustal) - any gaps will be removed from DNA sequences

Editing Alignments

- The automated alignment often contains misaligned regions.
  - the user should check the alignment carefully for biological relevance and edit the alignment if necessary.
  - This involves introducing or removing gaps to maximize biologically meaningful matches.
- *BioEdit* is a multifunctional sequence alignment editor for Windows.
- *Rascal* is a web-based program that automatically refines a multiple sequence alignment.



Courier New 11

8 total sequences

Mode: Edit Insert

Selection: 0

Position:

Sequence Mask: None

Numbering Mask: None

```
                   10        20        30        40        50        60
sars          SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDTVYCPRHVICTAEDMLNPNYEDLLIRKS
1p9uF         SGLRKMAQPSGLVEPCIVRVSYGNNVLNGLWLGDEVICPRHVIASDTTRV-INYENEMSSVR
HumanCoV229e  AGLRKMAQPSGFVEKCVVRVCYGNTVLNGLWLGDIVYCPRHVIASNTTSA-IDYDHEYSIMR
Feline IPV    SGLRKMAQPSGVVEPCIVRVAYGNNVLNGLWLGDEVICPRHVIASDTSRV-INYENELSSVR
Porcine EDV   AGLRKMAQPSGVVEKCIVRVCYGNMALNGLWLGDIVMCPRHVIASSTTST-IDYDYALSVLR
Murine HV     SGIVKMVSPTSKVEPCIVSVTYGNMTLNGLWLDDKVYCPRHVICSSADMTDPDYPNLLCRVT
Bovine CoV    SGIVKMVNPTSKVEPCIVSVTYGNMTLNGLWLDDKVYCPRHVICSASDMTNPDYTNLLCRVT
Avian IBV     SGFKKLVSPSSAVEKCIVSVSYRGNNLNGLWLGDTIYCPRHVLGKFSGDQ---WNDVLNLAN
```

References

- Mostly used:
  - Essential bioinformatics, Chapter 4 (Multiple Sequence Alignment)
- Second reference:
  - Bioinformatics and functional genomics, Chapter 6 (Multiple Sequence Alignment)

- IP notice: some slides were selected from Drena Dobbs’ slides.

Thanks for your attention