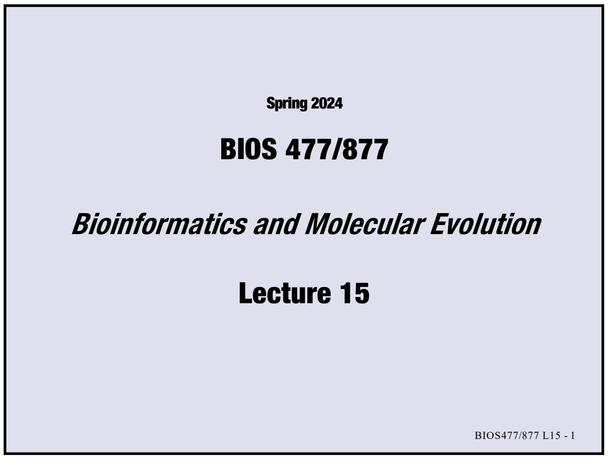


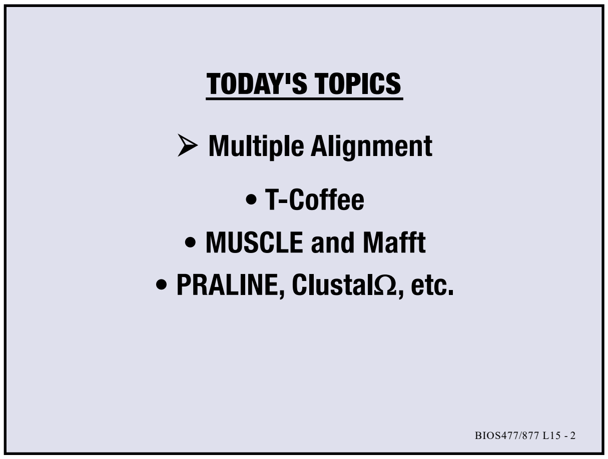
Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 15

1

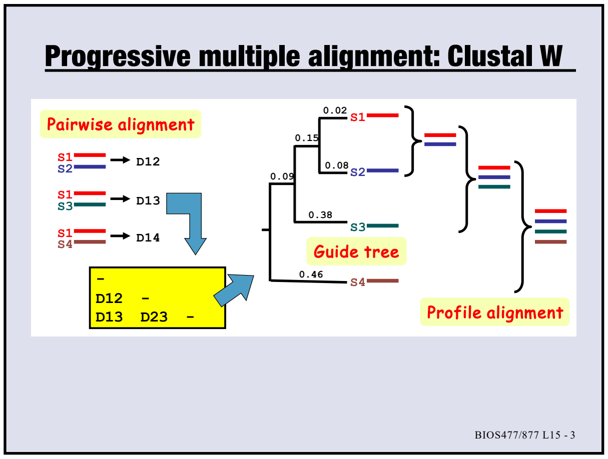


TODAY'S TOPICS

➢ Multiple Alignment

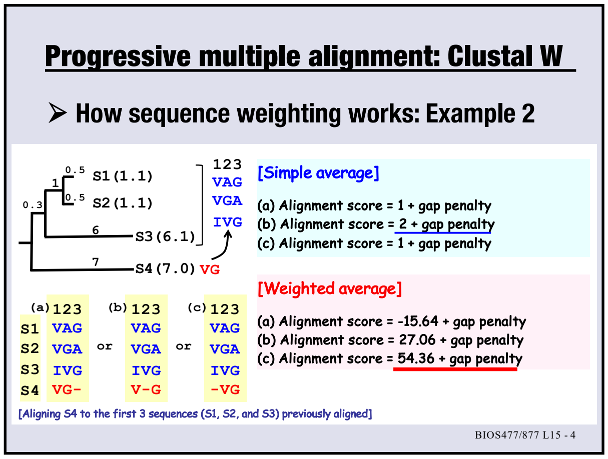
- T-Coffee
- MUSCLE and Mafft
- PRALINE, ClustalΩ, etc.

2



Progressive multiple alignment: Clustal W

3



Progressive multiple alignment: Clustal W

➢ How sequence weighting works: Example 2

[Simple average]

(a) Alignment score = 1 + gap penalty
(b) Alignment score = 2 + gap penalty
(c) Alignment score = 1 + gap penalty

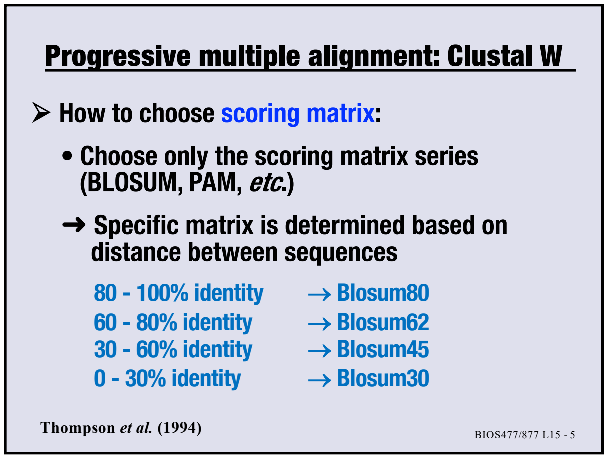
	(a) 123		(b) 123		(c) 123
S1	VAG		VAG		VAG
S2	VGA	or	VGA	or	VGA
S3	IVG		IVG		IVG
S4	VG-		V-G		-VG

[Weighted average]

(a) Alignment score = -15.64 + gap penalty
(b) Alignment score = 27.06 + gap penalty
(c) Alignment score = 54.36 + gap penalty

[Aligning S4 to the first 3 sequences (S1, S2, and S3) previously aligned]

4



Progressive multiple alignment: Clustal W

➢ How to choose scoring matrix:

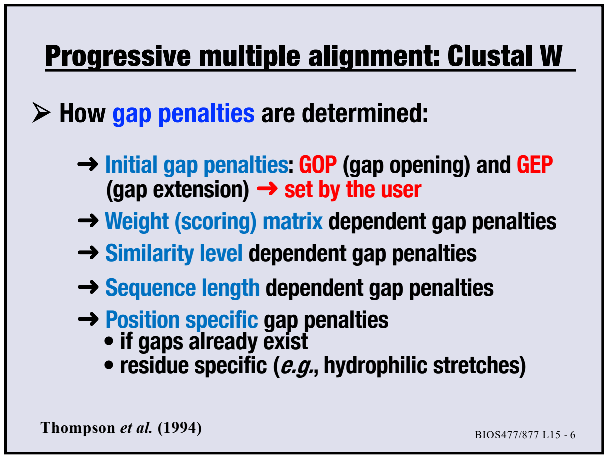
- Choose only the scoring matrix series (BLOSUM, PAM, *etc.*)

➜ Specific matrix is determined based on distance between sequences

80 - 100% identity	→ Blosum80
60 - 80% identity	→ Blosum62
30 - 60% identity	→ Blosum45
0 - 30% identity	→ Blosum30

Thompson *et al.* (1994)

5



Progressive multiple alignment: Clustal W

➢ How gap penalties are determined:

➜ Initial gap penalties: GOP (gap opening) and GEP (gap extension) ➜ set by the user

➜ Weight (scoring) matrix dependent gap penalties

➜ Similarity level dependent gap penalties

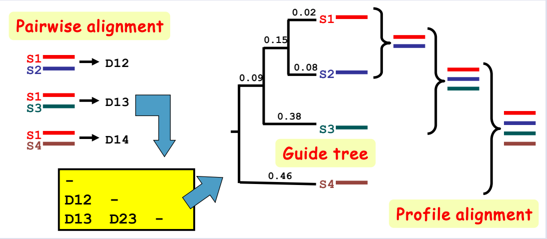
➜ Sequence length dependent gap penalties

➜ Position specific gap penalties
- if gaps already exist
- residue specific (*e.g.*, hydrophilic stretches)

Thompson *et al.* (1994)

6

Progressive multiple alignment: Clustal W

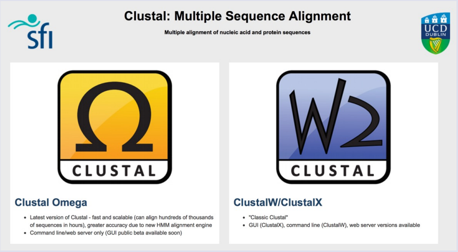


- Progressive alignment
  - ➜ Greedy (finds local optima, but no guarantee for global optima)
  - ➜ Errors (incorrect gap positions) in the early alignments cannot be rectified later
- Global alignment only (local similarity may be missed)

7

Clustal Web servers

http://www.clustal.org/ (Clustal original website)



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

Use Clustal Ω instead of Clustal W

[Classic version]

https://galaxy.pasteur.fr/ (part of Galaxy@Pasteur)

https://www.genome.jp/tools-bin/clustalw

8

To solve the progressive-alignment problems

- ➢ Incorporate more information to reduce early errors
  - Structural alignment (*e.g.*, Expresso, PROMALS3D, TM-Coffee, PRALINE, MAFFT-DASH)
  - Profile/profile-HMM alignment (*e.g.*, PRALINE, PSI-Coffee, PROMALS3D, ProbCons/CONTRAlign, ClustalΩ, MUSCLE5)
- ➢ Avoid the greedy-algorithm problem
  - Iterative refinement to search the global maxima
    - ➜ A good objective function is required (*e.g.*, MUSCLE/MUSCLE5, MAFFT, ProbCons/CONTRAlign)
- ➢ Global (or local) only alignment problem
  - Combine both methods (*e.g.*, T-Coffee)
- ➢ More accurate insertion/deletion placement
  - Phylogeny aware gap-placement (*e.g.*, PRANK, ProPIP, Bali-Phy, SATé)

9

Multiple alignment: T-Coffee

T-Coffee Home page
https://tcoffee.crg.eu/
https://tcoffee.org

https://tcoffee.crg.eu/apps/tcoffee/index.html
(mirror site)

https://www.ebi.ac.uk/jdispatcher/msa/tcoffee
(only T-Coffee, without other associated programs)

Notredame, Higgins & Heringa (2000); Taly *et al.* (2011)

10

Multiple alignment: T-Coffee

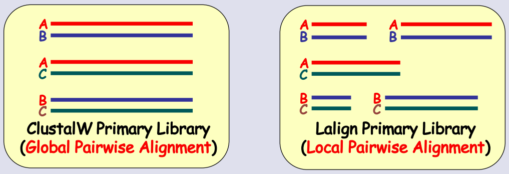
- T-Coffee: Tree-based Consistency Objective Function for alignment Evaluation
- Based on the progressive alignment algorithm
  - ➜ Uses a guide tree, fast
- Tries to avoid the greedy nature of the progressive algorithm
  - ➜ Using alignment libraries derived from a mixture of alignment programs (global, local, *etc.*)

11

Multiple alignment: T-Coffee

(Tree-based Consistency Objective Function for alignment Evaluation)

- ➢ Primary libraries of alignments

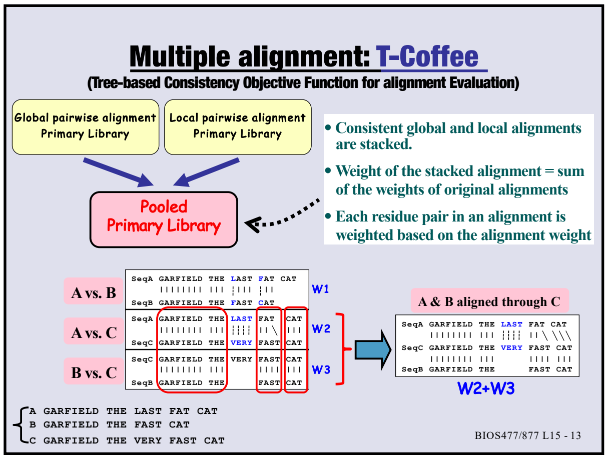


➜ Global and local, or any combination of pairwise alignment methods

➜ Each pairwise alignment is given a weight based on % identity ignoring gap sites

ATTCGG
ATAGCG ⇨ 3/6 = 50%

12



Multiple alignment: T-Coffee

(Tree-based Consistency Objective Function for alignment Evaluation)

- Consistent global and local alignments are stacked.
- Weight of the stacked alignment = sum of the weights of original alignments
- Each residue pair in an alignment is weighted based on the alignment weight

Global pairwise alignment Primary Library
Local pairwise alignment Primary Library
Pooled Primary Library
A vs. B
A vs. C
B vs. C
W1
W2
W3
A & B aligned through C
W2+W3

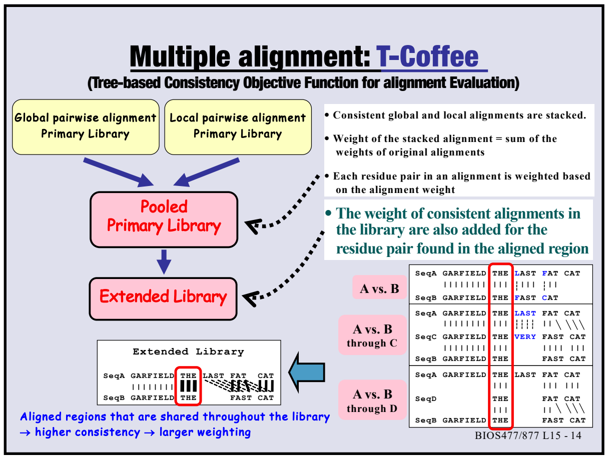
```
SeqA GARFIELD THE LAST FAT CAT
     |||||||| ||| |||| |||
SeqB GARFIELD THE FAST CAT

SeqA GARFIELD THE LAST FAT  CAT
     |||||||| ||| |||| || \ |||
SeqC GARFIELD THE VERY FAST CAT

SeqC GARFIELD THE VERY FAST CAT
     |||||||| |||      |||| |||
SeqB GARFIELD THE      FAST CAT
```

```
SeqA GARFIELD THE LAST FAT CAT
     |||||||| ||| |||| || \ \\\
SeqC GARFIELD THE VERY FAST CAT
     |||||||| |||      |||| |||
SeqB GARFIELD THE      FAST CAT
```

```
A GARFIELD THE LAST FAT CAT
B GARFIELD THE FAST CAT
C GARFIELD THE VERY FAST CAT
```

BIOS477/877 L15 - 13

13



Multiple alignment: T-Coffee

(Tree-based Consistency Objective Function for alignment Evaluation)

- Consistent global and local alignments are stacked.
- Weight of the stacked alignment = sum of the weights of original alignments
- Each residue pair in an alignment is weighted based on the alignment weight
- The weight of consistent alignments in the library are also added for the residue pair found in the aligned region

Global pairwise alignment Primary Library
Local pairwise alignment Primary Library
Pooled Primary Library
Extended Library
A vs. B
A vs. B through C
A vs. B through D

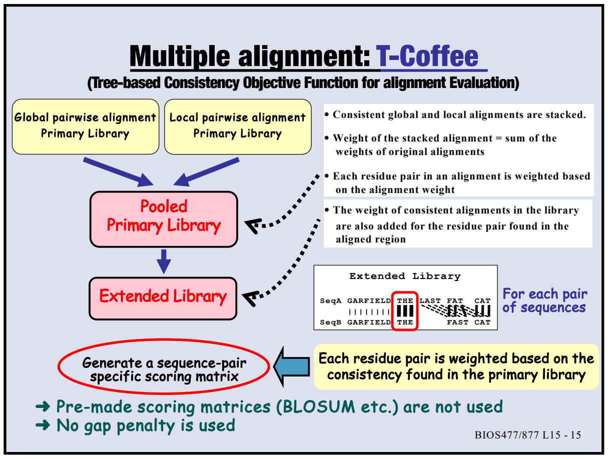
```
SeqA GARFIELD THE LAST FAT CAT
     |||||||| ||| |||| |||
SeqB GARFIELD THE FAST CAT

SeqA GARFIELD THE LAST FAT  CAT
     |||||||| ||| |||| || \ \\\
SeqC GARFIELD THE VERY FAST CAT
     |||||||| |||      |||| |||
SeqB GARFIELD THE      FAST CAT

SeqA GARFIELD THE LAST FAT  CAT
              |||      |||  |||
SeqD          THE      FAT  CAT
              |||      || \ \\\
SeqB GARFIELD THE      FAST CAT
```

```
        Extended Library

SeqA GARFIELD THE LAST FAT  CAT
     |||||||| |||
SeqB GARFIELD THE      FAST CAT
```

Aligned regions that are shared throughout the library → higher consistency → larger weighting

BIOS477/877 L15 - 14

14



Multiple alignment: T-Coffee

(Tree-based Consistency Objective Function for alignment Evaluation)

- Consistent global and local alignments are stacked.
- Weight of the stacked alignment = sum of the weights of original alignments
- Each residue pair in an alignment is weighted based on the alignment weight
- The weight of consistent alignments in the library are also added for the residue pair found in the aligned region

Global pairwise alignment Primary Library
Local pairwise alignment Primary Library
Pooled Primary Library
Extended Library
Generate a sequence-pair specific scoring matrix

```
        Extended Library

SeqA GARFIELD THE LAST FAT  CAT
     |||||||| |||
SeqB GARFIELD THE      FAST CAT
```

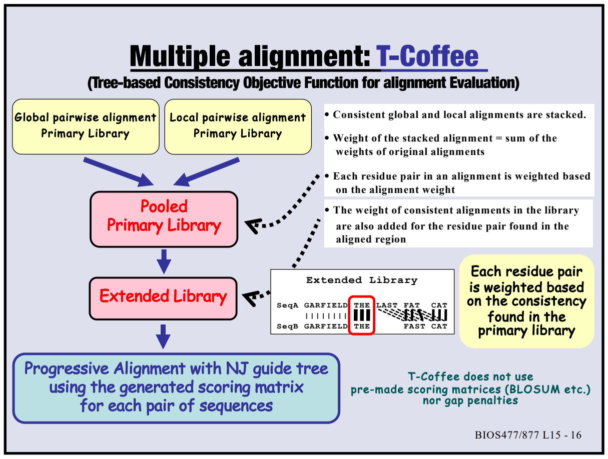
For each pair of sequences

Each residue pair is weighted based on the consistency found in the primary library

➜ Pre-made scoring matrices (BLOSUM etc.) are not used
➜ No gap penalty is used

BIOS477/877 L15 - 15

15



Multiple alignment: T-Coffee

(Tree-based Consistency Objective Function for alignment Evaluation)

- Consistent global and local alignments are stacked.
- Weight of the stacked alignment = sum of the weights of original alignments
- Each residue pair in an alignment is weighted based on the alignment weight
- The weight of consistent alignments in the library are also added for the residue pair found in the aligned region

Global pairwise alignment Primary Library
Local pairwise alignment Primary Library
Pooled Primary Library
Extended Library

```
        Extended Library

SeqA GARFIELD THE LAST FAT  CAT
     |||||||| |||
SeqB GARFIELD THE      FAST CAT
```

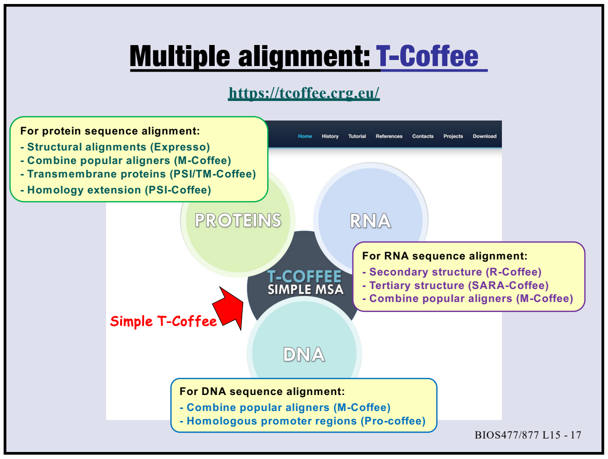
Each residue pair is weighted based on the consistency found in the primary library

Progressive Alignment with NJ guide tree using the generated scoring matrix for each pair of sequences

T-Coffee does not use pre-made scoring matrices (BLOSUM etc.) nor gap penalties

BIOS477/877 L15 - 16

16



Multiple alignment: T-Coffee

https://tcoffee.crg.eu/

For protein sequence alignment:
- Structural alignments (Expresso)
- Combine popular aligners (M-Coffee)
- Transmembrane proteins (PSI/TM-Coffee)
- Homology extension (PSI-Coffee)

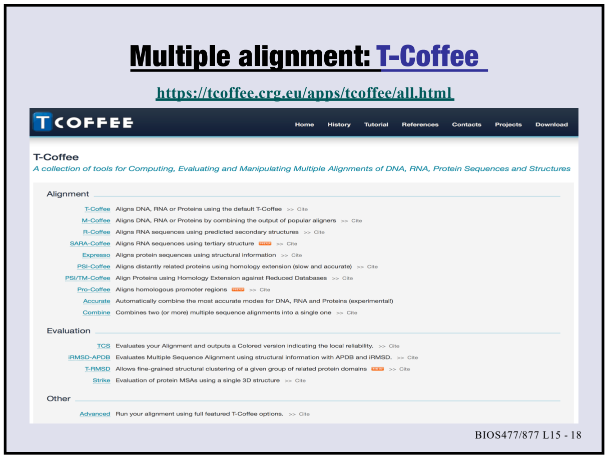
For RNA sequence alignment:
- Secondary structure (R-Coffee)
- Tertiary structure (SARA-Coffee)
- Combine popular aligners (M-Coffee)

For DNA sequence alignment:
- Combine popular aligners (M-Coffee)
- Homologous promoter regions (Pro-coffee)

Simple T-Coffee

BIOS477/877 L15 - 17

17



Multiple alignment: T-Coffee

https://tcoffee.crg.eu/apps/tcoffee/all.html

T-Coffee

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

Alignment

T-Coffee	Aligns DNA, RNA or Proteins using the default T-Coffee >> Cite
M-Coffee	Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite
R-Coffee	Aligns RNA sequences using predicted secondary structures >> Cite
SARA-Coffee	Aligns RNA sequences using tertiary structure NEW >> Cite
Expresso	Aligns protein sequences using structural information >> Cite
PSI-Coffee	Aligns distantly related proteins using homology extension (slow and accurate) >> Cite
PSI/TM-Coffee	Align Proteins using Homology Extension against Reduced Databases >> Cite
Pro-Coffee	Aligns homologous promoter regions NEW >> Cite
Accurate	Automatically combine the most accurate modes for DNA, RNA and Proteins (experimental!)
Combine	Combines two (or more) multiple sequence alignments into a single one >> Cite

Evaluation

TCS	Evaluates your Alignment and outputs a Colored version indicating the local reliability. >> Cite
iRMSD-APDB	Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite
T-RMSD	Allows fine-grained structural clustering of a given group of related protein domains NEW >> Cite
Strike	Evaluation of protein MSAs using a single 3D structure >> Cite

Other

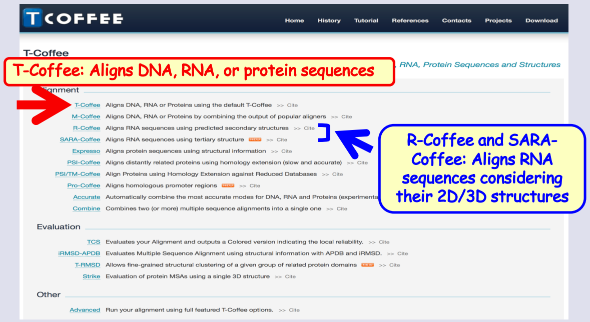
Advanced	Run your alignment using full featured T-Coffee options. >> Cite

BIOS477/877 L15 - 18

18

Multiple alignment: T-Coffee

https://tcoffee.crg.eu/apps/tcoffee/all.html



T-Coffee

T-Coffee: Aligns DNA, RNA, or protein sequences

RNA, Protein Sequences and Structures

gnment

T-Coffee Aligns DNA, RNA or Proteins using the default T-Coffee >> Cite

M-Coffee Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite

R-Coffee Aligns RNA sequences using predicted secondary structures >> Cite

SARA-Coffee Aligns RNA sequences using tertiary structure NEW >> Cite

Expresso Aligns protein sequences using structural information >> Cite

PSI-Coffee Aligns distantly related proteins using homology extension (slow and accurate) >> Cite

PSI/TM-Coffee Align Proteins using Homology Extension against Reduced Databases >> Cite

Pro-Coffee Aligns homologous promoter regions NEW >> Cite

Accurate Automatically combine the most accurate modes for DNA, RNA and Proteins (experimenta

Combine Combines two (or more) multiple sequence alignments into a single one >> Cite

R-Coffee and SARA-Coffee: Aligns RNA sequences considering their 2D/3D structures

Evaluation

TCS Evaluates your Alignment and outputs a Colored version indicating the local reliability. >> Cite

iRMSD-APDB Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite

T-RMSD Allows fine-grained structural clustering of a given group of related protein domains NEW >> Cite

Strike Evaluation of protein MSAs using a single 3D structure >> Cite

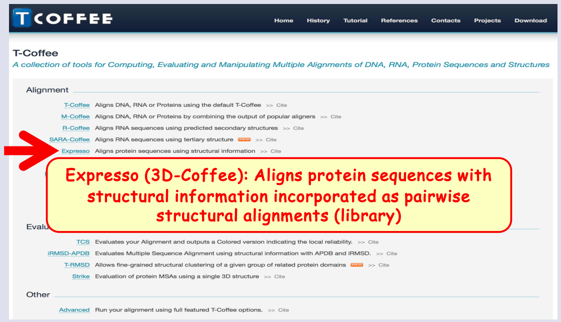
Other

Advanced Run your alignment using full featured T-Coffee options. >> Cite

19

Multiple alignment: T-Coffee

https://tcoffee.crg.eu/apps/tcoffee/all.html



T-Coffee

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

Alignment

T-Coffee Aligns DNA, RNA or Proteins using the default T-Coffee >> Cite

M-Coffee Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite

R-Coffee Aligns RNA sequences using predicted secondary structures >> Cite

SARA-Coffee Aligns RNA sequences using tertiary structure NEW >> Cite

Expresso Aligns protein sequences using structural information >> Cite

Expresso (3D-Coffee): Aligns protein sequences with structural information incorporated as pairwise structural alignments (library)

Evalu

TCS Evaluates your Alignment and outputs a Colored version indicating the local reliability. >> Cite

iRMSD-APDB Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite

T-RMSD Allows fine-grained structural clustering of a given group of related protein domains NEW >> Cite

Strike Evaluation of protein MSAs using a single 3D structure >> Cite

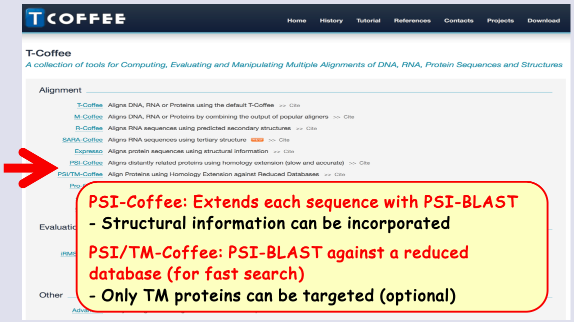
Other

Advanced Run your alignment using full featured T-Coffee options. >> Cite

20

Multiple alignment: T-Coffee

https://tcoffee.crg.eu/apps/tcoffee/all.html



T-Coffee

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

Alignment

T-Coffee Aligns DNA, RNA or Proteins using the default T-Coffee >> Cite

M-Coffee Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite

R-Coffee Aligns RNA sequences using predicted secondary structures >> Cite

SARA-Coffee Aligns RNA sequences using tertiary structure NEW >> Cite

Expresso Aligns protein sequences using structural information >> Cite

PSI-Coffee Aligns distantly related proteins using homology extension (slow and accurate) >> Cite

PSI/TM-Coffee Align Proteins using Homology Extension against Reduced Databases >> Cite

Pro

Evaluati

iRMS

Other

Adva

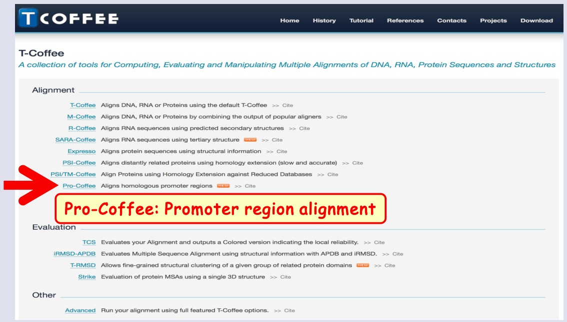
PSI-Coffee: Extends each sequence with PSI-BLAST
- Structural information can be incorporated

PSI/TM-Coffee: PSI-BLAST against a reduced database (for fast search)
- Only TM proteins can be targeted (optional)

21

Multiple alignment: T-Coffee

https://tcoffee.crg.eu/apps/tcoffee/all.html



T-Coffee

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

Alignment

T-Coffee Aligns DNA, RNA or Proteins using the default T-Coffee >> Cite

M-Coffee Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite

R-Coffee Aligns RNA sequences using predicted secondary structures >> Cite

SARA-Coffee Aligns RNA sequences using tertiary structure NEW >> Cite

Expresso Aligns protein sequences using structural information >> Cite

PSI-Coffee Aligns distantly related proteins using homology extension (slow and accurate) >> Cite

PSI/TM-Coffee Align Proteins using Homology Extension against Reduced Databases >> Cite

Pro-Coffee Aligns homologous promoter regions NEW >> Cite

Pro-Coffee: Promoter region alignment

Evaluation

TCS Evaluates your Alignment and outputs a Colored version indicating the local reliability. >> Cite

iRMSD-APDB Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite

T-RMSD Allows fine-grained structural clustering of a given group of related protein domains NEW >> Cite

Strike Evaluation of protein MSAs using a single 3D structure >> Cite

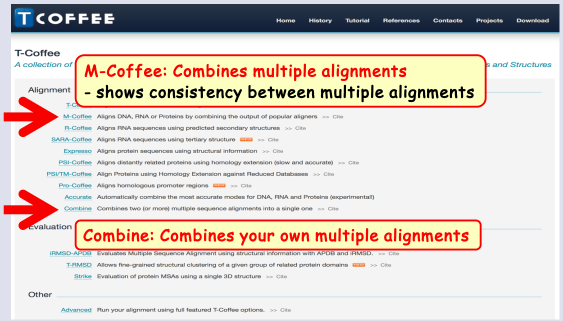
Other

Advanced Run your alignment using full featured T-Coffee options. >> Cite

22

Multiple alignment: T-Coffee

https://tcoffee.crg.eu/apps/tcoffee/all.html



T-Coffee

A collection of

M-Coffee: Combines multiple alignments
- shows consistency between multiple alignments

s and Structures

Alignment

T-C

M-Coffee Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite

R-Coffee Aligns RNA sequences using predicted secondary structures >> Cite

SARA-Coffee Aligns RNA sequences using tertiary structure NEW >> Cite

Expresso Aligns protein sequences using structural information >> Cite

PSI-Coffee Aligns distantly related proteins using homology extension (slow and accurate) >> Cite

PSI/TM-Coffee Align Proteins using Homology Extension against Reduced Databases >> Cite

Pro-Coffee Aligns homologous promoter regions NEW >> Cite

Accurate Automatically combine the most accurate modes for DNA, RNA and Proteins (experimental) >> Cite

Combine Combines two (or more) multiple sequence alignments into a single one >> Cite

valuation

Combine: Combines your own multiple alignments

iRMSD-APDB Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite

T-RMSD Allows fine-grained structural clustering of a given group of related protein domains NEW >> Cite

Strike Evaluation of protein MSAs using a single 3D structure >> Cite

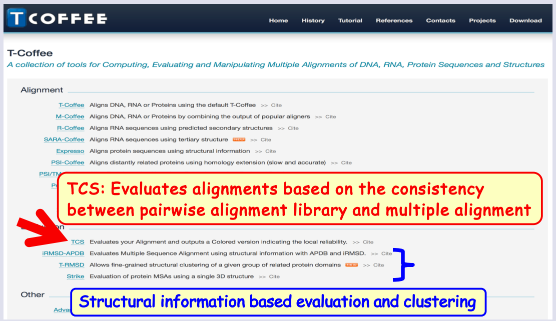
Other

Advanced Run your alignment using full featured T-Coffee options. >> Cite

23

Multiple alignment: T-Coffee

https://tcoffee.crg.eu/apps/tcoffee/all.html



T-Coffee

A collection of tools for Computing, Evaluating and Manipulating Multiple Alignments of DNA, RNA, Protein Sequences and Structures

Alignment

T-Coffee Aligns DNA, RNA or Proteins using the default T-Coffee >> Cite

M-Coffee Aligns DNA, RNA or Proteins by combining the output of popular aligners >> Cite

R-Coffee Aligns RNA sequences using predicted secondary structures >> Cite

SARA-Coffee Aligns RNA sequences using tertiary structure NEW >> Cite

Expresso Aligns protein sequences using structural information >> Cite

PSI-Coffee Aligns distantly related proteins using homology extension (slow and accurate) >> Cite

PSI/TM

P

TCS: Evaluates alignments based on the consistency between pairwise alignment library and multiple alignment

on

TCS Evaluates your Alignment and outputs a Colored version indicating the local reliability. >> Cite

iRMSD-APDB Evaluates Multiple Sequence Alignment using structural information with APDB and iRMSD. >> Cite

T-RMSD Allows fine-grained structural clustering of a given group of related protein domains NEW >> Cite

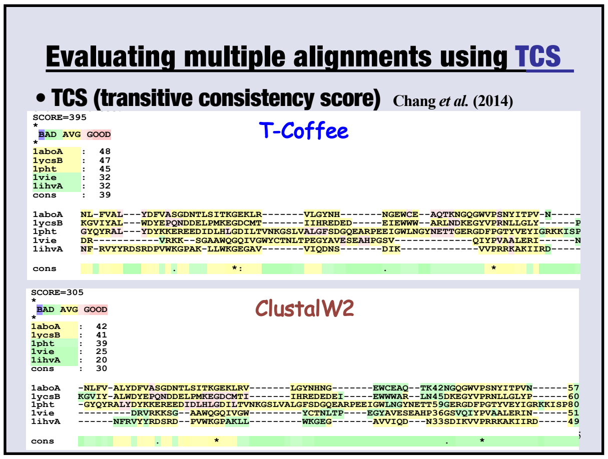
Strike Evaluation of protein MSAs using a single 3D structure >> Cite

Other

Adva

Structural information based evaluation and clustering

24



Evaluating multiple alignments using TCS

- **TCS (transitive consistency score)** Chang *et al.* (2014)

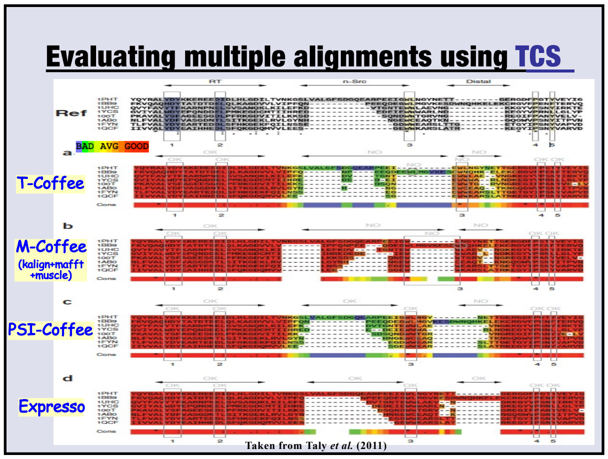
T-Coffee

```
SCORE=395
*
 BAD AVG GOOD
*
1aboA    :  48
1ycsB    :  47
1pht     :  45
1vie     :  32
1ihvA    :  32
cons     :  39

1aboA    NL-FVAL---YDFVASGDNTLSITKGEKLR-------VLGYNH-------NGEWCE--AQTKNGQGWVPSNYITPV-N-----
1ycsB    KGVIYAL---WDYEPQNDDELPMKEGDCMT-------IIHREDED-----EIEWWW--ARLNDKEGYVPRNLLGLY------P
1pht     GYQYRAL---YDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETTGERGDFPGTYVEYIGRKKISP
1vie     DR-----------VRKK--SGAAWQGQIVGWYCTNLTPEGYAVESEAHPGSV-------------QIYPVAALERI------N
1ihvA    NF-RVYYRDSRDPVWKGPAK-LLWKGEGAV-------VIQDNS-------DIK-------------VVPRRKAKIIRD-----

cons               .     *:                 .                 *
```

ClustalW2

```
SCORE=305
*
 BAD AVG GOOD
*
1aboA    :  42
1ycsB    :  41
1pht     :  39
1vie     :  25
1ihvA    :  20
cons     :  30

1aboA    -NLFV-ALYDFVASGDNTLSITKGEKLRV-------LGYNHNG-------EWCEAQ--TK42NGQGWVPSNYITPVN------57
1ycsB    KGVIY-ALWDYEPQNDDELPMKEGDCMTI-------IHREDEDEI-----EWWWAR--LN45DKEGYVPRNLLGLYP------60
1pht     -GYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETT59GERGDFPGTYVEYIGRKKISP80
1vie     ---------DRVRKKSG--AAWQGQIVGW---------YCTNLTP----EGYAVESEAHP36GSVQIYPVAALERIN------51
1ihvA    ------NFRVYYRDSRD--PVWKGPAKLL---------WKGEG-------AVVIQD---N33SDIKVVPRRKAKIIRD-----49

cons             .     *                                .      *
```

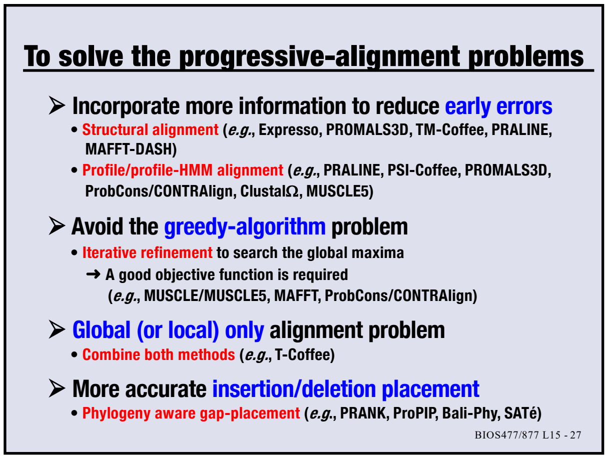
25



Evaluating multiple alignments using TCS

Taken from Taly *et al.* (2011)

26

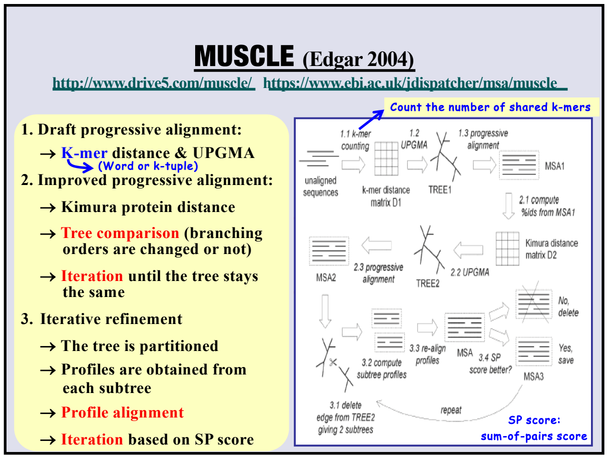


To solve the progressive-alignment problems

- ➢ Incorporate more information to reduce early errors
  - Structural alignment (*e.g.*, Expresso, PROMALS3D, TM-Coffee, PRALINE, MAFFT-DASH)
  - Profile/profile-HMM alignment (*e.g.*, PRALINE, PSI-Coffee, PROMALS3D, ProbCons/CONTRAlign, ClustalΩ, MUSCLE5)
- ➢ Avoid the greedy-algorithm problem
  - Iterative refinement to search the global maxima
    - → A good objective function is required (*e.g.*, MUSCLE/MUSCLE5, MAFFT, ProbCons/CONTRAlign)
- ➢ Global (or local) only alignment problem
  - Combine both methods (*e.g.*, T-Coffee)
- ➢ More accurate insertion/deletion placement
  - Phylogeny aware gap-placement (*e.g.*, PRANK, ProPIP, Bali-Phy, SATé)

BIOS477/877 L15 - 27

27



MUSCLE (Edgar 2004)

http://www.drive5.com/muscle/ https://www.ebi.ac.uk/jdispatcher/msa/muscle

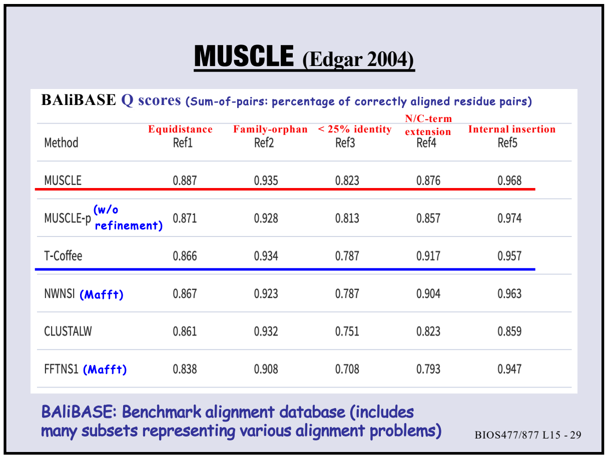
1. Draft progressive alignment:
   - → K-mer distance & UPGMA
     (Word or k-tuple)
2. Improved progressive alignment:
   - → Kimura protein distance
   - → Tree comparison (branching orders are changed or not)
   - → Iteration until the tree stays the same
3. Iterative refinement
   - → The tree is partitioned
   - → Profiles are obtained from each subtree
   - → Profile alignment
   - → Iteration based on SP score

Count the number of shared k-mers

1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
3.1 delete edge from TREE2 giving 2 subtrees
3.2 compute subtree profiles
3.3 re-align profiles
MSA
3.4 SP score better?
No, delete
Yes, save
MSA3
repeat

SP score: sum-of-pairs score

28



MUSCLE (Edgar 2004)

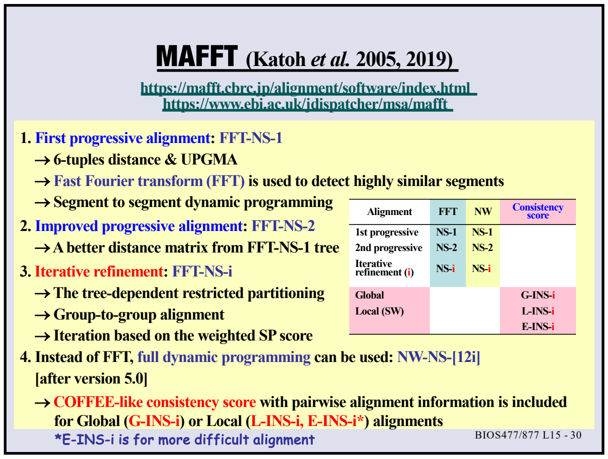
BAliBASE Q scores (Sum-of-pairs: percentage of correctly aligned residue pairs)

Method	Equidistance Ref1	Family-orphan Ref2	< 25% identity Ref3	N/C-term extension Ref4	Internal insertion Ref5
MUSCLE	0.887	0.935	0.823	0.876	0.968
MUSCLE-p (w/o refinement)	0.871	0.928	0.813	0.857	0.974
T-Coffee	0.866	0.934	0.787	0.917	0.957
NWNSI (Mafft)	0.867	0.923	0.787	0.904	0.963
CLUSTALW	0.861	0.932	0.751	0.823	0.859
FFTNS1 (Mafft)	0.838	0.908	0.708	0.793	0.947

BAliBASE: Benchmark alignment database (includes many subsets representing various alignment problems)

BIOS477/877 L15 - 29

29



MAFFT (Katoh *et al.* 2005, 2019)

https://mafft.cbrc.jp/alignment/software/index.html
https://www.ebi.ac.uk/jdispatcher/msa/mafft

1. First progressive alignment: FFT-NS-1
   - → 6-tuples distance & UPGMA
   - → Fast Fourier transform (FFT) is used to detect highly similar segments
   - → Segment to segment dynamic programming
2. Improved progressive alignment: FFT-NS-2
   - → A better distance matrix from FFT-NS-1 tree
3. Iterative refinement: FFT-NS-i
   - → The tree-dependent restricted partitioning
   - → Group-to-group alignment
   - → Iteration based on the weighted SP score
4. Instead of FFT, full dynamic programming can be used: NW-NS-[12i] [after version 5.0]
   - → COFFEE-like consistency score with pairwise alignment information is included for Global (G-INS-i) or Local (L-INS-i, E-INS-i*) alignments

Alignment	FFT	NW	Consistency score
1st progressive	NS-1	NS-1	
2nd progressive	NS-2	NS-2	
Iterative refinement (i)	NS-i	NS-i	
Global			G-INS-i
Local (SW)			L-INS-i
			E-INS-i

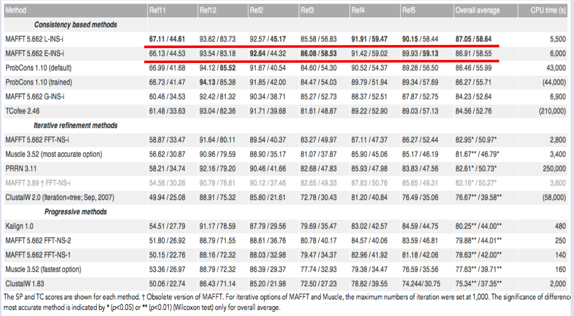
*E-INS-i is for more difficult alignment

BIOS477/877 L15 - 30

30

MAFFT

https://mafft.cbrc.jp/alignment/software/eval/accuracy.html



Method	Ref11	Ref12	Ref2	Ref3	Ref4	Ref5	Overall average	CPU time (s)
Consistency based methods								
MAFFT 5.662 L-INS-i	**67.11 / 44.61**	93.62 / 83.73	92.57 / **45.17**	85.58 / 56.83	**91.91 / 59.47**	**90.15** / 58.44	**87.05 / 58.64**	5,500
MAFFT 5.662 E-INS-i	66.13 / 44.53	93.54 / 83.18	**92.64** / 44.32	**86.08 / 58.53**	91.42 / 59.02	89.93 / **59.13**	86.91 / 58.55	6,000
ProbCons 1.10 (default)	66.99 / 41.68	94.12 / **85.52**	91.67 / 40.54	84.60 / 54.30	90.52 / 54.37	89.28 / 56.50	86.46 / 55.99	43,000
ProbCons 1.10 (trained)	66.73 / 41.47	**94.13** / 85.38	91.85 / 42.00	84.47 / 54.03	89.79 / 51.94	89.34 / 57.69	86.27 / 55.71	(44,000)
MAFFT 5.662 G-INS-i	60.46 / 34.53	92.42 / 81.32	90.34 / 38.71	85.27 / 52.73	88.37 / 52.51	87.87 / 52.75	84.23 / 52.64	6,900
TCofee 2.46	61.48 / 33.63	93.04 / 82.36	91.71 / 39.68	81.61 / 48.87	89.22 / 52.90	89.03 / 57.13	84.56 / 52.76	(210,000)
Iterative refinement methods								
MAFFT 5.662 FFT-NS-i	58.87 / 33.47	91.64 / 80.11	89.54 / 40.37	83.27 / 49.97	87.11 / 47.37	86.27 / 52.44	82.95* / 50.97*	2,800
Muscle 3.52 (most accurate option)	56.62 / 30.87	90.96 / 79.59	88.90 / 35.17	81.07 / 37.87	85.90 / 45.06	85.17 / 46.19	81.67** / 46.79*	3,400
PRRN 3.11	58.21 / 34.74	92.16 / 79.20	90.46 / 41.66	82.68 / 47.83	85.93 / 47.98	83.83 / 47.56	82.61* / 50.73*	250,000
MAFFT 3.89 † FFT-NS-i	54.56 / 30.26	90.78 / 78.61	90.12 / 37.46	82.65 / 49.33	87.83 / 50.76	85.65 / 49.31	82.16* / 50.27*	3,600
ClustalW 2.0 (Iteration=tree; Sep, 2007)	49.94 / 25.08	88.91 / 75.32	85.80 / 21.61	72.78 / 30.43	81.20 / 40.84	76.49 / 35.06	76.67** / 39.58**	(58,000)
Progressive methods								
Kalign 1.0	54.51 / 27.79	91.17 / 78.59	87.79 / 29.56	79.69 / 35.47	83.02 / 42.57	84.59 / 44.75	80.25** / 44.00**	480
MAFFT 5.662 FFT-NS-2	51.80 / 26.92	88.79 / 71.55	88.61 / 36.76	80.78 / 40.17	84.57 / 40.06	83.59 / 46.81	79.88** / 44.01**	250
MAFFT 5.662 FFT-NS-1	50.15 / 22.76	88.16 / 72.32	88.03 / 32.98	79.47 / 34.37	82.96 / 41.92	81.18 / 42.06	78.63** / 42.00**	140
Muscle 3.52 (fastest option)	53.36 / 26.97	88.79 / 72.32	86.39 / 29.37	77.74 / 32.93	79.38 / 34.47	76.59 / 35.56	77.63** / 39.71**	160
ClustalW 1.83	50.06 / 22.74	86.43 / 71.14	85.20 / 21.98	72.50 / 27.23	78.82 / 39.55	74.244/ 30.75	75.34** / 37.35**	2,000

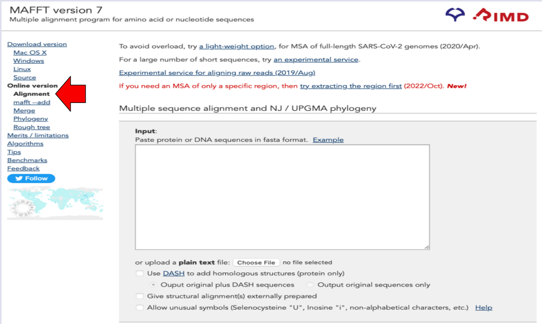
The SP and TC scores are shown for each method. † Obsolete version of MAFFT. For iterative options of MAFFT and Muscle, the maximum numbers of iteration were set at 1,000. The significance of differenc most accurate method is indicated by * ($p<0.05$) or ** ($p<0.01$) (Wilcoxon test) only for overall average.

Scores: SP (sum of pairs)/TC (total column)
tested using BAliBASE benchmarking alignment datasets

31

MAFFT website

https://mafft.cbrc.jp/alignment/software/



MAFFT version 7
Multiple alignment program for amino acid or nucleotide sequences

- Download version
  - Mac OS X
  - Windows
  - Linux
  - Source
- **Online version**
  - **Alignment**
  - mafft --add
  - Merge
  - Phylogeny
  - Rough tree
- Merits / limitations
- Algorithms
- Tips
- Benchmarks
- Feedback

To avoid overload, try a light-weight option, for MSA of full-length SARS-CoV-2 genomes (2020/Apr).

For a large number of short sequences, try an experimental service.

Experimental service for aligning raw reads (2019/Aug)

If you need an MSA of only a specific region, then try extracting the region first (2022/Oct). ***New!***

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:
Paste protein or DNA sequences in fasta format. Example

or upload a **plain text** file: Choose File no file selected

□ Use DASH to add homologous structures (protein only)

● Ouput original plus DASH sequences ○ Output original sequences only

□ Give structural alignment(s) externally prepared

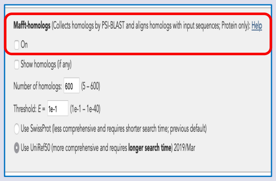
□ Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, *etc.*) Help

32

Mafft-homologs

Accuracy of an alignment of distantly related sequences can be improved when aligned with their close homologs

MAFFT options



Mafft-homologs (Collects homologs by PSI-BLAST and aligns homologs with input sequences; Protein only): Help

□ On

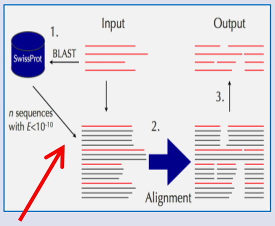
□ Show homologs (if any)

Number of homologs: 600 (5 – 600)

Threshold: *E* = 1e-1 (1e-1 – 1e-40)

○ Use SwissProt (less comprehensive and requires shorter search time; previous default)

● Use UniRef50 (more comprehensive and requires **longer search time**) 2019/Mar



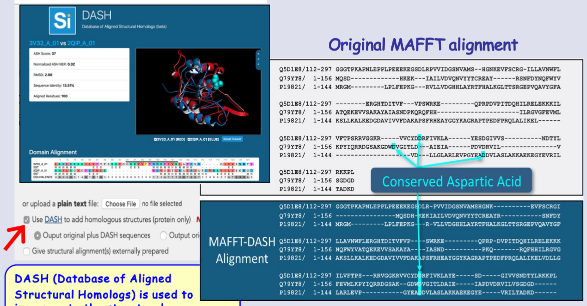
PSI-BLAST is used to collect similar sequences

Similar approach is used in PSI-Coffee, PRALINE, etc.

33

MAFFT-DASH (Rozewicki *et al.* 2019)

Integrated protein sequence and structural alignment



or upload a **plain text** file: Choose File no file selected

☑ Use DASH to add homologous structures (protein only)

● Ouput original plus DASH sequences ○ Output ori

□ Give structural alignment(s) externally prepared

Original MAFFT alignment

```
Q5D1E8/112-297 GGGTPKAPNLEPPLPEEEKEGSDLRPVVIDGSNVAMS--HGNKEVFSCRG-ILLAVNWFL
Q79YT8/  1-156 MQSD------------HKEK---IAILVDVQNVYYTCREAY------RSNFDYNQFWYV
P19821/  1-144 MRGM---------LPLFEPKG---RVLLVDGHHLAYRTFHALKGLTTSRGEPVQAVYGFA

Q5D1E8/112-297 --------ERGHTDITVF---VPSWRKE---------QPRPDVPITDQHILRELEKKKIL
Q79YT8/  1-156 ATQEKEVVSAKAYAIASNDPKQRQFHH---------------------ILRGVGFEVML
P19821/  1-144 KSLLKALKEDGDAVIVVFDAKAPSFRHEAYGGYKAGRAPTPEDFPRQLALIKEL------

Q5D1E8/112-297 VFTPSRRVGGKR-----VVCYDDRFIVKLA------YESDGIVVS---------NDTYL
Q79YT8/  1-156 KPYIQRRDGSAKGDWDVGITLD---AIEIA------PDVDRVIL---------------V
P19821/  1-144 ------------------VD---LLGLARLEVPGYEADDVLASLAKKAEKEGYEVRIL

Q5D1E8/112-297 RKKPL
Q79YT8/  1-156 SGDGD
P19821/  1-144 TADKD
```

Conserved Aspartic Acid

MAFFT-DASH Alignment

```
Q5D1E8/112-297 GGGTPKAPNLEPPLPEEEKEGSDLR-PVVIDGSNVAMSHGNK----------EVFSCRGI
Q79YT8/  1-156 ----------------MQSDH-KEKIAILVDVQNVYYTCREAYR----------SNFDY
P19821/  1-144 MRGM---------LPLFEPKG--R-VLLVDGHHLAYRTFHALKGLTTSRGEPVQAVYGF

Q5D1E8/112-297 LLAVNWFLERGHTDITVFVP----SWRKE---------QPRP-DVPITDQHILRELEKKK
Q79YT8/  1-156 NQFWYVATQEKEVVSAKAYA----IASND----------PKQ-------RQFHHILRGVG
P19821/  1-144 AKSLLKALKEDGDAVIVVFDAKAPSFRHEAYGGYKAGRAPTPEDFPRQLALIKELVDLLG

Q5D1E8/112-297 ILVFTPS---RRVGGKRVVCYDDRFIVKLAYE-----SD-----GIVVSNDTYLRKKPL
Q79YT8/  1-156 FEVMLKPYIQRRDGSAK--GDWDVGITLDAIE-----IAPDVDRVILVSGDGD------
P19821/  1-144 LARLEVP----------GYEADDVLASLAKKAEKEGYE-----VRILTADKD------
```

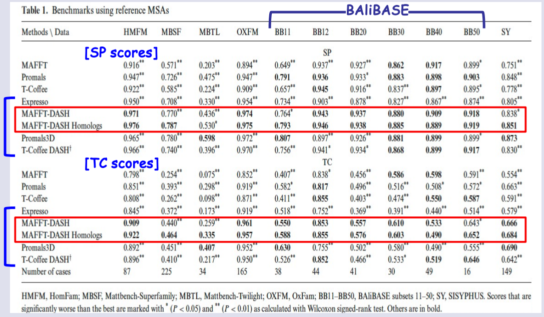
DASH (Database of Aligned Structural Homologs) is used to incorporate the structural information to improve the alignment

34

MAFFT-DASH (Rozewicki *et al.* 2019)

Integrated protein sequence and structural alignment

Table 1. Benchmarks using reference MSAs



Methods \ Data	HMFM	MBSF	MBTL	OXFM	BB11	BB12	BB20	BB30	BB40	BB50	SY
					BAliBASE						
[SP scores]						SP					
MAFFT	0.916^{**}	0.571^{**}	0.203^{**}	0.894^{**}	0.649^{**}	0.937^{**}	0.927^{**}	**0.862**	**0.917**	0.899^{*}	0.751^{**}
Promals	0.947^{**}	0.726^{**}	0.475^{**}	0.947^{**}	**0.791**	**0.936**	0.933^{*}	**0.883**	**0.898**	**0.903**	0.848^{**}
T-Coffee	0.922^{**}	0.585^{**}	0.224^{**}	0.909^{**}	0.657^{**}	**0.945**	0.916^{**}	0.837^{**}	**0.897**	0.895^{*}	0.778^{**}
Expresso	0.950^{**}	0.708^{**}	0.330^{**}	0.954^{**}	0.734^{**}	0.903^{**}	0.878^{**}	0.827^{**}	0.867^{**}	0.874^{**}	0.805^{**}
MAFFT-DASH	**0.971**	0.770^{**}	0.436^{**}	**0.974**	0.764^{*}	**0.943**	**0.937**	**0.880**	**0.909**	**0.918**	0.838^{*}
MAFFT-DASH Homologs	**0.976**	**0.787**	0.530^{*}	**0.975**	**0.793**	**0.946**	**0.938**	**0.885**	**0.889**	**0.919**	**0.851**
Promals3D	0.965^{**}	0.780^{**}	**0.598**	0.972^{**}	**0.807**	0.897^{**}	0.926^{**}	**0.881**	**0.899**	0.899^{*}	**0.873**
T-Coffee DASH†	0.966^{**}	0.740^{**}	0.396^{**}	0.970^{**}	0.756^{**}	0.941^{*}	0.934^{*}	**0.868**	**0.899**	**0.917**	0.830^{**}
[TC scores]						TC					
MAFFT	0.798^{**}	0.254^{**}	0.075^{**}	0.852^{**}	0.407^{**}	0.838^{*}	0.456^{**}	**0.586**	**0.598**	0.591^{**}	0.554^{**}
Promals	0.851^{**}	0.393^{**}	0.298^{**}	0.919^{**}	0.582^{*}	**0.817**	0.496^{**}	0.516^{**}	0.508^{*}	0.572^{*}	0.663^{**}
T-Coffee	0.808^{**}	0.262^{**}	0.098^{**}	0.871^{**}	0.411^{**}	**0.855**	0.403^{**}	0.474^{**}	**0.550**	**0.587**	0.591^{**}
Expresso	0.845^{**}	0.372^{**}	0.173^{**}	0.919^{**}	0.518^{**}	0.752^{**}	0.369^{**}	0.391^{**}	0.440^{**}	0.514^{**}	0.579^{**}
MAFFT-DASH	**0.909**	0.440^{**}	0.259^{**}	**0.961**	**0.550**	**0.853**	**0.557**	**0.610**	**0.533**	0.643^{*}	**0.666**
MAFFT-DASH Homologs	**0.922**	**0.464**	**0.335**	**0.957**	**0.588**	**0.855**	**0.576**	**0.603**	**0.490**	**0.652**	**0.684**
Promals3D	0.892^{**}	0.451^{**}	**0.407**	0.952^{**}	**0.630**	0.755^{**}	0.502^{**}	0.580^{**}	0.490^{**}	0.555^{**}	**0.690**
T-Coffee DASH†	0.896^{**}	0.410^{**}	0.217^{**}	0.950^{**}	0.526^{**}	**0.852**	0.466^{**}	0.533^{*}	**0.519**	**0.646**	0.642^{**}
Number of cases	87	225	34	165	38	44	41	30	49	16	149

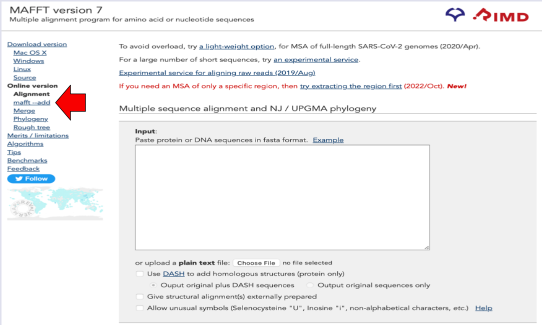
HMFM, HomFam; MBSF, Mattbench-Superfamily; MBTL, Mattbench-Twilight; OXFM, OxFam; BB11–BB50, BAliBASE subsets 11–50; SY, SISYPHUS. Scores that are significantly worse than the best are marked with * ($P < 0.05$) and ** ($P < 0.01$) as calculated with Wilcoxon signed-rank test. Others are in bold.

Incorporates 3D information

35

MAFFT website

https://mafft.cbrc.jp/alignment/software/



MAFFT version 7
Multiple alignment program for amino acid or nucleotide sequences

- Download version
  - Mac OS X
  - Windows
  - Linux
  - Source
- **Online version**
  - **Alignment**
  - mafft --add
  - Merge
  - Phylogeny
  - Rough tree
- Merits / limitations
- Algorithms
- Tips
- Benchmarks
- Feedback

To avoid overload, try a light-weight option, for MSA of full-length SARS-CoV-2 genomes (2020/Apr).

For a large number of short sequences, try an experimental service.

Experimental service for aligning raw reads (2019/Aug)

If you need an MSA of only a specific region, then try extracting the region first (2022/Oct). ***New!***

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:
Paste protein or DNA sequences in fasta format. Example

or upload a **plain text** file: Choose File no file selected

□ Use DASH to add homologous structures (protein only)

● Ouput original plus DASH sequences ○ Output original sequences only

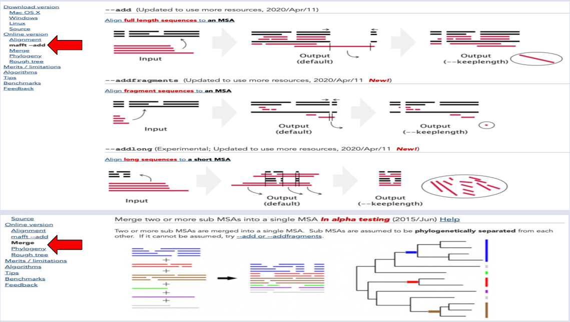
□ Give structural alignment(s) externally prepared

□ Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, *etc.*) Help

36

MAFFT website

https://mafft.cbrc.jp/alignment/server/add.html


Download version
Mac OS X
Windows
Linux
Source
Online version
Alignment
mafft --add
Merge
Phylogeny
Rough tree
Merits / limitations
Algorithms
Tips
Benchmarks
Feedback

--add (Updated to use more resources, 2020/Apr/11)

Align full length sequences to an MSA

--addfragments (Updated to use more resources, 2020/Apr/11 New!)

Align fragment sequences to an MSA

--addlong (Experimental; Updated to use more resources, 2020/Apr/11 New!)

Align long sequences to a short MSA

Merge two or more sub MSAs into a single MSA In alpha testing (2015/Jun) Help

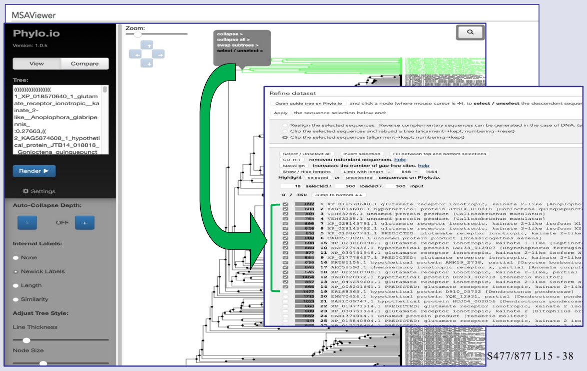
Two or more sub MSAs are merged into a single MSA. Sub MSAs are assumed to be phylogenetically separated from each other. If it cannot be assumed, try --add or --addfragments.

https://mafft.cbrc.jp/alignment/server/merge.html

BIOS477/877 L15 - 37

37

Alignment Refinement with MAFFT


MSAViewer

Phylo.io
Version: 1.0.k

Tree:
((((((((((((((((((
1_XP_018570640_1_glutamate_receptor_ionotropic__kainate_2-like__Anoplophora_glabripennis_:0.27663,((
2_KAG5874608_1_hypothetical_protein_JTB14_018818__Gonioctena_quinquepunct

Auto-Collapse Depth: OFF

Internal Labels:
○ None
● Newick Labels
○ Length
○ Similarity

Adjust Tree Style:
Line Thickness
Node Size

Zoom:

Refine dataset

Open guide tree on Phylo.io and click a node (where mouse cursor is +), to select / unselect the descendent sequen

Apply the sequence selection below and:

○ Realign the selected sequences. Reverse complementary sequences can be generated in the case of DNA. (a
○ Clip the selected sequences and rebuild a tree (alignment→kept; numbering→reset)
● Clip the selected sequences (alignment→kept; numbering→kept)

Select / Unselect all | Invert selection | Fill between top and bottom selections

CD-HIT removes redundant sequences. help

MaxAlign increases the number of gap-free sites. help

Show / Hide lengths | Limit with length : 545 – 1454

Highlight selected or unselected sequences on Phylo.io.

18 selected / 360 loaded / 360 input

0 / 360 Jump to bottom ↓↓

```
☑ 892  1 XP_018570640.1 glutamate receptor ionotropic, kainate 2-like [Anoplopho
☑ 603  2 KAG5874608.1 hypothetical protein JTB14_018818 [Gonioctena quinquepunct
☑ 891  3 VEN63256.1 unnamed protein product [Callosobruchus maculatus]
☑ 754  4 VEN63255.1 unnamed protein product [Callosobruchus maculatus]
☑ 896  7 XP_028145791.1 glutamate receptor ionotropic, kainate 2-like isoform X1
☑ 828  8 XP_028145792.1 glutamate receptor ionotropic, kainate 3-like isoform X2
☑ 810  5 XP_019867781.1 PREDICTED: glutamate receptor ionotropic, kainate 2-like
☑ 888  6 CAH0553020.1 unnamed protein product [Brassicogethes aeneus]
☑ 698 15 XP_023018098.1 glutamate receptor ionotropic, kainate 1-like [Leptinot
☑ 880 10 KAF7274436.1 hypothetical protein GWI33_012907 [Rhynchophorus ferrugin
☑ 877 11 XP_030751945.1 glutamate receptor ionotropic, kainate 2-like isoform X
☑ 884  9 XP_017778457.1 PREDICTED: glutamate receptor ionotropic, kainate 2-like
☑ 635 16 KRT85106.1 hypothetical protein AMK59_2738, partial [Oryctes borbonicu
☑ 845 17 AKC58590.1 chemosensory ionotropic receptor x, partial [Anomala corpul
☑ 545 18 XP_022910700.1 glutamate receptor ionotropic, kainate 2-like, partial
☑ 1454 12 KAH0820072.1 hypothetical protein GEV33_002718 [Tenebrio molitor]
☑ 867 13 XP_044259601.1 glutamate receptor ionotropic, kainate 2-like isoform X
☑ 865 14 XP_008201661.1 PREDICTED: glutamate receptor ionotropic, kainate 2-lik
☐ 1477 19 ERL88365.1 hypothetical protein D910_05752 [Dendroctonus ponderosae]
☐ 1712 20 ENN70426.1 hypothetical protein YQE_12931, partial [Dendroctonus ponde
☐ 1621 21 KAH1009747.1 hypothetical protein HUJ04_002056 [Dendroctonus ponderosa
☐ 860 22 XP_019771914.1 PREDICTED: glutamate receptor ionotropic, kainate 2 iso
☐ 909 23 XP_030751944.1 glutamate receptor ionotropic, kainate 2 [Sitophilus or
☐ 1662 24 CAH1374044.1 unnamed protein product [Tenebrio molitor]
☐ 911 25 XP_015840804.1 PREDICTED: glutamate receptor ionotropic, kainate 2 iso
```

S477/877 L15 - 38

38

To solve the progressive-alignment problems

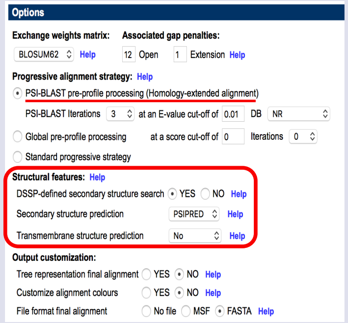
- **Incorporate more information to reduce early errors**
  - **Structural alignment (*e.g.*, Expresso, PROMALS3D, TM-Coffee, PRALINE, MAFFT-DASH)**
  - **Profile/profile-HMM alignment (*e.g.*, PRALINE, PSI-Coffee, PROMALS3D, ProbCons/CONTRAlign, ClustalΩ, MUSCLE5)**
- **Avoid the greedy-algorithm problem**
  - **Iterative refinement to search the global maxima**
    - **➔ A good objective function is required (*e.g.*, MUSCLE/MUSCLE5, MAFFT, ProbCons/CONTRAlign)**
- **Global (or local) only alignment problem**
  - **Combine both methods (*e.g.*, T-Coffee)**
- **More accurate insertion/deletion placement**
  - **Phylogeny aware gap-placement (*e.g.*, PRANK, ProPIP, Bali-Phy, SATé)**

BIOS477/877 L15 - 39

39

PRALINE (Simossis, Kleinjung & Heringa 2005)

https://www.ibi.vu.nl/programs/pralinewww/


Options

Exchange weights matrix: BLOSUM62 Help

Associated gap penalties: 12 Open 1 Extension Help

Progressive alignment strategy: Help

● PSI-BLAST pre-profile processing (Homology-extended alignment)

PSI-BLAST Iterations 3 at an E-value cut-off of 0.01 DB NR

○ Global pre-profile processing at a score cut-off of 0 Iterations 0

○ Standard progressive strategy

Structural features: Help

DSSP-defined secondary structure search ● YES ○ NO Help

Secondary structure prediction PSIPRED Help

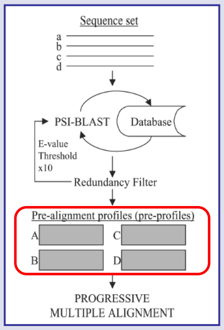
Transmembrane structure prediction No Help

Output customization:

Tree representation final alignment ○ YES ● NO Help

Customize alignment colours ○ YES ● NO Help

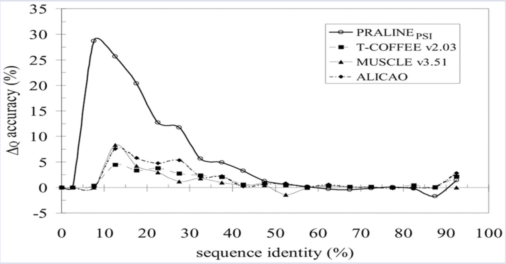
File format final alignment ○ No file ○ MSF ● FASTA Help



BIOS477/877 L16 - 40

40

PRALINE (Simossis, Kleinjung & Heringa 2005)



- From 624 HOMSTRAD pairwise alignments
- Q score: Sum-of-pairs (percentage of correctly aligned residue pairs)

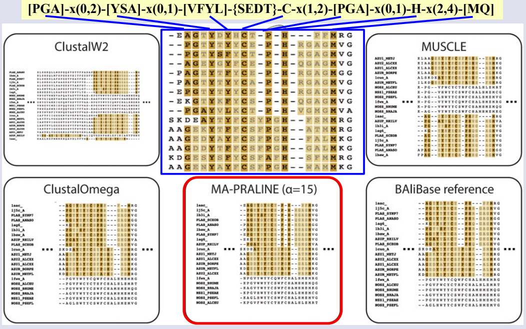
ΔQ: Q score difference from PRALINE without PSI-BLAST

BIOS477/877 L16 - 41

41

Motif-Aware PRALINE (Dijkstra *et al.* 2021)

Copper-binding site (PS00196):
[PGA]-x(0,2)-[YSA]-x(0,1)-[VFYL]-{SEDT}-C-x(1,2)-[PGA]-x(0,1)-H-x(2,4)-[MQ]



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

https://github.com/ibivu/MA-PRALINE

BIOS477/877 L16 - 42

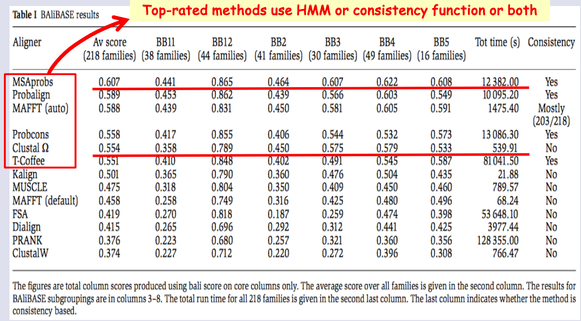
42

Clustal Ω http://www.clustal.org/omega/ https://www.ebi.ac.uk/jdispatcher/msa/clustalo

- **Progressive alignment** following the guide tree
- Features a fast method for making "guide trees"
  - ➜ calculates only distances to *n* references (mBed method)
  - ➜ scalable for very large datasets
- Alignment is done using HHalign (a profile hidden Markov model alignment)
  - ➜ highly accurate alignment
- Simple iterative refinement
  - ➜ Alignment is converted to hidden Markov model (HMM)
  - ➜ Realign input sequences against the HMM

Sievers *et al.* (2011, 2018)

Clustal Ω http://www.clustal.org/omega/ https://www.ebi.ac.uk/jdispatcher/msa/clustalo



Top-rated methods use HMM or consistency function or both

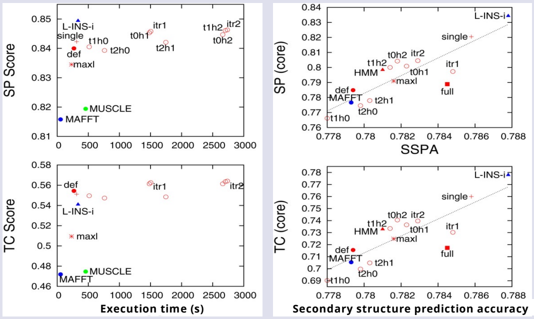
Table 1 BAliBASE results

Aligner	Av score (218 families)	BB11 (38 families)	BB12 (44 families)	BB2 (41 families)	BB3 (30 families)	BB4 (49 families)	BB5 (16 families)	Tot time (s)	Consistency
MSAprobs	0.607	0.441	0.865	0.464	0.607	0.622	0.608	12 382.00	Yes
Probalign	0.589	0.453	0.862	0.439	0.566	0.603	0.549	10 095.20	Yes
MAFFT (auto)	0.588	0.439	0.831	0.450	0.581	0.605	0.591	1475.40	Mostly (203/218)
Probcons	0.558	0.417	0.855	0.406	0.544	0.532	0.573	13 086.30	Yes
Clustal Ω	0.554	0.358	0.789	0.450	0.575	0.579	0.533	539.91	No
T-Coffee	0.551	0.410	0.848	0.402	0.491	0.545	0.587	81 041.50	Yes
Kalign	0.501	0.365	0.790	0.360	0.476	0.504	0.435	21.88	No
MUSCLE	0.475	0.318	0.804	0.350	0.409	0.450	0.460	789.57	No
MAFFT (default)	0.458	0.258	0.749	0.316	0.425	0.480	0.496	68.24	No
FSA	0.419	0.270	0.818	0.187	0.259	0.474	0.398	53 648.10	No
Dialign	0.415	0.265	0.696	0.292	0.312	0.441	0.425	3977.44	No
PRANK	0.376	0.223	0.680	0.257	0.321	0.360	0.356	128 355.00	No
ClustalW	0.374	0.227	0.712	0.220	0.272	0.396	0.308	766.47	No

The figures are total column scores produced using bali score on core columns only. The average score over all families is given in the second column. The results for BAliBASE subgroupings are in columns 3–8. The total run time for all 218 families is given in the second last column. The last column indicates whether the method is consistency based.

Sievers *et al.* (2011)

Clustal Ω http://www.clustal.org/omega/ https://www.ebi.ac.uk/jdispatcher/msa/clustalo



Sievers *et al.* (2018)