BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 **Lecture 16**

1

Today's topics

- Multiple Sequence Alignment
  - ClustalΩ, MUSCLE, Mafft
  - Phylogeny-aware gap placement methods (PRANK, etc.)
  - and more …
- Alignment Trimming/filtering

2

How to solve progressive-alignment problems

- Incorporate more information to reduce early errors
  - Structural alignment (e.g., Expresso, PROMALS3D, TM-Coffee, PRALINE, MAFFT-DASH, MUSCLE-3D)
  - 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)

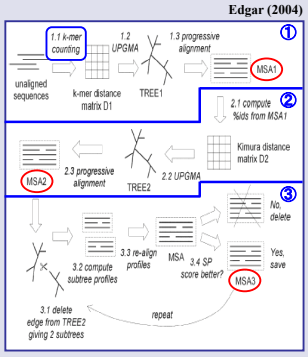
3

MUSCLE v3

Available at MUSCLE website (in v3 legacy page) and EBI Tools

Edgar (2004)

1. Draft progressive alignment:
   - Approximated distance based on frequencies of shared k-mers (k-tuples) & UPGMA
2. Improved progressive alignment:
   - Kimura protein distance (more accurate)
   - Alignment is done for subtrees where the branching patterns are changed between TREE1 and TREE2
3. Iterative refinement
   - The tree is partitioned
   - Profile is obtained from each subtree
   - Profile alignment
   - Iteration using Sum-of-Pairs (SP) score



4

MUSCLE v3 performance

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)

5

MAFFT

Available at MAFFT website or EBI Tools

Katoh *et al.* (2005, 2019)

1. First progressive alignment: NS-1
   - 6-tuple distance from pairwise alignment & modified UPGMA for guide tree
   - Alignment strategy:
     - FFT: Fast Fourier Transform approximation is used to detect highly similar segments → Segment to segment alignment
     - NW: full dynamic programming (global)

Alignment	FFT	NW
1st progressive	FFT-NS-1	NW-NS-1
2nd progressive	FFT-NS-2	NW-NS-2
Iterative refinement (i)	FFT-NS-i	NW-NS-i

2. Improved progressive alignment: NS-2
   - Better distance calculation from NS-1 alignment
3. Iterative refinement: NS-i
   - The tree-dependent restricted partitioning
   - Group-to-group (profile) alignment
   - Iteration based on the weighted SP score

Alignment	Consistency score
Global	G-INS-i
Local (SW)	L-INS-i
	E-INS-i

4. COFFEE-like consistency score from pairwise alignment information is used in G-INS-i (global), L-INS-i (local), and E-INS-i (local & more difficult alignment)

6

MAFFT performance

BAliBASE

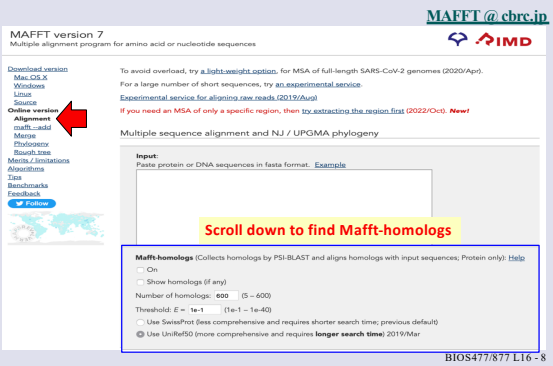
Method	Ref11	Ref12	Ref2	Ref3	Ref4	Ref5	Overall average	CPU time (s)
Consistency based methods								
MAFFT 5.662 L-INS-i	**67.11 / 44.61**	93.62 / 83.73	92.57 / **45.17**	85.58 / 56.83	**91.91 / 59.47**	**90.15** / 58.44	**87.05 / 58.64**	5,500
MAFFT 5.662 E-INS-i	66.13 / 44.53	93.54 / 83.18	**92.64** / 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**	91.87 / 40.54	84.60 / 54.30	90.52 / 54.37	89.28 / 56.50	86.48 / 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	86.37 / 52.51	87.87 / 52.75	84.23 / 52.64	6,900
TCoffee 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 / 21.38	72.50 / 27.23	78.82 / 39.55	74.24/ 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/TC
- **SP (sum-of-pairs) score: proportion of correctly aligned residue pairs**
- **TC (total column) score: proportion of fully correctly aligned columns**

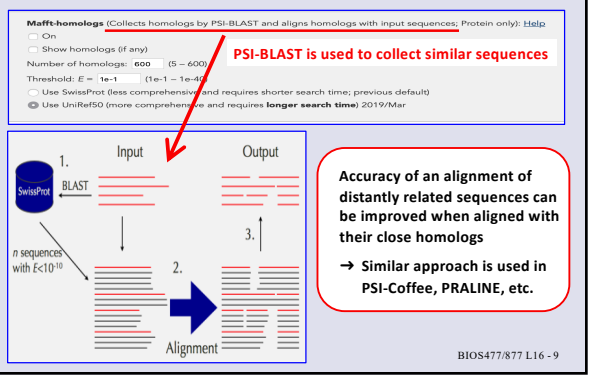
7

MAFFT website



8

Mafft-homologs



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

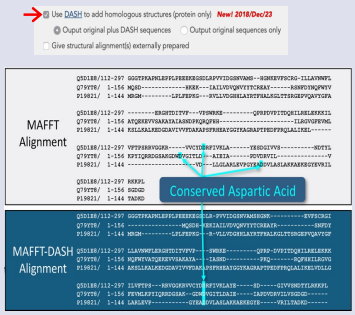
→ **Similar approach is used in PSI-Coffee, PRALINE, etc.**

9

MAFFT-DASH

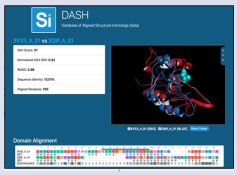
Rozewicki *et al.* (2019)

➢ **Integrated protein sequence and structural alignment**



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

→ **Similar approach is used in EXPRESSO, PROMALS3D, etc.**



10

MAFFT-DASH performance

Rozewicki *et al.* (2019)

Table 1. Benchmarks using reference MSAs

BAliBASE

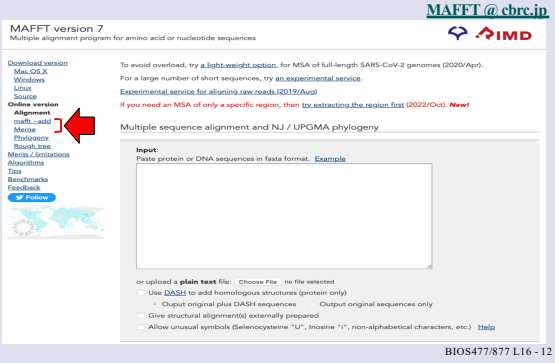
Methods \ Data	HMFM	MBSF	MBTL	OXFM	BB11	BB12	BB20	BB30	BB40	BB50	SY
[SP (sum-of-pairs) score: proportion of correctly aligned residue pairs]											
MAFFT	0.916**	0.571**	0.203**	0.894**	0.649**	0.937**	0.927**	**0.862**	**0.917**	0.899*	0.751**
Promals	0.947**	0.726**	0.475**	0.947**	**0.791**	**0.936**	0.933*	**0.883**	**0.898**	**0.903**	0.848**
T-Coffee	0.925**	0.585**	0.224**	0.909**	0.657**	**0.945**	0.916**	0.833**	**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**
PromalsID	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**	0.839**
[TC (total column) score: proportion of fully correctly aligned columns]											
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**
PromalsID	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.

3D information incorporated

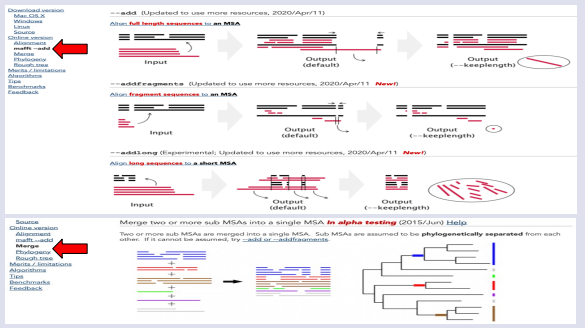
11

MAFFT: 'add' and 'merge'



12

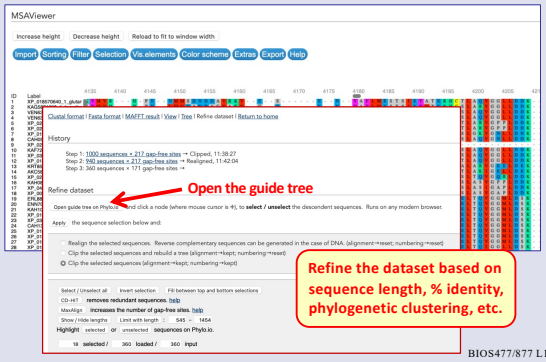
MAFFT: add/merge alignment



There are also --seed and --addprofile options. Read the "tips" section on the MAFFT website.

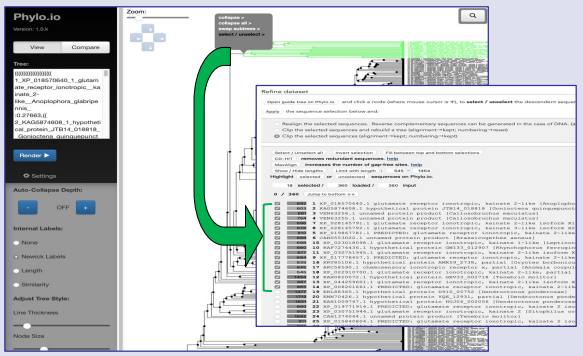
13

Refining alignment with MAFFT website



14

Refining alignment with MAFFT website



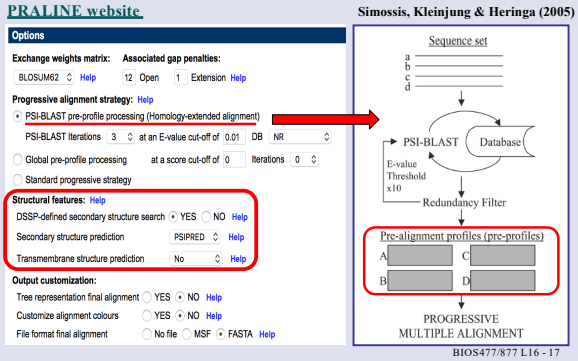
15

How to solve progressive-alignment problems

- Incorporate more information to reduce early errors
  - Structural alignment (e.g., Expresso, PROMALS3D, TM-Coffee, PRALINE, MAFFT-DASH, MUSCLE-3D)
  - 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)

16

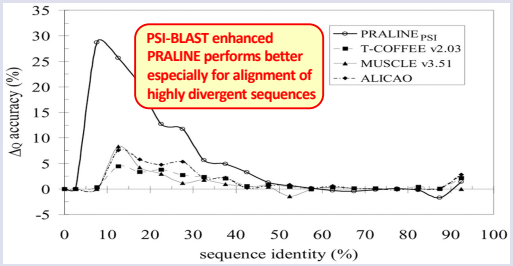
PRALINE



17

PRALINE performance

Simossis, Kleinjung & Heringa (2005)



- From 624 HOMSTRAD pairwise alignments
- Q score = SP (sum-of-pairs) score: % correctly aligned residue pairs
- ΔQ: Q score difference from PRALINE without PSI-BLAST

18

Clustal Ω

Available at EBI Tools Sievers *et al.* (2011, 2018)

- **Progressive alignment following the guide tree**
  - Features a fast method for making the "initial guide tree" (mBed)
    - → Calculates distances to only a small number (n=$\log_2 N$) of seed sequences: if $N = 1000$, $n = \log_2 1000 \approx 10$
    - → Sequences are represented as an n element vector, and clustered using Euclidean distances (UPGMA for subtrees)
    - → Scalable for very large datasets
  - Alignment is done using HHalign (a profile hidden Markov model alignment)
    - → Probabilistic, highly accurate alignment
  - Simple iterative refinement
    - → Alignment is converted to profile hidden Markov model (HMM)
    - → realigns input sequences against the profile HMM

19

Clustal Ω

Sievers *et al.* (2011)

Table 1 BAliBASE results

Top-rated methods use HMM or consistency function or both

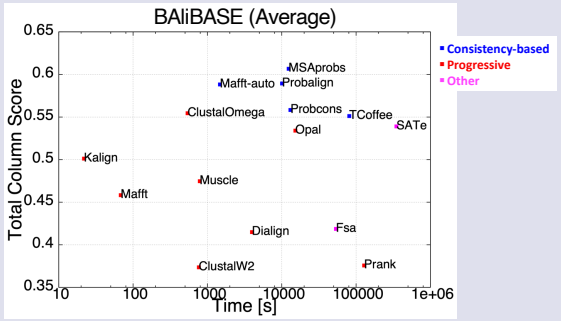
Aligner	Av score (218 families)	BB11 (38 families)	BB12 (44 families)	BB2 (41 families)	BB3 (30 families)	BB4 (49 families)	BB5 (16 families)	Tot time (s)	Consistency
MSAprobs	0.607	0.441	0.865	0.464	0.607	0.622	0.608	12 382.00	Yes
Probaligin	0.589	0.453	0.862	0.439	0.566	0.603	0.549	10 095.20	Yes
MAFFT (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.

20

Clustal Ω

Sievers *et al.* (2011)



21

Muscle5

Available at MUSCLE website or EBI Tools

Edgar (2022)

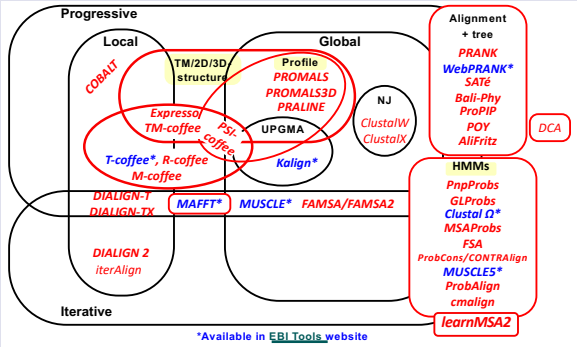
- **Improved re-implementation of ProbCons (slow but highly accurate MSA method based on hidden Markov model)**
  - Parallelization for rapid alignment
  - Can be used for large datasets
  - For both protein and DNA alignments

- More accurate than Clustal Ω, MAFFT, or ProbCons
- Ensemble MSAs can be generated for reference-free estimate of MSA accuracy

Method	Balibase (protein) Time (mins.)	Max mem (Mb)	TC	Bralibase (RNA) Time (mins.)	Max mem (Mb)	TC
Clustal-Omega	5	0.98	0.56	5	0.5	0.73
MAFFT	35	0.5	0.61	7	0.4	0.73
MUSCLE v5	27	2.1	0.63	6	0.8	0.78
MUSCLE v3	8	0.08	0.53	2	0.2	0.75
PROBCONS	182	0.54	0.62	n/a	n/a	n/a

22

Multiple sequence alignment methods



For reviews: Katoh (2021) and more available on CANVAS

23

PRANK, WebPRANK

Mind the gaps: Progress in progressive alignment

D. G. Higgins*, G. Blackshields, and I. M. Wallace
Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland

(2005 *PNAS* commentary)

"CLUSTALW attempts to compensate by using an elaborate scoring scheme to encourage gaps to end up on top of each other. ... results in alignments that are very "block-like"…"

"… there may be a price for this prettiness and detachment from phylogenetic reality. CLUSTALW (and other programs) may be guilty of "overalignment", that is where sequences that should not go together are forced into neat-looking blocks. These overaligned regions may be neat looking but misleading."

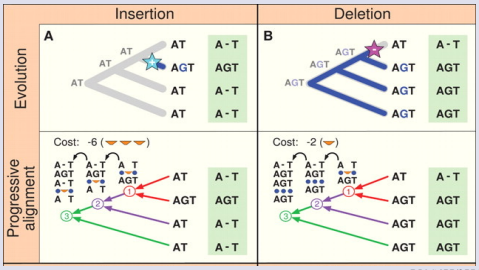
"There is an understandable tendency for users of multiple alignment software to want their residues neatly aligned in blocks and columns. This is fine when such blocks are biologically accurate as will happen in parts of protein alignments. In cases where insertions or deletions have happened in a less organized manner, as will happen in many noncoding DNA sequences and in less organized parts of protein sequences, such block-like alignments may be biologically meaningless. Perhaps we need to reeducate our eyes to see beauty in what actually happened rather than what looks nice on paper."

24

PRANK, WebPRANK

Löytynoja & Goldman (2008)

Insertions are more penalized than deletions in progressive sequence alignment.



25

PRANK, WebPRANK

Available at EBI website

Löytynoja & Goldman (2005, 2008, 2010)

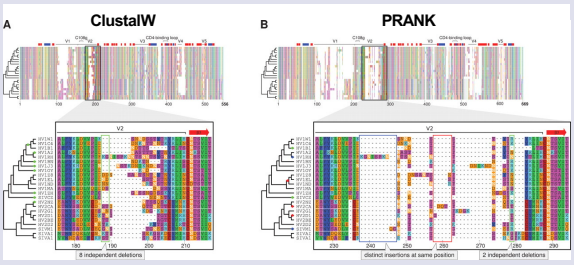
- PRANK: Probabilistic Alignment Kit
  - A probabilistic multiple alignment program for DNA, codon, and amino-acid sequences.
  - Treats insertions correctly.
  - Avoids over-estimation of the number of deletion events.
  - Not meant for the alignment of very diverged protein sequences.

26

PRANK, WebPRANK

Löytynoja & Goldman (2008)

Different sequence alignment approaches can give contradicting pictures of evolutionary mechanisms behind functional sequence changes.



MSAs generated by traditional methods show excess substitutions
➜ Can be erroneously thought to be under positive selection

27

ProPIP: Progressive MSA with Poisson Indel Process

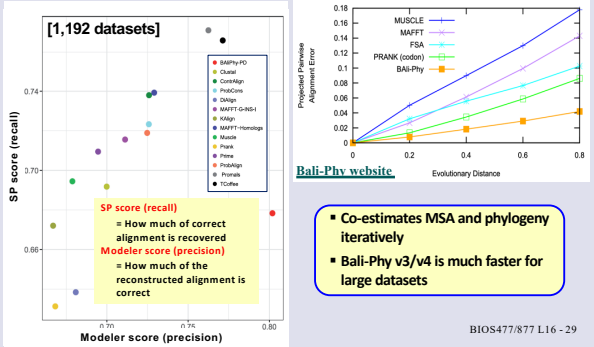
Maiolo *et al.* (2018, 2021)

- Rigorous mathematical indel model is incorporated
- Gaps can be inferred in a phylogenetically consistent way
- Over-alignment can be avoided

28

BAli-Phy: Statistical coestimation method

Nute *et al.* (2019), Redelings (2021), Gupta *et al.* (2021)



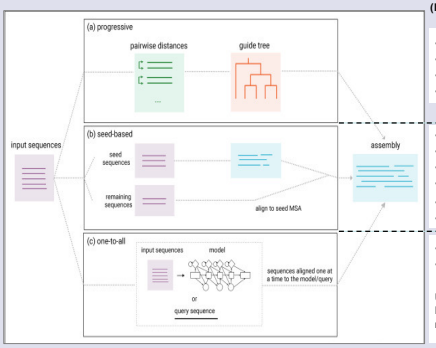
SP score (recall)
= How much of correct alignment is recovered
Modeler score (precision)
= How much of the reconstructed alignment is correct

Bali-Phy website

- Co-estimates MSA and phylogeny iteratively
- Bali-Phy v3/v4 is much faster for large datasets

29

Large-scale MSA



(large-scale MSA methods)

- MAFFT DPPartTree
- ClustalΩ
- FAMSA/FAMSA2
- Kalign3 ...
- PASTA
- UPP/UPP2
- MAGUS
- MAFFT/Sparsecore
- Regressive/T-Coffee
- MUSCLE5 ...
- MMseq2
- learnMSA/learnMSA2

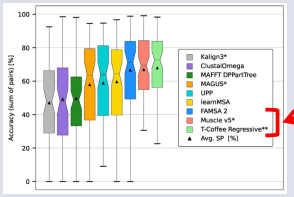
Using pHMM, deep learning protein language modeling, *etc.*

Santus *et al.* (2023)

30

Large-scale MSA

Santus *et al.* (2023)



[Data set size (N)]
- 25,000 ≈ 415,519
- Ultra-large Pfam family (N: 1.8M ≈ 3.5M)

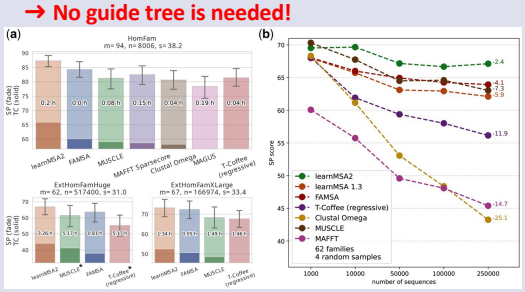
Algorithm	Response_reg. (PF00072). 1.8 M sequences				ABC_tran (PF00005). 3.5 M sequences			
	SP	TC	Time [hh:mm]	Memory [GB]	SP	TC	Time [hh:mm]	Memory [GB]
Kalign3	75.2	26.4	00:03	6.6	13.6	0.6	00:12	13.1
ClustalOmega	9.3	0.0	11:06	88.9	Failed at guide tree stage			
MAFFT DPPartTree	67.4	27.9	03:26	29.3	36.8	10.5	22:59	320.9
UPP	83.6	45.0	01:59	8.2	39.9	13.0	12:43	57.2
learnMSA	66.7	20.0	01:52	21.5	22.8	3.8	5:31	43.8
FAMSA	87.0	47.1	01:06	9.7	42.4	6.0	4:47	21.7
T-Coffee Regressive	75.1	35.0	17:22	16.1	Failed at ClustalOmega guide tree			

31

learnMSA2

Becker and Stanke (2024)

- Deep learning with profile HMM and language model
  - → No guide tree is needed!



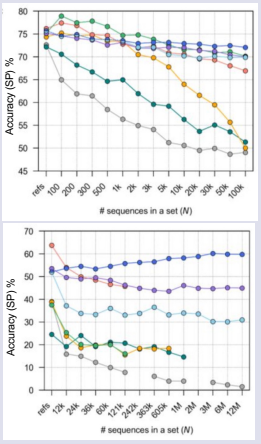
32

FAMSA2

- Progressive alignment at protein-universe scale

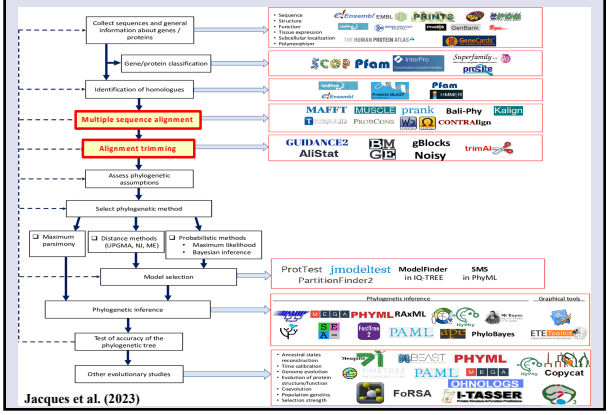
Gudyś *et al.* (2025)

	SP [%]	TC [%]	Total time [dd hh mm]	Memory [GB]
Kalign	61.9	43.2	1h 08m	140.2
MAFFT DPPartTree	62.9	44.3	8d 10h 57m	1067.7
Clustal Omega	63.8	46.9	7d 02h 44m	144.4
Muscle5	75.7	59.4	13d 03h 39m	225.2
T-Coffee regressive	78.1	61.3	36d 22h 58m	410.1
FAMSA	78.1	60.1	1d 07h 57m	46.9
FAMSA2	**79.6**	**61.8**	14h 57m	23.9
FAMSA2 medoid	76.5	58.5	**47m**	**17.7**



33

Protocol for reconstructing MSA and phylogeny



34

Alignment trimming

TABLE 1. Overview of filtering methods considered in this study

Filtering methods	Type of "undesirable" sites filtered out by the method	Accounts for tree structure?	Uses a substitution matrix or model of evolution?	Adapts parameters for particular data sets?	References
Gblocks	Gap-rich and variable sites	No	No	No	Talavera and Castresana (2007)
TrimAl	Gap-rich and variable sites	No	Yes	Yes	Capella-Gutiérrez et al. (2009)
Noisy	Homoplastic sites	In part	No	No	Dress et al. (2008)
Aliscore	Random-like sites	No	Indirectly	No	Kück et al. (2010)
BMGE	High entropy sites	No	Yes	No	Criscuolo and Gribaldo (2010)
Zorro	Sites with low posterior	Yes	Yes	No	Wu et al. (2012)
Guidance	Sites sensitive to the alignment guide tree	Yes	Indirectly	No	Penn et al. (2010)

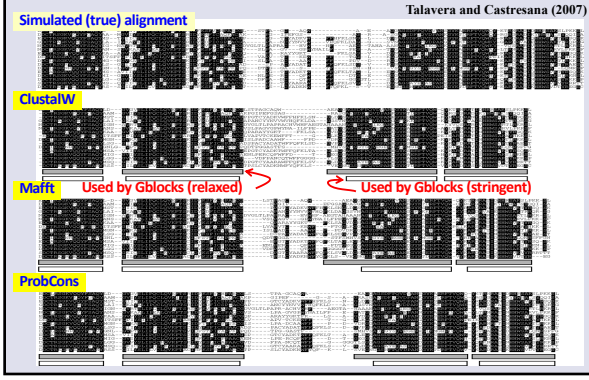
- **Gblocks**: selection of conserved blocks from multiple alignments (included in Phylogeny.fr)
- **trimAl**: a tool for automated alignment trimming in large-scale phylogenetic analysis (included in Phylemon2)

Tan *et al.* (2015); See also TCS paper by Chang *et al.* (2014)

35

Alignment trimming: Gblocks

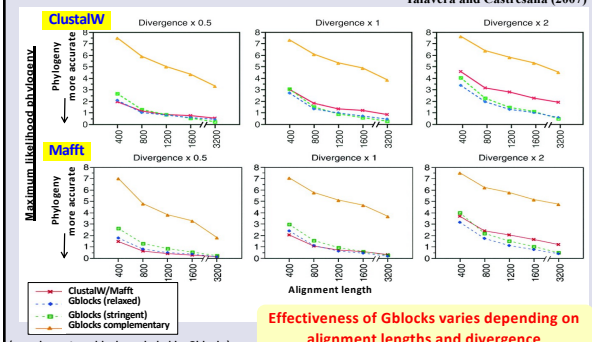
Talavera and Castresana (2007)



36

Gblocks performance

Talavera and Castresana (2007)



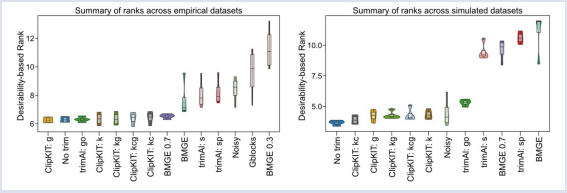
Effectiveness of Gblocks varies depending on alignment lengths and divergence

37

Alignment trimming: ClipKIT

Gblocks, trimAl: remove phylogenetically uninformative (gap-rich or highly variable) sites

ClipKIT: identifies and retain phylogenetically informative (parsimony-informative) sites



Desirability score (phylogeny-based accuracy): combines tree accuracy and average bootstrap support

Steenwyk *et al.* (2020)

38