Education

# Recent Evolutions of Multiple Sequence Alignment Algorithms

Cédric Notredame

An ever-increasing number of biological modeling methods depend on the assembly of an accurate multiple sequence alignment (MSA). These include phylogenetic trees, profiles, and structure prediction. Assembling a suitable MSA is not, however, a trivial task, and none of the existing methods have yet managed to deliver biologically perfect MSAs. Many of the algorithms published these last years have been extensively described [1–3], and this review focuses only on the latest developments, including meta-methods and template-based alignment techniques.

The purpose of an MSA algorithm is to assemble alignments reflecting the biological relationship between several sequences. Computing exact MSAs is computationally almost impossible, and in practice approximate algorithms (heuristics) are used to align sequences, by maximizing their similarity. The biological relevance of these MSAs is usually assessed by systematic comparison with established collections of structure-based MSAs ("gold standards"; for review see [4]). Since only a few sequences have known structures, the accuracy measured on the references is merely an estimation of how well a package may fare on standard datasets. Gold standards have had a considerable effect on the evolution of MSA algorithms, refocusing the entire methodological development toward the production of structurally correct alignments. Their use has also coincided with a notable algorithmic harmonization, most MSA packages being now based on the "progressive algorithm" [5]. This greedy heuristic assembly algorithm involves estimating a guide tree (rooted binary tree) from unaligned sequences, and then incorporating the sequences into the MSA with a pairwise alignment algorithm while following the tree topology. The progressive algorithm is often embedded in an iterative loop where the guide tree and the MSA are reestimated until convergence. Most MSA packages reviewed here [6–18] follow this canvas, albeit more or less extensively adapted for improved performances [1–3].

The scoring schemes used by the pairwise alignment algorithm are arguably the most influential component of the progressive algorithm. They can be divided in two categories: matrix- and consistency-based. Matrix-based algorithms such as ClustalW [14], MUSCLE [6], and Kalign [19] use a substitution matrix to assess the cost of matching two symbols or two profiled columns. Although profile statistics can be more or less sophisticated, the score for matching two positions depends only on the considered columns or their immediate surroundings. By contrast, consistency-based schemes incorporate a larger share of information into the evaluation. This result is achieved by using a recipe initially developed for T-Coffee [10] and inspired by Dialign overlapping weights [20]. Its principle is to compile a collection of pairwise global and local alignments (primary library) and to use this collection as a position-specific substitution matrix during a regular progressive alignment. The aim is to deliver a final MSA as consistent as possible with the alignments contained in the library. Many recent packages have built upon this initial framework. For instance, PCMA [15] decreases T-Coffee computational requirements by prealigning closely related sequences. ProbCons [7] uses Bayesian consistency and fills the primary library using the posterior decoding of a pair hidden Markov model. The substitution costs are estimated from this library using Bayesian statistics. MUMMALS [17] combines the ProbCons scoring scheme with the PCMA strategy, while including secondary structure predictions in its pair hidden Markov model. The most accurate flavors of MAFFT [8] (i.e., the GNS and LNS modes) use a T-Coffee–like evaluation. A majority of studies indicate that consistency-based methods are more accurate than their matrix-based counterparts [4], although they typically require an amount of CPU time *N* times higher than simpler methods (*N* being the number of sequences). Most of these methods are available online, either as downloadable packages or as online Web services (Table 1).

The wealth of available methods and their increasingly similar accuracies makes it harder than ever to objectively choose one over the others. Consensus methods such as M-Coffee [12] provide an interesting framework to address this problem. M-Coffee is a consensus meta-method based on T-Coffee. Given a sequence dataset, it fills up the library by using various MSA methods to compute alternative alignments. T-Coffee then uses this library to compute a final MSA consistent with the original alignments. When combining eight of the most accurate and distinct MSA packages, M-Coffee produces 67% of the time a better MSA than ProbCons, the best individual method [12]. Aside from its ease of extension, M-Coffee's main advantage is its ability to estimate the local consistency between the final alignment and the combined MSAs (CORE index [21]; Figure 1). This useful index has been shown to be well-correlated with the MSAs' structural correctness [21,22]. M-Coffee is not, however, the ultimate answer to the MSA problem, and its limited performances on remote homologs suggest that

**Editor:** Fran Lewitter, Whitehead Institute, United States of America

**Citation:** Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol 3(8): e123. doi:10.1371/journal.pcbi.0030123

**Abbreviations:** MSA, multiple sequence alignment

Cédric Notredame is with Information Génomique et Structurale, CNRS UPR2589, Institute for Structural Biology and Microbiology, Parc Scientifique de Luminy, Marseille, France. E-mail: cedric.notredame@europe.com

**Table 1.** Summary of the Methods Described in the Review

| Method | Score | Templates | Validation Values | | Server |
|---|---|---|---|---|---|
| | | | PreFab | HOMSTRAD | |
| ClustalW [14] | Matrix | — | 61.80 [12] | — | http://www.ebi.ac.uk/clustalw/ |
| Kalign | Matrix | — | 63.00 [18] | — | http://msa.cgb.ki.se/ |
| MUSCLE [6] | Matrix | — | 68.00 [16] | 45.0 [9] | http://www.drive5.com/muscle/ |
| T-Coffee [10] | Consistency | — | 69.97 [12] | 44.0 [9] | http://www.tcoffee.org/ |
| ProbCons [7] | Consistency | — | 70.54 [12] | — | http://probcons.stanford.edu/ |
| MAFFT [8] | Consistency | — | 72.20 [12] | — | http://align.genome.jp/mafft/ |
| M-Coffee [12] | Consistency | — | 72.91 [12] | — | http://www.tcoffee.org/ |
| MUMMALS [16] | Consistency | — | 73.10 [16] | — | http://prodata.swmed.edu/mummals/ |
| DbClustal [24] | Profiles | — | — | — | http://bips.u-strasbg.fr/PipeAlign/ |
| PRALINE [9] | Matrix | Profiles | — | 50.2 [9] | http://zeus.cs.vu.nl/programs/pralinewww/ |
| PROMALS [16] | Consistency | Profiles | 79.00 [16] | — | http://prodata.swmed.edu/promals/ |
| SPEM [28] | Matrix | Profiles | 77.00 [28] | — | http://sparks.informatics.iupui.edu/Softwares-Services_files/spem.htm |
| Expresso [13] | Consistency | Structures | — | 71.9 [11][a] | http://www.tcoffee.org/ |
| T-Lara [29] | Consistency | Structures | — | — | https://www.mi.fu-berlin.de/w/LiSA/ |

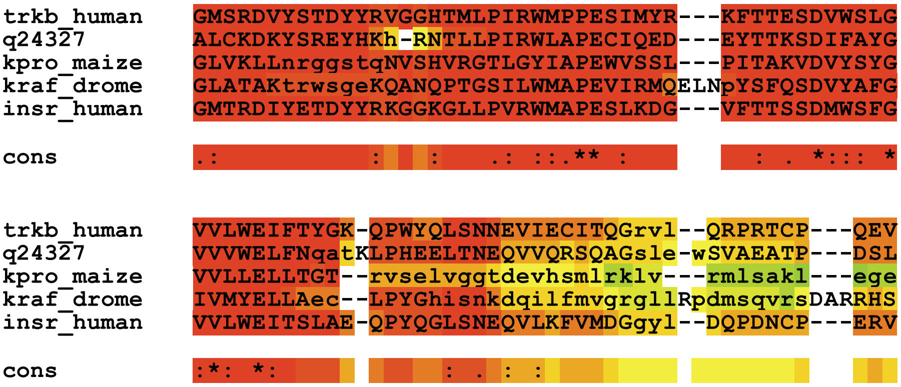
Validation values were compiled from several sources, and selected for comparability. PreFab validations were made using PreFab version 3. HOMSTRAD validations were made on datasets having less than 30% identity. The source of each value is indicated by the accompanying reference citation.

[a] The Expresso value comes from a slightly more demanding subset of HOMSTRAD (HOM39) made of sequences less than 25% identical.

doi:10.1371/journal.pcbi.0030123.t001

further improvement using only sequence information remains an elusive goal. Progress is nonetheless needed, and, at this point, the most promising approach is probably to incorporate within the datasets any information likely to improve the alignments, such as structural and homology data. Template-based alignment methods [13] follow this approach.

Structural extension was initially described by Taylor [23]. The principle is fairly straightforward (Figure 2) and involves identifying with BLAST a structural template in the Protein Data Bank for each sequence, aligning the templates using a structure superposition method, and mapping the original sequences onto their template's alignment. The resulting sequence alignments are compiled in the primary library and used by a consistency-based method to compute the final MSA. Homology extension was originally introduced in the DbClustal package [24] and works along the same lines, using a profile rather than a structure. PSI-BLAST is used to build a profile for each sequence, and these profiles are used as templates to generate better sequence alignments, thanks to the evolutionary information they contain. The only difference between homology and structure extension is the templates' nature and the associated alignment method. This generic approach can easily be extended to any kind of template. For instance, Expresso [13] uses SAP [25,26] and FUGUE [27] to align structural templates identified by a BLAST against the Protein Data Bank. PROMALS [17], PRALINE [9], and SPEM [28] make a profile–profile alignment with PSI-BLAST profiles used as templates. In PRALINE and PROMALS, the profile can be complemented with a secondary structure prediction in an attempt to improve the alignment accuracy. PROMALS uses ProbCons Bayesian consistency to fill its library with the posterior decoding of a pair hidden Markov model. T-Lara [29] uses

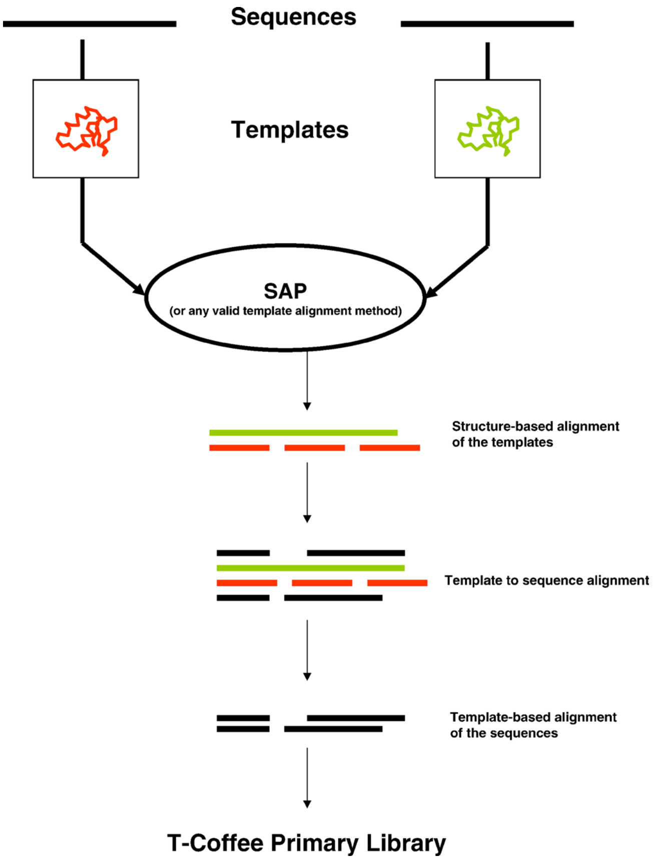


```
trkb_human   GMSRDVYSTDYYRVGGHTMLPIRWMPPESIMYR---KFTTESDVWSLG
q24327       ALCKDKYSREYHKh-RNTLLPIRWLAPECIQED---EYTTKSDIFAYG
kpro_maize   GLVKLLnrggstqNVSHVRGTLGYIAPEWVSSL---PITAKVDVYSYG
kraf_drome   GLATAKtrwsgeKQANQPTGSILWMAPEVIRMQELNpYSFQSDVYAFG
insr_human   GMTRDIYETDYYRKGGKGLLPVRWMAPESLKDG---VFTTSSDMWSFG

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


trkb_human   VVLWEIFTYGK-QPWYQLSNNEVIECITQGrvl--QRPRTCP---QEV
q24327       VVVWELFNqatKLPHEELTNEQVVQRSQAGsle-wSVAEATP---DSL
kpro_maize   VVLLELLTGT--rvselvggtdevhsmlrklv---rmlsakl---ege
kraf_drome   IVMYELLAec--LPYGhisnkdqilfmvgrgllRpdmsqvrsDARRHS
insr_human   VVLWEITSLAE-QPYQGLSNEQVLKFVMDGgyl--DQPDNCP---ERV

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

doi:10.1371/journal.pcbi.0030123.g001

**Figure 1.** Typical Output of M-Coffee

This output was obtained on the kinase1_ref5 BaliBase dataset, by combining MUSCLE, MAFFT, POA, Dialign-T, T-Coffee, ClustalW, PCMA, and ProbCons with M-Coffee. Correctly aligned residues (as judged from the reference) are uppercase; noncorrect ones are lowercase. The color of each residue indicates the agreement of the individual MSAs with respect to the alignment of that specific residue. Red indicates residues aligned in a similar fashion among all the individual MSAs; blue indicates very low agreement between MSAs. Dark yellow, orange, and red residues can be considered to be reliably aligned.


Sequences
Templates
SAP
(or any valid template alignment method)
Structure-based alignment of the templates
Template to sequence alignment
Template-based alignment of the sequences
T-Coffee Primary Library

doi:10.1371/journal.pcbi.0030123.g002

**Figure 2.** Framework of a Template-Based Method

Structural templates are first identified, mapped onto the sequences, and aligned using SAP. The sequence–template mapping is then used to guide the alignment of the original sequences. This alignment is integrated into the library that is used to compute the final MSA.

RNA secondary structure predictions as templates and fills a T-Coffee library with the Lara pairwise algorithm. With the exception of PRALINE and SPEM, which use a regular progressive algorithm, most template-based methods described here are consistency-based (some of them taking advantage of T-Coffee modular structure). Their main advantage is increased accuracy. Recent benchmarks on PROMALS (Table 1) show that homology extension results in a ten-point improvement over existing methods. Likewise, structure-based methods such as Expresso produce alignments much closer to the structural references than do any of their sequence-based counterparts. One must, however, be careful not to over-interpret validation values like that given for Expresso in Table 1, since both the reference and the Expresso alignments were computed using the same structural information.

This last point raises the important issue of method validation and benchmarking. A recent study [4] shows that with the exception of artificial datasets, benchmarks carried out on most reference databases tend to deliver compatible results. It also suggests that the best methods have become indistinguishable, except when considering remote homologs (less than 25% identity). Unfortunately, remote homologs are poorly suited to generating reference alignment, owing to the fact that their superposition often yields alternative sequence alignments that are structurally equivalent [30]. However, one can bypass the reference alignment stage by directly comparing the evaluated alignment to some idealized 3-D superposition. Such an alignment-independent evaluation has been described and used by several authors [17,31,32]. Another trend, not well accounted for by current reference collections, is the alignment of very large datasets. While many new methods incorporate special algorithms for aligning several hundred sequences [6,8,18], current reference databases do not allow the evaluation of very large datasets, thus making it unclear how the published accuracies scale with the number of sequences. While this last issue could probably be satisfyingly solved in the current benchmarking framework, another problem remains that is much harder to address. All the existing validation approaches have in common their reliance on the "one size fits all" assumption that structurally correct alignments are the best possible MSAs for modeling any kind of biological signal (evolution, homology, or function). A report on profile construction [33] has recently challenged this view by showing that structurally correct alignments do not necessarily result in better profiles. Likewise, it may be reasonable to ask whether better alignments always result in better phylogenetic trees, and, more systematically, to question and quantify the relationship between the accuracy of MSAs and the biological relevance of any model drawn upon them.

In this review, I have presented some of the latest additions to the MSA computation arsenal. An interesting milestone has been the development of meta-methods able to seamlessly combine the output of several methods. Aside from easing the user's work, the main advantage of these consensus methods is probably the local estimation of reliability they provide (Figure 1). Using this estimation to filter out unreliable regions has already proven useful in homology modeling [34] and could probably be used further. The main improvement reported here, however, is probably the notion of template-based alignment. Template-based alignment is more than a trivial extension of consistency-based methods. Under this new model, the purpose of an MSA is not to squeeze a dataset and extract all the information it may contain, but rather to use the dataset as a starting point for exploring and retrieving all the related information contained in public databases. This information is to be used not only for mapping purposes, but also for driving the MSA computation. Such a usage of sequence information makes template-based methods a real paradigm shift and a major step toward global biological data integration. ■

## Acknowledgments

The author thanks the two anonymous reviewers for suggesting several missing references.

**Author contributions.** CN analyzed the data and wrote the paper.

**Funding.** CN is funded and supported by the Centre National de la Recherche Scientifique, France.

**Competing interests.** The author has declared that no competing interests exist.

### References

1. Edgar RC, Batzoglou S (2006) Multiple sequence alignment. Curr Opin Struct Biol 16: 368–373.
2. Wallace IM, Blackshields G, Higgins DG (2005) Multiple sequence alignments. Curr Opin Struct Biol 15: 261–266.

3. Gotoh O (1999) Multiple sequence alignment: Algorithms and applications. Adv Biophys 36: 159–206.
4. Blackshields G, Wallace IM, Larkin M, Higgins DG (2006) Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol 6: 321–339.
5. Hogeweg P, Hesper B (1984) The alignment of sets of sequences and the construction of phylogenetic trees. An integrated method. J Mol Evol 20: 175–186.
6. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
7. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15: 330–340.
8. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511–518.
9. Simossis VA, Heringa J (2005) PRALINE: A multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res 33: W289–W294.
10. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–217.
11. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C (2004) 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. J Mol Biol 340: 385–395.
12. Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: Combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34: 1692–1699.
13. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, et al. (2006) Expresso: Automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 34: W604–W608.
14. Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4690.
15. Pei J, Sadreyev R, Grishin NV (2003) PCMA: Fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 19: 427–428.
16. Pei J, Grishin NV (2007) PROMALS: Towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23: 802–808.
17. Pei J, Grishin NV (2006) MUMMALS: Multiple sequence alignment improved by using hidden Markov models with local structural information. Nucleic Acids Res 34: 4364–4374.
18. Lassmann T, Sonnhammer EL (2005) Kalign—An accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6: 298.
19. Lassmann T, Sonnhammer EL (2006) Kalign, Kalignvu and Mumsa: Web servers for multiple sequence alignment. Nucleic Acids Res 34: W596–W599.
20. Morgenstern B, Dress A, Wener T (1996) Multiple DNA and protein sequence based on segment-to-segment comparison. Proc Natl Acad Sci U S A 93: 12098–12103.
21. Notredame C, Abergel C (2003) Using multiple alignment methods to assess the quality of genomic data analysis. In: Andrade M, editor. Bioinformatics and genomes: Current perspectives. Wymondham (United Kingdom): Horizon Scientific Press. pp. 30–50.
22. Lassmann T, Sonnhammer EL (2005) Automatic assessment of alignment quality. Nucleic Acids Res 33: 7120–7128.
23. Taylor WR (1986) Identification of protein sequence homology by consensus template alignment. J Mol Biol 188: 233–258.
24. Thompson JD, Plewniak F, Thierry J, Poch O (2000) DbClustal: Rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res 28: 2919–2926.
25. Taylor WR, Orengo CA (1989) Protein structure alignment. J Mol Biol 208: 1–22.
26. Taylor WR (1999) Protein structure comparison using iterated double dynamic programming. Protein Sci 8: 654–665.
27. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310: 243–257.
28. Zhou H, Zhou Y (2005) SPEM: Improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 21: 3615–3621.
29. Bauer M, Klau G, Reinert K (2005) Multiple structural RNA alignment with Lagrangian relaxation. Lect Notes Comput Sci 3692: 303–314.
30. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS (2000) ProSup: A refined tool for protein structure alignment. Protein Eng 13: 745–752.
31. O'Sullivan O, Zehnder M, Higgins D, Bucher P, Grosdidier A, et al. (2003) APDB: A novel measure for benchmarking sequence alignment methods without reference alignments. Bioinformatics 19: I215–I221.
32. Armougom F, Moretti S, Keduas V, Notredame C (2006) The iRMSD: A local measure of sequence alignment accuracy using structural information. Bioinformatics 22: e35–e39.
33. Griffiths-Jones S, Bateman A (2002) The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. Bioinformatics 18: 1243–1249.
34. Claude JB, Suhre K, Notredame C, Claverie JM, Abergel C (2004) CaspR: A web server for automated molecular replacement using homology modelling. Nucleic Acids Res 32: W606–W609.