ELSEVIER

Multiple sequence alignment

Robert C Edgar[1] and Serafim Batzoglou[2]

Multiple sequence alignments are an essential tool for protein structure and function prediction, phylogeny inference and other common tasks in sequence analysis. Recently developed systems have advanced the state of the art with respect to accuracy, ability to scale to thousands of proteins and flexibility in comparing proteins that do not share the same domain architecture. New multiple alignment benchmark databases include PREFAB, SABMARK, OXBENCH and IRMBASE. Although CLUSTALW is still the most popular alignment tool to date, recent methods offer significantly better alignment quality and, in some cases, reduced computational cost.

Addresses
[1] 45 Monterey Drive, Tiburon, CA, USA
[2] Department of Computer Science, Stanford University, Stanford, CA 94305-9025, USA

Corresponding author: Edgar, Robert C (bob@drive5.com)

Current Opinion in Structural Biology 2006, **16**:368–373

This review comes from a themed issue on
Sequences and topology
Edited by Nick V Grishin and Sarah A Teichmann

Available online 5th May 2006

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2006.04.004

Introduction

A multiple sequence alignment (MSA) arranges protein sequences into a rectangular array with the goal that residues in a given column are homologous (derived from a single position in an ancestral sequence), superposable (in a rigid local structural alignment) or play a common functional role. Although these three criteria are essentially equivalent for closely related proteins, sequence, structure and function diverge over evolutionary time and different criteria may result in different alignments. Manually refined alignments continue to be superior to purely automated methods; there is therefore a continuous effort to improve the biological accuracy of MSA tools. Additionally, the high computational cost of most naive algorithms motivates improvements in speed and memory usage to accommodate the rapid increase in available sequence data. In this review, we describe the state of the art in MSA software and benchmarking, and offer our recommended procedures for creating multiple alignments from typical types of input data.

Computational approaches to multiple sequence alignment

MSA algorithm development is an active area of research two decades after the first programs were written. The standard computational formulation of the pairwise problem is to identify the alignment that maximizes protein sequence similarity, which is typically defined as the sum of substitution matrix scores for each aligned pair of residues, minus some penalties for gaps. The mathematically — though not necessarily biologically — exact solution can be found in a fraction of a second for a pair of proteins. This approach is generalized to the multiple sequence case by seeking an alignment that maximizes the sum of similarities for all pairs of sequences (the sum-of-pairs, or SP, score).

The SP score is the foundation of many MSA algorithms, but has a number of drawbacks. The minimum possible computational time and memory required to maximize the SP score has been shown to scale exponentially with the number of sequences [1] and is not practical for more than a handful of sequences on current computers. Heuristic or approximate alternatives are therefore required for typical input data. The most widely used approach to construct a multiple alignment is 'progressive alignment' [2], whereby a set of N proteins are aligned by performing N–1 pairwise alignments of pairs of proteins or pairs of intermediate alignments, guided by a phylogenetic tree connecting the sequences.

In contrast to the pairwise case, the SP score has no rigorous theoretical foundation and, in particular, fails to exploit phylogeny or incorporate an evolutionary model. SP, like most other scores in common use, assumes that the input sequences are globally alignable, that is to say, substitutions and small insertions and deletions are the only mutational events separating the sequences. If full-length sequences are used, this implies that all proteins must have the same domain organization (the same domains in the same order); otherwise, the user is required to identify globally alignable subsequences, such as a common domain, before creating an MSA. For known domains, tools such as PFAM [3] can be used; progress towards an automated solution is demonstrated by the recently released ProDA program (http://proda.stanford.edu).

A methodology that has been successfully used as an improvement of progressive alignment based on the SP formulation is 'consistency-based' scoring [4–6]. Given three sequences, A, B and C, the pairwise alignments A-B and B-C imply an alignment of A and C that may be different from the directly computed A-C alignment.

This motivates a search for an MSA that maximizes agreement ('consistency') with a set of pairwise alignments computed for the input sequences [7,8,9••].

Benchmarks

Validation of an MSA program typically uses a benchmark data set of reference alignments. An alignment produced by the program is compared with the corresponding reference alignment, giving an accuracy score. Alignments of protein structures can be generated without considering sequence and can therefore be used as independent references for sequence-based methods. Unfortunately, multiple structure alignment is also a hard problem, so, in practice, pairwise structure alignments are often used.

Before 2004, the *de facto* standard benchmark was BAliBASE [10], a database created by a combination of automated and manual methods. Recently, several new benchmarks have appeared, including OXBENCH [11], PREFAB [12•], SABmark [13], IRMBASE [14] and a new, extended version of BALIBASE (http://www-bio3d-igbmc.u-strasbg.fr/balibase/). The new benchmarks are largely constructed by automated means, in contrast to the labor-intensive protocol used for BAliBASE. As a result, reference alignments have varying quality and the accuracy of results for a given alignment is often questionable; however, the relative ranking of MSA programs can be reliably achieved by averaging over a large set.

Most of the reference alignments in these databases contain globally alignable sequences, and measure sensitivity (the number of correctly aligned positions) but not specificity (i.e. there is no penalty for aligning non-homologous regions). These are significant issues in practice, as sequences collected by local similarity search methods are often not globally alignable. To assess the specificity of an alignment tool, the f_M measure, as used by SABmark, identifies the proportion of matched residues predicted that also appear in a reference alignment [15]. For sequences of known structure, some regions are clearly alignable and some are clearly not alignable; however, there are usually also intermediate cases, whereby an arbitrary structure divergence cutoff is needed, or sequence and structure have diverged to the point at which homology is not reliably detectable. As a result, the f_M score, at best, provides a noisy assessment of alignment tool specificity, one that becomes increasingly less reliable as one considers sequences of increasing structural divergence. IRMBASE uses simulated sequence data to test both sensitivity and specificity; however, the relationship of these simulations to evolutionary models of real biological sequences is not well understood.

Methods

CLUSTALW [16] was introduced in 1994 and quickly became the method of choice for biologists, as it represented dramatic progress in alignment sensitivity combined with speed compared with other existing tools. CLUSTALW is still the most widely used MSA program. However, to the best of our knowledge, no significant improvements have been made to the algorithm since 1994 and several modern methods achieve better performance in accuracy, speed or both.

In the category of global alignment tools that are directly comparable to CLUSTALW, we consider the best current programs to be MAFFT [17,18], MUSCLE [12•,19], T-COFFEE [7] and PROBCONS [9••]. MAFFT and MUSCLE have a similar design, building on work done by Gotoh in the 1990s that culminated in the PRRN and PRRP programs [20,21], which achieved the best accuracy of their time but were relatively slow and were not widely adopted. T-COFFEE is the prototypical consistency-based method; it is still among the most accurate available programs. More recently, PROBCONS introduced a consistency-based approach using a probabilistic model and maximum expected accuracy scoring [22]. For divergent sequences, consistency-based methods often have an advantage in terms of accuracy, but frequently this comes at the expense of computational resources. Because of speed and memory requirements, PROBCONS and T-COFFEE have a practical limit of around 100 sequences on current desktop computers. MAFFT and MUSCLE offer significant improvements in scalability with comparable accuracy, and thus provide reasonable starting points for general alignment problems.

The recently published methods ALIGN-M [23], DIALIGN [8,14,24], POA [25,26] and SATCHMO [27] have relaxed the requirement for global alignability by allowing both alignable and non-alignable regions. Although these methods are sometimes described as 'local', alignable regions must still be co-linear (i.e. appear in the same order in each sequence). This model is appropriate for protein families with well-conserved core blocks surrounded by variable regions, but not when input sequences have different domain organizations. DIALIGN takes an 'all-or-nothing' view: a column is either alignable or is not, whereas POA and SATCHMO allow the extent of alignable regions to vary, permitting longer alignments between closely related subfamilies and shorter alignments for the complete set of sequences. ALIGN-M produces an all-or-nothing multiple alignment and a set of pairwise alignments guided by consistency. These methods perform relatively poorly on global benchmarks, which, as noted earlier, measure only sensitivity; nonetheless, they provide alternatives that might be useful when the 'overalignment' common to regular global alignment methods is undesirable.

Until recently, no MSA program was truly local, so as to be capable of producing multiple alignments of homologous regions in proteins with different domain organizations. A

conceptual advance in this direction was made with ABA [28••], which produces a graphical representation of the relationships between a set of sequences, but stops short of providing explicit alignments of similar regions. A new truly local method, ProDA (P Tu-Minh, CB Do, RC Edgar, S Batzoglou, unpublished), has recently been made available online (http://proda.stanford.edu/); the tool, which is still largely experimental, provides the first step towards dealing with this challenging new breed of alignment tasks.

Improvements in alignment accuracy can be achieved by incorporating additional data beyond the input sequences. Recent examples include 3DCOFFEE [29], which exploits one or more protein structures, PRALINE [30], which uses PSI-BLAST to collect homologs and build a profile for each sequence, and SPEM [31], which builds profiles and also uses predicted secondary structure. Because these last two tools rely on PSI-BLAST queries, performing an alignment takes several orders of magnitude longer than the standalone applications described above; however, these tools can be extremely useful to a biologist interested in a single protein, as they automate the process of identifying homologs for improvement of alignment quality.

Finally, CONTRAlign [30] is a newly developed experimental protein alignment tool that uses discriminative learning techniques recently introduced in the machine learning literature. On carefully cross-validated benchmark tests, CONTRAlign has demonstrated substantial improvements in low-identity pairwise alignment accuracy; the effectiveness of generalizing the algorithm to the multiple sequence case is currently unknown.

Choosing a program

There are three main considerations in choosing a program: biological accuracy, execution time and memory usage (Tables 1 and 2). Biological accuracy is generally the most important concern. The most accurate programs according to benchmark tests are MAFFT, MUSCLE, PROBCONS and T-COFFEE. On most benchmarks, PROBCONS achieves the best performance; recent versions of the MAFFT tool achieve comparable results by incorporating consistency-based scoring.

In practice, accuracy claims can be difficult to validate due to the frequent practice of parameter tuning to optimize performance on one or more benchmarks. Some methods are, in principle, more immune to that caveat, because of unsupervised training, as in PROBCONS, or rigorous cross-validation, as in the experimental CONTRAlign tool. Furthermore, many benchmarking databases contain over-represented sequence families, thus invalidating significance tests that assume all test samples to be independent. Regardless, on nearly all benchmarks, new methods outperform the CLUSTALW tool in terms of average accuracy; for the practitioner, using any of the new methods may give significant gains.

Benchmark scores are typically based on averages over many alignments; on any given test, the rankings may be different. For example, PROBCONS v1.08 aligns an average of 91% of positions correctly on tests in BAliBASE v2.0 compared with 86% for CLUSTALW v1.8; on test 1aboA [1], CLUSTALW aligns 76% correctly compared with 67% for PROBCONS. Thus far, attempts to predict which method will work best on a given set of sequences have not been successful. When accuracy of a particular protein alignment is paramount, we recommend using two or three programs based on distinctively different algorithms (e.g. T-COFFEE, PROBCONS and MUSCLE) and comparing the outputs using a tool such as the ALTAVIST web server [32]. Regions of agreement are more likely to be correctly aligned.

Another consideration is computational expense. Whereas T-COFFEE and PROBCONS may be good choices for multiple alignment of up to 100 protein sequences because of their high accuracy, they don't

Table 1

Summary of MSA programs that we consider to be the best currently available

Program	Advantages	Cautions
CLUSTALW	Uses less memory than other programs	Less accurate or scalable than modern programs
DIALIGN	Attempts to distinguish between alignable and non-alignable regions	Less accurate than CLUSTALW on global benchmarks
MAFFT, MUSCLE	Faster and more accurate than CLUSTALW; good trade-off of accuracy and computational cost. Options to run even faster, with lower average accuracy, for high-throughput applications.	For very large data sets (say, more than 1000 sequences) select time- and memory-saving options
PROBCONS	Highest accuracy score on several benchmarks	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)
ProDA	Does not assume global alignability; allows repeated, shuffled and absent domains.	High computational cost and less accurate than CLUSTALW on global benchmarks
T-COFFEE	High accuracy and the ability to incorporate heterogeneous types of information	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)

Table 2

Typical alignment tasks and our recommended procedures

Input data	Recommendations
2–100 sequences of typical protein length (maximum around 10,000 residues) that are approximately globally alignable	Use PROBCONS, T-COFFEE, and MAFFT or MUSCLE, compare the results using ALTAVIST. Regions of agreement are more likely to be correct. For sequences with low percent identity, PROBCONS is generally the most accurate, but incorporating structure information (where available) via 3DCoffee (a variant of T-COFFEE) can be extremely helpful.
100–500 sequences that are approximately globally alignable	Use MUSCLE or one of the MAFFT scripts with default options. Comparison using ALTAVIST is possible, but the results are hard to interpret with larger numbers of sequences unless they are highly similar.
>500 sequences that are approximately globally alignable	Use MUSCLE with a faster option (we recommend maxiters-2) or one of the faster MAFFT scripts
Large numbers of alignments, high-throughput pipeline.	Use MUSCLE with faster options (e.g. maxiters-1 or maxiters-2) or one of the faster MAFFT scripts
2–100 sequences with conserved core regions surrounded by variable regions that are not alignable	Use DIALIGN
2–100 sequences with one or more common domains that may be shuffled, repeated or absent.	Use ProDA
A small number of unusually long sequences (say, >20,000 residues)	Use CLUSTALW. Other programs may run out of memory, causing an abort (e.g. a segmentation fault).

scale much beyond that on modern desktops. For high-throughput applications, MAFFT and MUSCLE offer options for creating alignments at very high speeds with accuracies comparable to that of CLUSTALW.

As a final note, many of the MSA algorithms above were designed in the context of protein sequence alignment, although their algorithms transfer naturally to the domain of small-scale DNA or RNA multiple alignment as well (for which issues such as large-scale genome rearrangements are less problematic). Interestingly, recent attempts at benchmarking some of these methods on RNA structural alignments [33•] demonstrated good overall performance by CLUSTALW; also, the authors noted that tuning parameters for MAFFT gave considerable accuracy benefits. In general, the biological accuracy of MSA methods on non-protein DNA sequences is hard to determine because of the lack of trusted reference alignments, hindering algorithm development and evaluation.

Future directions

Multiple alignment of protein sequences will remain an important application in the foreseeable future. The number of newly available protein sequences still far outpaces the number of determined protein three-dimensional structures, and therefore sequence homology remains the main method by which to infer protein structure, function, active sites and evolutionary history.

In recent years, protein MSA tools have improved rapidly in both scalability and accuracy. Future improvements are likely to come by combining sequence alignment with other information, such as known structures of some of the proteins being aligned or homology to a larger pool of proteins. Parameter selection for alignment tools remains an important problem, as demonstrated by the sensitivity of RNA benchmarking results to parameter choice. Algorithmically, consideration of all sequences at once as an alternative to progressive alignment (consistency-based methods are a step in this direction) has been shown to be an effective strategy. Finally, better utilization of phylogenetic relationships and incorporation of models of protein sequence evolution also hold promise for improved alignment performance.

More broadly, organization of protein space will become increasingly relevant, as new sequences, structures and functional information become available. Given a newly obtained set of proteins, automated methods should be capable of placing them in detailed databases that will infer domain organization, structure, evolutionary relationships and function, including enzymatic activity, and protein–ligand and protein–protein interactions. The hypothesis that a limited number of folds account for all proteins, introduced 15 years ago [34], continues to be confirmed as our repository of three-dimensional structures expands [35]. Recently, protein classification methods have improved dramatically [36]. Protein evolution is complicated because, in addition to point mutations, it involves duplications, horizontal transfers, fusions and other events. Sequencing of environmental samples [37–39] is a new source of protein populations with intriguing evolutionary relationships. Multiple protein sequence comparison methods promise to continue to be central to the study of molecular biology and evolution.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Wang L, Jiang T: **On the complexity of multiple sequence alignment**. *J Comput Biol* 1994, **1**:337-348.

2. Feng DF, Doolittle RF: **Progressive sequence alignment as a prerequisite to correct phylogenetic trees**. *J Mol Evol* 1987, **25**:351-360.

3. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al.*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-D141.

4. Notredame C, Holm L, Higgins DG: **COFFEE: an objective function for multiple sequence alignments**. *Bioinformatics* 1998, **14**:407-422.

5. Gotoh O: **Consistency of optimal sequence alignments**. *Bull Math Biol* 1990, **52**:509-525.

6. Vingron M, von Haeseler A: **Towards integration of multiple alignment and phylogenetic tree construction**. *J Comput Biol* 1997, **4**:23-34.

7. Notredame C, Higgins DG, Heringa J: **T-Coffee: a novel method for fast and accurate multiple sequence alignment**. *J Mol Biol* 2000, **302**:205-217.

8. Morgenstern B, Frech K, Dress A, Werner T: **DIALIGN: finding local similarities by multiple sequence alignment**. *Bioinformatics* 1998, **14**:290-294.

9. •• Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: **ProbCons: probabilistic consistency-based multiple sequence alignment**. *Genome Res* 2005, **15**:330-340.

This paper describes a new approach to MSA based on a probabilistic model and a maximum-accuracy objective score. A program implementing these ideas, PROBCONS, is shown to achieve better accuracy on current benchmarks than the best current programs.

10. Thompson JD, Plewniak F, Poch O: **BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs**. *Bioinformatics* 1999, **15**:87-88.

11. Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ: **OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy**. *BMC Bioinformatics* 2003, **4**:47.

12. • Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**:1792-1797.

The authors introduce a new program, MUSCLE, and a new accuracy benchmark, PREFAB. MUSCLE is shown to achieve accuracy equal to or better than the best current programs and is typically much faster.

13. Van Walle I, Lasters I, Wyns L: **SABmark-a benchmark for sequence alignment that covers the entire known fold space**. *Bioinformatics* 2005, **21**:1267-1268.

14. Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B: **DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment**. *BMC Bioinformatics* 2005, **6**:66.

15. Sauder JM, Arthur JW, Dunbrack RL Jr: **Large-scale comparison of protein sequence alignment algorithms with structure alignments**. *Proteins* 2000, **40**:6-22.

16. Thompson JD, Higgins DG, Gibson TJ: **CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673-4680.

17. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment**. *Nucleic Acids Res* 2005, **33**:511-518.

18. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform**. *Nucleic Acids Res* 2002, **30**:3059-3066.

19. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.

20. Gotoh O: **Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments**. *J Mol Biol* 1996, **264**:823-838.

21. Gotoh O: **A weighting system and algorithm for aligning many phylogenetically related sequences**. *Comput Appl Biosci* 1995, **11**:543-551.

22. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological Sequence Analysis*. Cambridge University Press; 1998.

23. Van Walle I, Lasters I, Wyns L: **Align-m-a new algorithm for multiple alignment of highly divergent sequences**. *Bioinformatics* 2004, **20**:1428-1435.

24. Morgenstern B: **DIALIGN: 2 improvement of the segment-to-segment approach to multiple sequence alignment**. *Bioinformatics* 1999, **15**:211-218.

25. Grasso C, Lee C: **Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems**. *Bioinformatics* 2004, **20**:1546-1556.

26. Lee C, Grasso C, Sharlow MF: **Multiple sequence alignment using partial order graphs**. *Bioinformatics* 2002, **18**:452-464.

27. Edgar RC, Sjölander K: **SATCHMO: sequence alignment and tree construction using hidden Markov models**. *Bioinformatics* 2003, **19**:1404-1411.

28. •• Raphael B, Zhi D, Tang H, Pevzner P: **A novel method for multiple alignment of sequences with repeated and shuffled elements**. *Genome Res* 2004, **14**:2336-2346.

A formulation for the multiple alignment of proteins with repeated, shuffled and absent domains.

29. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C: **3DCoffee: combining protein sequences and structures within multiple sequence alignments**. *J Mol Biol* 2004, **340**:385-395.

30. Simossis VA, Heringa J: **PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information**. *Nucleic Acids Res* 2005, **33**:W289-W294.

31. Zhou H, Zhou Y: **SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures**. *Bioinformatics* 2005, **21**:3615-3621.

32. Morgenstern B, Goel S, Sczyrba A, Dress A: **AltAVisT: comparing alternative multiple sequence alignments**. *Bioinformatics* 2003, **19**:425-426.

33. • Gardner PP, Wilm A, Washietl S: **A benchmark of multiple sequence alignment programs upon structural RNAs**. *Nucleic Acids Res* 2005, **33**:2433-2439.

This paper describes the first RNA alignment benchmark and a comparison of alignment programs on this benchmark. CLUSTALW, MAFFT and MUSCLE are all shown to have good performance.

34. Chothia C: **Proteins. One thousand families for the molecular biologist**. *Nature* 1992, **357**:543-544.

35. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536-540.

36. Weston J, Leslie CS, Zhou D, Elisseeff A, Noble WS: **Semi-supervised protein classification using cluster kernels**. *Bioinformatics* 2005, **21**:3241-3247.

37. Tringe SG, Rubin EM: **Metagenomics: DNA sequencing of environmental samples**. *Nat Rev Genet* 2005, **6**:805-814.

38. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC *et al.*: **Comparative metagenomics of microbial communities**. *Science* 2005, **308**:554-557.

39. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al.*: **Environmental genome shotgun sequencing of the Sargasso sea**. *Science* 2004, **304**:66-74.