Full text provided by www.sciencedirect.com

Multiple sequence alignments

Iain M Wallace, Gordon Blackshields and Desmond G Higgins

Multiple sequence alignments are very widely used in all areas of DNA and protein sequence analysis. The main methods that are still in use are based on 'progressive alignment' and date from the mid to late 1980s. Recently, some dramatic improvements have been made to the methodology with respect either to speed and capacity to deal with large numbers of sequences or to accuracy. There have also been some practical advances concerning how to combine three-dimensional structural information with primary sequences to give more accurate alignments, when structures are available.

Addresses
The Conway Institute of Biomolecular and Biomedical Research,
University College Dublin, Ireland

Corresponding author: Higgins, Desmond G (des.higgins@ucd.ie)

Current Opinion in Structural Biology 2005, **15**:261–266

This review comes from a themed issue on
Sequences and topology
Edited by Steven E Brenner and Anna Tramontano

Available online 30th April 2005

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2005.04.002

Introduction

Multiple alignments are among the most useful objects in bioinformatics. They are needed whenever sets of homologous sequences are compared and are an essential precursor to numerous further analyses. Up to the mid 1980s, they were routinely generated by hand because generalisations of the traditional dynamic programming algorithm for two sequences [1] proved to be prohibitively slow when used on more than three sequences. The MSA program [2] used a branch and bound technique to make small multiple alignments practical in reasonable time, but this was restricted to small sets of moderately similar sequences. Stochastic optimisation techniques were tried at various stages over the past 20 years (e.g. [3]), but are sometimes either slow or require that the sequences are already very similar. Iteration has been successfully used and now appears to be a key optimisation technique for multiple alignment, either alone [4] or in combination with other methods [5,6].

By far the majority of multiple alignments are generated using a heuristic that Feng and Doolittle [7] referred to as 'progressive alignment'. This was described in various ways by several different groups and resulted in a series of programs in the mid to late 1980s that are still in use today [8–12]. Progressive alignment allows large alignments of distantly related sequences to be constructed quickly and simply. It is based on building the full alignment up progressively, using the branching order of a quick approximate tree (called the guide tree) to guide the alignments. It is implemented in the most widely used programs (ClustalW [13] and ClustalX [14]), but is also used as the optimiser for other programs, such as T-Coffee [15]. The latter is unusual in that it uses the maximum weight trace [16] or 'Coffee' [17] objective function, rather than a more conventional dynamic programming sequence distance score. T-Coffee is of great interest not only because of the way it allows heterogeneous data to be merged in alignments (see 3D-Coffee below) but also as a precursor to the probabilistic-based program PROBCONS [18••], which is the most accurate method available.

In this review, we briefly mention some recently published multiple alignment programs (MAFFT, PSI-PRALINE, PROBCONS and MUSCLE) that we feel represent useful developments. We include a short description of 3D-Coffee, which allows structural information to be mixed with normal sequence data. We also give a very brief mention of some of the methods that are used to measure the success or otherwise of various programs and algorithmic developments (see Table 1 for some useful URLs).

3D-Coffee

3D-Coffee [19•] is designed to create protein sequence alignments that incorporate three-dimensional structural information, when appropriate structures exist. Within a data set, it is common to find Protein Data Bank (PDB) entries for one or more of the input sequences. In principle, utilizing three-dimensional structures facilitates the alignment of distantly related sequences. Structural elements are generally more conserved than primary sequences, retaining their alignability well into the twilight zone (≤25% sequence identity). In Figure 1, we give the example of a set of diverse globins (this example was used by Barton and Sternberg [9] to test a multiple alignment method) whose primary sequences have diverged to less than 10% identity for the most divergent pairs but whose structures can still clearly be superimposed. In practice, using this information can be complicated, although various combinations of threaders and sequence aligners have been described in the past that accomplish this (e.g. [20]).

Table 1

Table of useful URLs.

3D-Coffee	http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi
MUSCLE	http://www.drive5.com/muscle/
PROBCONS	http://probcons.stanford.edu/
MAFFT	http://timpani.genome.ad.jp/~mafft/server/
PRALINE	http://ibivu.cs.vu.nl/programs/pralinewww/
VAST, Cn3D	http://www.ncbi.nlm.nih.gov/Structure
ClustalW	http://www.ebi.ac.uk/clustalw/

3D-Coffee is a fast, simple and accurate method that exploits T-Coffee's ability to incorporate heterogeneous information to incorporate structural data into an alignment and to improve its accuracy, even when only one or two structures are available. The T-Coffee [15] program generates a pairwise alignment library of weighted pairs of residues from the input sequences and finds a multiple alignment that is most compatible with these. When given just one structure, 3D-Coffee matches each sequence to the structure using an external threading algorithm and converts the output to a two-sequence alignment. These alignments contain information about how each sequence aligns to the structure and so, indirectly, how the sequences align to each other.

3D-Coffee incorporates a link to the FUGUE [21] threading package, which carries out sequence alignment using local structural information. If two or more structures are available, these can also be matched to each other using a full structure superposition package such as SAP [22], the output of which is then passed to T-Coffee. The more structures included, the more accurate will be the alignments (a reported increase of approximately 10% for every 20% increase in the structure:sequence ratio in the most difficult cases). Furthermore, you are not restricted to using just SAP and FUGUE, although the 3D-Coffee web server uses these by default; you can use your choice of software to convert a pair of structures or a sequence and a structure into an alignment or a set of alignments.

Measuring multiple alignment accuracy

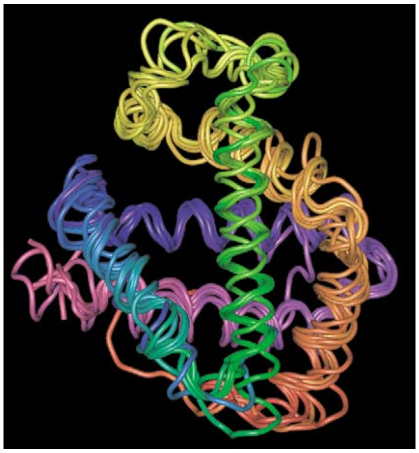
Over the past 15 years, dozens of multiple alignment programs and algorithms have been proposed. How can these different packages be compared and how can we tell which work best in which situations? Several benchmarks exist for this purpose, taking the form of databases of precompiled alignments to which the alignments generated by test algorithms are compared. These comparisons are usually done by calculating the fraction of columns of aligned residues that are identical in both test and reference alignments (the so-called column score [23]), although several alternative measures exist. The use of sets of test cases is attractive due to the simplicity of the tests and evaluations. Care must be taken, however, to avoid over fitting or over training. It is possible to

Figure 1

```
hba_horse   ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-
hba_human   ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
hbb_horse   --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
hbb_human   --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
glb5_petma  PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
myg_phyca   ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
lgb2_luplu  --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
                      *:  :   :   *  . .         :  .:   * :   *  :   .


hba_horse   ----HGSAQVKAHGKKVGDALTLAVGHLDD-----LPGALSNLSDLHAHKLRVDPVNFKL
hba_human   ----HGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHKLRVDPVNFKL
hbb_horse   PGAVMGNPKVKAHGKKVLHSFGEGVHHLDN-----LKGTFAALSELHCDKLHVDPENFRL
hbb_human   PDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDKLHVDPENFRL
glb5_petma  ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
myg_phyca   EAEMKASEDLKKHGVTVLTALGAILKKKGH------HEAELKPAQSHATKHKIPIKYLEF
lgb2_luplu  VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
                  . .:: *.  :   .                      .  *  .  :    : .


hba_horse   LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
hba_human   LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
hbb_horse   LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
hbb_human   LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
glb5_petma  LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
myg_phyca   ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
lgb2_luplu  VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
            :   :  .:      .     ...       .   :     . . ..
```



Current Opinion in Structural Biology

ClustalW sequence alignment (left) and VAST structural alignment (right) of seven globins, visualized using Cn3D. The coloured regions within the sequence alignment correspond to the coloured regions within the superposition and represent the SCOP-defined structural core of six α helices. Both VAST and Cn3D are available on the NCBI web site (http://www.ncbi.nlm.nih.gov/Structure).

optimize any particular alignment program to a test set by searching for the optimal set of alignment parameters. This will artificially inflate the performance on the test sets while doing nothing to help or possibly even degrading the performance in general.

Small sets of test cases have been used to demonstrate the power of particular programs since the first multiple alignment papers [2,7,9,13]. A small set of test cases was used by McClure *et al.* [24] to systematically compare a range of the then available packages. These test cases were all either small, thus preventing statistical analysis, or lacked coverage and were not representative. BAliBASE [25] was the first large-scale purpose-built benchmark and it continues to be used regularly today as a proving ground for new methods. The data set is subdivided into several manually refined reference sets, each dealing with a different type of alignment problem (global/local alignments of different lengths and sequence identity, long internal gaps, etc.). These alignments make use of structural information when it is available, but otherwise the alignments are manual. This allows varied coverage of alignment situations, but comes at the cost of some subjectivity.

HOMSTRAD [26] comprises over 1000 alignments, each based on a particular protein family. HOMSTRAD is exclusively based on sequences with known three-dimensional structures and PDB files. Within each entry, a structural alignment of the proteins is automatically generated. These alignments can then be used as alignment benchmarks. This has the advantage of greater automation and objectivity compared to BAliBASE, but comes at the cost of less coverage.

PREFAB [27••] is also automatically generated. Two proteins with known structures are structurally aligned and their sequences used to query a database, from which high-scoring hits are collected. The queries and their hits are combined, and then aligned using the software that is being tested. Accuracy is assessed on the original pair alone, by comparison with their structural alignment/superposition.

The SABmark [28] database contains two large subdatasets of alignments of up to 25 input sequences each. *Superfamily* comprises 425 groups, each representing a protein family or superfamily that possesses 25–50% sequence identity, whereas *Twilight_Zone* contains 209 groups, each representing a particular SCOP-defined fold comprising input sequences that share less than 25% identity.

Rather than using real-life biological sequence data, IRMbase [29] entries are created by the implantation of ROSE-generated [30] motifs into otherwise random sequences. This data set provides several reference groups of alignments of varying lengths and differing numbers of implants.

Unlike all of the above test methods, APDB [31] does not rely on comparison with a pre-existing reference alignment. Rather, alignment quality is measured based on the superposition that is implied by a test multiple alignment of known PDB structures of the input sequences. Any pair of sequences with known structures will have a superposition that is implied from the alignment. Alignment quality is then a matter of assessing the quality of that superposition by measuring the agreement of the two structures. Given an alignment containing at least two sequences with PDB structures/entries, APDB returns an assessment for each pair of structures and a global measure from comparison with all the available structures that is scaled like a percentage match or column score. Independence from a reference alignment is an advantage, as it eliminates any bias that may exist within the reference and the software or expert used to generate it. In any case, there is no guarantee that a suitable reference even exists. Clearly, APDB still suffers from the constraint that there must be PDB entries for at least two of the sequences. This constraint is difficult to overcome without manual alignments and is even more severe in the case of DNA or RNA alignments.

New alignment packages

MAFFT

MAFFT [32••] uses a fast Fourier transform to quickly generate a guide tree for progressive alignment. A fast tree-based iteration strategy is then used to refine the alignment by optimising the weighted sum of pairs (WSP) objective function. This protocol resulted in very accurate and very fast alignments, as benchmarked by BAliBASE [33]. The latest version of MAFFT (version 5.3) incorporates pairwise alignment information into the objective function. Three different algorithms are described. The G-INS-i method incorporates global pairwise alignment information, the H-INS-i algorithm incorporates local pairwise information from the fasta34 program in FASTA [34] and, finally, the F-INS-i option uses the fasta34 program without the Smith–Waterman [35] optimisation.

These algorithms were tested using three benchmarks: HOMSTRAD, SABmark and PREFAB. The new algorithms outperformed the original MAFFT algorithm (FFT-NS-i) by up to 6% on all benchmarks. However, they are also more computationally demanding. The original strategy scales as O(N), but all of the new methods scale as O(N^2), where N is the number of sequences to be aligned.

PSI-PRALINE

Katoh *et al.* [32••] also extended the HOMSTRAD and SABmark alignments, in a similar manner to PREFAB, by incorporating a large number of close homologues, as found by a BLAST search. It had been shown that including more sequences in an alignment tended to increase the accuracy of the alignment [25]. Sadreyev and Grishin [36] demonstrated that including confident homologues increased the accuracy of profile alignments using the COMPASS program. Katoh *et al.* showed that, by including up to 100 close homologues in the alignment, the accuracy of most methods increased noticeably. The new algorithms saw the most dramatic improvement; the improvement was almost as good as including structural information. An alternative multiple alignment program, PSI-PRALINE, includes homologues, this time found by PSI-BLAST, to improve the performance of PRALINE [37•].

PROBCONS

PROBCONS [18••] is currently the most accurate multiple alignment method, as benchmarked using BAliBASE. It performs best on all five of the reference sets that make up BAliBASE. It also finds the unique best alignment in 46.1% of the BAliBASE cases, as well as the joint best alignment in 66.7% of cases. Broadly, PROBCONS is like T-Coffee, but it uses probabilities instead of the heuristic residue pair weights of the latter program.

Initially, all of the sequences are aligned with each other using a pair-HMM (hidden Markov model) generated with the maximum expected accuracy objective function [38]. This allows the calculation of the posterior probability, $P(x_i \sim y_j \mid x,y)$, for each pair of amino acids. This is the probability that residue i in sequence x is matched with residue j in sequence y in the final alignment.

Next, a consistency transformation is applied. If a third homologous sequence, z, is available, a better estimate of the posterior probability $P(x_i \sim y_j)$ can be obtained using information about how x and y align with z. This is defined as $P(x_i \sim y_j \mid x,y,z)$ and is calculated by the following heuristic, which can be solved in approximately constant time:

$$P(x_i \sim y_j \mid x, y, z) \approx \sum_k P(x_i \sim z_k \mid x, z) P(y_j \sim z_k \mid y, z)$$

The multiple alignment is then generated by using a progressive alignment scheme. The guide tree is calculated by clustering the sequences based on their expected accuracy (the sum of posterior probabilities). Subalignments are combined using a sum of pairs scheme, in which the score of the multiple alignment is calculated by summing all posterior probabilities for all pairs of sequences present. The final alignment is then subjected to an iterative refinement protocol.

MUSCLE

MUSCLE (multiple sequence comparison by log expectation) [27••,39•] is a new progressive alignment package that is extremely fast and accurate. The first step in MUSCLE is to rapidly generate a rough draught of the alignment using a very crude guide tree. Distances between pairs of input sequences are estimated by k-mer (short exact matches of fixed length) counting using a compressed alphabet [40]. These distances are clustered to give an initial tree, which is then used to construct a progressive alignment of the sequences. MUSCLE implements the log expectation (LE) score to align profiles during the progressive alignment; this has been shown to outperform other scoring functions in homology searches [41].

The next stage in the process is to refine the rough draught by generating a more accurate guide tree, which is based on the initial alignment. A second progressive alignment is generated using this improved tree. For increased speed, new pairwise profile alignments are calculated only for those subtrees that changed relative to the initial tree. An optional tree-based iteration step [6] is included to further improve the alignment quality.

During development, MUSCLE was assessed using several alignment databases, including the BAliBASE benchmark, on which it achieved the highest ranking of any method at the time of publication. The speed of MUSCLE was also demonstrated, by aligning 5000 sequences on a PC in 7 min. The latest version of MUSCLE, version 6, is a collaboration between the developers of MUSCLE and PROBCONS, and uses a new refinement strategy based on the PROBCONS algorithm. It gives a significant increase in accuracy at a modest computational cost.

Conclusions

Multiple alignments are so widely used that any improvements to software or algorithms can have a significant impact on the scientific community. Nonetheless, there must be a limit to the accuracy that straightforward sequence-based algorithms can reach. With protein alignments, the most useful methods in the long term will be those that sensibly and flexibly incorporate information from mixtures of sources, such as existing multiple alignment collections or protein structures. The alignment of RNA and/or DNA is a different matter. In this case, there are no obvious general benchmarks that are available to compare methods. Furthermore, if the DNA sequences are long enough, the problem changes to one of genome alignment [42–44], motif finding [45–47] or best local alignments [48], all of which are very active areas of research.

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. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J Mol Biol* 1970, **48**:443-453.

2. Lipman DJ, Altschul SF, Kececioglu JD: **A tool for multiple sequence alignment**. *Proc Natl Acad Sci USA* 1989, **86**:4412-4415.

3. Notredame C, Higgins DG: **SAGA: sequence alignment by genetic algorithm**. *Nucleic Acids Res* 1996, **24**:1515-1524.

4. 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.

5. Wallace IM, O'Sullivan O, Higgins DG: **Evaluation of iterative alignment algorithms for multiple alignment**. *Bioinformatics* 2004, doi:10.1093/bioinformatics/bti159.

6. Hirosawa M, Totoki Y, Hoshida M, Ishikawa M: **Comprehensive study on iterative algorithms of multiple sequence alignment**. *Comput Appl Biosci* 1995, **11**:13-18.

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

8. Hogeweg P, Hesper B: **The alignment of sets of sequences and the construction of phyletic trees: an integrated method**. *J Mol Evol* 1984, **20**:175-186.

9. Barton GJ, Sternberg MJ: **A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons**. *J Mol Biol* 1987, **198**:327-337.

10. Higgins DG, Sharp PM: **CLUSTAL: a package for performing multiple sequence alignment on a microcomputer**. *Gene* 1988, **73**:237-244.

11. Corpet F: **Multiple sequence alignment with hierarchical clustering**. *Nucleic Acids Res* 1988, **16**:10881-10890.

12. Taylor WR: **A flexible method to align large numbers of biological sequences**. *J Mol Evol* 1988, **28**:161-169.

13. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: 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.

14. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools**. *Nucleic Acids Res* 1997, **25**:4876-4882.

15. 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.

16. Kececioglu JD: **The maximum weight trace problem in multiple sequence alignment**. In *Lecture Notes In Computer Science*, vol 684. Edited by Apostolico A, Crochemore M, Galil Z, Manber U. Springer-Verlag; 1993:106-119.

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

18. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S:
•• **ProbCons: probabilistic consistency-based multiple alignment of amino acid sequences**. *Genome Res* 2005, **15**:330-340.

PROBCONS is like T-Coffee but uses probabilities for residue scores. T-Coffee is based on finding a multiple alignment that best matches a library of input pieces of information. These pieces are usually pairs of amino acids, each pair with a score that reflects how much we wish to align the pair in the final alignment. In the original T-Coffee, these scores were simply the percent identity of the parent pairwise alignment. In PROBCONS, these are posterior probabilities for pair-HMMs. PROBCONS is a very accurate alignment program.

19. 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.

3D-Coffee is a practical method for mixing sequences and structures.

20. Al-Lazikani B, Sheinerman FB, Honig B: **Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases**. *Proc Natl Acad Sci USA* 2001, **98**:14796-14801.

21. Shi J, Blundell TL, Mizuguchi K: **FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties**. *J Mol Biol* 2001, **310**:243-257.

22. Taylor WR, Orengo CA: **Protein structure alignment**. *J Mol Biol* 1989, **208**:1-22.

23. Karplus K, Hu B: **Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set**. *Bioinformatics* 2001, **17**:713-720.

24. McClure MA, Vasi TK, Fitch WM: **Comparative analysis of multiple protein-sequence alignment methods**. *Mol Biol Evol* 1994, **11**:571-592.

25. Thompson JD, Plewniak F, Poch O: **A comprehensive comparison of multiple sequence alignment programs**. *J Mol Biol* 1999, **27**:2682-2690.

26. Mizuguchi K, Deane CM, Blundell TL, Overington JP: **HOMSTRAD: a database of protein structure alignments for homologous families**. *Protein Sci* 1998, **7**:2469-2471.

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

MUSCLE is faster than Clustal and as accurate as T-Coffee, making it a very slick and useful piece of work.

28. Walle IV, Lasters I, Wyns L: **SABmark - a benchmark for sequence alignment that covers the entire known fold space**. *Bioinformatics* 2004. doi:10.1093/bioinformatics/bth493.

29. 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.

30. Stoye J, Evers D, Meyer F: **Rose: generating sequence families**. *Bioinformatics* 1998, **14**:157-163.

31. O'Sullivan O, Zehnder M, Higgins D, Bucher P, Grosdidier A, Notredame C: **APDB: a novel measure for benchmarking sequence alignment methods without reference alignments**. *Bioinformatics* 2003, **19(suppl 1)**:i215-i221.

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

This paper describes the latest improvements to the very fast and accurate MAFFT programs.

33. 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.

34. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison**. *Proc Natl Acad Sci USA* 1988, **85**:2444-2448.

35. Smith TF, Waterman MS: **Identification of common molecular subsequences**. *J Mol Biol* 1981, **147**:195-197.

36. Sadreyev RI, Grishin NV: **Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs**. *Bioinformatics* 2004, **20**:818-828.

37. Simossis VA, Kleinjung J, Heringa J: **Homology-extended**
• **sequence alignment**. *Nucleic Acids Res* 2005, **33**:816-824.

The authors show how the multiple sequence alignment program PRALINE can now include information obtained from database searches to dramatically improve the quality of an alignment.

38. Holmes I, Durbin R: **Dynamic programming alignment accuracy**. *J Comput Biol* 1998, **5**:493-504.

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

The algorithms used in the new MUSCLE program are described in detail in this paper.

40. Edgar RC: **Local homology recognition and distance measures in linear time using compressed amino acid alphabets**. *Nucleic Acids Res* 2004, **32**:380-385.

41. van Ohsen N, Zimmer R: **Improving profile-profile alignments via log average scoring**. In *Algorithms in Bioinformatics, First International Workshop, WABI 2001; Aarhus, Denmark, August 28–31, 2001*. *Lecture Notes In Computer Science*, vol 2149. Edited by Gascuel O, Moret BME. Springer; 2001:11–26.

42. Bray N, Pachter L: **MAVID: constrained ancestral alignment of multiple sequences**. *Genome Res* 2004, **14**:693-699.

43. Brudno M, Chapman M, Gottgens B, Batzoglou S, Morgenstern B: **Fast and sensitive multiple alignment of large genomic sequences**. *BMC Bioinformatics* 2003, **4**:66.

44. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: **LAGAN and multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA**. *Genome Res* 2003, **13**:721-731.

45. Sinha S, Blanchette M, Tompa M: **PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences**. *BMC Bioinformatics* 2004, **5**:170.

46. Frith MC, Hansen U, Spouge JL, Weng Z: **Finding functional sequence elements by multiple local alignment**. *Nucleic Acids Res* 2004, **32**:189-200.

47. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.

48. Zhang Y, Waterman MS: **An Eulerian path approach to local multiple alignment for DNA sequences**. *Proc Natl Acad Sci USA* 2005, **102**:1285-1290.