Briefings in Bioinformatics Advance Access published March 27, 2008

doi:10.1093/bib/bbn013

Recent developments in the MAFFT multiple sequence alignment program

Kazutaka Katoh and Hiroyuki Toh

Submitted: 14th January 2008; Received (in revised form): 26th February 2008

Abstract

The accuracy and scalability of multiple sequence alignment (MSA) of DNAs and proteins have long been and are still important issues in bioinformatics. To rapidly construct a reasonable MSA, we developed the initial version of the MAFFT program in 2002. MSA software is now facing greater challenges in both scalability and accuracy than those of 5 years ago. As increasing amounts of sequence data are being generated by large-scale sequencing projects, scalability is now critical in many situations. The requirement of accuracy has also entered a new stage since the discovery of functional noncoding RNAs (ncRNAs); the secondary structure should be considered for constructing a high-quality alignment of distantly related ncRNAs. To deal with these problems, in 2007, we updated MAFFT to Version 6 with two new techniques: the PartTree algorithm and the Four-way consistency objective function. The former improved the scalability of progressive alignment and the latter improved the accuracy of ncRNA alignment. We review these and other techniques that MAFFT uses and suggest possible future directions of MSA software as a basis of comparative analyses. MAFFT is available at http://align.bmr.kyushu-u.ac.jp/mafft/software/.

Keywords: *large-scale sequence alignment; fast tree building; ncRNA structural alignment; amino acid sequence alignment*

INTRODUCTION

Multiple sequence alignment (MSA) is an important step in various types of comparative studies of biological sequences. MSA is used in phylogenetic inference, conserved region detection, structure prediction of noncoding RNAs (ncRNAs) and proteins and many other situations. For an easy MSA problem, such as an alignment consisting of a small number ($<\sim$100) of short ($<\sim$5000) sequences with global and high similarity (percent identity of $>\sim$40% for protein cases and $>\sim$70% for nucleotide cases), most of the current programs return a correct MSA, and no special consideration is needed. However, if all three of these conditions are not met, then the construction of an MSA can be a difficult task from both computational and biological viewpoints.

There is an established method based on the Dynamic Programming (DP) algorithm for calculating a pairwise alignment (an alignment between two sequences) [1–3] with a time complexity of $O(L^2)$, where L is the sequence length. However, when more than two sequences must be aligned, the situation is somewhat complicated. Theoretically, the DP algorithm can be extended for cases of more than two sequences, but the time and space complexities of the naively extended algorithm, $O(L^N)$, are impossibly large, where N is the number of sequences. Finding the exactly optimum MSA quickly becomes computationally intractable when the number of sequences increases [4]. Considerable efforts have been made to obtain the optimum MSA of $\sim$10 sequences [5–10], which is still substantially smaller than the alignment size biologists now need. Therefore, some sort of heuristics are inevitable.

Even if the optimal MSA is successfully obtained, it is not always the correct solution from a biological viewpoint [11, 12]. This suggests that we should pay attention to a biologically relevant objective function, as well as to algorithmic techniques for obtaining the optimum solution. This is one of the reasons why various multiple sequence alignment schemes have

Corresponding author. Kazutaka Katoh, Digital Medicine Initiative, Kyushu University, Fukuoka 812-8582, Japan. Tel: 81-92-642-6967; E-mail: katoh@bioreg.kyushu-u.ac.jp

Kazutaka Katoh received his PhD from Kyoto University in 2001. He is an Associate Professor of Digital Medicine Initiative at Kyushu University. His research interests are in bioinformatics and molecular evolution.

Hiroyuki Toh received his PhD from Kyushu University in 1989. He is a Professor of Medical Institute of Bioregulation, Kyushu University, and his research focus is bioinformatics.

been extensively studied to date, but there is no definitive one. Moreover, the accuracy of multiple alignment is improved by adding homologs or profiles [13–15]. This is because homologs make family-specific information available and enrich the profiles used in the multiple alignment processes [16]. Recent protein MSA studies indeed tended to use external sequence information [17–19]. Therefore, for an alignment program, the ability to handle many sequences is an important factor for yielding accurate results, as well as for large-scale analyses.

The MAFFT sequence aligner [20] was originally developed to perform the rapid calculation of an MSA consisting of a large number of sequences. A fast group-to-group alignment algorithm based on fast Fourier Transform (FFT) [20] and an approximate distance calculation method (the 6mer method) [20–23] facilitate the rapid calculation. Due to the increasing necessity for MSA of distant homologs, in 2005, we sought to improve the accuracy of MAFFT, and released Version 5 [14], which adopted a new objective function, the summation of a traditional weighted sum-of-pairs (WSP) score [24], and a consistency score similar to COFFEE [25] calculated from all-to-all pairwise alignments before constructing an MSA.

As a result, the current version of MAFFT has several options, as listed in Table 1, and covers various types of MSA problems, ranging from a small alignment consisting of distantly related sequences to a large-scale alignment. Recent benchmark studies under various conditions [26–30] consistently concluded that MAFFT is one of the best choices.

However, MAFFT does not completely cover all of the situations that biologists encounter. Especially for distantly related sequences, the use of multiple independent methods is important. The different MSAs computed by independent methods can be subjected to meta-aligners such as M-Coffee [31], to generate a more accurate MSA than those yielded by individual tools. The consensus among the different MSAs also provides information about which sites were reliably aligned [32, 33]. It should also be noted that different alignments sometimes result in quite different trees in phylogenetic analyses [34–36]. Such contradiction is partially (but not completely) avoided, by subjecting only reliably aligned sites to a phylogenetic inference.

In this article, after reviewing the general multiple alignment algorithms implemented in the MAFFT sequence aligner, we describe two new techniques introduced in Version 6: (i) a new tree-building algorithm, PartTree, for handling even larger numbers of sequences and (ii) a multiple ncRNA alignment framework incorporating structural information. We also describe some utility options that were added in Version 6 and provide tips to produce a reasonable alignment efficiently. For situations outside the scope of MAFFT, we introduce alternative tools developed by other groups.

GENERAL ALGORITHMS

Terms and basic concepts

Sequence, alignment, homology and gap

A sequence alignment is a set of corresponding residues among a collection of nucleotide or amino acid sequences. The sequences can be protein- or RNA-coding sequences or noncoding nucleotide sequences, such as introns or spacers. The sequences involved in an alignment are assumed to be homologous; that is, derived from a single common ancestral sequence. Aligned residues are usually interpreted as sharing their evolutionary origin. When a sequence has no corresponding residue because of an insertion or deletion event, the position is displayed as '−' or another symbol and is called a 'gap'. Most alignment programs do not attempt to filter out nonhomologous sequences, leaving the decision of what sequences to include in the MSA as an external decision for the user. However, this problem is sometimes important in actual analyses.

Global homology and local homology

Some MSA methods assume that all of the input sequences are globally alignable; that is, the entire regions of the sequences are assumed to be homologous, but this assumption does not necessarily agree with real analyses. Local alignment methods avoid the assumption of global homology. Some MSA methods, such as DIALIGN [37–39] and T-Coffee, have a facility to incorporate a local alignment algorithm to detect short patches of strong sequence similarity.

Alignment of genomic sequences

Unlike database-search programs [40, 41], most MSA programs try to include all of the residues in the input sequences, even when a local alignment algorithm is employed in a part of the calculation process. This policy makes the programs impractical when there are large nonhomologous regions within the sequences. We sometimes encounter such

Table 1: Options of MAFFT Version 6.5

Option name	Command	
For a large-scale alignment ($N > \sim$10 000). Progressive methods with the PartTree algorithm		
NW-NS-PartTree1	`mafft --parttree --retree 1`	Distance is by the 6mer method
NW-NS-PartTree2	`mafft --parttree --retree 2`	Distance is by the 6mer method. Guide tree is rebuilt
NW-NS-DPPartTree1	`mafft --dpparttree --retree 1`	Distance is estimated based on DP
NW-NS-DPPartTree2	`mafft --dpparttree --retree 2`	Distance is estimated based on DP. Guide tree is re-built
NW-NS-FastaPartTree1	`mafft --fastaparttree --retree 1`	Requires FASTA [40] to estimate distances
NW-NS-FastaPartTree2	`mafft --fastaparttree --retree 2`	Requires FASTA [40]. Guide tree is rebuilt
For a medium-scale alignment ($\sim$10 000 $> N > \sim$200). Progressive methods		
FFT-NS-1	`mafft --retree 1`	Approximately two times faster than the default
FFT-NS-2	`mafft`	Default
For a small-scale alignment ($N < \sim$200, $L < \sim$10 000). Iterative refinement methods		
FFT-NS-i	`mafft-fftnsi`	Fastest of the four in this category. Uses WSP score only
G-INS-i	`mafft-ginsi`	Uses WSP score and consistency score from global alignments
L-INS-i	`mafft-linsi`	Uses WSP score and consistency score from local alignments
E-INS-i	`mafft-einsi`	Uses WSP score and consistency score from local alignments with a generalized affine gap cost
For a small-scale RNA alignment ($N < \sim$50, $L < \sim$1 000). Structural alignment methods		
Q-INS-i	`mafft-qinsi`	Requires no external structural alignment programs
X-INS-i-scarnapair	`mafft-xinsi --scarnapair`	Requires MXSCARNA (Tabei *et al.*, submitted for publication)
X-INS-i-larapair	`mafft-xinsi --larapair`	Requires LaRA [78]
X-INS-i-foldalignlocalpair	`mafft-xinsi --foldalignlocalpair`	Requires FOLDALIGN [79]. Uses the local alignment option
X-INS-i-foldalignglobalpair	`mafft-xinsi --foldalignglobalpair`	Requires FOLDALIGN [97]. Uses the global alignment option
If not sure which option to use		
Automatic	`mafft --auto`	Selects an appropriate option from FFT-NS-2, FFT-NS-i and L-INS-i, according to the size of input data

N is the number of sequences and L is the sequence length.

situations when aligning genomic sequences. In such cases, MAFFT consumes a large amount of time. Instead, MLAGAN [42] and MAVID [43] are useful.

Order of residues to be aligned

When we are handling rearranged genomic sequences or mosaic proteins with rearranged multiple domains, the order of alignable residues can differ among the sequences. In such a case, tools without the assumption of the conservation of the order of aligned residues should be used. ABA [44] (for protein/DNA), ProDA [45] (for protein), TBA [46] (for genomic DNA) and MAUVE [47] (for genomic DNA) are available.

Progressive method—FFT-NS-2

The progressive method [48, 49] is the most commonly used multiple alignment algorithm. Clustal W [50, 51], MAFFT, POA [52], Kalign [53] and many other MSA packages use this method with various modifications. The procedure of the progressive method implemented in MAFFT is schematically illustrated in Figure 1A. A guide tree, a tentative tree only used for constructing an alignment, is created based on all-to-all pairwise comparisons, and an MSA is constructed using a group-to-group alignment algorithm at each node of the guide tree.

To achieve a reasonable balance between speed and accuracy, MAFFT [20] adopts, by default, a two-cycle progressive method, called FFT-NS-2, in which low-quality all pairwise distances are rapidly calculated, a tentative MSA is constructed, refined distances are calculated from the MSA, and then the second progressive alignment is performed, as shown in Figure 1A. In addition, MAFFT uses two key techniques, an FFT-based group-to-group alignment algorithm [20] (Figure 2) and the 6mer method [20–23] for all pairwise comparisons, to reduce the CPU time of progressive methods. The time complexity of the progressive method implemented in MAFFT is basically $O(N^2L) + O(NL^2)$, where L is the sequence length and N is the number of sequences. The first term corresponds to the guide tree calculation and the second term corresponds to the group-to-group alignment stage. When the input sequences are highly similar to each other, it is reduced to $O(N^2L) + O(NL) = O(N^2L)$, because of the FFT-based alignment method (See [20] for details).

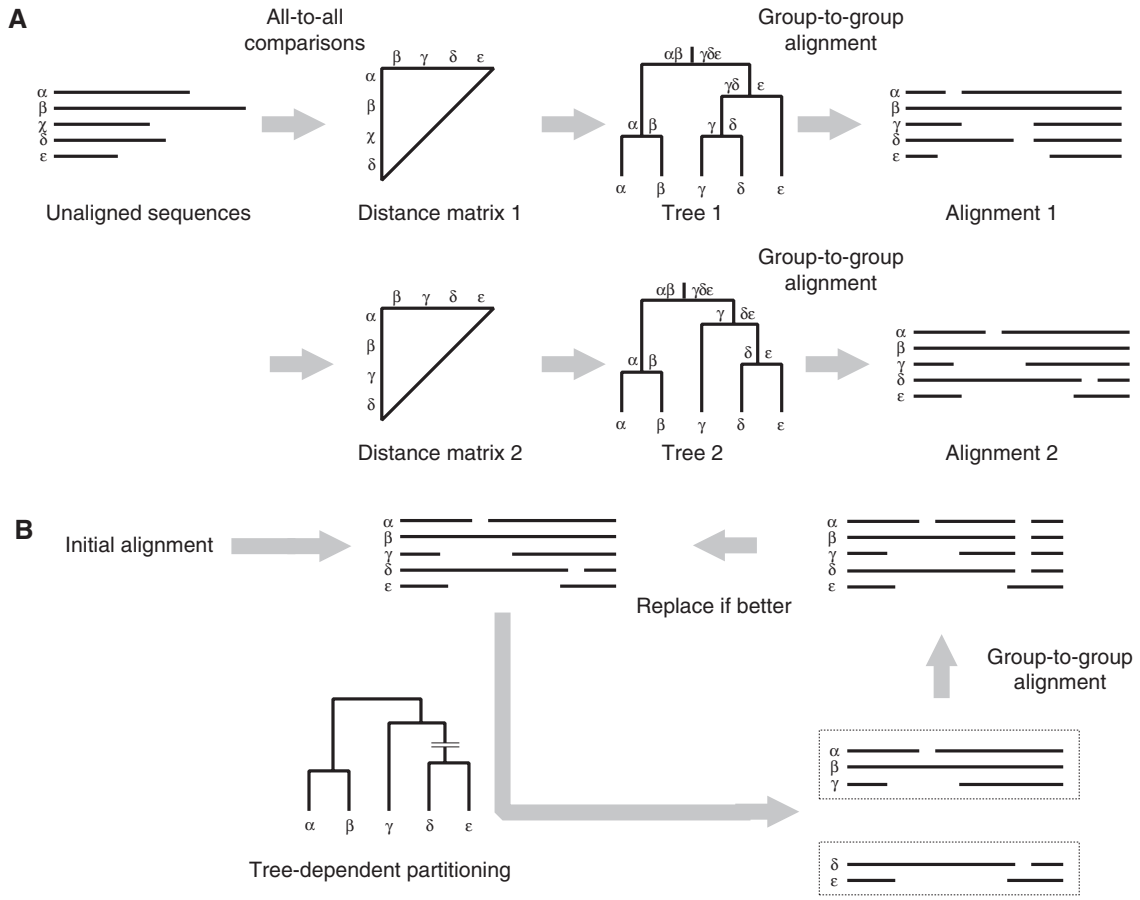

Figure 1: Calculation procedures of the progressive method (**A**) and the iterative refinement method (**B**).

Its space complexity is basically $O(N^2) + O(L^2) + O(NL)$. When the sequence length exceeds the threshold (set as 10 000 residues at present), FFT-NS-2 automatically switches the DP algorithm to a memory saving one [54] and the space complexity becomes $O(N^2) + O(NL)$. On a current desktop computer, this method can be applied to an MSA consisting of up to ~10 000 sequences. The maximum length depends on the similarity level: ~10 000 residues for distantly related sequences or ~500 000 residues for closely related sequences with global homology.

The progressive method has a drawback in that once a gap is incorrectly introduced at a step, the gap is never removed in later steps. To overcome this drawback, there are two types of solutions, the iterative refinement method [55–61] and the consistency-based method [25, 62–64]. These two procedures are quite different: the former tries to correct mistakes in the initial alignment, whereas the latter tries to avoid mistakes in advance, but both work well to improve the alignment accuracy.

Iterative refinement method with the WSP score—FFT-NS-i

In the iterative refinement method, an objective function that represents the 'goodness' of the MSA is explicitly defined. An initial MSA, calculated by the progressive or another method, is subjected to an iterative process and is gradually modified so that the objective function is maximized, as shown in Figure 1B. Various combinations of objective functions and optimization strategies have been proposed to date [55–61]. Among them, Gotoh's iterative refinement method, PRRN [16], is the most successful one, and it forms the basis of recent methods, including MAFFT, MUSCLE [23, 65] and PRIME [66]. The iterative alignment option of MAFFT, called FFT–NS–i, uses the weighted sum–of–pairs (WSP) objective function [24]. As shown in Figure 1B, an MSA is partitioned into two groups, which are then realigned using an approximate group-to-group alignment algorithm [20]. The new MSA replaces the old one if it has a higher score. This process is repeated until no more improvements

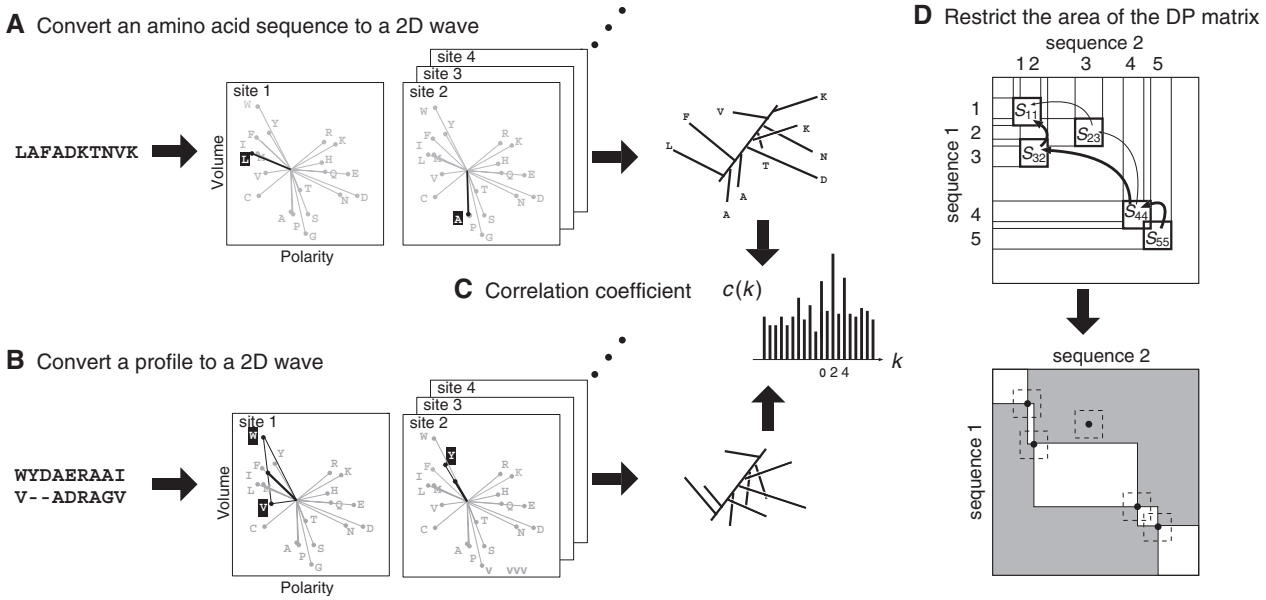

Figure 2: Outline of a fast group-to-group alignment algorithm based on FFT (reprinted from [113]). (**A**) A sequence is converted to a two-dimensional (2D) wave, arrangement of vectors. (**B**) A set of aligned sequences is also converted to a 2D wave. (**C**) The correlation between the two waves can be rapidly computed with FFT. (**D**) The highly conserved regions detected by FFT are used as anchors, and the area of the DP matrix can be restricted.

are made. To save computation time, the partitions of the MSA are restricted to those corresponding to the branches of a tree among the sequences [67]. The time complexity of this method is $O(N^2L) + O(NL^2)$ and the space complexity is $O(N^2) + O(L^2) + O(NL)$ or $O(N^2) + O(NL)$, depending on the sequence length, as in the case of the progressive option. On a current desktop computer, this method can be applied to an MSA consisting of up to ~500 sequences. The maximum length depends on the similarity level: ~10 000 residues for distantly related sequences or ~500 000 residues for closely related sequences with global homology.

Iterative refinement method with consistency and WSP scores—G-INS-i

T-Coffee [64], ProbCons [68] and other methods [69, 70] take an entirely different approach to overcome the drawback of the progressive method by using a consistency criterion, in which an MSA consistent with pairwise alignments is judged to be relevant. Several types of consistency criteria were described previously [25, 62, 63], but T-Coffee achieved a great improvement in accuracy by using the COFFEE criterion [25] together with the library extension technique [64] in the progressive method. In 2005, MAFFT adopted a consistency criterion into the iterative refinement method [14]. Instead of the library extension procedure, which plays an important role in T-Coffee but requires considerable computing power, we took a iterative strategy with an objective function of the summation of the WSP score [24] and a consistency-based score like COFFEE [25]. This method was implemented as the G-INS-i option. Its time complexity is $O(N^2L^2)$. Its space complexity is at least $O(N^2) + O(L^2) + O(NL)$ but greatly depends on the similarity level. The G-INS-i option assumes that the input sequences are globally alignable. This option is suitable when the lengths of the input sequences are similar to each other, as in Figure 3A. On a current desktop computer, this method can be applied to an MSA consisting of up to ~200 sequences. The maximum length is ~5000 residues.

NEW FEATURES IN VERSION 6

Variants of G-INS-i—L-INS-i and E-INS-i for large gaps

After the publication of MAFFT Version 5 [20], we added the L-INS-i and E-INS-i options, which are variants of G-INS-i, to the MAFFT package. Their basic procedures are the same as that of G-INS-i, but different algorithms are used in the pairwise alignment stage. L-INS-i uses a local pairwise alignment [3] with the affine gap cost [2], while

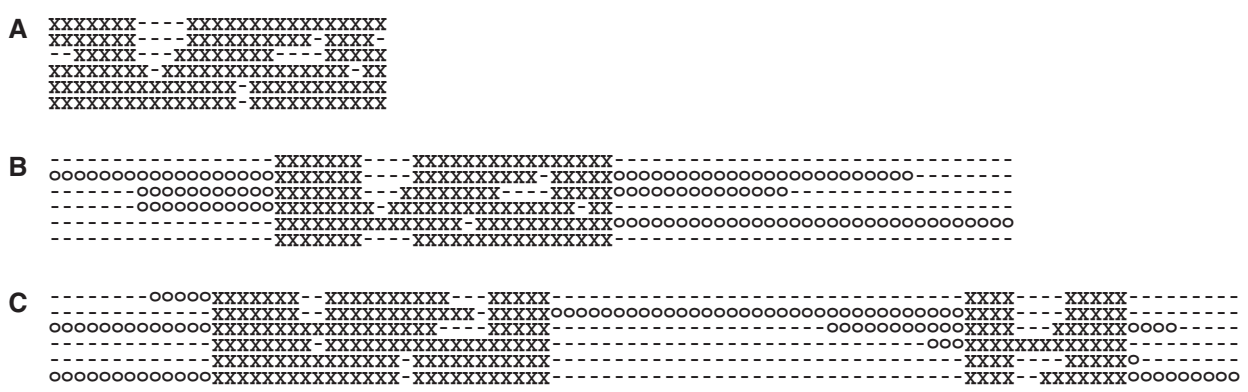

Figure 3: Globally alignable (**A**), locally alignable (**B**), and long internal gaps (**C**). '-' represents a gap, 'X' represents an aligned residue and 'o' is an unalignable residue.

E-INS-i uses a local pairwise alignment with the generalized affine gap cost [71], in which the unalignable region is left unaligned at the pairwise alignment stage. L-INS-i and E-INS-i can be applied to the cases where alignments those in Figure 3B–C are expected, respectively. Note that E-INS-i also includes all of the residues in the alignment during the rest of its procedure, and the resulting alignment is always a full-length alignment. Moreover, it is better to see how MSA programs work and to try some independent methods with various parameters, particularly when a set of distantly related sequences is aligned [33, 72]. T-Coffee, ProbAlign and DIALIGN can be alternative choices for such situations requiring large gaps.

PartTree

Since increasing amounts of sequence data are being generated from large-scale sequencing projects, scalability is now critical in many situations. As noted above, the time complexity of the progressive method is $O(N^2L) + O(NL^2)$. The first term corresponds to all-to-all comparisons of input sequences and guide tree building by the UPGMA method [23, 73]. As this term can be the time-limiting factor when large numbers (10 000 or more) of sequences are aligned, it is desirable to omit the two steps. Without a guide tree, however, the resulting alignment highly depends on the input order, and the quality is not acceptable in most cases.

Hence, we developed a scalable tree-building algorithm, PartTree [74], to generate a guide tree from a set of unaligned sequences with a time complexity of $O(N \log N)$. PartTree is a divisive clustering algorithm. In summary, n representative sequences are randomly selected from the input sequences and then the other sequences are grouped with the n representatives, according to the similarity. The calculation of similarity is performed only nN times at this time. The UPGMA tree among n representatives is calculated. This step is recursively repeated for each of the n groups, unless the group has only one sequence. The n UPGMA trees returned by child processes are combined into a single tree. In total, the similarity calculation is performed $N \log N$ times, on average. Thus, this algorithm is faster than the conventional UPGMA algorithm, which requires all pairwise similarity calculations with a time complexity of $O(N^2)$.

The PartTree option implemented in MAFFT Version 6 (Table 1) can successfully align a dataset consisting of a large number (~60 000) of homologous sequences, at the cost of an accuracy loss of ~2%. Various combinations of distance estimation methods (FASTA-based, DP-based or 6mer-based), a parameter n (the number of representatives) and the number of re-estimations of the guide tree (as shown in Figure 1A) can be selected, according to the needs for balance between accuracy and speed. See the original paper [74] for the benchmark results.

RNA alignment

The importance of RNA alignment is increasing, since the discovery of functional ncRNAs. MAFFT Version 6 has two new options, Q-INS-i and X-INS-i, for RNA alignment. Both methods consider secondary structure information, as a form of base-pairing probability, predicted by either the McCaskill algorithm [75] or the CONTRAfold algorithm [76]. In Q-INS-i, the base-pairing probability is incorporated into the resulting alignment with a new objective function, Four-way Consistency (Katoh and Toh, submitted). In X-INS-i, the structural information is also used

Table 2: Comparison of aligners for multiple RNAs using 52 Rfam alignments as references

				Accuracy of predicted structure (MCC)			
Method	**Time (s)**	**SPS**	**SCI**	**Pfold**	**McCaskill-MEA**	**RNAalifold**	**(intrinsic)**
FFT-NS-2	1.2	0.832	0.674	0.678	0.663	0.669	
Clustal W v2 (Default)	2.6	0.795	0.646	0.640	0.641	0.648	
Clustal W v2 (Iteration = tree)	22	0.798	0.641	0.649	0.641	0.652	
G-INS-i	3.5	0.866	0.719	0.710	0.684	0.681	
ProbConsRNA	16	**0.874**	0.721	0.708	0.689	0.684	
StrAl (2006)	18	0.809	**0.699**	0.662	0.662	0.675	
LaRA 1.3l (June 2007)	5200	0.835	**0.741**	0.708	0.687	0.683	
Murlet (November 2006)	4800	**0.875**	**0.737**	**0.732**	0.702	0.705	
MXSCARNA (May 2007)	47	0.856	0.732	**0.731**	**0.708**	0.705	
Q-INS-i (May 2007)	54	**0.877**	**0.741**	0.730	0.701	0.695	
RNA Sampler (May 2007)	6900	0.809	**0.789**	**0.733**	**0.700**	**0.725**	0.699
MASTR (August 2007)	5400	0.824	**0.748**	0.677	0.685	0.692	0.700
X-INS-i-scarnapair (December 2007)	390	**0.880**	**0.769**	**0.736**	**0.708**	**0.731**	

The methods above the dashed line are purely sequence-based alignment methods. RNA structural alignment methods are listed below the dashed line. The names of MAFFT options are shaded. The benchmark dataset in the MASTR paper [87] was used. The alignment accuracies were assessed with two criteria, SPS and SCI [110, 111], using the compalign and scif programs distributed with the BRAliBASE Version 2.1 benchmark dataset [29]. The SPS SCI values were computed for each alignment and then averaged across all the alignments. The alignment by each method was subjected to three external prediction programs, Pfold [112], McCaskill-MEA [90] and RNAalifold [89], and then the differences from the Rfam curated structure were calculated with Matthews correlation coefficient (MCC) criterion. The accuracy values for secondary structure internally predicted by RNA Sampler and MASTR are shown in the (intrinsic) column. The MCC values were computed for each sequence and then averaged across all the sequences. The highest accuracy values are underlined for each column. The accuracy values close to the highest ($P > 0.01$ in the Wilcoxon test) are shown in bold. McCaskill-MEA was run with the default α value of 0.91. RNA Sampler was run with the `-i 15 -S 100` arguments. See Katoh and Toh (submitted for publication) for benchmark results using larger datasets. See the MASTR paper [87] for the results of other methods, including FoldalignM, LocARNA and RNAcast, that are not listed here.

at the pairwise alignment stage, in addition to the Four-way Consistency objective function. At present, three different external structural alignment programs, MXSCRANA [77], LaRA [78] and FOLDALIGN [79], are supported as the source of pairwise structural alignments for X-INS-i. Although MXSCARNA and LaRA are multiple RNA alignment programs themselves, only their pairwise alignment functions are used.

A benchmark result (Table 2) indicates that the performances of structural alignment methods for multiple RNAs have rapidly improved in 2007, as a result of intensive studies by many groups [78, 80–88], and that the combination of X-INS-i and SCARNA (denoted as X-INS-i-scarnapair in Table 2) is the most accurate method, according to most benchmark criteria. Moreover, the calculation time of X-INS-i-scarnapair is shorter than those of other accurate methods, such as RNA Sampler [84], MASTR [87] and Murlet [83]. The difference in accuracy between X-INS-i-scarnapair and MXSCARNA reflects the improvement gained by the X-INS-i framework, because these two methods use the same pairwise structural alignment algorithm, SCARNA [77].

Group-to-group alignment and seed alignment

MAFFT Version 6 has the `mafft-profile` program, which functions like the profile alignment option of Clustal W. When two alignments are given, the `mafft-profile` program converts each alignment into a profile and returns an alignment between the two alignments.

```
% mafft-profile aligned_group1
aligned_group2 > output
```

This is sometimes useful in actual analyses, but needs consideration of the phylogenetic relationship between the two groups. The profile alignment assumes that the two input alignments are phylogenetically separated, as in Figure 4A or B.

When another phylogenetic relationship is expected, as in Figure 4C, the profile alignment could introduce misalignments. In fact, we sometimes encounter such a situation when we want to add

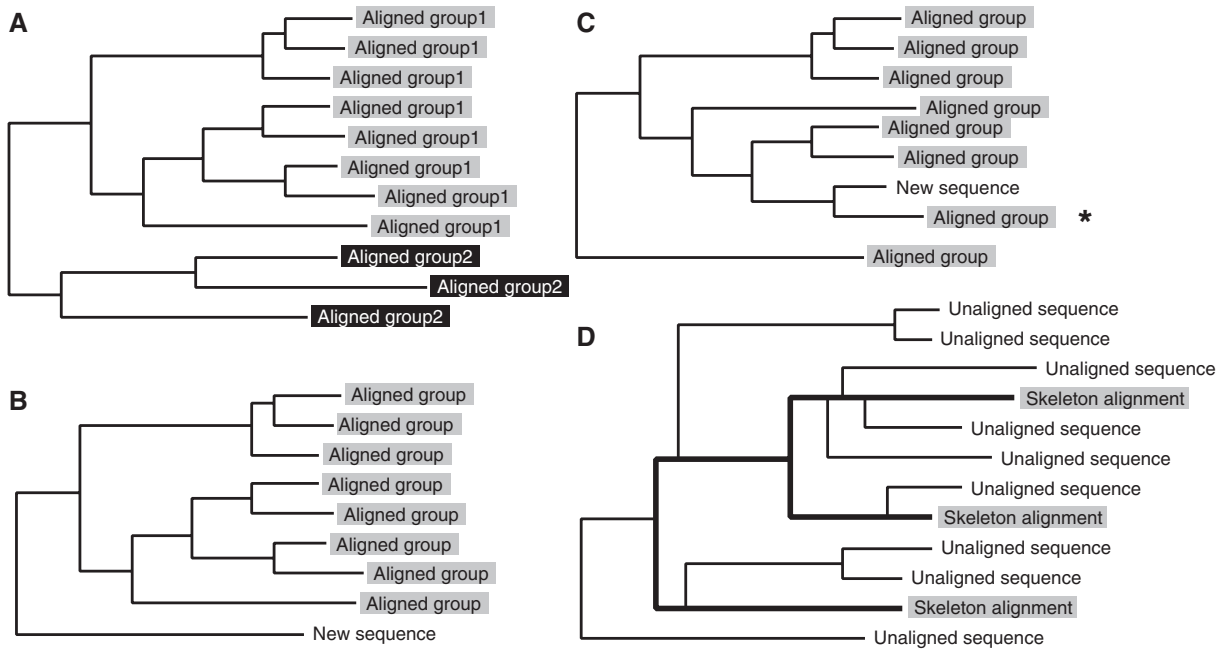

Figure 4: Possible relationships between a group of aligned sequences and new sequence(s). The profile alignment method is applicable to cases **A** and **B**, whereas the application of the method to cases **C** and **D** should be avoided.

a newly determined sequence into an established alignment, which was already adjusted by eye or taken from an annotated database. For such a case, MAFFT Version 6 provides another type of group-to-group alignment option based on the consistency criterion,

```
% mafft-linsi --seed aligned_
sequences new_sequence > output
```

which discards all of the gaps and then makes a new alignment consisting of all of the members of `aligned_sequences` and `new_sequence`, in which the alignment within `aligned_sequences` is exactly reconstructed. Thus, `new_sequence` is first aligned with the nearest sequence (marked with `*`) in `aligned_sequences` and then is aligned with the other members. This can be applied to a situation like that in Figure 4C.

The seed alignment option can be used in a more complex situation like that in Figure 4D, in which we already have a skeleton alignment, based on structural information or manual annotation, and we have to add multiple unaligned homologs into the alignment. In such a case,

```
% mafft-linsi --seed aligned_
sequences unaligned_sequences >
output
```

makes an entire alignment while preserving the skeleton alignment.

AVAILABILITY

The source code of MAFFT Version 6 is available at http://align.bmr.kyushu-u.ac.jp/mafft/software/. The code for McCaskill routine was taken from the Vienna RNA package Version 1.5 [89] and the McCaskill-MEA package [90]. The binaries for Macintosh, Windows and Linux are also available at the same site. We provide alignment and phylogenetic inference services (Figure 5) at http://align.bmr.kyushu-u.ac.jp/mafft/online/server/. For the phylogenetic inference, users can chose either the NJ [91] or UPGMA [73] method. For the NJ method, several methods for distance estimation can be selected: the Poisson correction, the maximum-likelihood (ML) estimation assuming the JTT [22] or WAG model [92] for protein alignment; and the Jukes-Cantor correction [93] for nucleotide alignment. We modified the MOLPHY package [94] to consider the variable substitution rate across sites with the discrete Γ model [95] and use it in the ML distance estimation from a protein alignment. Bootstrap analysis is also supported. Alignments and phylogenetic trees are visualized with the Jalview [96] and ATV [97] viewers, respectively.

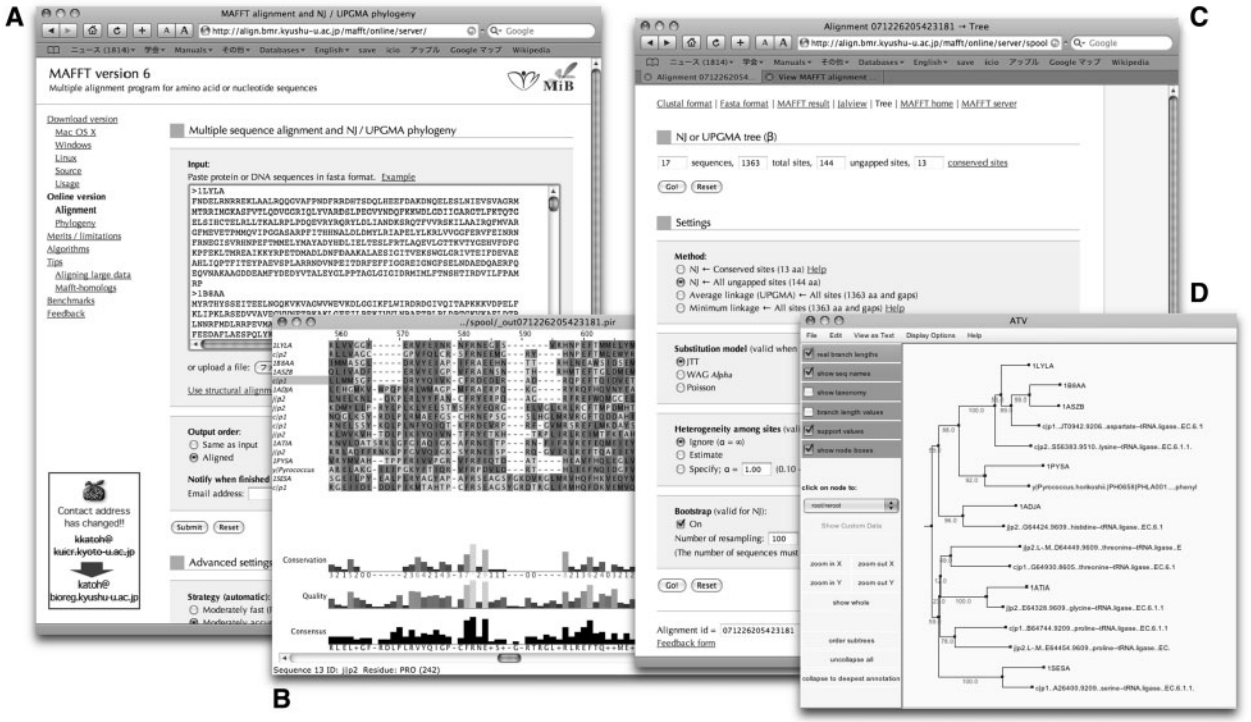

Figure 5: MAFFT web server. (**A**) Interface for sequence input, (**B**) visualization of an MSA with Jalview [96], (**C**) interface for phylogenetic inference and (**D**) visualization of a phylogenetic tree with ATV [97].

FUTURE DIRECTIONS

Consideration of RNA structure

As of December 2007, X-INS-i-scarnapair is one of the most accurate methods for multiple structural RNA alignment. With the current implementation with a time complexity of $O(N^2L^3)$, X-INS-i-scarnapair is already faster than other accurate methods, such as RNA Sampler [84], MASTR [87] and Murlet [83], but the time complexity of X-INS-i-scarnapair can be further reduced to $O(NL^3) + O(N^2L^2)$ if the SCARNA source becomes open [see Katoh and Toh, submitted for publication for details].

Many research groups are now working on the RNA alignment issue [77–88], and the accuracy and speed of ncRNA aligners have rapidly improved in the last several months, as shown in Table 2. Many of them are based on the Sankoff algorithm [98], which simultaneously performs alignment and secondary structure prediction with a time complexity of $O(L^{3N})$. For pairwise structural alignment ($N = 2$), several successful methods are becoming available [77, 79], which reduced the time complexity from $O(L^6)$ to $O(L^3)$ or so, by introducing various approximations. However, it does not seem fruitful to directly extend the Sankoff algorithm to multiple alignment, for the reason explained in the Introduction section, as well as the problem of time complexity.

This is one of the motivations behind the development of the X-INS-i framework for multiple structural RNA alignment based on the Four-way Consistency (Katoh and Toh, submitted for publication). We are ready to support any pairwise structural alignment algorithm, regardless of whether it is Sankoff-based or not, to be extended to the multiple alignment problem using our framework, which was designed to accept various types of pairwise structural alignments and combine them into a single multiple structural alignment.

Consideration of protein structure

The consideration of structural information is also important for protein alignment, and thus many efforts have been made. SPEM [17], MUMMALS [70], PROMALS [18] and other methods incorporate predicted structural information into an alignment, like the RNA aligners noted above. In contrast, 3DCoffee [99], Expresso [100] and other methods incorporate experimentally determined protein structure information into an alignment by using external structural alignment algorithms, such

as SAP [101], and structure-sequence alignment algorithms, such as FUGUE [102]. Both of these approaches achieved considerable improvement in the accuracy [33]. The latter approach seems promising, because many protein structures are being determined along with the progress of structural genomics. We are planning to explore a combination of MAFFT with the ASH [103–105] structural alignment algorithm.

Scalability

The FFT-based alignment algorithm, the PartTree algorithm and other techniques successfully improved the scalability of MSA. However, these are only applicable to a progressive method, FFT-NS-2, but not to the most accurate methods, G-INS-i, L-INS-i and E-INS-i. The maximum size of the sequence data for these options is currently ~200 sequences × ~5000 sites or so, on a typical desktop computer. We are planning to parallelize the pairwise alignment part of these three options. The maximum data size can be extended by parallelization, although the order of time complexity does not change.

Determining an appropriate set of sequences and positions to be included within an MSA

For phylogenetic analyses, we sometimes encounter a serious problem, in terms of which sequences should be included in an MSA and a phylogenetic tree, and which sequences should be excluded. As a large number of homologs are available from databases, such a problem becomes quite bothersome. There are several types of unusual sequences that degrade the accuracies of alignment and phylogenetic inference: (i) fragment sequences, (ii) amino acid sequences incorrectly translated from genomic data and (iii) nonhomologous sequences, etc.

Gouveia-Oliveira *et al.* [106] recently described a tool, MaxAlign, that deletes unusual sequences from a given MSA to maximize the size of 'alignment area', the number of residues in gap-free columns. MaxAlign seems to be an interesting approach and it may be more useful if an MSA method itself automatically determines the sequences to be included within the alignment. One possible way is to extract a commonly aligned region by all-to-all pairwise local alignments. However, such a method may miss a considerable part of alignable residues, because pairwise alignment is usually less sensitive than multiple alignment. A iterative application of multiple alignment and MaxAlign may be worth trying.

There can be different situations where unusual sequences should not be excluded. For example, an MSA itself is useful to identify misidentified genes and other unusual sequences. Therefore, an alignment algorithm that is robust to unusual data is also an important issue.

Incorporation and extraction of biological knowledge in an MSA

In order to construct a biologically relevant MSA, we have to consider the structural, functional and evolutionary information, as well as the optimality with respect to a given scoring system. Manual inspection based on biological knowledge will thus remain important [35], although it is becoming difficult with the increasing number of available sequences. In such a situation, the use of databases of annotated alignments [107, 108] will be a fruitful and practical way to construct an accurate MSA, as well as to extract solid information from an MSA [109]. Hence, more flexible frameworks and tools to build an MSA combining various types of alignment-related data, including structural alignments and manually annotated information, will become important, in addition to more relevant objective functions and faster algorithms.

Key Points

- MAFFT Version 6 has two major new features, the PartTree algorithm for handling a large number (>~10 000) of sequences and the Four-way Consistency objective function for multiple structural alignment of ncRNAs.
- PartTree is a divisive recursive clustering algorithm with a time complexity of $O(N \log N)$. It is more scalable than the conventional UPGMA algorithm with a time complexity of $O(N^2)$. The PartTree option can create a large alignment composed of ~60 000 sequences, at the cost of an accuracy loss of ~2%.
- The X-INS-i-scarnapair, which is a combination of an external pairwise structural RNA alignment method, SCARNA, and the Four-way Consistency objective function, is one of the most accurate methods for multiple RNA structural alignment. It requires less CPU time than other accurate structural alignment methods, such as RNA Sampler, MASTR and Murlet.
- Two different types of group-to-group alignment methods, the profile alignment option and the seed option, were implemented, in order to deal with the various possible phylogenetic relationships between two groups.
- MAFFT Version 6 has L-INS-i and E-INS-i options, which are variants of G-INS-i, the iterative refinement method with WSP and consistency scores. L-INS-i allows large terminal gaps, while E-INS-i is applicable to a dataset with internal unalignable regions.

Acknowledgements
We thank Hiroshi Suga, Kei-ichi Kuma and Go Sasaki for providing a part of the tree inference programs, which were developed at Takashi Miyata's laboratory, Kyoto University, during 1991–2004. We also thank Kazuharu Misawa for a key idea on the FFT-based alignment algorithm. This work was supported by a Grant-in-Aid for 'Comparative Genomics' from MEXT of Japan.

References

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**:443–53.
2. Gotoh O. An improved algorithm for matching biological sequences. *J Mol Biol* 1982;**162**:705–8.
3. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
4. Just W. Computational complexity of multiple sequence alignment with SP-score. *J Comput Biol* 2001;**8**:615–23.
5. Murata M, Richardson JS, Sussman JL. Simultaneous comparison of three protein sequences. *Proc Natl Acad Sci USA* 1985;**82**:3073–7.
6. Carrillo H, Lipman D. The multiple sequence alignment problem in biology. *SIAM J Appl Math* 1988;**48**:1073–82.
7. Lipman DJ, Altschul SF, Kececioglu JD. A tool for multiple sequence alignment. *Proc Natl Acad Sci USA* 1989;**86**: 4412–15.
8. Althaus E, Caprara A, Lenhof HP, *et al*. Multiple sequence alignment with arbitrary gap costs: computing an optimal solution using polyhedral combinatorics. *Bioinformatics* 2002; **18 (Suppl. 2)**:S4–16.
9. Gupta SK, Kececioglu JD, Schaffer AA. Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *J Comput Biol* 1995;**2**:459–72.
10. Zhang X, Kahveci T. QOMA: quasi-optimal multiple alignment of protein sequences. *Bioinformatics* 2007;**23**: 162–8.
11. Notredame C. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics* 2002;**3**:131–44.
12. Heringa J. Local weighting schemes for protein multiple sequence alignment. *Comput Chem* 2002;**26**:459–77.
13. Thompson JD, Plewniak F, Thierry J, *et al*. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res* 2000;**28**: 2919–26.
14. Katoh K, Kuma K, Toh H, *et al*. MAFFT Version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 2005;**33**:511–18.
15. Simossis VA, Kleinjung J, Heringa J. Homology-extended sequence alignment. *Nucleic Acids Res* 2005;**33**:816–24.
16. 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**:823–38.
17. Zhou H, Zhou Y. SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics* 2005;**21**:3615–21.
18. Pei J, Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics* 2007;**23**:802–8.
19. Papadopoulos JS, Agarwala R. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* 2007;**23**:1073–9.
20. Katoh K, Misawa K, Kuma K, *et al*. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66.
21. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 1988;**73**:237–44.
22. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 1992;**8**:275–82.
23. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;**5**:113.
24. Gotoh O. A weighting system and algorithm for aligning many phylogenetically related sequences. *Comput Appl Biosci* 1995;**11**:543–51.
25. Notredame C, Holm L, Higgins DG. COFFEE: an objective function for multiple sequence alignments. *Bioinformatics* 1998;**14**:407–22.
26. Nuin PA, Wang Z, Tillier ER. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* 2006;**7**:471.
27. Ahola V, Aittokallio T, Vihinen M, *et al*. A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics* 2006;**7**:484.
28. Golubchik T, Wise MJ, Easteal S, *et al*. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Mol Biol Evol* 2007;**24**:2433–42.
29. Wilm A, Mainz I, Steger G. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol* 2006;**1**:19.
30. Carroll H, Beckstead W, O'Connor T, *et al*. DNA reference alignment benchmarks based on tertiary structure of encoded proteins. *Bioinformatics* 2007;**23**:2648–9.
31. Wallace IM, O'Sullivan O, Higgins DG, *et al*. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 2006;**34**:1692–9.
32. Lassmann T, Sonnhammer EL. Automatic extraction of reliable regions from multiple sequence alignments. *BMC Bioinformatics* 2007;**8 (Suppl. 5)**:S9.
33. Notredame C. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol* 2007;**3**:e123.
34. Lake JA. The order of sequence alignment can bias the selection of tree topology. *Mol Biol Evol* 1991;**8**:378–85.
35. Morrison D. Multiple sequence alignment for phylogenetic purposes. *Aust Syst Bot* 2006;**19**:479–539.
36. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. *Science* 2008;**319**:473–6.
37. Morgenstern B, Dress A, Werner T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc Natl Acad Sci USA* 1996;**93**: 12098–103.
38. Morgenstern B. DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* 1999;**15**:211–18.

39. Subramanian AR, Weyer-Menkhoff J, Kaufmann M, *et al*. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics* 2005;**6**:66.
40. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;**85**: 2444–8.
41. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search pro grams. *Nucleic Acids Res* 1997;**25**:3389–402.
42. Brudno M, Do CB, Cooper GM, *et al*. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 2003;**13**:721–31.
43. Bray N, Pachter L. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* 2004;**14**:693–9.
44. Raphael B, Zhi D, Tang H, *et al*. A novel method for multiple alignment of sequences with repeated and shuffled elements. *Genome Res* 2004;**14**:2336–46.
45. Phuong TM, Do CB, Edgar RC, *et al*. Multiple alignment of protein sequences with repeats and rearrangements. *Nucleic Acids Res* 2006;**34**:5932–42.
46. Blanchette M, Kent WJ, Riemer C, *et al*. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 2004;**14**:708–15.
47. Darling AC, Mau B, Blattner FR, *et al*. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004;**14**:1394–403.
48. Hogeweg P, Hesper B. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *J Mol Evol* 1984;**20**:175–86.
49. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* 1987; **25**:351–60.
50. 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**:4673–80.
51. Larkin MA, Blackshields G, Brown NP, *et al*. Clustal W and Clustal X Version 2.0. *Bioinformatics* 2007;**23**:2947–8.
52. 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**:1546–56.
53. Lassmann T, Sonnhammer EL. Kalign – an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 2005;**6**:298.
54. Myers EW, Miller W. Optimal alignments in linear space. *Comput Appl Biosci* 1988;**4**:11–7.
55. 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**: 327–37.
56. Berger MP, Munson PJ. A novel randomized iterative strategy for aligning multiple protein sequences. *Comput Appl Biosci* 1991;**7**:479–84.
57. Gotoh O. Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput Appl Biosci* 1993;**9**:361–70.
58. Ishikawa M, Toya T, Hoshida M, *et al*. Multiple sequence alignment by parallel simulated annealing. *Comput Appl Biosci* 1993;**9**:267–73.
59. Notredame C, Higgins DG. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res* 1996;**24**:1515–24.
60. Wallace IM, O'Sullivan O, Higgins DG. Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics* 2005;**21**:1408–14.
61. Chakrabarti S, Lanczycki CJ, Panchenko AR, *et al*. Refining multiple sequence alignments with conserved core regions. *Nucleic Acids Res* 2006;**34**:2598–606.
62. Vingron M, Argos P. A fast and sensitive multiple sequence alignment algorithm. *Comput Appl Biosci* 1989;**5**:115–21.
63. Gotoh O. Consistency of optimal sequence alignments. *Bull Math Biol* 1990;**52**:509–25.
64. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 2000;**302**:205–17.
65. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**:1792–7.
66. Yamada S, Gotoh O, Yamana H. Improvement in accuracy of multiple sequence alignment using novel group-to-group sequence alignment algorithm with piecewise linear gap cost. *BMC Bioinformatics* 2006;**7**:524.
67. Hirosawa M, Totoki Y, Hoshida M, *et al*. Comprehensive study on iterative algorithms of multiple sequence alignment. *Comput Appl Biosci* 1995;**11**:13–8.
68. Do CB, Mahabhashyam MS, Brudno M, *et al*. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* 2005;**15**:330–40.
69. Roshan U, Livesay DR. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* 2006;**22**:2715–21.
70. Pei J, Grishin NV. MUMMALS: multiple sequence alignment improved by using hidden markov models with local structural information. *Nucleic Acids Res* 2006;**34**: 4364–74.
71. Altschul SF. Generalized affine gap costs for protein sequence alignment. *Proteins* 1998;**32**:88–96.
72. Edgar RC, Batzoglou S. Multiple sequence alignment. *Curr Opin Struct Biol* 2006;**16**:368–73.
73. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 1958; **28**:1409–38.
74. Katoh K, Toh H. PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. *Bioinformatics* 2007;**23**:372–74.
75. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 1990;**29**:1105–19.
76. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* 2006;**22**:e90–8.
77. Tabei Y, Tsuda K, Kin T, *et al*. SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments. *Bioinformatics* 2006;**22**:1723–9.
78. Bauer M, Klau GW, Reinert K. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics* 2007;**8**:271.
79. Havgaard JH, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput Biol* 2007;**3**: 1896–908.

80. Reeder J, Giegerich R. Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics* 2005;**21**:3516–23.
81. Dalli D, Wilm A, Mainz I, *et al*. StrAl: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics* 2006;**22**:1593–9.
82. Yao Z, Weinberg Z, Ruzzo WL. CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics* 2006;**22**:445–52.
83. Kiryu H, Tabei Y, Kin T, *et al*. Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics* 2007;**23**:1588–98.
84. Xu X, Ji Y, Stormo GD. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics* 2007;**23**: 1883–91.
85. Torarinsson E, Havgaard JH, Gorodkin J. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics* 2007;**23**:926–32.
86. Kruspe M, Stadler PF. Progressive multiple sequence alignments from triplets. *BMC Bioinformatics* 2007;**8**:254.
87. Lindgreen S, Gardner PP, Krogh A. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics* 2007;**23**:3304–11.
88. Will S, Reiche K, Hofacker IL, *et al*. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* 2007;**3**:e65.
89. Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction for aligned RNA sequences. *J Mol Biol* 2002;**319**: 1059–66.
90. Kiryu H, Kin T, Asai K. Robust prediction of consensus secondary structures using averaged base pairing probability matrices. *Bioinformatics* 2007;**23**:434–41.
91. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;**4**:406–25.
92. Olsen R, Loomis WF. A collection of amino acid replacement matrices derived from clusters of orthologs. *J Mol Evol* 2005;**61**:659–65.
93. Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN (ed). *Mammalian protein metabolism*. New York: Academic Press, 1969;21–132.
94. Adachi J, Hasegawa M. Molphy Version 2.3: Programs for molecular phylogenetics based on maximum likelihood. *Computer Science Monograph*. Tokyo: The Institute of Statistical Mathematics, 1996.
95. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 1994;**39**:306–14.
96. Clamp M, Cuff J, Searle SM, *et al*. The Jalview Java alignment editor. *Bioinformatics* 2004;**20**:426–7.
97. Zmasek CM, Eddy SR. ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* 2001;**17**: 383–4.
98. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math* 1985;**45**:810–25.
99. O'Sullivan O, Suhre K, Abergel C, *et al*. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol* 2004;**340**:385–95.
100. Armougom F, Moretti S, Poirot O, *et al*. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res* 2006;**34**:W604–8.
101. Taylor WR, Orengo CA. Protein structure alignment. *J Mol Biol* 1989;**208**:1–22.
102. 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**:243–57.
103. Toh H. Introduction of a distance cut-off into structural alignment by the double dynamic programming algorithm. *Comput Appl Biosci* 1997;**13**:387–96.
104. Standley DM, Toh H, Nakamura H. GASH: an improved algorithm for maximizing the number of equivalent residues between two protein structures. *BMC Bioinformatics* 2005;**6**: 221.
105. Standley DM, Toh H, Nakamura H. Ash structure alignment package: sensitivity and selectivity in domain classification. *BMC Bioinformatics* 2007;**8**:116.
106. Gouveia-Oliveira R, Sackett PW, Pedersen AG. MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics* 2007;**8**:312.
107. Letunic I, Copley RR, Pils B, *et al*. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 2006;**34**:D257–60.
108. Finn RD, Tate J, Mistry J, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2008;**36**:D281–8.
109. Thompson JD, Muller A, Waterhouse A, *et al*. MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics* 2006;**7**:318.
110. Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* 2005;**33**:2433–9.
111. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA* 2005;**102**:2454–9.
112. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res* 2003;**31**:3423–8.
113. Katoh K, Misawa K. Multiple sequence alignments: the next generation. *Seibutsubutsuri* 2006;**46**:312–7.