Briefings in Bioinformatics, 17(6), 2016, 1009–1023

doi: 10.1093/bib/bbv099
Advance Access Publication Date: 27 November 2015
Software Review

Multiple sequence alignment modeling: methods and applications

Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb and Cedric Notredame

Corresponding author: Cedric Notredame, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain. Tel.: +34 933160271; Fax: +34 933160099; E-mail: cedric.notredame@crg.eu

Abstract

This review provides an overview on the development of Multiple sequence alignment (MSA) methods and their main applications. It is focused on progress made over the past decade. The three first sections review recent algorithmic developments for protein, RNA/DNA and genomic alignments. The fourth section deals with benchmarks and explores the relationship between empirical and simulated data, along with the impact on method developments. The last part of the review gives an overview on available MSA local reliability estimators and their dependence on various algorithmic properties of available methods.

Key words: multiple sequence alignments; protein alignments; alignment benchmarks; alignment reliability measures; RNA alignments; DNA alignments

Introduction

Multiple sequence alignment (MSA) methods refer to a series of algorithmic solution for the alignment of evolutionarily related sequences, while taking into account evolutionary events such as mutations, insertions, deletions and rearrangements under certain conditions. These methods can be applied to DNA, RNA or protein sequences. A recent study in Nature [1] reveals MSA to be one of the most widely used modeling methods in biology, with the publication describing ClustalW [2] pointing at #10 among the most cited scientific papers of all time. Indeed, a large number of *in silico* analyses depend on MSA methods. These include domain analysis, phylogenetic reconstruction, motif finding and a whole range of other applications, extensively described in [3–4].

MSA is indeed an important modeling tool whose development has required addressing a complex combination of computational and biological problems. The computation of an accurate MSA has long been known to be an NP-complete problem, a situation that explains why over 100 alternative methods have been developed these past three decades [4]. Original MSA methods (MSAMs) and their applications have been extensively covered by several reviews [3–5]. To avoid redundancy, we will focus here on the main developments that have taken place over these past 10 years and put them in a broader historical

Maria Chatzou holds a BSc in Computer Science and Biomedical Informatics and an MSc in Bioinformatics. Currently, she is working at the Centre for Genomic Regulation, in Barcelona, Spain, conducting her doctoral studies in the field of Comparative Bioinformatics, with **Dr Cedric Notredame** as her supervisor. Her main research is about designing and deploying tools and methods that will facilitate the analysis of Big Biomedical Data, allow for biological discoveries and promote personalized medicine.
Cedrik Magis, PhD, is a Research Technician in Comparative Bioinformatics Group at Center for Genomic Regulation (CRG) in Barcelona, Spain.
Jia Ming Chang, PhD, is a Postdoctoral Researcher at the Institute of Human Genetics, CNRS, in Montpelier, France.
Carsten Kemena, PhD, is a Postdoctoral Researcher in the Institute for Evolution and Bioidiversity at University of Munster, in Munster, Germany.
Giovanni Bussotti, PhD, is a Postdoctoral Researcher in Enright Research Group at EMBL-EBI, in Hixton, Cambridge, UK. His research activities focuses on developing and evaluating bioinformatics tools for sequencing data and comparative genomics.
Ionas Erb has a PhD in mathematics and a background in statistical physics. His work in the Center for Genomic Regulation (CRG) in Barcelona, Spain, focuses on multivariate statistical methods and their applications to the analysis of biological sequences, gene expression and behavioral data.
Cedric Notredame, PhD, is a Senior Principal Investigator in the Center for Genomic Regulation (CRG) in Barcelona, Spain, where he leads the Comparative Bioinformatics Group.
Submitted: 10 September 2015; **Received (in revised form):** 16 October 2015

context when needed. The three first sections will detail the general algorithmic framework of MSAMs and show how it relates to the newest methods and their application to all sorts of biological sequences (proteins, RNA, DNA). The fourth part will cover method validation and available benchmarks, with a special emphasis on the newest generation designed to cater for evolutionary and structural modeling. The last part of this review will deal with the quantification of local reliability within MSAs. This task had long been identified as instrumental, and possibly more important than the computation of the models—necessarily approximate. It is, however, only recently that systematic approaches have been developed with the explicit aim of quantifying local reliability, thus allowing a systematic filtering and weighting for downstream modeling. We will review these methods in the light of the latest reports.

Algorithmic frameworks for MSA computation

Despite their wide diversity, MSAMs all share a major key property: their reliance on approximate and usually greedy heuristics, imposed by the NP-complete nature of the problem. These heuristics all depend, more or less explicitly, on specific data properties, such as size, nature of the homology, relatedness, length and so on. As a consequence, any change—even minor—on the kind of data being modeled requires the development of novel heuristic strategies. Such changes have recently included the need of upscaling under the high-throughput sequencing pressure and the need for more complex sequence descriptors, including non-coding RNA or non-transcribed genomic sequences. Shifting modeling needs can also drive the developments of novel heuristics, a fact well illustrated by the recent development of phylogeny-aware aligners. Another driving force behind the development of new heuristics has been the increasing availability of structural data that has fueled the development of hybrid methods able to simultaneously deal with sequences and secondary (RNA) or tertiary (RNA and proteins) structures. Likewise, the explosion of available genomic data has put a lot of pressure on the development of a new generation of non-coding/non-transcribed DNA aligners.

Commonly used algorithms

Given a set of biological sequences (RNA, proteins, DNA), the purpose of a MSA method is to align the sequences in a way that will either reflect their evolutionary, functional or structural relationship (Figure 1). This purpose is achieved by inserting gaps of varying length within the sequences, allowing homologous positions to be aligned with each other—just like one would align beads of identical color in an abacus. In an evolutionary context, these gaps represent insertions and deletions (indels) within the genome that are hypothesized to have occurred during the evolution from a common ancestor. In a correct MSA, aligned residues should be maximally similar according to some specified criteria. For instance, if the alignment is meant for evolutionary reconstruction, the residues should be homologous, that is to say, correspond to the same residue in the last unique common ancestor of the considered sequences. If the alignment is meant to be a structural model, aligned residues should have comparable positions in their respective 2D or 3D structures. If the alignment is functional, as may happen when analyzing genomic data, aligned positions are expected to support similar functions. Even though it is reasonable to expect a significant overlap between these criteria, it must be stressed that the complexity of evolutionary forces is such that their full agreement cannot be taken for granted. For instance, two structures may be similar as a consequence of convergent evolution but non-homologous from an evolutionary point of view.

To build an MSA, one needs a scoring function (objective function) able to quantify the relative merits of any alternative alignment with respect to the modeled relationship. The MSA can then be estimated by computing an optimally scoring model. The objective function is a critical parameter, as it precisely defines the modeling accuracy of an MSA and its predictive capacity. When it comes to evolutionary reconstructions, the most commonly used objective functions involve maximizing weighted similarities (as provided by a PAM or BLOSUM substitution matrix) while using an affine gap penalty to estimate indels costs. The substitution cost can be adjusted using tree-based weighting schemes that reflect the independent information contribution of each sequence, and the score of columns is estimated by considering the total all-against-all (sums-of-pairs) substitution cost. It is well known that the sum-of-pairs functions are unlikely to be modeling biological relationships accurately enough [6], but they have been shown to provide a reasonable trade-off between structural correctness and computability, that is to say, the possibility to rapidly estimate a reasonable MSA.

Under their most common formulations, the optimization of sums-of-pairs evaluation schemes is NP-complete. One therefore needs to rely on heuristics, the most common one being the progressive alignment algorithm initially described by Hogeweg and Hesper [7]. This algorithm involves incorporating the input sequences one by one into the final model, following an inclusion order defined by a pre-computed guide tree. At each node, a pairwise alignment is carried out between either a pair of sequences, a sequence and a profile or two profiles. The pairwise alignments taking place at each node are estimated using more or less sophisticated adaptations of the Needlman and Wunsch global dynamic programming alignment algorithm [8]. The combination between a tree-based progressive strategy and a global pairwise alignment algorithm forms the backbone of most available methods (Figure 1), including ClustalW [2], T-Coffee [9] and ProbCons [10]. It is also particularly well adapted for the design of iterative strategies (Figure 1), involving reestimating trees and alignments until both converge [11], as implemented in MUSCLE [12], MAFFT [13] and Clustal Omega [14].

Aside from the objective function, the main algorithmic component of the progressive alignment is the guide tree estimation procedure. This tree, that decides in which order the sequences will be incorporated, can be obtained using a wide variety of methods, the most standard being Neighbor Joining (NJ) [15] and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [16]. The interaction between the objective function (substitution scheme and gap penalties), the weighting scheme and the tree is complex and was extensively explored by Wheeler [17] who showed how the proper tuning of these various components can take a standard method up to the level of the most accurate ones. It is therefore unsurprising to observe that the latest algorithmic developments have been focused on guide trees and objective function improvements.

The main caveat of the progressive alignment approach is the existence of local minima (high level of similarity between a subset of sequences resulting from an artifact). For instance, if the guide tree induces the alignment of two distantly related sequences, it often happens that the optimal alignment of these two sequences will not correspond to the pairwise projection

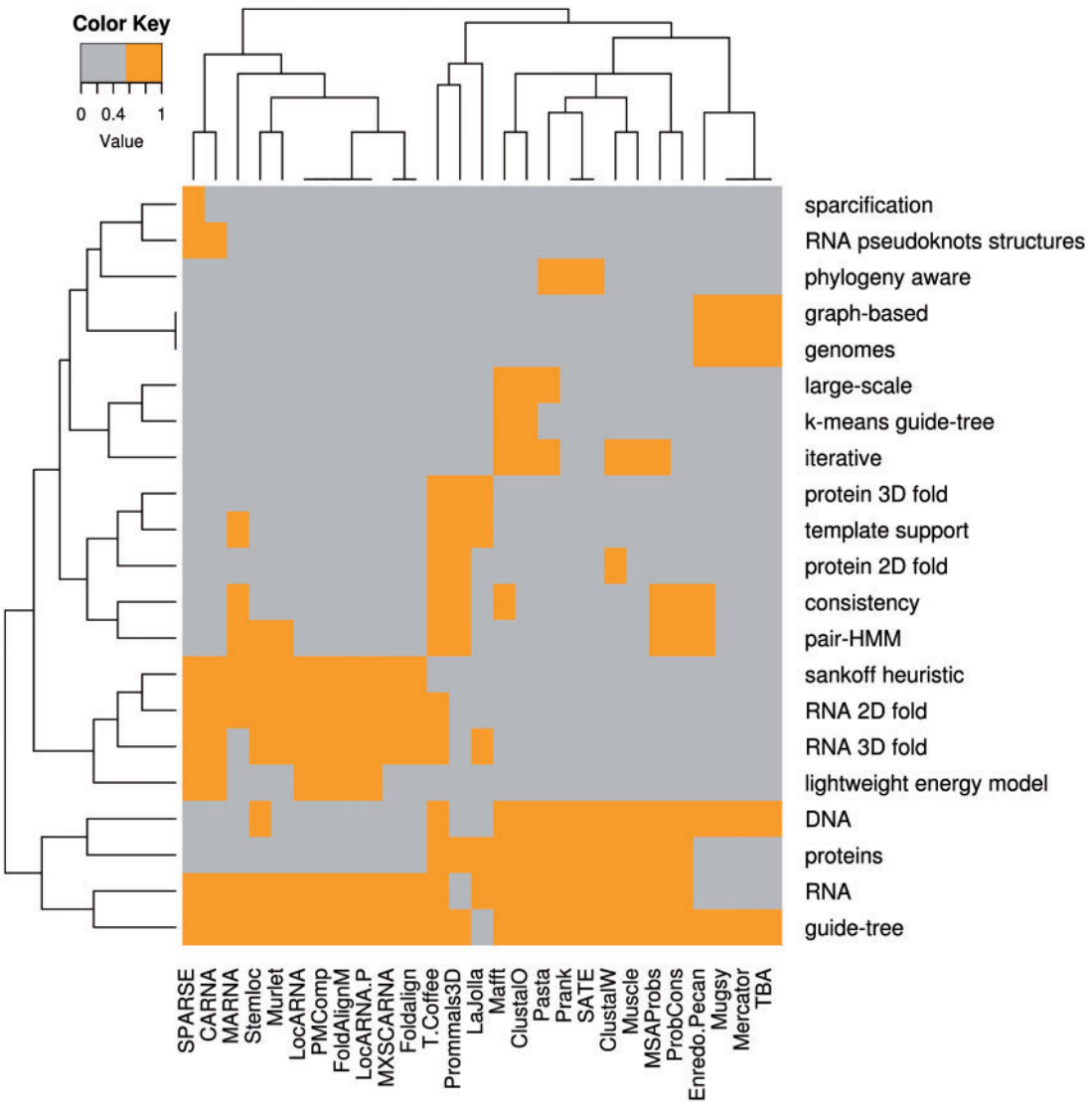

Figure 1. Main algorithmic components of the most widely used multiple aligners. On the heatmap, orange entries indicate a feature implemented in the considered method. Both the aligners and the components were clustered by similarity using the R-package. A colour version of this figure is available online at BIB online: https://academic.oup.com/bib.

one would get from the optimal MSA of the entire data set (i.e. within the MSA the alignment of the two sequences will be slightly suboptimal so as to allow global optimality at the MSA level). This situation is common when dealing with low-identity or low-complexity sequences. When this occurs, the early computation of the first pairwise alignment may prevent the computation of a globally optimal MSA.

The most common strategy to avoid local minima during a progressive alignment is the use of consistency, as originally described by [9]. The rationale of consistency is relatively straightforward: given a set of sequences and their associated pairwise alignments, treated as constraints, scores for matching pairs of residues are reestimated so as to deliver pairwise alignments more likely to be compatible with a globally optimal MSA. The first strategy involving such a reestimation of match costs was reported by Morgenstern as overlapping weights [18]. This scheme later inspired the T-Coffee scoring scheme that has become the archetypical progressive consistency-based aligner [9]. Optimizing an alignment against a set of predefined constraint is known as the Maximum Weight Trace problem. It is NP-complete under its most common formulations and can only be solved for small instances [19, 20]. The T-Coffee algorithm is a heuristic approach that involves reestimating the initial costs of every potential pairwise match by taking into account its compatibility with the rest of the pairwise alignment. The resulting scoring scheme makes it more likely to assemble consistent sub-alignments during the progressive MSA procedure. The main strength of this approach is to allow the computation of MSAs even when an objective function is only available to be optimized at the pairwise level. Consistency-based methods and their relationships have been extensively reviewed in [4]. Since then, the consistency-based approach has become one of the most popular algorithmic frameworks for the development of novel methods (Figure 1).

In a consistency-based algorithm, the most critical parameter is the primary library. Given a set of sequences, the primary library is a collection of all possible pairwise sequence comparisons. This library is used to define the consistency-based objective function. In the original T-Coffee [9], the library was a compilation of all pairs of residues found aligned in the entire

pairwise local and global alignments. These residue pairs were weighted according to the estimated reliability of their source alignments. Later on, a variation of the T-Coffee algorithm named ProbCons [10] established the superiority of pair-HMM-based libraries. In ProbCons, libraries are compiled using a pair-HMM to estimate the posterior probability of all possible pairs of residues (between distinct sequences) to be aligned. The use of a pair-HMM soon became popular among other alignment methods (Figure 1). The main novel features of ProbCons over T-Coffee were the use of a more formal probabilistic framework, thanks to the HMM and the implementation of a biphasic gap penalty when estimating pairwise alignments. Algorithms relying on a similar combination are often referred to as probabilistic consistency algorithms; they include the PECAN multiple genome aligner [21], which uses the Durbin [22] forward only divide and conquer pairwise alignment and MSAProbs [23], which relies on a partition function to achieve more informative posterior probabilities when compiling the library. Some degree of consistency was also incorporated in the MAFFT 'linsi' algorithm.

When benchmarked on structure-based reference alignments, consistency-based aligners have long been shown to yield the most accurate MSAs [4, 23]. This accuracy comes, however, at a significant memory and CPU cost, with most implementations being cubic in CPU and quadratic in memory with the number of sequences. Three strategies have been proposed to address this problem. The simplest one involves faster library computation. For instance, FM-Coffee, the fast implementation of T-Coffee computes its library using three fast aligners, and eventually extracts the resulting pairwise projections. The high correlation between the various projections then makes it possible to band the consistency extension and significantly lower time and memory complexity at a near-quadratic level. Even though the resulting alignments are not as accurate as those obtained using the default procedure, they tend to be more accurate than those produced individually by the combined methods. The second strategy involves parallelization. Two such schemes have been recently published Cloud-Coffee [24] and MSAProbs [23], both which involve parallelizing library computation and the relaxation step during which pairwise costs are reestimated when the progressive alignment assembly is taking place. The last step, which involves splitting computation according to the tree topology, is highly dependent on the guide tree symmetry, best performances being achieved with perfectly balanced guide trees. The third strategy is more sophisticated and involves tuning the library granularity by considering sequence segments rather than single residues. This implementation, available in SeqAn [25], is especially well suited to long closely related sequences, in which long identical segments can be identified.

Large-scale multiple aligners

Even in their most optimized forms, consistency-based methods cannot deal with more than a few hundred sequences. This limit is rather severe in a context where the explosion of genomic sequence availability has resulted in unprecedented large homologous families that can require aligning up to 1.5 million members (number of ABC transporters in Pfam) and soon many more. While the biological relevance of large MSAs can be questioned, recent analysis indicates that important results can be established from such large models [26], thus making the accurate and efficient building of large MSAs one of the current grand challenge of modern biology. Three methods are currently able to deal with such large data sets: the PartTree mode of MAFFT [13], Clustal Omega [14] and PASTA [27], the newest version of SATé [28] (Figure 1). These methods share a common characteristic: their reliance on a fast pre-clustering step (sub-quadratic in time) that makes it possible to rapidly determine the order in which sequences should be aligned.

In the original progressive methods, the guide tree was estimated by comparing all the sequences against one another to estimate a distance matrix. This comparison can be based on a slow Needleman and Wunsch [8] alignment or on a fast *k*-tuple vector comparison as implemented in MAFFT [13], MUSCLE [29] and T-Coffee [9]. The fast comparison does not, however, solve the issue of quadratic time and space requirements for the matrix computation followed by the cubic time complexity of tree estimation when using either UPGMA or NJ. These requirements become prohibitive when processing over 10 000 sequences. Recent clustering methods have been designed to address this issue. In Clustal Omega [14], the guide tree is estimated using the mBed method [30]. The principle of mBed is to first estimate the distance between each sequence and a tiny subset of sequences selected on the basis of their length. For each sequence, the result is a distance vector that can be used to run a hierarchical *k*-means clustering (Figure 1), whose relatively low complexity (*NlogN* under the most common heuristic implementations) allows large data sets of 10 000 sequences or more to be aligned. PartTree relies on a slightly different procedure that also involves using a small set of seed sequences to rapidly pre-cluster the sequences. In both mBed and PartTree, the pre-cluster step is followed by the computation of sub-trees that are eventually combined together to form the guide tree. The PartTree approach was recently improved in the SATé algorithm, which involves an extra iterative tree-refinement step. The latest attempt at aligning large data sets is an adapted version of the T-Coffee algorithm that involves combining *k*-means clustering with consistency-based MSAs at a lower level [26]. These approaches scale well, but at the cost of significantly lower accuracies when aligning >1000 sequences, as shown in the Clustal Omega benchmark analysis [14]. A probable side effect of this decreased accuracy has been the report of high alignment inconsistencies between MAFFT, Clustal Omega and T-Coffee when dealing with large data sets of relatively similar orthologous mitochondrial sequences. When considering full data sets, the authors report average agreement levels as low as 60% [26].

Phylogeny-aware multiple aligners

A major milestone in the development of MSAMs has been the introduction of structure-based reference alignments that can be used to compare the relative capacities of various methods to reconstruct structurally correct alignments from sequence only. The choice of structure seems rather natural because 3D features are known to be more evolutionary resilient than the underlying sequences. On the other hand, this approach relies on the unproven rationale that structurally and evolutionary correct alignments are identical. No proof exists that this assumption may be correct, and a simple reasoning suggests it may not be the case. Indeed, while there can be only one correct way of matching homologous residues—the one that perfectly reflects the unique evolutionary history of the considered sequences and matches—there can be as many structurally correct alignments as there are ways to superpose the sequences with equivalent 3D compactness. Another major potential discrepancy between structural and evolutionary alignments results from convergent evolution. Whenever such a process has

shaped some portions of a sequence data set, the resulting alignment matching convergent regions will be structurally correct and evolutionary false—and reciprocally.

This issue has recently been addressed by a series of works aiming at evaluating multiple aligners' accuracy on the basis of the quality of the phylogenetic models they support. These aligners are referred to as phylogeny-aware aligners (Figure 1). PRANK [31] was one of the first. It relies on the idea that correct MSAs must have indels patterns properly reflecting the underlying phylogenetic tree. PRANK was rapidly followed by SATé [28], an iterative multiple aligner derived from MAFFT that attempts to estimate the MSA supporting the highest-scoring maximum likelihood tree. An important merit of this approach is to depart from the long-held assumption that the best MSA is the one maximizing similarity between sequences. In the context of phylogeny-aware aligners, the best MSA is defined as the one yielding the best phylogenetic model [32]. The consequences on the resulting alignments are rather significant and were particularly well illustrated in a recent analysis by Blackburne and Whelan who found that 'similarity-based' MSAMs (e.g. ClustalW, MUSCLE, ProbCons, MAFFT, T-Coffee) and 'evolution-based' MSAMs (e.g. PRANK and BAliPhy) tend to form discrete clusters under the multidimensional scaling based on their own similarity measures between pairs of alternative MSAs [33]. These authors found that the selected aligners have a substantial impact on downstream phylogenetic inference and report the tree topologies and branch length to depend on the aligner category. Aligners also have a clear impact when quantifying positive selection, with different readouts associated with various aligners as reported on the analysis of several *Drosophila* genomes [34]. Morrison suggests that phylogeneticists are usually dissatisfied with similarity-based alignments and tend to manually edit their MSAs to produce alignments more likely to reflect homology from a true evolutionary stand-point [32]. This observation may also explain why results and method ranking achieved on evolutionarily simulated data sets significantly differ from those measured on structure-based empirical data [4]. However, a recent study by Chang [35] shows that the same reliability index can be used to select both the most phylogenetically informative positions and the positions most likely to contain structurally analogous residues. It is worth mentioning that this study gave contrasting results with respect to phylogeny-aware aligners, and while SATé appears to perform well both on structure and evolutionary benchmarks, PRANK was found to return poor structural alignments, while being able to produce alignments supporting trees with an accuracy comparable with the other MSAMs—phylogeny aware and similarity based.

Structure-based MSAs

Thanks to its high evolutionary resilience, structural information can help produce high-quality models, especially in situations where one aims at modeling structural and functional relationships. This section briefly reviews some of the methods able to combine sequence and structural information when aligning RNA or protein sequences.

Protein sequence/structure multiple alignments using template-based protein aligners

Structural information has long been known to be more resilient than its underlying sequence counterpart [36]. Yet, it is only recently that the corpus of available structural information has made it worthy to develop methods able to combine sequence and structural information within a single model. While the first generation of methods used to rely on protein structure threading and related methods, the newer generation of aligners takes advantage of the availability of multiple experimental structures within an increasing number of protein families. It has become common practice to combine structural aligners output (pairwise or multiple) using a consistency-based framework (Figure 1). The principle is fairly straightforward and involves associating each sequence with a template that can be either a bona-fide structure or a sequence with a known structure related closely enough to the sequence of interest so that no ambiguity exists in the template/target sequence alignment. When no structure is available, sequences are replaced with profiles from which one can either generate a conservation profile (as done in PSI-Coffee or TM-Coffee [37]) or a secondary structure prediction using PSIPRED [38], or both as done in Prommals3D [39], 3D-Coffee [40] and Expresso [41]. The library is then built by aligning the sequences in pairs, using the pairwise method best suited for the considered templates. In this way, alternative methods can be combined seamlessly. This approach is especially convenient when dealing with pairwise structural alignment methods lacking a multiple alignment implementation. The possibility of combining several alternative structural aligners also provides a simple way to address the difficulty of objectively telling alternative structure-based sequence alignment models apart. In this context, the consistency-based approach makes it possible to identify the portion of a model best supported by all the considered methods. This approach has been implemented in the Expresso package, which supports three of the most commonly used structural aligners and can easily accommodate any other third-party aligner.

RNA multiple sequence aligners

The low-complexity alphabet of RNA molecules makes their alignment more challenging than that of protein sequences, with biologically meaningful alignments difficult to estimate $<$60% identity [42]. Whenever secondary structures are evolutionarily conserved, covariation often becomes the strongest available signal. However, standard aligners, like ClustalW, MAFFT or T-Coffee, assume site independence and cannot take this information into account, at least in their default usage. In fact, for these standard aligners, covariation is more of a confounding factor as it decreases sequence identity. More specialized aligners are therefore needed, able to simultaneously recognize similarity at the sequence and secondary-structure level. These algorithms are all heuristic approximations, more or less explicitly related to the Sankoff dynamic programming algorithm [43], which simultaneously folds and aligns RNAs at a prohibitive computational cost $O(N^{3m})$, with m being the number of sequences and N their length. Several banded implementations of the algorithm have been reported (Figure 1). These enforce restrictions on the size or shape of substructures; they can be pairwise aligners such as Consan [44], Dynalign [45, 46], Stemloc [47] and Foldalign [48, 49] or multiple aligners such as MXSCARNA [50], a progressive multiple aligner based on SCARNA [51], a pairwise alignment method based on fixed-length stem fragments defined by means of McCaskill's algorithm [52]. Murlet [53] is another such aligner that first estimates the base pairing and match probabilities before running the Sankoff algorithm with these probabilities to estimate the final alignment. In MARNA [54], the structural information is

used for pairwise RNA comparisons before joining them into a MSA with T-Coffee. PMcomp [55] is a method for progressive multiple alignments based on a McCaskill's algorithm to generate and then compare base pairing probability matrices, which enables lightweight computation. For this purpose, it uses a base pair-based energy model instead of the original loop-based energy model.

PMcomp simplifies Sankoff's model by predicting only a single consensus structure and has been an important source of inspiration of the development of many RNA aligners such as LocARNA [56], FoldAlignM [57] and LocARNA-P [58] (Figure 1) that use additional heuristics to further restrict the folding space to be explored, thus resulting in an $O(n^4)$ time complexity. CARNA [59] extends the PMcomp model to pseudoknot structures (Figure 1). RAF [60] combined the ideas of [61] and [55], resulting in a lightweight Sankoff-variant with sequence-based speed up. SPARSE [62] is one of the newest algorithm in the Sankoff-style category. It is reliable. It runs in quadratic time thanks to its reliance on 'sparsified' prediction and RNA alignments based on their structure ensembles. Compared with LocARNA, SPARSE achieves similar alignment and better folding quality in significantly less time (speedup: 3.7). Another approach that StrAl [63] implements is a scoring scheme that combines sequence similarity with pairing probability. This fast heuristic allows a runtime similar to ClustalW. T-Lara [64] implements a graph-based representation of sequence-structure alignments modeled using integer linear programming. The resulting alignments are then further integrated into a T-Coffee style library using Lagrangian relaxation and eventually resolved into an MSA model using T-Coffee. RNAsampler is a sampling-based algorithm able to find common RNA structures in multiple RNA sequences [65]. The program probabilistically samples aligned RNA stems based on inter-sequence base alignment probabilities and stem conservation calculated from intra-sequence base-pairing probabilities. Another example is RNAcast [66], which for each sequence predicts structure profiles within a defined minimum free-energy threshold and then computes the optimal consensus structure that is shared by all the RNAs.

More recently, following up on T-Lara, systematic attempts were made to apply the consistency paradigm to secondary structure predictions. One can do so by considering libraries made of pairs of pairing residues. This principle has been developed in R-Coffee [67], which adopts a pre-folding approach, predicting with RNAplfold [68] the shape of the individual RNA sequences in an early step. Subsequently, the program estimates the MSA with the highest agreement between structures and sequences. A similar approach was later developed in the RNA compliant version of MAFFT [69], where consistency is measured by combining pairs of paired residues across combination of triplets. Both packages achieve comparable levels of accuracy, the main strength of R-Coffee being its capacity to combine complex pairwise RNA aligners like Consan into highly accurate multiple aligners.

The scarcity of RNA 3D information probably explains why so little attention has so far been given to the generation of accurate 3D structure-based multiple RNA alignments. The situation is slowly changing with several novel algorithms recently described to deal with this problem. Existing tools include pairwise aligners like ARTS [70], SARA [71], DIAL [72] and R3D Align [73], and multiple ones like SARSA [74], LaJolla [75] and SARA-Coffee [76]. The heuristic nature of these algorithms tends to make them error prone, hence the importance of RNA-specific MSA editors. Many such tools are available (4SALE [77], CONSTRUCT [78], JPHYDIT [79], RALEE [80], SARSE [81]) and able to dynamically display secondary and compensatory information while editing RNA MSAs.

It is important to note that these algorithms only work well when dealing with RNA-containing evolutionary conserved secondary structure. In their validation of the SARA-Coffee algorithm, Kemena *et al.* [76] reported that when <70% of the nucleotides are involved in evolutionary-conserved Watson and Crick base pairs, structure-aware aligners like those listed above tend to degrade alignment accuracy. This degradation is a mechanical consequence of the explicit algorithmic attempt to seek and match secondary structures under the assumption that these should be homologous. In practice, however, the structures may not be conserved, or not properly predicted, especially in cases where protein/RNA interaction play a role on the *in vivo* RNA fold. This problem is especially important when considering the issue of aligning long noncoding RNA (lncRNA), the most recently described class of RNA genes [82]. So far, no indication of extensively conserved secondary structure has been reported for these genes, which makes it increasingly likely that this new category of transcripts will require a new generation of aligners in the years to come, possibly motif biased and drawing on the recent report that dinucleotide information can help improve lncRNA alignments [83, 84].

Multiply aligning non-transcribed sequences

The increasing availability of complete genomes makes it a pressing need to develop non-transcribed intergenic sequence alignment tools (Figure 1). Indeed, these sequences come with challenges of their own: extreme length, poor conservation, order variations (inversions, translocations and duplications) and the extreme molecular clock heterogeneity resulting from the wide range of functions supported in different ways by the untranslated part of the genome. This last issue is likely to become increasingly important as novel genomic functions, often associated with epigenetics, keep being reported [85, 86].

Multiple genome alignments

While standard sequence aligners usually imply the modeling of three evolutionary operations, insertion, deletion and substitutions, genome-scale alignments must incorporate at least three more operations: inversions, translocations and duplications. In general, multiple genome aligners achieve this through two separate steps. In a first step, homologous genomic fragments are sorted into bins, and in a second step, these bins are turned into standard MSA models. This last step usually depends on standard progressive aligners, algorithmically similar to the ones described in the first part of this review.

While the first alignment step could rely on a simple clustering/segmentation approach, such a procedure would yield disconnected MSA blocks, giving few insights into genomic evolution. For this reason, most new-generation genome aligners rely on the sorting by reversal algorithm for the segmentation step. Sorting by reversal is an NP-complete problem that amounts to reconstructing the minimum chain of events that would edit one genome into another using a series of translocations and inversions [87]. It is not necessary to solve this problem to align genomes, but it helps quantifying the evolutionary cost of alternative alignments. In practice, most algorithms start by seeking colinear segments, often relying on anchor points (usually proteins) gathered using an all-against-all BLAST

procedure. The most popular procedures include Mercator [88] that uses protein anchors, MUMS/Mems (Mugsy [89] or the systematic use of local alignments [90].

TBA [91] was one of the first algorithm to consider a multiple genome alignment (MGA) as a set of separate blocks rather than a continuous sequence, thus making data processing a necessary prerequisite (Figure 1). In the newest generation of MGAs, the pre-processing has become tightly integrated with the alignment process, as in Mercator-Mavid [88] or Enredo/Pecan [92], which uses graph structures (Figure 1) to identify the different genome rearrangements, splits the multiple genomes accordingly and feeds the resulting bins of multiple sequences to Pecan, a space-efficient consistency-based aligner using the Durbin forward-only linear space dynamic programming procedure [22]. Other graph structures (e.g. A-Brujin graph [93], Cactus graph [94]) have been used for this purpose (see Kehr *et al.* [95] for a comparison of different graph structures). Another alternative is to simultaneously carry out alignment and segmentation in a progressive way. This procedure developed by Brudno [90] uses the equivalent of consistency to identify rearrangements most likely to be supported by the whole data set.

MGA method development has, however, been hampered by the difficulty to objectively assess the relative merits of each aligner. In contrast with proteins or RNA sequences, no such thing as a structure or its equivalent is available for genomes, and when the Alignathon [96] contest proposed to compare the capacities of MGAs on eukaryotic data, the benchmarking was eventually carried out using the PSAR objective function [97], a sequence-based estimator relying on probabilistic sampling. The PSAR objective function was initially developed to evaluate genomic MSAs. Its principle is somehow similar to the consistency-based approach of T-Coffee, though more complete and more computationally demanding. In PSAR, given a data set, all sequences are removed in turn, the remaining sequences realigned and the removed sequence realigned to the sub-alignment. The stability of the realignment with respect to the input MSA is then used to estimate the reliability of each residue positioning within the final alignment model. This procedure is generic with no constraint limiting it to nucleotide alignments. It has, however, so far only been tested and benchmarked on simulated genomic data sets.

The Alignathon contest remains the only generic attempt to compare the reliability of multiple genome aligners. 13 MGA packages were compared on either *Drosophila* genomes or artificially generated mammalian genomes. As pointed out by the authors themselves, a major issue in this work is the design of an acceptable standard of truth. The Alignathon coordinators took the decision to use the PSAR objective function as a standard of truth. Such a decision comes with important caveat, possibly reflected in the clear dominance of PSAR-align—a package explicitly optimizing this function—over most alternative aligners. Of more relevance is certainly the measure made by the authors of the agreement between aligners. In the corresponding Jaccard index analysis, they found that on the non-simulated fly genomes, over 50% of the aligned positions—at the nucleotide levels—are inconsistent between pair of methods (Figures 8B and C in the considered report). Such dispersion should be taken as a measure of the complexity one faces when trying to develop a generic DNA aligner. In contrast, more focused effort on well-defined genomic regions can be used to deliver high-quality alignments of functionally homologous regions. This approach has been successfully developed and used to study eukaryotic genome promoters.

Multiple promoter alignments

MGAs aim at using the genome reordering information so as to better understand evolutionary relationships and possibly identify functional constraints associated with gene organization conservation. In this context, promoter multiple comparisons are probably the best example of functional multiple alignments, aiming at uncovering common regulatory patterns between related sequences. These patterns are used to reveal transcription factor binding sites (TFBS). From an algorithmic point of view, the problem can be separated into two distinct categories: motif discovery among unaligned non-homologous (or distantly related sequences) and regular MSAs. The motif-finding techniques relevant for promoter analysis have been extensively reviewed in (see, e.g. [98, 99] for reviews), and their description is beyond the scope of this review.

Methods for the discovery and comparison of homologous promoter regions are more recent. They were initially reported for the discovery of TFBS, through a process often referred to as evolutionary foot printing. Several methods have been described for that purpose. For instance, in [100], potential binding sites are first predicted on single sequences and then used as anchors during the alignment process. Another strategy is to use an alternative scoring scheme on the positions within a sequence known to fit a regulatory element [101]. This more or less amounts to dressing up a sequence with profile weight matrices that define a position-specific scoring scheme. The main limitation, however, of these motif-based methods is their reliance on pre-computed sets of reference motifs. As an alternative, one can simultaneously identify the motifs and align the sequences as proposed in [102, 103]. Other methods can also model inversions and translocation, thus taking into account the fast motif turnover reported in promoter regions [106]. All these methods are computationally too intensive to scale-up over a few (usually two) sequences, and scalable alternatives have been proposed for multiple sequence analysis [104]. It is also possible to fine-tune existing methods for multiple promoter alignments, as shown by Erb *et al.* [105]. In this work, the authors optimized three popular methods (MAFFT, Muscle, T-Coffee) for their capacity to effectively align experimentally proven homologous TFBS. The tuning also tooks into account the discriminative capacity between alignments of orthologous and paralogous gene regions.

Benchmarking multiple aligners accuracy

Quantifying the accuracy of multiple aligners is just as critical as aligning sequences, especially when considering the aligners approximate nature. This seemingly obvious aspect has been generally overlooked by the community as reflected by the relative lack of correlation between the packages overall usage and their reported accuracy. ClustalW, for instance—whose 42 000 citations suggest a global usage level higher than all other packages put together—has not been consistently reported as the most accurate method. This surprising observation probably reflects on a combination of factors. The most obvious is the relationship between benchmarks rankings and day-to-day usability. It is likely that ClustalW, even though it does not rank #1 on all benchmarks, is nonetheless sufficiently accurate for many modeling activities, especially when dealing with orthologous data sets. One may also speculate on the existence of a strong methodological inertia within the biological community, where tool usage tends to snowball through protocol recycling.

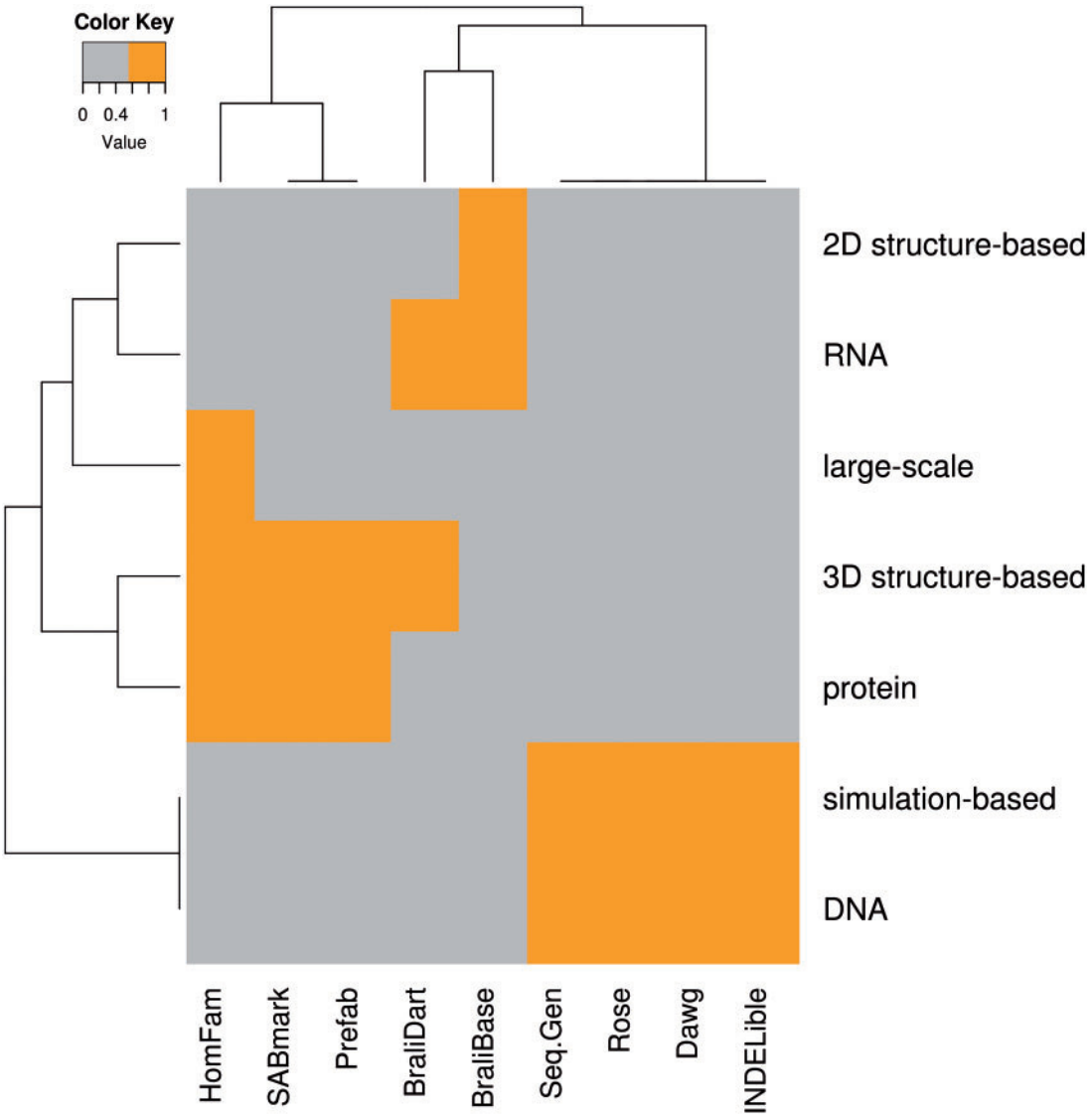

Figure 2. Main benchmark methods and their most relevant properties. On the heatmap, orange entries indicate a property describing a given method. Both properties and benchmarks were clustered by similarity. A colour version of this figure is available online at BIB online: https://academic.oup.com/bib.

The most critical component of an MSA is its scoring/objective function, the mathematical formula that quantifies the total score and therefore defines optimality, given a set of sequences. The rest of the algorithm is an optimization procedure attempting to generate an MSA model that maximizes the objective function. It is well established that even the best objective functions are merely approximations trying to model the behavior of biological sequences [107]. As a consequence, there is no guarantee that a perfectly optimized MSA will systematically result in the most biologically meaningful MSA. This is the reason why multiple aligners also need to be evaluated/benchmarked for their capacity to produce correct alignments. A benchmarking procedure relies on existing collections of reference alignments considered as gold standards. These reference MSAs are routinely used as predictors for the accuracy of a given aligner on a given type of data sets and have had a major influence on methodological developments. Existing protein benchmark collections were recently extensively and critically reviewed in [108] and [109] where the authors propose to group benchmarks in four categories: simulation based, consistency based, structure based and phylogeny based. The latter three categories meet the criterion of reference data sets, in that they can be pre-compiled and used to quantify the relative merits of one aligner over another. The simulation-based benchmarks, however, define an objective function rather than a benchmark procedure and cannot be considered a benchmark measure in the same sense as the others.

Structure-based protein benchmarks

Multiple aligners benchmarking has largely been driven by the use of structure-based reference MSAs, BAliBASE [110] being the most widely used. These benchmarks all rely on structure-based reference alignments to evaluate any aligner able to handle their sequences (Figure 2). It has become customary to report new aligners along with the benchmark readouts established on at least two available structure-based reference data sets. This probably owes to the surprisingly low consistency between benchmarks. Indeed, as shown in [4], aligner's rankings established on the basis of the most common benchmarks are

on average <50% consistent. This means that if a given benchmark suggests that method *A* is more accurate than method *B*, there is less than one chance of two that the same ranking is supported by another benchmark collection. In his detailed analysis of available benchmarks, Edgar suggested SABmark [111] to be the most complete and informative, but only when using a subset of SABmark made of compatible pairwise structural alignments.

The growing need for large-scale aligners has resulted in the development of a new benchmark generation able to estimate alignment accuracy when assembling large data sets. The main issue when doing so is the scarcity of structural information. Of the 16 230 Pfam families with experimental structural information, about 50% merely have one member with a known 3D structure, and 25% have two members only. To accommodate this limitation, reference data sets were built by embedding sequences with a known structure within larger data sets made of sequences with unknown structure. This approach already used in PREFAB [12]—with two sequences of known structure embedded within a data set of 50 sequences—has been extended in HomFam [14, 112], so as to define much larger data sets of up to 100 000 sequences in which an average of 10 sequences with known structures are embedded. When doing so, accuracy is estimated by first aligning the large data sets. The projections of sequences with known structures are then extracted and accuracy is quantified by comparing these projections with the reference. In this procedure, the main caveat lies in the assumption that the seed sequence accuracy reflects well the global data set. This assumption is, however, only correct if the sequences with known structures are evenly distributed within the considered data set.

Structure-based benchmarking does not necessarily depend on a reference alignment, and alternative methods have also been designed that rely on structural superposition rather than structural superposition-induced alignments. These developments were mostly the consequence of work by Lackner [114], who reported on situations where the structure-based superposition is ambiguous enough to support equally well several alternative sequence alignments. When this occurs, the reference alignment becomes the arbitrary prioritization of one reference over another, thus biasing the benchmark process. Most reference benchmarks deal with this problem by specifying core regions in which the reference alignment is expected to be less ambiguous, but this procedure remains dependent on the way in which core regions are defined. A more general alternative exists that involves comparing intra-molecular distances between pairs of aligned residue pairs. This measure, named iRMSD [115], makes it possible to quantify the structural fit implied by an alignment without having to rely on a reference.

Structure based RNA alignment benchmarks

Structural benchmarks have also been developed for RNA alignment evaluation (Figure 2). Three such benchmarks exist. BAliBASE [116] is the most commonly used. It makes it possible to evaluate the accuracy of a multiple aligner on RNA sequences by considering the modeling capacity of the evaluated aligner with respect to some reference secondary structure. This dependence on sequence (on which the secondary structure estimation is based) slightly limits its scope, as it implies common dependencies between the reference compilation and the evaluation procedure. BraliDart [76], a newer data set, that is only based on structural information and contains sets of homologous RNA families with known experimental structures, has been recently reported. This data set is limited by the relative scarceness of experimental RNA 3D structures. Another specificity of BraliDart is its non-reliance on a reference structural alignment but rather on the structural fit implied by the sequence alignment using a distance RMSD measure, as defined by the iRMSD method. The third main category of RNA benchmark is made of ribosomal RNA reference alignments [113]. They have not been assembled for benchmarking purposes, but rather as a consequence of the importance of accurate ribosomal RNA (rRNA) alignments when estimating the tree of life. These alignments have been done manually while taking into account highly conserved rRNA secondary structures that play critical roles in the ribosome functional capacities. At the time we write this review, no reference data set has yet been published to validate the MSAs of long non-coding RNAs, a recently described population of transcripts.

Simulated data sets for evolutionary analysis

Although empirical data benchmarks are the most commonly used strategies to evaluate alignment methods, they remain limited by their dependence on structural data and the lack of such data for the evaluation of certain kinds of alignments—such as non-transcribed DNA. Furthermore, it remains to be established to which extent structure-based alignments can be expected to be evolutionarily correct. This question is especially critical considering that phylogenetic modeling is one of the main applications of MSA modeling. A major issue of the most popular aligner methods is their systematic reliance, and possible tuning on structurally correct sequence alignments. These methods are, however, often used to carry out phylogenic reconstruction. This inconsistency has long been pointed out by the evolutionary community, which routinely relies on simulated data sets rather than empirical ones [117].

Simulated data sets rely on models mimicking evolution to generate sequences whose diversity is expected to represent a true evolutionary process. The main strength of this approach is to provide a perfectly traceable model, in which the relationship between nucleotides or amino acids is explicitly known. Their most obvious drawback is to rely on evolutionary models assumed to be correct, while the true extent to which they represent biologically realistic scenarios remains unknown. In any case, these approaches are useful when estimating the impact of extreme conditions on modeling capacity, for instance accelerated evolution, long-branch attraction and similar effects that can confound standard analysis. Several packages have been designed to generate simulated data sets (Figure 2), the most widely used being Rose [118], Seq-Gen [119], Dawg [120] or INDELible [121].

When using these packages, the simulated alignments are considered as 'true' alignment, thus making it possible to use the same scoring system (Sum of Pairs Score, SP, or Column Score, CS [122]) as for empirical benchmarks. It is worth noting that whenever simulated and structure-based reference data sets have been used to validate similar algorithms for alignment accuracy, the rankings were found to differ significantly between these two groups of benchmarks, a clear indication that different alignment characteristics are being evaluated [4, 123]. All phylogeny-aware aligners are currently evaluated using these simulated data sets. When doing so, the evaluation is often done on tree modeling capacity rather than on the MSA itself. Such algorithms include [117, 124–126].

Resolving the apparent discrepancies between structure-based and simulated reference data sets will probably require a

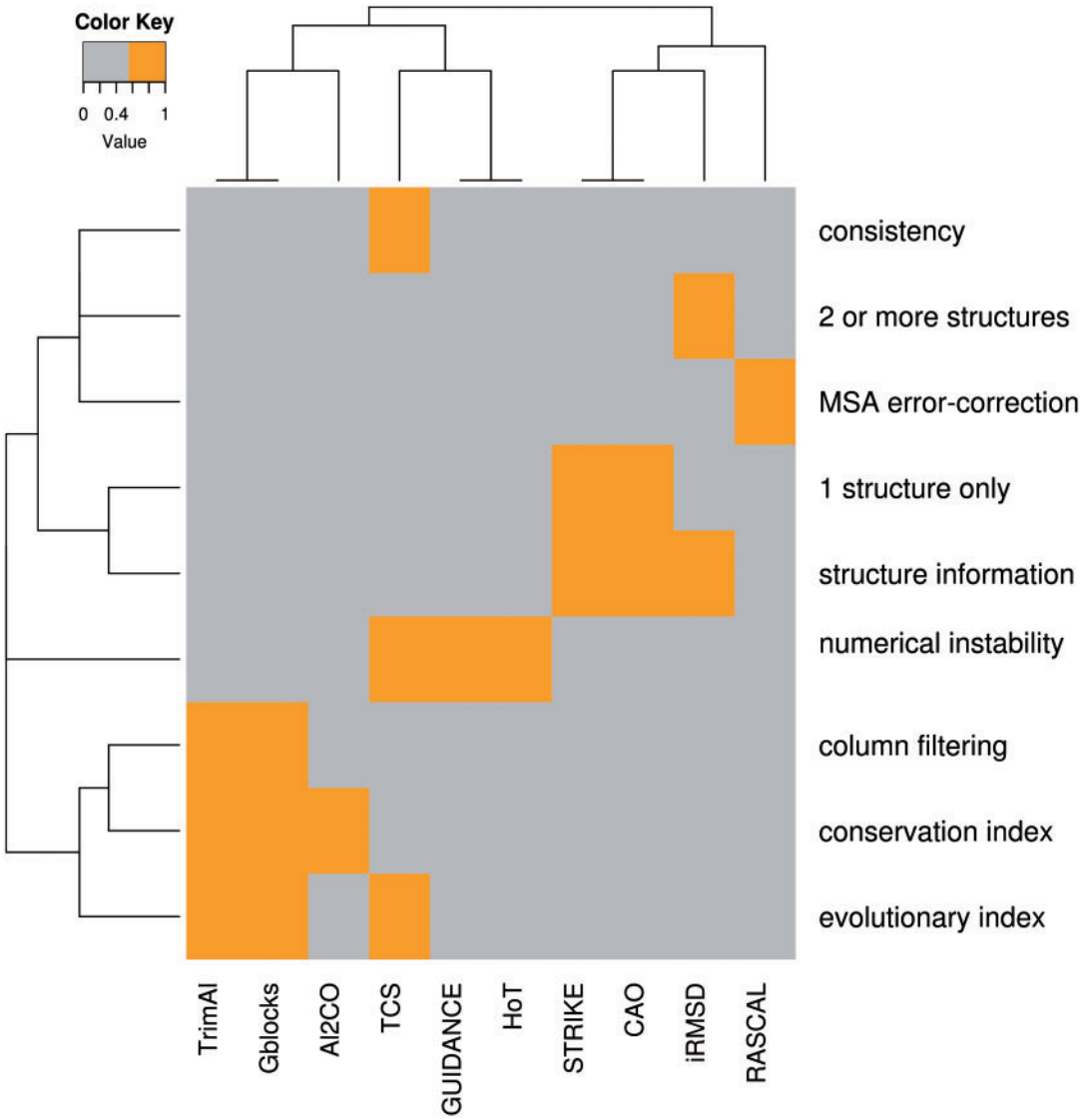

Figure 3. MSA quality indexes and their features. Features with zero are not used by the specific quality index. A colour version of this figure is available online at BIB online: https://academic.oup.com/bib.

better understanding of the complex relation between alignment accuracy and trustworthy phylogenetic reconstruction. Moving one step in this direction, Dessimoz and Gil recently introduced tree-based tests of alignment accuracy, which not only use large and representative samples of real biological data, but also enable the evaluation of the effect of gap placement on phylogenetic inference [127]. In an unrelated work [35], Chang and coauthors proposed the use of empirical data sets obtained by enriching collections of orthologous genes in families likely to support the Tree of Life. When using such data sets, the discrepancy between phylogenetic and structural correctness appears to be less marked.

Quality indexes for the estimation of MSA reliability

Given the approximate nature of all available aligners, identifying the trustworthy portions of an alignment is probably of higher practical importance than knowing the overall expected accuracy. Over the past few years, new methods have been reported aiming precisely at this (Figure 3). They can be roughly divided in three categories: those using structural information to assess protein or RNA accuracy, those depending on a conservation index to identify the positions most likely to be correct and methods depending on some form of local numerical instability to identify the most stable portions of an MSA model.

Structural conservation indexes

With increasingly available structural data, the systematic use of 3D information for the monitoring of MSA accuracy is slowly becoming a realistic prospect. The first such methods [41, 128] were designed using the structural accuracy measured on all possible pairs of sequences with a known 3D structure as a proxy for global accuracy. This approach is useful but suffers from the major limitation that is the limited number of protein and RNA families for which more than one structure is available (about 25% of all PFAM families with a known structure, and <1% in RFAM). Recent efforts were therefore focused toward the

use of single structures to estimate MSA accuracy. The CAO contact substitution matrix [129] is one of the earliest work in this direction. The principle is to embed a sequence with a known structure in the MSA. This structure is then used to identify putative amino-acid contacts and the corresponding columns are then reevaluated using the CAO, a 400×400 substitution matrix assigning a score to every possible contact substitution. Unfortunately, the estimation of this matrix is limited by the lack of available data. This problem was addressed by the STRIKE algorithm [130], in which the contact substitution matrix is replaced with a contact potential metrics that considers the score of all potential contacts, as obtained from structural data. When using this matrix to evaluate an MSA, column contacts—as implied by at least one embedded structure—are evaluated by summing the contact score found in the contact log-odd matrix. This approach was shown to be significantly superior to CAO as a mean to discriminate between alternative alignments.

Sequence conservation indexes

Sequence conservation is one of the most straightforward ways of estimating MSA accuracy. A large number of tools have been developed for this purpose that roughly fall in two main categories: structural (i.e. structural estimates using sequence information) and evolutionary indexes. The evolutionary indexes aim at identifying within an MSA all positions likely to hamper phylogenetic reconstruction. These indexes are usually focused on the removal of diverse columns or indel-enriched regions. The most commonly used tools are Gblocks [131,132] and trimAl [136], a re-implementation of Gblocks using an automated parameterization procedure to adjust the filtering level. While these tools are extremely popular and form part of many large-scale phylogenetic pipelines, the actual value of column filtering remains a point of discussion. Two recent reports suggest that filtering could decrease MSA phylogenetic modeling potential [28, 35]. Similar tools have been developed to estimate the structural correctness of protein MSAs. The simplest ones like AL2CO [133] merely measure conservation according to various physicochemical criterions. Columns and residues eventually get assigned an index value that can be used when doing modeling. More sophisticated variations include RASCAL [137], an MSA scanning procedure meant to identify spurious regions within an MSA.

Alignment stability indexes

The most widely used MSA packages rely on a combination between the progressive algorithm and more or less sophisticated dynamic programming implementations, allowing pairwise alignments of sequences or profiles. These dependencies make these algorithms inherently unstable. Over the past few years, the development of methods able to quantify this instability to estimate local reliability has become a fast growing trend. The idea of using robustness as an indicator of biological accuracy is not new and had already been used as early as 1996 [138] in a procedure that involved removing in turn every pair of amino acid in a pair of sequences before realigning them, so as to assess local alignment stability. Later on, the T-Coffee objective function [107] was used to show the predictive power of consistency. In general, any procedure that may be used to perturbate an alignment lends itself to the definition of a robustness index. Such indexes can then be evaluated for their correlation with structural or phylogenetic modeling potential. The Head or Tail (HoT) procedure [134] is a good example of a simple method (sequences are simply inverted), yielding useful information at the cost of a moderate computational overhead. Other similar procedure albeit more costly have been described. PSAR is one of them [97]. It is a method that involves generating several alternative MSAs while removing each sequence in turn. Another such procedure is named GUIDANCE [135, 139], where the MSA is reestimated several times using guide trees estimated from bootstrap replicates of the original MSA. The main issue with these two approaches is their relatively high computational cost. These methods are, however, much more informative than their sequence conservation alternatives. In a recent report, the T-Coffee consistency score [35] has been shown to outperform both HoT and GUIDANCE for the identification of structurally correct portions within MSAs and both trimAL and Gblocks for the construction of accurate phylogenetic trees.

Conclusion

This review is an attempt to put in context and cover the developments that have taken place in the field of MSAs over the past decade or so. The unprecedented pace of development makes it difficult to be truly exhaustive. We have nonetheless tried to provide the reader with an overview of the main aspects, and how they connect to one another. As shown in Figure 1, the progressive alignment framework (aligning the sequences following a tree-order) is the main algorithmic heuristic that has been adopted by almost all existing alignment methods. Further, we can observe a clear clustering of the methods based on the type of sequences they are designed to align (RNA, DNA/genomes or proteins). It is also worth noting that the current inflation in the number of available methods merely reflects the growing pace of data accumulation. MSA modeling is one of the most powerful ways to make sense of biological sequences. MSAMs, by their approximate nature, are doomed to follow a red-queen evolutionary strategy and will need to keep evolving, faster and faster, to keep up with the processing of standard biological data.

Key Points

- This review provides an overview on the development of Multiple Sequence Alignment (MSA) methods and their main applications.
- MSA method is one of the most powerful and widely used modeling methods in biology, and a series of algorithmic solutions has been proposed over the years for the alignment of evolutionarily related sequences, while taking into account evolutionary events such as mutations, insertions, deletions and rearrangement under certain conditions.
- We report on the main development of these past 10 years that include: the development of consistency based methods, the development of sequence/structure alignment methods, the development of structure based RNA aligners and the development of index-based filtering methods.
- The main challenges for multiple sequence aligners will be to keep up with growing data set sizes and effectively deal with nucleic acid alignments.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness (grant no. BFU2014-55062-P); the Secretariat of Universities and Research, Dept. of

Economy and Knowledge of the Government of Catalonia (2014 SGR 1114); the "la Caixa" International Fellowship Programme for a predoctoral fellowship at the CRG to M.C.; and the Spanish Ministry of Economy and Competitiveness, "Centro de Excelencia Severo Ochoa 2013-2017", SEV-2012-0208.

References

1. Van Noorden R, Maher B, Nuzzo R. The top 100 papers. *Nature* 2014;**514**:550–3.
2. Thompson J, Higgins D, Gibson T. 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**:4673–90.
3. Thompson JD, Linard B, Lecompte O, *et al*. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* 2011;**6**:e18093.
4. Kemena C, Notredame C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* 2009;**25**:2455–65.
5. Edgar RC, Batzoglou S. Multiple sequence alignment. *Curr Opin Struct Biol* 2006;**16**:368–73.
6. Notredame C, Higgins DG. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res* 1996;**24**:1515–24.
7. Hogeweg P, Hesper B. The alignment of sets of sequences and the construction of phylogenetic trees: an integrated method. *J Mol Evol* 1984;**20**:175–86.
8. 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.
9. 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.
10. Do CB, Mahabhashyam MS, Brudno M, *et al*. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res* 2005;**15**:330–40.
11. Wallace IM, O'Sullivan O, Higgins DG. Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics* 2005;**21**:1408–14.
12. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**:1792–7.
13. 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.
14. Sievers F, Wilm A, Dineen D, *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011;**7**:539.
15. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;**4**:406–25.
16. Murtagh F. Complexities of hierarchic clustering algorithms: state of the art. *Comput Stat Q* 1984;**1**:101–13.
17. Wheeler TJ, Kececioglu JD. Multiple alignment by aligning alignments. *Bioinformatics* 2007; **23**:i559–68.
18. Morgenstern B, Frech K, Dress A, *et al*. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* 1998;**14**:290–4.
19. Kececioglu JD. The maximum weight trace problem in multiple sequence alignment. *Lect Notes Comput Sci* 1983;**684**:106–19.
20. Kececioglu JD, Lenhof HP, Mehlhorn K, *et al*. A polyhedral approach to sequence alignment problems. *Discret Appl Math* 2000;**104**:143–86.
21. Paten B, Herrero J, Beal K, *et al*. Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics* 2009;**25**:295–301.
22. Durbin R, Eddy S, Krogh A, *et al*. Probabilistic models of proteins and nucleic acids. *Biol Seq Anal* 1998;**14**:164–73.
23. Liu Y, Schmidt B, Maskell DL. MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* 2010;**26**:1958–64.
24. Di Tommaso P, Orobitg M, Guirado F, *et al*. Cloud-Coffee: implementation of a parallel consistency-based multiple alignment algorithm in the T-Coffee package and its benchmarking on the Amazon Elastic-Cloud. *Bioinformatics* 2010;**26**:1903–4.
25. Rausch T, Emde AK, Weese D, *et al*. Segment-based multiple sequence alignment. *Bioinformatics* 2008;**24**:i187–92.
26. Breen MS, Kemena C, Vlasov PK, *et al*. Epistasis as the primary factor in molecular evolution. *Nature* 2012;**490**:535–8.
27. Mirarab S, Nguyen N, Guo S, *et al*. PASTA: ultra-large multiple sequence alignment for Nucleotide and Amino-acid sequences. *J Comput Biol* 2015;**22**:377–86.
28. Liu K, Raghavan S, Nelesen S, *et al*. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 2009;**324**:1561–4.
29. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;**5**:113.
30. Blackshields G, Sievers F, Shi W, *et al*. Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms Mol Biol* 2010;**5**:21.
31. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA* 2005;**102**:10557–62.
32. Morrison DA. Why would phylogeneticists ignore computerized sequence alignment? *Syst Biol* 2009; **58**:150–8.
33. Blackburne BP, Whelan S. Class of multiple sequence alignment algorithm affects genomic analysis. *Mol Biol Evol* 2013;**30**:642–53.
34. Markova-Raina P, Petrov D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res* 2011;**21**:863–74.
35. Chang JM, Tommaso P, Notredame C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol Biol Evol* 2014;**31**:1625–37.
36. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**:85–94.
37. Chang JM, Di Tommaso P, Taly JF, *et al*. Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. *BMC Bioinformatics* 2012;**13**(Suppl 4):S1.
38. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;**292**:195–202.
39. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* 2008;**36**:2295–300.
40. 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.
41. 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.
42. Capriotti E, Marti-Renom MA. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics* 2010;**11**:322.

43. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J Appl Math* 1985;**45**:810–25.
44. Dowell RD, Eddy SR. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics* 2006;**7**:400.
45. Mathews DH, Turner DH. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol* 2002;**317**:191–203.
46. Mathews DH. Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics* 2005;**21**:2246–53.
47. Holmes I. Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics* 2005;**6**:73.
48. Gorodkin J, Heyer LJ, Stormo GD. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res* 1997;**25**:3724–32.
49. Havgaard JH, Lyngso RB, Stormo GD, *et al*. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics* 2005;**21**:1815–24.
50. Tabei Y, Kiryu H, Kin T, *et al*. A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics* 2008;**9**:33.
51. 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.
52. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 1990;**29**:1105–19.
53. Kiryu H, Tabei Y, Kin T, *et al*. Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics* 2007;**23**:1588–98.
54. Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics* 2005;**21**:3352–9.
55. Hofacker IL, Bernhart SHF, Stadler PF. Alignment of RNA base pairing probability matrices. *Bioinformatics* 2004;**20**:2222–7.
56. 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.
57. Torarinsson E, Havgaard JH, Gorodkin J. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics* 2007;**23**:926–32.
58. Will S, Joshi T, Hofacker IL, *et al*. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA* 2012;**18**:900–14.
59. Sorescu DA, Möhl M, Mann M, *et al*. CARNA–alignment of RNA structure ensembles. *Nucleic Acids Res* 2012;**40**:W49–53.
60. Do CB, Foo CS, Batzoglou S. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics* 2008;**24**:i68–76.
61. Harmanci AO, Sharma G, Mathews DH. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics* 2007;**8**:130.
62. Will S, Otto C, Miladi M, *et al*. SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics* 2015;**31**:2489–96.
63. 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.
64. Bauer M, Klau GW, Reinert K. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics* 2007;**8**:271.
65. 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.
66. Reeder J, Giegerich R. Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics* 2005;**21**:3516–23.
67. Wilm A, Higgins DG, Notredame C. R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res* 2008;**36**:e52.
68. Bernhart SH, Hofacker IL, Stadler PF. Local RNA base pairing probabilities in large sequences. *Bioinformatics* 2006;**22**:614–15.
69. Katoh K, Toh H. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics* 2008;**9**:212.
70. Dror O, Nussinov R, Wolfson H. ARTS: alignment of RNA tertiary structures. *Bioinformatics* 2005;**21**(Suppl 2):ii47–53.
71. Capriotti E, Marti-Renom MA. RNA structure alignment by a unit-vector approach. *Bioinformatics* 2008;**24**:i112–18.
72. Ferrè F, Ponty Y, Lorenz WA, *et al*. DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Res* 2007;**35**:W659–68.
73. Rahrig RR, Leontis NB, Zirbel CL. R3D Align: global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics* 2010;**26**:2689–97.
74. Chang Y-F, Huang YL, Lu CL. SARSA: a web tool for structural alignment of RNA using a structural alphabet. *Nucleic Acids Res*. 2008;**36**:W19–24.
75. Bauer RA, Rother K, Moor P, *et al*. Fast structural alignment of Biomolecules using a Hash table, N-Grams and string descriptors. *Algorithms* 2009;**2**:692–709.
76. Kemena C, Bussotti G, Capriotti E, *et al*. Using tertiary structure for the computation of highly accurate multiple RNA alignments with the SARA-Coffee package. *Bioinformatics* 2013;**29**:1112–19.
77. Seibel PN, Müller T, Dandekar T, *et al*. 4SALE–a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics* 2006;**7**:498.
78. Lück R, Gräf S, Steger G. ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res* 1999;**27**:4208–17.
79. Jeon YS, Chung H, Park S, *et al*. jPHYDIT: a JAVA-based integrated environment for molecular phylogeny of ribosomal RNA sequences. *Bioinformatics* 2005;**21**:3171–3.
80. Griffiths-Jones S. RALEE–RNA ALignment editor in Emacs. *Bioinformatics* 2005;**21**:257–9.
81. Andersen ES, Lind-Thomsen A, Knudsen B, *et al*. Semiautomated improvement of RNA alignments. *RNA* 2007;**13**:1850–9.
82. Derrien T, Johnson R, Bussotti G, *et al*. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012;**22**:1775–89.
83. Bussotti G, Raineri E, Erb I, *et al*. BlastR–fast and accurate database searches for non-coding RNAs. *Nucleic Acids Res* 2011;**39**:6886–95.
84. Lindgreen S, Gardner PP, Krogh A. MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics* 2007;**23**:3304–11.
85. Orom UA, Derrien T, Beringer M, *et al*. Long noncoding RNAs with enhancer-like function in human cells. *Cell* 2010;**143**:46–58.
86. Tilgner H, Nikolaou C, Althammer S, *et al*. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* 2009;**16**:996–1001.

87. Sankoff D, Blanchette M. Multiple genome rearrangement and breakpoint phylogeny. *J Comput Biol* 1998;**5**:555–70.
88. Dewey CN. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol Biol* 2007;**395**:221–36.
89. Angiuoli S V, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 2011;**27**:334–42.
90. 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.
91. Blanchette M, Kent WJ, Riemer C, *et al*. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 2004;**14**:708–15.
92. Paten B, Herrero J, Beal K, *et al*. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* 2008;**18**:1814–28.
93. 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.
94. Paten B, Earl D, Nguyen N, *et al*. Cactus: algorithms for genome multiple sequence alignment. *Genome Res* 2011;**21**:1512–28.
95. Kehr B, Trappe K, Holtgrewe M, *et al*. Genome alignment with graph data structures: a comparison. *BMC Bioinformatics* 2014;**15**:99.
96. Earl D, Nguyen N, Hickey G, *et al*. Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res* 2014;**24**:2077–89.
97. Kim J, Ma J. PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. *Nucleic Acids Res* 2011;**39**:6359–68.
98. Su J, Teichmann SA, Down TA. Assessing computational methods of cis-regulatory module prediction. *PLoS Comput Biol* 2010;**6**:e1001020.
99. Aerts S. Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Curr Top Dev Biol* 2012;**98**:121–45.
100. Berezikov E, Guryev V, Plasterk RHA, *et al*. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res* 2004;**14**:170–8.
101. Sinha S, He X. MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. *PLoS Comput Biol* 2007;**3**:e216
102. Satija R, Pachter L, Hein J. Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. *Bioinformatics* 2008;**24**:1236–42.
103. Satija R, Novák A, Miklós I, *et al*. BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. *BMC Evol Biol* 2009;**9**:217.
104. Majoros WH, Ohler U. Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. *PLoS Comput Biol* 2010;**6**:e1001037.
105. Erb I, Gonzalez-Vallinas JR, Bussotti G, *et al*. Use of ChIP-Seq data for the design of a multiple promoter-alignment method. *Nucleic Acids Res* 2012;**40**:e52.
106. Ettwiller L, Paten B, Souren M, *et al*. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol* 2005;**6**:R104.
107. Notredame C, Holm L, Higgins DG. COFFEE: an objective function for multiple sequence alignments. *Bioinformatics* 1998;**14**:407–22.
108. Edgar RC. Quality measures for protein alignment benchmarks. *Nucleic Acids Res* 2010;**38**:2145–53.
109. Iantorno S, Gori K, Goldman N, *et al*. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods Mol Biol* 2014;**1079**:59–73.
110. Bahr A, Thompson JD, Thierry JC, *et al*. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res* 2001;**29**:323–6.
111. Van Walle I, Lasters I, Wyns L. SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics* 2005;**21**:1267–8.
112. Sievers F, Dineen D, Wilm A, *et al*. Making automated multiple alignments of very large numbers of protein sequences. *Bioinformatics* 2013;**29**:989–95.
113. Yilmaz P, Parfrey LW, Yarza P, *et al*. The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks. *Nucleic Acids Res* 2014;**42**:D643–8.
114. Lackner P, Koppensteiner WA, Sippl MJ, *et al*. ProSup: a refined tool for protein structure alignment. *Protein Eng* 2000;**13**:745–52.
115. Armougom F, Moretti S, Keduas V, *et al*. The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics* 2006;**22**:e35–9.
116. Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* 2005;**33**:2433–9.
117. Ogden TH, Rosenberg MS. Multiple sequence alignment accuracy and phylogenetic inference. *Syst Biol* 2006;**55**:314–28.
118. Stoye J, Evers D, Meyer F. Rose: generating sequence families. *Bioinformatics* 1998;**14**:157–63.
119. Strope CL, Abel K, Scott SD, *et al*. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. *Mol Biol Evol* 2009;**26**:2581–93.
120. Cartwright RA. DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics* 2005;**21**(Suppl 3):iii31–8.
121. Fletcher W, Yang Z. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol* 2009;**26**:1879–88.
122. Thompson JD, Plewniak F, Poch O. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* 1999;**15**:87–8.
123. Wallace IM, Blackshields G, Higgins DG. Multiple sequence alignments. *Curr Opin Struct Biol* 2005;**15**:261–6.
124. Hall BG. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol Biol Evol* 2005;**22**:792–802.
125. Rosenberg MS. Multiple sequence alignment accuracy and evolutionary distance estimation. *BMC Bioinformatics* 2005;**6**:278.
126. Wang L-S, Leebens-Mack J, Kerr Wall P, *et al*. The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM Trans Comput Biol Bioinform* 2011;**8**:1108–19.
127. Dessimoz C, Gil M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol* 2010;**11**:R37.
128. O'Sullivan O, Zehnder M, Higgins D, *et al*. APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics* 2003;**19**(Suppl 1):i215–21.
129. Lin K, Kleinjung J, Taylor WR, *et al*. Testing homology with Contact Accepted mutatiOn (CAO): a contact-based Markov model of protein evolution. *Comput Biol Chem* 2003;**27**:93–102.
130. Kemena C, Taly JF, Kleinjung J, *et al*. STRIKE: evaluation of protein MSAs using a single 3D structure. *Bioinformatics* 2011;**27**:3385–3391.
131. Hickson RE, Simon C, Perrey SW. The performance of several multiple-sequence alignment programs in relation to

secondary-structure features for an rRNA sequence. *Mol Biol Evol* 2000;**17**:530–9.
132. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 2007;**56**:564–577.
133. Pei J, Grishin N V. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 2001;**17**:700–12.
134. Landan G, Graur D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol* 2007;**24**:1380–3.
135. Penn O, Privman E, Landan G, *et al*. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol* 2010;**27**:1759–67.
136. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;**25**:1972–3.
137. Thompson JD, Thierry JC, Poch O. RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics* 2003;**19**:1155–61.
138. Mevissen HT, Vingron M. Quantifying the local reliability of a sequence alignment. *Protein Eng* 1996;**9**:127–32.
139. Sela I, Ashkenazy H, Katoh K, *et al*. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res* 2015;**43**:W7–14.