BIOINFORMATICS REVIEW Vol. 25 no. 19 2009, pages 2455–2465
doi:10.1093/bioinformatics/btp452

Sequence analysis

Upcoming challenges for multiple sequence alignment methods in the high-throughput era

Carsten Kemena and Cedric Notredame*

Centre For Genomic Regulation (Pompeus Fabre University), Carrer del Doctor Aiguader 88, 08003 Barcelona, Spain

Received on April 3, 2009; revised on June 24, 2009; accepted on July 16, 2009

Advance Access publication July 30, 2009

Associate Editor: Jonathan Wren

ABSTRACT

This review focuses on recent trends in multiple sequence alignment tools. It describes the latest algorithmic improvements including the extension of consistency-based methods to the problem of template-based multiple sequence alignments. Some results are presented suggesting that template-based methods are significantly more accurate than simpler alternative methods. The validation of existing methods is also discussed at length with the detailed description of recent results and some suggestions for future validation strategies. The last part of the review addresses future challenges for multiple sequence alignment methods in the genomic era, most notably the need to cope with very large sequences, the need to integrate large amounts of experimental data, the need to accurately align non-coding and non-transcribed sequences and finally, the need to integrate many alternative methods and approaches.

Contact: cedric.notredame@crg.es

1 INTRODUCTION

An ever increasing number of biological modeling methods depend on the assembly of accurate multiple sequence alignments (MSAs). Traditionally, the main applications of sequence alignments have included phylogenetic tree reconstruction, Hidden Markov Modeling (profiles), secondary or tertiary structure prediction, function prediction and many minor but useful applications such as PCR primer design and data validation. With the notable exception of ribosomal RNA, a large majority of these applications are based on the analysis of protein sequences, possibly back-translated into nucleic acid sequences in the context of phylogenetic analysis. While this type of approaches still constitutes the vast majority of published applications for MSAs, recent biological discoveries coupled with the massive delivery of functional, structural and genomic data are rapidly expanding the potential scope of alignment methods. In order to make the best of the available data, sequence aligners will have to evolve and become able to deal with a very large number of sequences or integrate highly heterogeneous information types such as evolutionary, structural and functional data. Merely aligning all the known orthologs of a given gene will soon require aligning several thousand sequences, and the massive re-sequencing effort currently underway (Siva, 2008) could even mean that within a few decades, multiple comparison methods may be required to align billions of closely related sequences.

An MSA is merely a way to organize data so that similar sequence features are aligned together. A feature can be any relevant biological information: structure, function or homology to the common ancestor. The goal is either to reveal patterns that may be shared by many sequences, or identify modifications that may explain functional and phenotypic variability. The features one is interested in and the way in which these features are described ultimately define the correct alignment, and in theory, given a set of sequences, each feature type may define a distinct optimal alignment. For instance, a structurally correct alignment is an alignment where aligned residues play similar role in the 3D structure. Given a set of distantly related sequences, there may be more than one alignment equally optimal from a structural point of view. An alternative to structural conservation is homology (meant in a phylogenetic sense). In that case, the alignment of two residues is a statement that these two residues share a similar relation to their closest common ancestor. Aside from evolutionary inertia, there is no well defined reason why a structure and a homology-based alignment of the same sequences should be identical. Likewise, in a functionally correct alignment, residues having the same function need to be aligned, even if their similarity results from convergent evolution. Overall, a multiple sequence alignment is merely a way of confronting and organizing the specific data one is interested in. Until recently, this notion was mostly theoretical, sequence information being almost the only available data. Alignments were optimized for their sequence similarity, in the hope that this one-size-fits all approach would fare well for most applications. The situation has now changed dramatically and the amount of data that could be integrated when building an MSA is rising by the day. It includes new sequences coming from large scale genome sequencing, with a density of information that will make it more and more possible to reconstruct evolutionary correct alignments (Frazer *et al*., 2007). Other high-throughput-based projects are delivering functional data in the form of transcript structure (Birney *et al*., 2007) and structural data is following a similar trend thanks to coordinated efforts like targetDB (Chandonia *et al*., 2006). Another ongoing trend is the delivery of large-scale functional data, resulting from to the use of robotic techniques. These make it possible to gather large amounts of functional information associated with homologous sequences (Fabian *et al*., 2005). This data is usually applied to Quantitative Structure and Activity Relationships analysis, but it could just as well

*To whom correspondence should be addressed.

be used when comparing protein sequences. Finally, the massive use of ChIp-Chip data makes it possible to reveal important protein/DNA interaction, thus allowing the enrichment of genomic data with functional data, an extra layer of information that could certainly be incorporated in sequence alignment strategies such as the ones we report here.

These trends have not gone un-noticed and over the last years, regular efforts have been made at developing and improving multiple sequence alignments methods so that they could take advantage of newly available data. Three areas have been actively explored: (i) accuracy improvement, achieved through the use of consistency-based methods (Do *et al*., 2005; Notredame *et al*., 2000); (ii) an expansion of MSA methods scope, thanks to the development of template-based approaches (Armougom *et al*., 2006b; Pei and Grishin, 2007; Pei *et al*., 2008; Wallace *et al*., 2006; Wilm *et al*., 2008), a natural development of consistency-based methods that makes it possible to efficiently integrate alternative methods and alternative types of data; and (iii) large-scale alignments (Edgar, 2004a; Katoh and Toh, 2008; Lassmann and Sonnhammer, 2005b). Most of the MSA methods currently available have been described and compared at length in several very complete reviews (Edgar and Batzoglou, 2006; Notredame, 2007; Pei 2008; Wallace *et al*., 2005a). In this specific review, we will mostly focus on the latest developments in an attempt to identify the main trends of this very active research field. We will also discuss an important challenge: the development of adequate benchmarking techniques, informative enough with respect to all potential applications of MSA methods, especially the reconstruction of accurate phylogenies. The urgency of this issue recently received a striking illustration with two high-impact papers dealing with the complex relationship that exist between MSA reconstruction and accurate phylogenetic estimation (Loytynoja and Goldman, 2008; Wong *et al*., 2008).

2 TRADITIONAL ISSUES OF ACCURATE MULTIPLE SEQUENCE ALIGNMENT COMPUTATION

Multiple sequence alignment computation stands at a cross-road between computation and biology. The computational issue is as complex to solve as it is straightforward to describe: given any sensible biological criterion, the computation of an exact MSA is NP-Complete and therefore impossible for all but unrealistically small datasets (Wang and Jiang, 1994). MSA computation therefore depends on approximate algorithms or heuristics and it is worth mentioning that almost every conceivable optimization technique has been adapted into a heuristic multiple sequence aligner. Over the last 30 years, >100 multiple sequence alignment methods have been published, based on all kind of heuristics, including simulated annealing (Abhiman *et al*., 2006), genetic algorithms (Gondro and Kinghorn, 2007; Notredame and Higgins, 1996), Tabu search (Riaz *et al*., 2005), branch and bound algorithms (Reinert *et al*., 1997), Hidden Markov Modeling (Eddy, 1995) and countless agglomerative approaches including the progressive alignment algorithm (Hogeweg and Hesper, 1984), by far the most widely used nowadays. The biological issue surrounding MSAs is even more complex: given a set of sequences, we do not know how to estimate similarity in a way that will guaranty the biological correctness of an alignment, whether this correctness is defined in evolutionary, structural or functional terms. In fact, one could argue that being able to compare the biological features coded by a DNA sequence implies having solved most of the *ab initio* problems associated with genetic information interpretation, including protein structure prediction. But, these problems are not solved and in practice multiple alignments are estimated by maximizing identity, in the hope that this simplistic criterion will be sufficiently informative to yield models usable for most type of biological inference. The objective function thus maximized is usually defined with a substitution matrix and a gap penalty scheme. The substitution matrix is relatively sophisticated when it comes to proteins, but barely more than an identity matrix for DNA and RNA. For a long time, the maximization was carried out using a combination of dynamic programming (DP) and log odds scoring scheme, but over the last year, Bayesian techniques have been implemented that rely on pair-Hidden Markov Models (HMMs) and take advantage of a better defined statistical framework (Durbin *et al*., 1998). While DP- and HMM-based approaches are mostly interchangeable, the last ones make it easier to explore the parameter space using off-the-shelves statistical tools such as Baum–Welch and Viterbi training. HMM modeling also offers easy access to a wider range of scoring possibilities, thanks to posterior decoding, thus making it possible to assess complex alignment scoring schemes. For instance, a significant share of the improvements measured in the ProbCons (Do *et al*., 2005) algorithm over other consistency-based packages seems to result from the use of a bi-phasic penalty scheme (Table 1), pre-defined as a finite state automata (FSA) and parameterized by applying the Baum–Welsch algorithm on BaliBase. Sequence identity is only a crude substitute to biological homology, and in practice, it has often been argued that structurally correct alignments are those more likely to be useful for further biological modeling. Similarity-based MSA methods have therefore been carefully tuned in order to produce structurally correct MSAs. This tuning (or validation) has relied on the systematic usage of structure-based reference multiple sequence alignments. This procedure has now been in use for more than a decade and has been a major shaping force on this entire field of research. We will now review the most common validation procedures with their associated databases.

3 ACCURACY ESTIMATION USING STRUCTURE-BASED REFERENCE ALIGNMENTS

The first systematic validation of a multiple sequence alignment using reference alignments was carried out by McClure (1994). McClure was evaluating her alignments by assessing the correct alignment of pre-defined functional motifs. Shortly after, Notredame and Higgins (1996) made the first attempt to systematically use structure-based alignments while evaluating the biological accuracy of the SAGA package. The validation was carried out on a relatively small dataset named 3D-ali (Pascarella *et al*., 1996). A few years later, Thompson developed a purpose built dataset named BaliBase I (Thompson *et al*., 1999). The main specificity of BaliBase was to address a wide range of different issues related to multiple sequence alignments. This included the alignment of distantly related homologues, the ability of alternative methods to deal with long insertions/deletions and their ability to properly integrate outliers. Its main weakness was the questionable accuracy of some alignments and the relatively small size (82) of the dataset. Most of

Table 1. Benchmarking of a selection of methods on the RV11 Balibase dataset. BaliBase/RV11 is made of 38 datasests consisting of seven or more highly divergent protein sequences (<20% pair-wise identity on the reference alignment)

Method	Version	Score	Mode	Templates	RV11	Sever
3DPSI-Coffee	7.05	Consistency	Accurate	Profile + Structure	61.00	www.tcoffee.org
PROMAL-3D	Server	Consistency	Default	Profile + Structure	58.66	prodata.swemd.edu/promals3d
PROMALS	Server	Consistency	Default	Profile	55.80	prodata.swemd.edu/promals3d
PSI-Coffee	7.05	Consistency	Psicoffee	Profile	53.71	www.tcoffee.org
M-Coffee4	7.05	Consistency	Muscl+Kal. + ProbC + TC	–	41.63	www.tcoffee.org
T-Coffee	7.05	Consistency	Default	–	42.30	www.tcoffee.org
ProbCons	1.1	Consistency	Default	–	40.80	probcons.stanford.edu
ProbCons	1.1	Consistency	Monophsic Penalty	–	37.53	probcons.stanford.edu
Kalign	2.03	It + Matrix	Default	–	33.82	msa.cgb.ki.se
MUSCLE	3.7	It + Matrix	Default	–	31.37	www.drive5.com/muscle
Mafft	6.603b	It + Matrix	Default	–	26.21	align.genome.jp/mafft
Prank	0.080715	Matrix	Default	–	26.18	www.ebi.ac.uk
Prank	0.080715	Matrix	+F	–	24.82	www.ebi.ac.uk
ClustalW	2.0.9	Matrix	Default	–	22.74	www.ebi.ac.uk/clustalw

All packages were ran using the default parameters. Servers were ran in August 2008.

these issues have been addressed in the latest version of BaliBase (BaliBase 3) (Thompson *et al*., 2005) and this database is now one of the most widely used reference standard. Nonetheless, BaliBase remains a handmade dataset, with potential arbitrary and uneven biases resulting from human intervention. The main alternative to BaliBase is Prefab (Edgar, 2004b), a very extensive collection of over a 1000 pairs of homologous structures, each embedded in a collection of ∼50 homologs (25 for each structure) gathered by PSI-BLAST. In Prefab, the reference alignment is defined as the portions of alignments consistently aligned by two structural aligners: CE (Shindyalov and Bourne, 1998) and DALI (Holm and Sander, 1995). Prefab, however, is not a multiple sequence alignment collection since each dataset only contains a pair of structures thus making it a less stringent than BaliBase where accuracy can be tested on entire multiple alignment columns rather than pairs of residues. Other commonly used databases for protein multiple sequence alignments include OXBench (Raghava *et al*., 2003), HOMSTRAD (Stebbings and Mizuguchi, 2004) and SABmark (Van Walle *et al*., 2005). One may ask why so many resources for addressing an apparently simple question. The answer probably lies in the complexity of structural alignments. While reasonably accurate structure-based alignments are easy enough to generate, owing to the strength of the structural signal, it is nonetheless very hard to objectively assess the relative merits of alternative structure-based alignments (Kolodny *et al*., 2005). Several alternative references are therefore available and no simple way exists to objectively evaluate their relative merits. In practice, the authors have taken the habit of running their methods on two or three datasets, verifying trend agreement. Recently, Blackshield and Higgins (Blackshields *et al*., 2006) produced an extensive benchmarking, comparing the 10 main MSA methods using six available datasets. The main trend uncovered by this analysis is that all the empirical reference datasets tend to yield similar results, quite significantly distinct from those measured on artificial datasets such as IRMbase (Subramanian *et al*., 2005, 2008), a collection of artificially generated alignments with local similarity. We checked by re-analyzing some of the Blackshield and Higgins benchmark data (Table 2) in the context of this review. The methodology is very straightforward: each reference dataset is divided in sub-categories, and altogether the six datasets make a total of 43 sub-categories (34 for the empirical datasets, 9 for the artificial). Given two MSA methods A and B, we counted how many times the ranking suggested by one sub-category is in agreement with the ranking suggested by another sub-categories (agreement in Table 2). We then compared all the sub-categories of a dataset against all the sub-categories of the other datasets and reported the average figure in Table 2. We also computed the average agreement within every dataset by measuring the agreement

Table 2. Comparison of alternative reference datasets (adapted from Blackshield and Higgins)

Dataset	#Categories	Agreement (%)	Self-agreement
BaliBase	11	71.4	82.9
RV11	1	77.4	83.3
RV50	1	76.8	80.6
SabMark	4	69.8	81.3
Oxbench	10	65.0	70.8
Prefab	5	64.6	72.3
Homstrad	4	66.8	76.9
IRMdb	9	58.1	88.1
Empirical datasets	**34**	**72.4**	–
All datasets	**43**	**66.1**	–

Blackshield and Higgins published the average accuracy of 10 MSA packages (Mafft, Muscle, POA, Dialign-T, Dialign2, PCMA, align_m, T-Coffee, Clustalw, ProbCons) on six reference databases. This table shows a new analysis of the original data. 'Dataset' indicates the considered dataset. In this column, 'RV11' and 'RV50' are two BaliBase categories, 'Empirical Dataset' refers to the five empirical datasets (BaliBase3, SabMark, Oxbench and Prefab). 'All datasets' includes IRMdb as well. '#Categories' indicates the number of sub-categories contained in the considered datasets. 'Agreement': average agreement between all the considered categories of a given dataset and all the categories of the other databases. The agreement is defined as the number of times two given databases sub-categories agree on the relative accuracy of two methods. The 'Empirical dataset' average is obtained by considering all possible pairs of methods across all possible pairs of categories within the empirical datasets (i.e. all datasets except IRMdb). 'Self-agreement': same measure but restricted to a single database (i.e. each category in turn against all the other categories of the considered database). The last two rows show the average agreement between all respectively all empirical datasets.

across different categories within a dataset. The results on Table 2 suggest that the five main empirical datasets are on average 72.4 % consistent with one another. It means that any prediction of accuracy made on the basis of a single reference dataset is likely to be supported by 72.4% of similar measurements made on the five other empirical reference datasets. A striking observation is the lower agreement between the artificial dataset (IRMdb) and the empirical ones. Observations made on IRMdb are on average only supported by 58.1% of the observations made on the empirical datasets. Two factors could explain this discrepancy: the local nature of IRMdb, mostly designed for assessing local alignment capacities, or its artificial nature. The fact that empirical datasets biased toward local similarity (BaliBase RV50, long indels, 76.8% agreement) do not show a similar trend suggest that the discrepancy between IRMdb and the empirical datasets owes much to its simulated component. Furthermore, at least three other studies reported similar findings, with results established on artificial datasets conflicting with empirical ones (Lassmann and Sonnhammer, 2002, 2005b; Loytynoja and Goldman, 2008)

While there is no clear consensus on this matter, we would argue here that the discrepancy between artificial and empirical datasets pleads in favor on not using the artificial ones. The use of artificial dataset should probably be restricted to situations where the process responsible for the sequence generation is well known and properly modeled, as happens in sequence assembly for instance. It is interesting to note that some sub-categories of BaliBase are extremely informative albeit relatively small. RV11 for instance is 77.4% consistent with the entire collection of empirical dataset which makes it one of the most compact and informative dataset. This is not so surprising if one considers the nature of RV11, a dataset made of highly divergent sequences with $<25\%$ sequence identity in the reference alignment. So far, this dataset has proven fairly resistant to heavy tuning and over-fitting and it is a striking observation that ProbCons, the only package explicitly trained on BaliBase is not the most accurate (as shown on Table 1). Table 1 shows a systematic benchmarking of most methods discussed here on the RV11 dataset. Results are in broad agreement with those reported in most benchmarking studies published over these last 10 years, but the challenging nature of the dataset makes it easier to reveal significant difference in accuracy that are otherwise blurred by other less challenging datasets.

BaliBase has had a strong influence on the field, prompting the design of novel reference datasets for sequences other than proteins. Similar to BaliBase, a reference dataset exists to validate ncRNA alignment methods, called BraliBase (Wilm *et al*., 2006). BraliBase works along the same lines as BaliBase and relies on a comparison between an RNA alignment and its structure-based counterpart. There is, nonetheless, a clear difference between these two reference datasets: in BraliBase, the reference structures are only predicted, and the final evaluation combines a comparison with the reference and an estimation of the predictive capacity of the new alignment. As such, BraliBase is at the same time more sophisticated than BaliBase (because it evaluates the prediction capacity of the alignment) and less powerful because it is not based on a sequence-independent method (unlike BaliBase that uses structural comparison). This limitation results from the relative lack of RNA 3D structures in databases. We will see in the last section of this review that the current benchmarking strategies have many short comings and cannot address all the situations relevant to MSA evaluation. These methods have nonetheless been used to validate all the currently available multiple sequence alignment packages and can certainly be credited (or blamed …) for having re-focused the entire methodological development toward the production of structurally correct alignments. Well standardized reference datasets have also gradually pushed the MSA field toward becoming a fairly codified discipline, where all contenders try to improve over each other's methods by developing increasingly sophisticated algorithms, all tested in the same arena. Given the increased accuracies reported these last years, one may either consider the case closed or suspect that time has come to change arena.

4 THE MOST COMMON ALGORITHMIC FRAMEWORKS FOR MSA COMPUTATION

An interesting consequence of the systematic use of benchmarking methods has been the gradual phase-off of most packages not based on the 'progressive algorithm' (Hogeweg and Hesper, 1984). With the exception of POA (Lee *et al*., 2002), most of the methods commonly used nowadays are built around the progressive alignment. This popular MSA assembly algorithm is a straightforward agglomerative procedure. Sequences are first compared two by two in order to fill up a distance matrix, containing the percent identity. A clustering algorithm (UPGMA or NJ) is then applied onto this distance matrix to generate a rooted binary tree (guide tree). The agglomerative algorithm follows the tree topology thus defined and works its way from the leaf to the root, aligning two by two each sequence pair (or profile) associated with each encountered node. The procedure can be applied using any algorithm able to align two sequences or two alignments. In most packages, this algorithm is the Needleman and Wunsch (1970) or more recently the Viterbi algorithm (Durbin *et al*., 1998).

As simple as it may seem, the progressive alignment strategy affords many possible adjustments, the most notable ones being the tree computing algorithm, the sequence weighting method and the gap weighting scheme. In recent work (Wheeler and Kececioglu, 2007), the authors have shown that a proper tuning of these various components can take a standard method up to the level of the most accurate ones. ClustalW (Thompson *et al*., 1994) is often considered to be the archetype of progressive alignments. It is a bit paradoxical since its implementation of the progressive alignment significantly differs from the canonical one, in that it delays the incorporation of the most distantly related sequences until the second and unique iteration. This delaying procedure was incorporated in ClustalW in order to address the main drawback of the progressive alignment strategy: the greediness. When progressing from the leaves toward the root, a progressive aligner ignores most of the information contained in the dataset, especially at the early stage. Whenever mistakes are made on these initial alignments, they cannot be corrected and tend to propagate in the entire alignment, thus affecting the entire process. With a large number of sequences, the propagation and the resulting degradation can have extreme effects. This is a well known problem, usually addressed via an iterative strategy. In an iterative scheme, groups of sequences are realigned a certain number of time, using either random splits or splits suggested by the guide tree. The most sophisticated iterative strategies [incorporated in Muscle and PRRP (Gotoh, 1996)], involve two nested iterative loops, an inner one in which the alignment is optimized with the respect to a guide tree, and an outer one in which the current

MSA is used to re-estimate the guide tree. The procedure keeps going until both the alignment and the guide tree converge. It was recently shown that these iterations almost always improve the MSA accuracy (Wallace *et al*., 2005b), especially when they are deeply embedded within the assembly algorithm.

5 CONSISTENCY-BASED MSA METHODS

The greediness of progressive aligners limits their accuracy, and even when using sophisticated iteration schemes, it can be very hard to correct mistakes committed early in the alignment process. In theory, these mistakes could easily be avoided if all the information contained in the sequences was simultaneously used. Unfortunately, this goal is computationally unrealistic, a limitation that has prompted the development of consistency-based methods. In their vast majority, algorithms based on consistency are also greedy heuristics (with the exception of the maximum weight trace (MWT) problem formulation of Kececioglu (1993), but even so, they have been designed to incorporate a larger fraction of the available information at a reasonable computational cost. The use of consistency for improved alignment accuracy was originally described Gotoh (1990) and later refined by Vingron and Argos (1991). Kececioglu provided an exact solution to this problem, reformulated as a MWT problem. This exact approach is limited to small datasets but was further expanded by Morgenstern who proposed the first heuristic to solve this problem for large instances, thanks to the concept of overlapping weights (Morgenstern *et al*., 1996). While the notions developed in these four approaches are not totally identical, they have in common the idea of evaluating pair-wise alignments through the comparison of a third sequences (i.e. considering an intermediate sequence). In practice, Gotoh did not use consistency to construct alignments, but rather to evaluate them, and only considering three sequences. The consistency described by Vingron is very strict because it results from dot-matrices multiplications, therefore requiring strict triplet consistency in order to deliver an alignment. The overlapping weights described by Morgenstern also involve considering the support given by an intermediate sequence to a pair-wise alignment, but in this context, the goal is to help guiding the incorporation of pair-wise segments into the final MSA. While the overlapping weights bear a strong resemblance to the most commonly used definition of consistency, it is important to point out that Morgenstern also uses the term consistency but gives it a different meaning to describe the compatibility of a pair of matched segments within the rest of a partly defined multiple sequence alignments. The first combination of a consistency-based scoring scheme with the progressive alignment algorithm was later developed in the T-Coffee package (Notredame *et al*., 2000). The main feature of a consistency-based algorithm is its scoring scheme, largely inspired by the Dialign overlapping weights. Regular scoring schemes are based on a substitution matrix, used to reward identities and penalize mismatches. In a consistency-based algorithm, the reward for aligning two residues is estimated from a collection of pair-wise residue alignments named the library. Given the library, any pair of residues receives an alignment score equal to the number of time these two residues have been found aligned, either directly or indirectly through a third residue (Fig. 1). The indirect alignments are estimated by combining every possible pair of pair-wise alignments (i.e. $XY + YZ = X - Y - Z$). Each observation can be weighted with a score reflecting the expected accuracy of

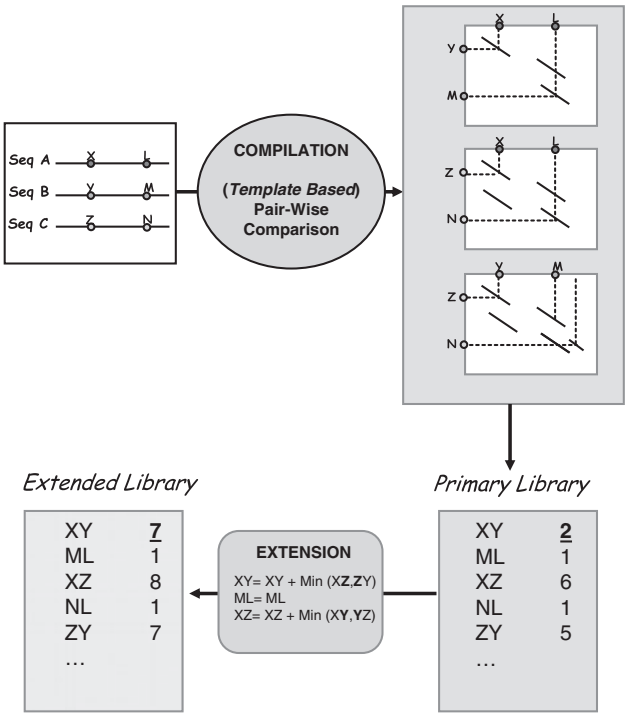

Fig. 1. Generic overview for the derivation of a consistency-based scoring scheme. The sequences are originally compared two by two using any suitable methods. The second box shows the projection of pair-wise comparisons. These projections may equally come from multiple sequence alignments, pair-wise comparison or any method able to generate such projections, including posterior decoding of an HMM. They may also come from a template-based comparison such as the one described in Figure 2. Pairs thus identified are incorporated in the primary library. These pairs are then associated with weights used during the extension. The figure shows the T-Coffee extension protocol. When using probabilistic consistency, the probabilities are treated as weights and triplet extension is made by multiplying the weights rather than taking the minimum. See Supplementary Material for color version of the figure.

the alignment on which the observation was made. In the original T-Coffee, the residue pairs contained in the library were generated using a global (ClustalW) and a local (Lalign) method applied on each pair of sequences. At the time, the T-Coffee protocol resulted in a significant improvement over all alternative methods. This protocol was later brought into a probabilistic framework with the package ProbCons. In ProbCons, the sequences are compared using a pair HMM with a bi-phasic gap penalty (i.e. a gap extension penalty higher for short gaps than long gaps). A posterior HMM decoding of this HMM is then used to identify the high-scoring pairs that are incorporated in the library, using their posterior probability as a weight. The library is then used to score the alignment with the T-Coffee triplet extension. Because it uses a library generated with a probabilistic method, this protocols is often referred to as 'probabilistic consistency' and has been incorporated in several packages, including SPEM (Zhou and Zhou, 2005), MUMMALS and PROMMAL (Pei and Grishin, 2006, 2007) as well as the latest version of T-Coffee (version 6.00 and higher). Interestingly, the improvement is usually considered to be a consequence of the probabilistic framework when in fact it seems to result mostly from

the use of a more appropriate gap penalty scheme at the pair-wise level. For instance, Table 1 shows the effect of applying a regular gap penalty scheme (monophasic) when compared with the bi-phasic gap penalty scheme that ProbCons uses by default. This improvement has also been observed when incorporating the bi-phasic scheme in T-Coffee. Consistency-based methods are typically 40 % accurate when considering the column score measured on the RV11 dataset. This makes consistency-based aligners ∼10 points more accurate than regular iterative progressive aligners like ClustalW, Kalign, Muscle or Mafft. This increased accuracy comes at a cost and consistency-based methods require on average *N* times more CPU time (*N* being the number of sequences) than a regular progressive aligner.

Aside from improved accuracy, an important aspect of consistency-based scheme is the conceptual separation it defines between the computation of the original alignments, merged into a library and the final transformation of this library into a multiple sequence alignment. This procedure made it straightforward to combine seemingly heterogeneous algorithms, such as ClustalW and Lalign in the original T-Coffee package, but it also opened the way towards a more generic combination of aligners. For instance, the latest version of T-Coffee (Version 6.00) is able to combine up to 15 different alignment methods, including pair-wise structural aligners, regular multiple sequence alignment methods and even RNA alignment methods such as Consan (Dowell and Eddy, 2006). From the start, the T-Coffee framework made it possible to turn any pair-wise method into a multiple alignment method, thus opening the way to two major developments undergone by multiple aligners these last years: meta alignment methods and template-based alignments.

6 META-METHODS AS AN ALTERNATIVE TO REGULAR MSA METHODS

The wealth of available methods and the lack of a globally accepted solution make it harder than ever for biologists to choose a specific method. This dilemma is real and has recently received some renewed attention with a high-impact report establishing the tight dependency of phylogenetic modeling on the chosen aligner. According to Wong and collaborators, phylogenetic trees may significantly vary depending on the methods used to compute the underlying alignment (Wong *et al*., 2008). In a similar way, several editions of the CASP (Battey *et al*., 2007) contest have revealed that a proper multiple alignment is an essential component of any successful structural modeling approach. A commonly advocated strategy is to use the method performing best on average, as estimated by benchmarking against structure-based reference datasets. It is a reasonable martingale, like betting on the horse with the best odds. One wins on average, but not always. Unsurprisingly, benchmarks also make it clear that no method outperforms all the others, and that it is almost impossible to predict with enough certainty which method will outperform all the others on a specific dataset. It is quite clear that the chosen method is irrelevant on datasets made of sufficiently similar sequences (>50% pair-wise identity). Yet, whenever remote homologs need to be considered, the accuracy drops and one would like to run all the available methods before selecting the best resulting alignment. This can be achieved when enough structural data is available (by selecting the alignment supporting the best structural superposition), or when functional information is at hand (by evaluating the alignment of similar features, such as catalytic residues). Unfortunately, experimental data is rarely available in sufficient amount, and when using several packages, one is usually left with a collection of alignments whose respective value is hard to assess in absolute terms. Meta-methods constitute an attempt to address this issue. So far, M-Coffee (Wallace *et al*., 2006) has been the only package explicitly engineered to be used as a meta-method, although in theory all consistency-based packages could follow suit. Given a multiple sequence dataset, M-Coffee computes alternative MSAs using any selected method. Each of the alignments thus produced is then turned into a primary library and merged to the main T-Coffee library. The resulting library is used to compute an MSA consistent with the original alignments. This final MSA may be considered as some sort of average of all the considered alignments. When combining eight of the most accurate and distinct MSA packages, M-Coffee produces alignments that are on average better than any of the individual methods. The improvement is not very high (1–2-point percent) but relatively consistent since the meta-method outperforms the best individual method (ProbCons) on ∼2/3 of the 2000 considered datasets (HOMSTRAD, Prefab and BaliBase) (Wallace *et al*., 2006). On a dataset like RV11, the improvement is much less marked (M-Coffee8 delivered alignments having an average accuracy of 37.5%) and one needs to restrict the combination to the four best non-template-based methods in order to obtain alignments with accuracy comparable to the best methods (Table 1). Yet, as desirable as it may be, the improved accuracy is not the main goal of M-Coffee and one may argue that rather than its accuracy, M-Coffee's main advantage is its ability to provide an estimate of local consistency between the final alignment and the combined MSAs. This measure (the CORE index; Notredame and Abergel, 2003) not only estimates the agreement among the various methods (Fig. 2) in a graphical way but it also gives precious indication on the local structural correctness (Lassmann and Sonnhammer, 2005a; Notredame and Abergel, 2003) and can therefore be considered as a good predictor of alignment accuracy. Previous benchmarking made on the original CORE measure suggest that a position with a consistency score of 50% or higher (i.e. 50% of the methods agreeing on a position) is 90% likely to be correct from a structural point of view. These results are consistent with those reported by Lassmann and Sonnhammer (2005a) who recently re-implemented this measure while basing it on libraries made of alternative multiple sequence alignments. Even though these predictions are only restricted to a subset of the alignment, they can be an invaluable asset whenever a modeling process is very sensitive to alignment accuracy. For instance, the CORE index is used by the CASPER server to guide molecular replacement (Claude *et al*., 2004). From a computational point of view, meta-methods are relatively efficient. Provided fast methods are used to generate the original alignment, the meta-alignment procedure of M-Coffee can use a sparse DP procedure that takes advantage of the strong agreement between the considered alignments. A recent re-implementation of M-Coffee in the SeqAn (Doering *et al*., 2008) alignment library shows that the multiple alignment step of M-Coffee is about twice faster than standard consistency-based aligners based on pair-wise alignments like ProbCons or Promals (Rausch *et al*., 2008).

Yet, all things considered, meta-methods only offer a marginal improvement over single methods, and they even suggest that the current state of the art aligners are reaching a limit that may hard

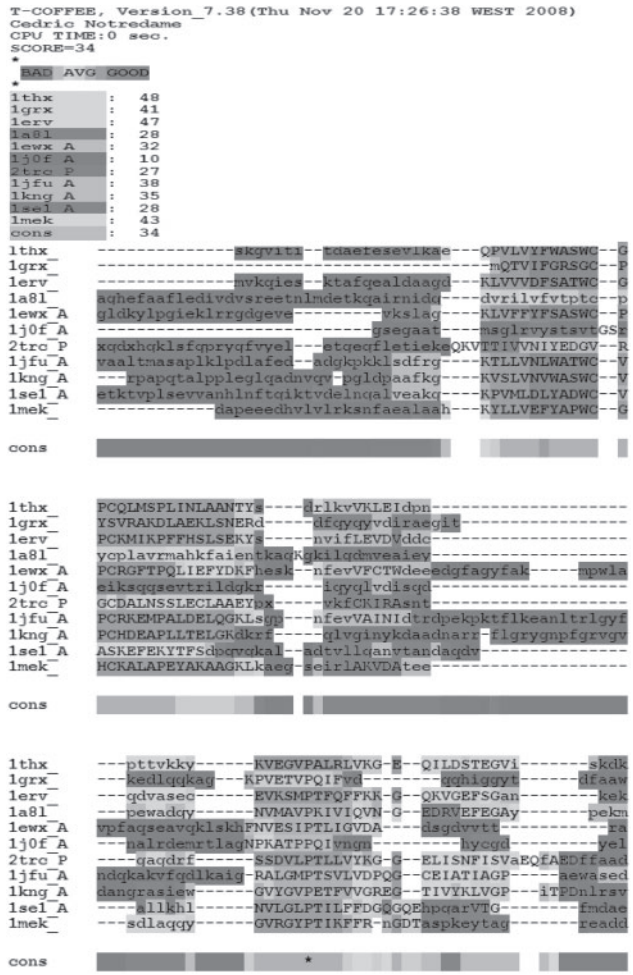

Fig. 2. Typical colored output of M-Coffee. This output was obtained on the RV11033 BaliBase dataset, made of 11 distantly related bacterial NADH dehydrogenases. The alignment was obtained by combining Muscle, T-Coffee, Kalign and Mafft with M-Coffee. Correctly aligned residues (correctly aligned with 50% of their column, as judged from the reference) are in upper case, non-correct ones are in lower case. In this colored output, each residue has a color that indicates the agreement of the four initial MSAs with respect to the alignment of that specific residue. Dark red indicates residues aligned in a similar fashion among all the individual MSAs, blue indicates a very low agreement. Dark yellow, orange and red residues can be considered to be reliably aligned. See Supplementary Material for color version of the figure.

to break without some novel development in the field of sequence alignment. While waiting for a method able to accurately align two remote homologs in an *ab initio* fashion (i.e. without using any other information than the sequences themselves), the best alternative is to use extra information, evolutionary, structural or functional. Template-based MSA methods have been design to precisely address this aspect of data integration.

7 TEMPLATE-BASED MSA METHODS

The word template-based alignment was originally coined by Taylor (Taylor, 1986) with reference to sequence/structure alignments. The

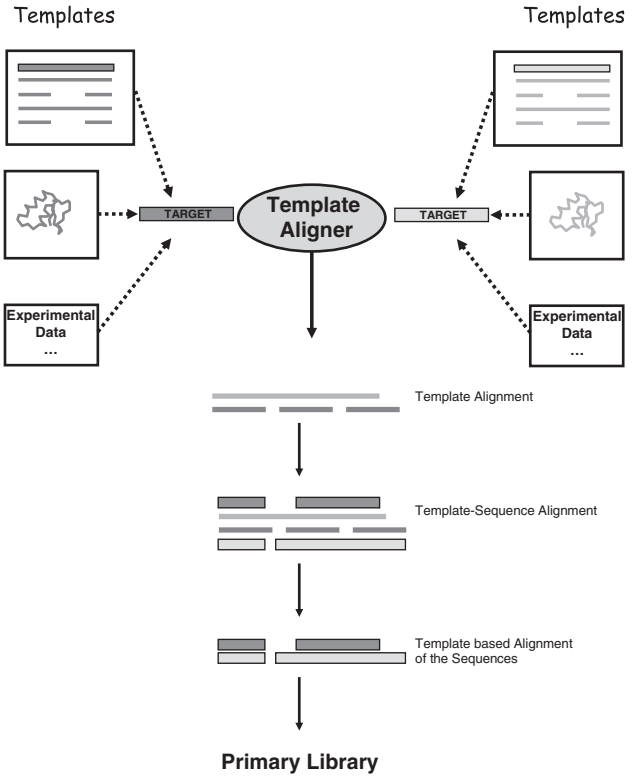

Fig. 3. Overview of template-based protocols. Templates are identified and mapped onto the target sequences. The figure shows three possible types of templates: homology extension, structure and functional annotation. The templates are then compared with a suitable method (profile aligner, structural aligner, etc.) and the resulting alignment (or comparison) is mapped onto the final alignment of the original target sequences. The residue pairs thus identified are then incorporated in the primary library. See Supplementary Material for color version of the figure.

notion was later extended within the T-Coffee package in a series of publications dedicated to protein and RNA alignments (Armougom *et al*., 2006b; Notredame and Higgins, 1996; O'Sullivan *et al*., 2004; Wilm *et al*., 2008). Template-based alignment refers to the notion of enriching a sequence with the information contained in a template (Fig. 3). The template can either be a 3D structure, a profile or a prediction of any kind. Once the template is precisely mapped onto the sequence, its information content can be used to guide the sequence alignment in a sequence independent fashion. Depending on the nature of the template one refers to its usage as 'structural extension' or 'homology extension' (sequence profile).

'Structural extension' is the most straightforward protocol. It takes advantage of the increasing number of sequences with an experimentally characterized homolog in the PDB database. Given two sequences with a homolog in PDB, one can accurately superpose the PDB structures (templates) and map the resulting alignment onto the original sequences. Provided the sequence/template alignment is unambiguous, this protocol yields an alignment of the original sequences having all the properties of a structure-based sequence alignment. This approach only defines pair-wise alignments, but the alignment thus compiled can be integrated into a T-Coffee library and turned into a consistency-based multiple sequence alignment (Figs 1 and 3). Structural extension was initially implemented

in 3D Coffee (O'Sullivan *et al*., 2004). EXPRESSO, a special mode of 3D Coffee was then designed so that templates could be automatically selected via a BLAST against the PDB database. This protocol has recently been re-implemented in the PROMAL-3D (Pei *et al*., 2008) package. Structural extension is not limited to proteins, and recently several approaches have been described using RNA secondary structures as templates, these include T-Lara (Bauer *et al*., 2005), MARNA (Siebert and Backofen, 2005) and R-Coffee (Wilm *et al*., 2008). In all these packages, sequences are associated with a predicted structural template (RNA secondary structure). The templates are then used by *ad hoc* algorithms to accurately align the sequences while taking into account the predicted structures (templates). The resulting pair-wise alignments are combined into a regular T-Coffee library and fed to T-Coffee.

'Homology extension' works along the same principle as structural extension but uses profiles rather than structures. In practice, each sequence is replaced with a profile containing homologs. The profiles could be built using any available techniques although fast methods like PSI-BLAST have been favored. The first homology extension protocol was described by Heringa and implemented in the PRALINE package (Simossis and Heringa, 2005). PROMALS was described shortly afterwards (Pei and Grishin, 2007). PROMALS is a consistency-based aligner, using libraries generated with the ProbCons pair-HMM posterior decoding strategy. PROMALS also uses secondary structure predictions in order to increase the alignment accuracy, although this extra information seems to only have a limited effect on the alignment accuracy. In Praline and PROMALS, sequences are associated with a PSI-BLAST profile. A similar mode is also available in T-Coffee (Version 6.00+, mode = psicoffee) based on BLAST profiles (Table 1).

The use of structural and homology-extended templates results in increased accuracy in all cases. For instance, the combination of RNAplfold (Bernhart *et al*., 2006) predicted secondary structures made R-Coffee more accurate at aligning RNA sequences than any of the alternative regular aligners, with a 4-point net improvement as estimated on BraliBase (Wilm *et al*., 2008). The improvements resulting from homology extension on proteins are even more significant. On Prefab, the authors of PROMALS reported nine points of improvement over the next best method (ProbCons). A similar usage of PROMALS or PSI-Coffee on category RV11 (distant homologs) of BaliBase resulted in >10 points of improvement over the next best regular non-template-based aligner (Table 1). Of course, the most accurate alignments are obtained when using structural extension. In a recent work, Grishin and collaborators reported an extensive validation using a combination of structure and homology extension (Pei *et al*., 2008). Their results suggest that template-based alignments achieve the best results when using structural extension. They also indicate that the choice of the structural aligner can make a difference, with DALI-Lite possibly more accurate than SAP. Given the same structural extension protocol, the authors report similar results between 3D Coffee and PROMALS-3D, suggesting that the structural aligner is the most important component of the protocol. The improvement is very significant, and on Prefab for instance, the combined use of DaliLite with homology extension resulted in nearly 30 points improvement over alternative non-template-based protocols. Results in Table 1 confirm these claims and suggest that the use of structural extension is the best way to obtain highly accurate alignments.

This very high accuracy, obtained when using structural information is, however, to be interpreted with some caution. On the one hand, these high figures suggest a broad agreement between PROMALS-3D or 3D Coffee alignments with the references. On the other hand, one should not forget that these methods use 3D information. As such, they are not any different from the methods used to derive the reference benchmarks themselves. It therefore means that PROAMLS-3D or 3D Coffee/Expresso alignments may be seen as new reference datasets, generated with a different structural alignment protocol. Whether these are more or less accurate than the benchmarks themselves is open to interpretations, as it amounts to comparing alternative multiple structure-based sequence alignments.

7.1 New issues with the validation of template-based methods

As reported by Kolodny *et al*. (2005), the task of comparing alternative structure-based alignments is complex. In order to address it, authors have recently started using alignment free evaluation methods. These methods consider the target alignment as a list of structurally equivalent residues and estimate how good would be the resulting structural superposition. These measures are either based on the RMSD (root mean squared deviation: average squared distance between homologous alpha carbons) or the dRMSD (distance RMSD: average square difference of distances between equivalent pairs of amino acids) like the DALI score (Holm and Sander, 1995), APDB (O'Sullivan *et al*., 2003) or the iRMSD (Armougom *et al*., 2006a). So far, three extensive studies (Armougom *et al*., 2006a; O'Sullivan *et al*., 2003; Pei *et al*., 2008) have suggested that the results obtained with these alignment-free benchmarking methods are in broad agreement with those reported when using regular benchmarks. The main drawback of these alignment-free evaluation methods is their reliance on distance measures strongly correlated with the methodology used by some structural aligners (Dali in particular) thus raising the question whether they might be biased toward this particular structural aligner. A simpler and not yet widely used alternative would be to evaluate the modeling potential of the alignments, by measuring the accuracy of structural predictions based upon it. This could probably achieved by recycling some components of the CASP evaluation pipelines.

8 ALIGNMENT OF VERY LARGE DATASETS

Accuracy has been a traditional limitation of multiple sequence alignments for the last 20 years, and it is no surprise that this issue has been the most actively addressed, if only because inaccurate alignments are simply useless. The other interesting development has been the increase of the number of sequences. Traditionally, the length of the sequences (L) was greater than the number of sequences (N), and most methods were tuned so that they could deal with any value of L, assuming N would not be a problem. This is especially true of consistency-based methods that are cubic in complexity with N, but only quadratic with L. With $N << L$, the extra-cost incurred by consistency remains manageable, but things degrade rapidly when N becomes big. Yet, it is now clear that L is bounded, at most by the average length of a genome. N, on the other hand, has no foreseeable limit and could reflect the total number

of species or the total number of individuals (past and present) in a population or even the total number of haplotypes in a system. Dealing with large values of L should therefore be considered a prime goal. In the context of a progressive algorithm, the first easy step is to speed up the guide tree estimation, for instance using a ktup-based method, as most packages currently do (-quicktree option in ClustalW). The second step is to use an efficient tree reconstruction algorithm. The default UPGMA and NJ algorithms are cubic with the number of sequences, but these algorithms can be adapted in order to become quadratic, as is the case with the current ClustalW implementation. Even so, quadratic algorithms will not be efficient enough when dealing with very large datasets and more efficient data compression methods (such as those used to decrease redundancy in databases) will probably need to be used in the close future (Blackshields *et al*., 2008). The next step for decreasing CPU requirements is to use an efficient DP strategy. This is the strategy used by MAFFT that relies on a very efficient DP. Consistency-based methods have a disadvantage because of the N-cubic requirement of consistency. Yet, the protocol is relatively flexible and heuristics can probably be designed to estimate the original library more efficiently. For instance, PCMA (Pei *et al*., 2003) starts by identifying sub-group of sequences closely related enough to be pre-aligned. SeqAn (Rausch *et al*., 2008) takes advantage of the sparse matrix defined by the extended library and only does the minimum required computation to guarantee optimality. SeqAn also makes an attempt to treat the sequences as a chain of segments rather than a chain of residues thus considerably reducing the CPU requirements for closely related sequences. The SeqAn library has been designed to be linked with any of the consistency-based aligners. Even so, the complexity of most consistency-based aligners remains too high to deal with the very large datasets that are expected to come. Currently, the most promising approaches are those implemented in Muscle and Mafft. Yet, it should be stressed that so far no dataset has been designed to evaluate the accuracy of very large number of sequences and it remains unclear how these methods scale and whether accuracy figures established on relatively small datasets can be safely extended to larger ones. It is therefore urgent to establish reference datasets suitable for the validation of large scale aligners (1000 sequences and more). Phylogeny being one of the main applications of large scale alignments, it will also be worth evaluating the phylogenetic potential of these large scale methods. Doing so is far from trivial as it connects with the delicate issue of establishing reference tree collections. More generally, it addresses the problem of predicting accurate trees from multiple sequence alignments.

9 PHYLOGENETIC RELEVANCE OF MULTIPLE SEQUENCE ALIGNMENTS

The pace of accumulation of new entire bacterial genomes (and to a lesser extend eukaryotic genomes) can only be compared with the discovery of new species along the nineteenth century. Never have we had so much molecular data at hand to reconstruct the natural history of life, a real challenge for intelligent design supporters. Multiple sequence alignments constitute the ideal compost on which to grow these trees, and although there have been a few reports of alignment free tree reconstruction methods (Ferragina *et al*., 2007), the difficulty of aligning distantly related sequences probably means that unless a breakthrough happens in the field of sequence alignments and guarantees error free pair-wise alignments, MSAs will remain the starting point for most phylogenetic analysis. An interesting paradox of the whole MSA field is that although most methods are defined within some sort of phylogenetic framework (progressive alignment), they are only evaluated for their capacity of producing structurally correct MSAs. As a consequence, we do not really know how MSA methods fare with respect to phylogenetic reconstruction and, assuming the current structural benchmarks reflect well enough the evolutionary relation among proteins, we do not really know if this analysis can be safely extrapolated to ncRNAs. Recent work suggest (Katoh and Toh, 2008) that the accuracy ranking of the best packages is roughly the same when benchmarking on RNA sequences (BraliBase) or protein sequences, but little is known about the accurate reconstruction of RNA-based phylogenetic tree. This is a paradoxical situation when considering that most trees of life are derived from a multiple sequence alignment of ribosomal RNA sequences.

This passed year, two high-impact publications have made an attempt to raise the attention of the community on the issue of phylogenetic reconstruction (Loytynoja and Goldman, 2008; Wong *et al*., 2008). The work by Wong shows that phylogenetic reconstruction can be very sensitive to the MSA method used to deliver the alignment. The authors stopped short of proposing a way for selecting the best phylogenetic trees, but they make it clear that various methods can lead to different models, a new concept in a field where MSAs had always been considered to be data rather than models. It is a context where meta-methods could certainly provide an element of answer, mostly by helping selecting the sites on the basis of their expected accuracy, using the CORE index or any related method. In this context, the main advantage of the CORE index is to provide a filter independent from sequence conservation, as opposed to other accuracy predictors. An MSA region can have a low level of conservation but a high CORE index, provided all the pair-wise alignments are in agreement with respect to the considered position. Regions where conservation is low and consistency high may be considered prime targets for phylogenetic reconstruction. The PRANK+F (Loytynoja and Goldman, 2008) algorithm was described shortly afterward and also addresses the issue of accurate phylogenetic reconstruction seen from an MSA perspective. PRANK+F is a novel attempt to model the gap insertion required by the alignment process in a phylogenetically meaningful way. This new approach opens up the possibility of incorporating the indel signal in the reconstruction of evolutionary scenarios, but it also raises an equally important question: given that alternative aligners lead to different trees, and given that the signal contained in the alignments can be used in many different ways, how are we going to evaluate the phylogenetic potential of multiple sequence alignment methods? Building reference datasets is a very difficult task in phylogeny where an objective, independent source of information for establishing the correct history of a set of sequences is usually lacking. Fossil records provide little help when it comes to selecting true orthologous sequences. Given a sequence dataset, it is therefore very hard, and may be impossible to establish a correct reference tree. So far, the validation of tree reconstruction methods has therefore focused on the method's ability to optimize a mathematical model (Guindon and Gascuel, 2003). Even when this optimization is highly successful, the only guarantee is the mathematical correctness of the final tree with no clear guarantee on its biological relevance, except that provided by expert diagnostic

of the tree (i.e. the observation that related species are grouped by the tree in a biologically meaningful way). This situation is very similar to that encountered with MSA computation where one has on the one hand the mathematical correctness of a method, estimated by its capacity to optimize a given objective function (sums-of-pairs, viterbi, etc.) and on the other hand, the biological accuracy, estimated by comparison with a reference alignment. In the context of MSA analysis, the use of structure made it clear that there could be a significant discrepancy between mathematical correctness and biological accuracy. Unfortunately, the equivalent of structural information is not available in phylogeny, and most current strategies, including Prank+F are validated on simulated data. The simplest approaches simulate both the data and the trees using generators like ROSE (Stoye *et al*., 1997). As pointed out earlier, the results obtained on simulated data differ significantly from those measured on empirical data, and for instance, PRANK outperforms all alternative packages on phylogenetic simulated data, but performs poorly when it comes to reconstructing structural alignments. Assuming the relevance of results established on the simulated data, this suggests there could major differences between phylogenetically accurate alignments and structurally accurate ones, an hypothesis that remains to be further tested and confirmed.

10 CONCLUSIONS

In this review, we have tried to give an overview of some of the newest aspects in the multiple sequence alignment field. We have also tried to describe some of the challenges lying ahead now that we have entered what will probably be known as the genomic era of biology. This review is not meant to be exhaustive and should be seen as a partly biased point of view on the direction in which we feel multiple sequence comparisons may develop over the next years. Multiple comparison is the essence of biology and provides us with a very powerful observation tool, especially when one lacks a precise idea on the nature of forces that shaped the observed diversity. One may argue that a multiple comparison of species formed the basis of Darwin's work. Likewise, it is certainly not a coincidence if the publication describing ClustalW has become one of the most widely cited paper in Biology (35 000 citations to this day). As long as new data will accumulate, there will be an increasing need for informative multiple comparison methods. In this review, we have outlined four major development directions: (i) the use of template-based methods that make it possible to combine heterogenous experimental data; (ii) meta-methods and the systematic use of consistency-based methods that make it possible to combine heterogeneous data but also to combine very different methods within a unified framework; (iii) the development of large-scale methods, necessary in a context where information is growing by the day; (iv) phylogentic reconstruction. Accurate phylogenetic reconstruction is probably one of the most pressing issues, since such modeling is bound to play an increasing role in our data analysis routines. Developing MSA methods that lend themselves to accurate phylogeny reconstruction should therefore be considered a prime goal. The difference of behavior between methods like Prank+F and Promals3D or 3D Coffee/Expresso underlines dramatically the possible difference that may exist between structural and evolutionary analysis. In our opinion, the recent progress on these two aspects have only started touching the problem, and one may expect that a sizeable amount of forthcoming work will be dealing with understanding whether there really is a discrepancy between structurally and phylogenetically accurate alignments and whether this discrepancy, if verified, can be turned into a usable signal for making sense of biological information.

The problem of multiple genome alignments and in general, the problem of multiply aligning non-transcribed sequences, has voluntarily been excluded from the scope of this review. There is currently a very clear gap between the multiple alignment of genomic sequences, and the multiple alignment of transcribed sequences. A good illustration of this separation is the relatively low overlap of authorship across these two neighbor fields of research. It is probably a safe bet that over the coming years, this gap will gradually close, thanks to the development of a continuous algorithmic framework [like SeqAn or Pecan (Paten *et al*. 2009)] bridging the gap. It is also an easy guess that the accurate alignment of non-coding DNA will become increasingly prominent field of research.

Conflict of Interest: none declared.

REFERENCES

Abhiman,S. *et al*. (2006) Prediction of function divergence in protein families using the substitution rate variation parameter alpha. *Mol. Biol. Evol.*, **23**, 1406–1413.

Armougom,F. *et al.* (2006a) The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, **22**, e35–e39.

Armougom,F. *et al.* (2006b) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.*, **34**, W604–W608.

Battey,J.N. *et al.* (2007) Automated server predictions in CASP7. *Proteins*, **69** (Suppl. 8), 68–82.

Bauer,M. *et al.* (2005) Multiple structural RNA alignment with Lagrangian relaxation. *Lect. Notes Comput. Sci.*. 303–314.

Bernhart,S.H. *et al.* (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **1**, 614–615.

Birney,E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Blackshields,G. *et al.* (2006) Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biol.*, **6**, 321–339.

Blackshields,G. *et al.* (2008) Fast embedding methods for clustering tens of thousands of sequences. *Comput. Biol. Chem.*, **32**, 282–286.

Chandonia,J.M. *et al.* (2006) Target selection and deselection at the Berkeley Structural Genomics Center. *Proteins*, **62**, 356–370.

Claude,J.B. *et al.* (2004) CaspR: a web server for automated molecular replacement using homology modelling. *Nucleic Acids Res.*, **32**, W606–W609.

Do,C.B. *et al.* (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15**, 330–340.

Dowell,R.D. and Eddy,S.R. (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, **7**, 400–417.

Doering,A. *et al.* (2008) SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Eddy,S.R. (1995) Multiple alignment using hidden Markov models. In Rawlings,C. *et al.* (eds) *Third International Converence on Intelligent Systems for Molecular Biology (ISMB)*. AAAI Press, Cambridge, England/Menlo Park, CA, pp. 114–120.

Edgar,R.C. (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edgar,R.C. (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Edgar,R.C. and Batzoglou,S. (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**, 368–373.

Fabian,M.A. *et al.* (2005) A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.*, **23**, 329–336.

Ferragina,P. *et al.* (2007) Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. *BMC Bioinformatics*, **8**, 252.

Frazer,K.A. *et al*. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Gondro,C. and Kinghorn,B.P. (2007) A simple genetic algorithm for multiple sequence alignment. *Genet. Mol. Res.*, **6**, 964–982.

Gotoh,O. (1990) Consistency of optimal sequence alignments. *Bull. Math. Biol.*, **52**, 509–525.

Gotoh,O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinements as assessed by reference to structural alignments. *J. Mol. Biol.*, **264**, 823–838.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Hogeweg,P. and Hesper,B. (1984) The alignment of sets of sequences and the construction of phylogenetic trees. An integrated method. *J. Mol. Evol.*, **20**, 175–186.

Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.

Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.*, **9**, 286–298.

Kececioglu,J.C. (1993) The maximum weight trace problem in multiple sequence alignment. In *Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching, CPM '93*. Springer, London, UK, pp. 106–119

Kolodny,R. *et al.* (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.

Lassmann,T. and Sonnhammer,E.L. (2002) Quality assessment of multiple alignment programs. *FEBS Lett.*, **18**, 126–130.

Lassmann,T. and Sonnhammer,E.L. (2005a) Automatic assessment of alignment quality. *Nucleic Acids Res.*, **33**, 7120–7128.

Lassmann,T. and Sonnhammer,E.L. (2005b) Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.

Lee,C. *et al*. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.

Loytynoja,A. and Goldman,N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635.

McClure,M.A. *et al.* (1994) Comparative analysis of multiple protein-sequence alignmnent methods. *Mol. Biol. Evol.*, **11**, 571–592.

Morgenstern,B. *et al.* (1996) Multiple DNA and Protein sequence based on segment-to-segment comparison. *Proc. Natl Acad. Sci. USA*, **93**, 12098–12103.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Notredame,C. (2007) Recent evolutions of multiple sequence alignment. *PLoS Comput. Biol.*, **3**, e123.

Notredame,C. and Abergel,C. (2003) Using multiple alignment methods to assess the quality of genomic data analysis. In Andrade,M. (ed.) *Bioinformatics and Genomes: Current Perspectives*. Horizon Scientific Press, Wymondham, UK, pp. 30–50.

Notredame,C. and Higgins,D.G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.*, **24**, 1515–1524.

Notredame,C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

O'Sullivan,O. *et al.* (2003) APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics*, **19** (Suppl. 1), i215–i221.

O'Sullivan,O. *et al.* (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.

Pascarella,S. *et al.* (1996) A databank (3D-ali) collecting related protein sequences and structures. *Protein Eng.*, **9,** 249–251.

Paten,B. *et al*. (2009) Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. *Bioinformatics*, **25**, 295–301.

Pei,J. (2008) Multiple protein sequence alignment. *Curr. Opin. Struct. Biol.*, **18**, 382–386.

Pei,J. and Grishin,N.V. (2006) MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. *Nucleic Acids Res.*, **34**, 4364–4374.

Pei,J. and Grishin,N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**, 802–808.

Pei,J. *et al.* (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.

Pei,J. *et al.* (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.

Raghava,G.P. *et al.* (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.

Rausch,T. *et al.* (2008) Segment-based multiple sequence alignment. *Bioinformatics*, **24**, i187–i192.

Reinert,K. *et al.* (1997) A branch-and-cut algorithm for multiple sequence alignment. In *Recomb97*, ACM Press, Santa Fe, NM, pp. 241–249.

Riaz,T. *et al*. (2005) A tabu search algorithm for post-processing multiple sequence alignment. *J. Bioinform. Comput. Biol.*, **3**, 145–156.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Siebert,S. and Backofen,R. (2005) MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, **21**, 3352–3359.

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

Siva,N. (2008) 1000 Genomes project. *Nat. Biotechnol.*, **26**, 256.

Stebbings,L.A. and Mizuguchi,K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res.*, **32**, D203–D207.

Stoye,J. *et al*. (1997) Generating benchmarks for multiple sequence alignments and phylogenetic reconstructions. *Ismb.*, **5**, 303–306.

Subramanian,A.R. *et al*. (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.

Subramanian,A.R. *et al*. (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol. Biol.*, **3**, 6.

Taylor,W.R. (1986) Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.*, **188**, 233–258.

Thompson,J. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4690.

Thompson,J.D. *et al*. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.

Thompson,J.D. *et al*. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.

Van Walle,I. *et al*. (2005) SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267–1268.

Vingron,M. and Argos,P. (1991) Motif recognition and alignment for many sequences by comparison of dot-matrices. *J. Mol. Biol.*, **218**, 33201343.

Wallace,I.M. *et al*. (2005a) Multiple sequence alignments. *Curr. Opin. Struct. Biol.*, **15**, 261–266.

Wallace,I.M. *et al*. (2005b) Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics*, **21**, 1408–1414.

Wallace,I.M. *et al*. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.

Wang,L. and Jiang,T. (1994) On the complexity of multiple sequence alignment. *J. comput. Biol.*, **1**, 337–348.

Wheeler,T.J. and Kececioglu,J.D. (2007) Multiple alignment by aligning alignments. *Bioinformatics*, **23**, i559–i568.

Wilm,A. *et al*. (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.*, **1**, 19.

Wilm,A. *et al*. (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, **36**, e52.

Wong,K.M. *et al*. (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.

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