Influence of alignment uncertainty on homology and phylogenetic modeling

Jia-Ming Chang

TESI DOCTORAL UPF / 2012

DIRECTOR DE LA TESI

Dr. Cedric Notredame

Bioinformatics and Genomics

Centre for Genomic Regulation

DEPARTAMENT EXPERIMENTAL AND HEALTH SCIENCES

UNIVERSITAT POMPEU FABRA

CRG Centre for Genomic Regulation

Obra Social Fundación "la Caixa"

Acknowledgement

This thesis is not only my work, but it also involves the contribution of many people. Without them, everything would not have been possible. I would like to acknowledge them according to their geography locations:

- BARCELONA: I would like to thank my supervisor Cedric Notredame for all the support I received over the last five years. Every discussion pushes me further in science discovery. It is impossible to come out those fruitful results without those interactions with him. I would also thank Cedrik Magis for giving a lot of useful suggestions for the first draft as well as Jean-Francois Taly, Sascha Meiers, Ionas Erb, Emily Chia-Yu Su and Heng-Chang Chen for the proof reading and Irantzu Anzar for translating “Abstract” into Spanish. Further thanks go to my thesis committee, Toni Gabaldon, Mar Alba and Matthieu Louis, who accompanied me on PhD way. Certainly I also thank my previous and current colleagues with whom I have been working, Romina Garrido Enamorado for always organizing everything so perfectly and our system administrators for computational support.
- CAMBRIDGE: I would like to thank Javier Herrero for providing an opportunity to be a trainee in his group at European Bioinformatics Institute for three months (May~July 2011). I was looking forward to work during the stay. I also thank the landlord, Muriel Ambourhouet, for a so lovely flat.

- 台灣: 感謝李文雄老師跟蕭孟昕博士，讓我有機會回中研院，進行兩個月的訪問 (二，三月, 2013)。許聞廉老師和宋定懿老師，每次回台灣的拜訪， 我總有種心得 - 要努力做研究。盧錦隆老師如平輩般的討論，是我研究的榜樣。我的生物資訊啓蒙 - 唐傳義老師，雖貴為校長，忙碌之時，仍不忘給予提攜的機會。不知道這是離家的第幾個年頭，自私的我，選擇自己的夢想，而你們只能透過網路視訊，遙遠地得知我的訊息，你們給予的自由，照就這一切的可能，謝謝爸爸和媽媽 - 張辰雄和李淑貴。老弟每個節日，你得回家陪爸媽，辛苦了!很開心看到凱凱繼承我們家的好傳統。讀博士班彷彿是走在看不見盡頭的隧道，而妳是另一端的一盞燈，沒有這五年的陪伴，是不可能完成這旅程地，謝謝妳 - 千瑩。

- EARTH: Finally, I would like to thank my friends for supporting me the whole time. When I face the nice Mediterranean Sea, I know one day I will leave because they are not beside me.

Abstract

Most evolutionary analyses are based upon pre-estimated multiple sequence alignment models. From a computational point of view, it is too complex to estimate a correct alignment, as it is to derive a correct tree from that alignment. Several works have recently reported on the influence of alignment on downstream analysis, and on the uncertainty inherent to their estimation. Chapter 1 develops the notion of alignment uncertainty as either inherent to the data (internal) or resulting from methodological biases (external). Chapter 2 presents two contributions of mine for the improvement of MSA methods through the use of homology extension (TM-Coffee) and thanks to an improved word-matching algorithm (SymAlign). In Chapter 3, I show how alignment uncertainty can be used to improve the trustworthiness of phylogenetic analysis. Chapter 4 shows how a similar improvement can be obtained through a simple adaptation of the T-Coffee transitive score, thus allowing downstream analysis to take into account internal alignment uncertainty. The final chapter contained a discussion of our current results and possible future work.

Resumen

La mayoría de los análisis evolutivos están basados en modelos establecidos de alineamiento de secuencia múltiple. Desde un punto de vista computacional, es igual de complejo la estimación de un alineamiento correcto, como la obtención de un árbol correcto a partir del alineamiento. Recientemente varios trabajos han informado sobre la influencia del alineamiento en los análisis posteriores, y en la incertidumbre inherente a su estimación. El Capítulo 1 desarrolla el concepto de incertidumbre de alineación, tanto inherente a los datos (internos), como resultante de los sesgos metodológicos (externo). El Capítulo 2 presenta dos contribuciones mías para la mejora de los métodos de MSA a través del uso de la extensión de homología (TM-Coffee) y gracias a un algoritmo de coincidencia de palabra mejorado (SymAlign). En el capítulo 3, se muestra cómo la incertidumbre de alineación puede ser utilizada para mejorar la confiabilidad del análisis filogenético. El capítulo 4 nos muestra como se puede obtener una mejora similar por medio de una simple adaptación de la puntuación transitiva del T-Coffee, lo cual permite un análisis posterior para tener en cuenta la incertidumbre de alineación interna. El último capítulo contiene un análisis de los resultados actuales y los posibles futuros trabajos.

Preface

Molecular biology research has changed a lot in the last 10 years. Next Generation Sequencing technique now makes it practical to sequence whole genome in reasonable time and cost. New areas of research have been opened, prompting the revision of many concepts taken for granting in our understanding of the molecular basis of life. For example, transcriptome analysis from ENCODE project discovers that non-coding RNAs occupy the large proportion of genome. These methodologies produce huge amounts of data that need to be analyzed, thus turning accurate computational approaches into an essential limiting step of this new big data biology. When I started my master degree in Taiwan, few people had an idea about bioinformatics and its utility. Today, many universities have provided the graduate courses of bioinformatics. In this thesis I contribute to the improvement of the two most popular methods in sequence analysis, the multiple sequence alignment and the phylogenetic reconstruction, which due to its difficulty, still provides challenges.

Contents

List of Figures

List of Tables

1. INTRODUCTION

Multiple sequence alignment (MSA) is probably the most widely used bioinformatics method in biology and is subsequently used for many more applications, such as structure prediction, motifs/patterns recognition, SNP analysis and phylogeny inference (Thompson and Poch 2006). Over the last 30 years, more than 100 papers relevant to MSA have been published (Kemena and Notredame 2009) and so far none has been proven to perform better than others in all different situations. One of the main focuses in the field over the last years has been to identify and propose the most accurate method. However, the focus on alignment uncertainty has intensified over the last few years, with several recent high impact papers have reporting on uncertainties in MSAs computation (Wong et al. 2008; Markova-Raina and Petrov 2011; Jordan and Goldman 2012) as being yet one more confounding factor when doing phylogenetic and evolutionary analyses due. To address this issue, many new confidence measures have been proposed. In this chapter, we start by reviewing the difficulty of defining a correct MSA such that alignment methods based on different criteria may give various results (section 1.1). When modeling a given dataset, variations arising across methods may be considered a manifestation of alignment uncertainty, especially in the absence of criteria able to tell the relative merits of the considered alignments. We then discuss the causes of such uncertainty. The next section is a review of available methods for the automated assessment of MSA reliability and accuracy (section 1.2). The last section explores the problem of post-processing of MSA including summarizing sample alignments and trimming the problematic region of MSA (section 1.3). We hope that this thesis can draw the attention of the wider molecular evolution research community to the importance of alignment uncertainty.

1.1 Computing and evaluating multiple sequence alignments

The simultaneous comparison of evolutionary related or functionally biological sequences usually starts with a sequence alignment. Such alignments may be considered as a summary of the relationships existing among the considered sequences. The variations within the alignment itself are a direct reflection of the exploratory process followed by natural evolution through a complex combination of mutations and fixation of novel alleles. Given a set of sequences, an MSA has not absolute definition, and its correctness will depend on the nature of the characters one wishes to model, these may be structural similarity, evolutionary similarity, and sequence similarity (Table 1-1). In the next section we review these various characters and discuss the potential consequences of divergent evaluations. We then report on the suitability of new kinds of experimental evidences, based on NGS data, for the evaluation of nucleotide alignments.

a) Various criteria for MSA evaluation

The easiest criterion one can use to estimate the biological accuracy of an alignment is sequence similarity. Under this scheme, the best possible MSA (optimal) is defined as the one yielding the highest level of similarity within aligned columns. Similarity can be defined as identity, or as weighted identity, using popular substitution matrices like BLOSUM62 or PAM250. The total estimation of the MSA cost is obtained by summing up the substitution cost of every pair of aligned residue. For this reason, the objective function is known as the Sum-of-Pairs (SoP). In many of its implementations, the SoP objective function is associated with a weighting scheme meant to reflect the information content of each considered sequence (Altschul et al. 1989; Sibbald and Argos 1990; Thompson et al. 1994). The goal of such weights is to insure that highly similar sequences do not end up dominating the MSA computation, and eventually prevent the correct alignment of more remote homologues.

When applying the SoP scoring scheme, an important assumption is that the score of each column of the alignment is independent from those of the rest. This scheme is very well suited to pairwise sequence alignment, where an optimal alignment can be computed using a dynamic programming approach (Bellman 1953; Bellman 1954). Interestingly, the dynamic programming approach used in biology and commonly referred to as Needlman and Wunsch was independently re-invented by these same authors (Needleman and Wunsch 1970) and only later recognized to be related to the dynamic programming approach of Bellman. Although the optimal pairwise alignment can be found through dynamic programming, possible paths with identical SoP scores needed tiebreak are the source of alignment uncertainty (section 1.2.a). Optimization becomes NP-hard when multiple sequence case is considered (Wang and Jiang 1994; Bonizzoni and Vedova 2001; Just 2001; Elias 2006). This makes it impossible to use brute force multiple dynamic programming to align more than three sequences. One can, however, use a bounded approach to define a smaller multi-dimensional envelop allowing the alignment of a larger number of sequences (up to 20) using the Carillo and Lipman algorithm implemented in "MSA" (Lipman et al. 1989). In practice, few MSA software produce a provably optimal alignment and most algorithms work by either applying heuristics to solve the original NP-complete optimization problem using delivering only an approximate solution or by replacing the SoP objective function with another objective whose optimization is tractable. The heuristics approach includes progressive aligners and consistency aligners, representing many methodological innovations for sequence alignment.

More sophisticated objective functions can be defined using structural information. In that case, MSAs are evaluated for their capacity to align equivalent residues as estimated from the global comparison of their respective folds. This approach has been used to establish several collections of structure based reference alignments, routinely used for the systematic benchmark of novel protein alignment methods (BAliBASE, PREFAB, HOMSTRAD, and SABmark (Edgar 2004a; Stebbings and Mizuguchi 2004; Thompson et al. 2005b; Van Walle et al. 2005b)). More recently, a similar approach has been applied to the RNA analysis (Kemena et al. 2013). The main rationale for using structure based alignments to evaluate sequence based procedure stems from the observation that structural folds appear to be evolutionary more resilient than their underlying sequences (Chothia and Lesk 1986). In order to best use this property, O'Sullivan O. *et al.* developed 3D-Coffee, a method for combining protein sequences and structures in order to generate high-quality MSAs. They found a linear correlation between MSA accuracy and the proportion of sequences with structure information (O'Sullivan et al. 2004).

The third criterion is functional similarity. The biological activity of a protein typically depends on the presence of a small number of functional residues. These residues are often remarkably conserved, as a consequence of the purifying selection under which they evolve. Functional similarity can also result from convergent evolution, as observed in the case of serine proteases, where the triad of catalytic residues appears to have been discovered at least twice by evolution (Casari et al. 1995). In this context, functional and evolutionary alignments may disagree, and therefore reflect convergent evolution. Functional analysis can also be used to track subtler processes such as the evolution of substrate specificity and affinity, materialized by positions differently conserved within subfamilies. A commonly accepted scenario is that whereas fully conserved positions related to functional features are common to all the members of the family, these other residues related to functional specificity are common to the members of the sub-family (e.g., binding of different cofactors) (Rausell et al. 2010). Magis *et al.* have recently shown that structure based analysis can be combined with evolutionary based inference in order to disentangle the complex interaction between genetic drift, purifying selection and convergent evolution. Using a novel tree reconstruction algorithm (T-RMSD) based on the comparison of intra-molecular distances, they show a process of convergent evolution occurring between a ligand and its receptor in the Tumor Necrosis Receptor family (Magis et al. 2010; Magis et al. 2012).

All things considered, the most widely used framework for the reconstruction of multiple sequence alignment is an evolutionary framework. In this context, the MSAs are assembled as a substrate for the estimation of the scenario underlying the divergence of the considered sequences. In this context, each aligned item in the MSA may be viewed as an evolutionary hypothesis. The columns of MSA explicitly support the hypothesis that all align residues correspond to the same unique ancestral residue in the last unique common ancestor of the considered sequences. While the most common aligners do not explicitly try to optimize an evolutionary scenario and merely optimize similarity, a more recent class of aligners have recently developed with the explicit purpose of generating the more evolutionary probable MSAs. They include phylogenetically aware alignment and statistical alignment. In contrast to the previous generation aligners, they are not benchmarked using structural information, but rather using simulated phylogenies (Hall 2005; Rosenberg 2005; Nuin et al. 2006; Ogdenw and Rosenberg 2006; Kumar and Filipski 2007; Landan and Graur 2009; Wang et al. 2011). Of course, one may argue that such aligners heavily depend on a priori simulated models, and it is a matter of fact that strong discrepancies remain with respect to the performances of various packages on structure based or simulated benchmarks. Resolving this issue will probably require a 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 (Dessimoz and Gil 2010).

The four criterion discussed in the above section that may be used to reconstruct multiple sequence alignments reflect well the complexity of the problem. On the one hand, protein sequences are part of living entities, and as such they undergo an evolutionary process that includes neutral processes. These same proteins evolve under various levels of selection. Some are only constrained not to harm their host and must therefore retain a functional fold. Others must retain precise functions or embark for survival on a red-queen race in order to follow their ligand, outpace their competitors or acquire novel functions. In theory, models could be built that address each of these aspects separately. Unfortunately, our knowledge of biology is not complete enough. We do not understand well the relation between sequences and structures, the rules governing evolution and selection, and even less the function/sequence relationship. This results in a necessary attempt to combine these four aspects of biological information in the hope that they will complement each other and lead to more accurate models at all levels. However, this seemingly obvious approach has now become less of evidence, especially when considering the growing number of conflicting reports on structure and evolutionary based MSA evaluation. I will now discuss in more details the nature of these inconsistencies and suggest how a better use of homology information may help address the question.

Table 1-1 Main criteria for building a multiple sequence alignment (adapted from (Claverie and Notredame 2003))

Similarity	Meaning	Dataset	Aligner
Sequence	Amino acids in the same column are those that yield an alignment with maximum similarity. Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, structure, evolutionary and functional similarities are equivalent to sequence similarity.		MAFFT (Thompson and Poch 2006), MUSCLE (Edgar 2004b), T-Coffee (Notredame et al. 2000), ProbCons (Do et al. 2005)
Structure	Amino acids that play the same role in each structure are in the same column. Structure superposition programs are the only ones that use this criterion	BAliBASE 3 (Thompson et al. 2005a), PREFAB 4 (Edgar 2004b), SABRE (Van Walle et al. 2005a), OXBENCH(Raghava et al. 2003)	TM-align (Zhang and Skolnick 2005), SAP (Taylor 2000),3D-Coffee (O'Sullivan et al. 2004)
Functional	Amino acids or nucleotides with the same function are in the same column. No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it or you can edit your alignment manuallv.		
Evolutionary	Amino acids or nucleotides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column. No automatic program explicitly used this criterion, but they all try to deliver an alignment that respects it.	Species-tree discordance, minimum duplication (Dessimoz and Gil 2010)	SATe (Liu et al. 2012), PRANK (Loytynoja and Goldman 2008), BaliPhy (Suchard and Redelings 2006), AliFritz (Fleissner et al. 2005), StatAlign (Novak et al. 2008), POY (Varón et al. 2010)

b) Inconsistencies among mathematical, structural and evolutionary based objective functions

The reason why it is difficult to compute a MSA is not only because of the time-consuming issue, but also because it is hard to precisely define, in mathematical terms, what an accurate MSA really is. This combined complexity of defining and estimating MSAs results in a crossroad problem between biology (what is "A" good MSA) and computer science (Given the definition of "A" good MSA, how can this be computed). As previously discussed, and even when one uses a well defined criteria (Structure, Evolution, Function), it remains difficult to objectively evaluate alternative alignment procedures (Kjer et al. 2007). The difference between alignments based on different criteria will be discussed in the following paragraphs. "Similarity" versus "Structure" will be addressed first since the largest proportion of MSA methods based on the similarity-based and the most popular MSA benchmarks based on structure perspectives. In the last section, "Similarity" versus "Evolutionary" is discussed because they are like the two ends of the spectrum (Anisimova et al. 2010).

The relation between a similarity optimality (using the so called SoP) function and structural accuracy was initially explored using a genetic algorithm (Notredame and Higgins 1996). The authors then concluded on the lack of a strong correlation between mathematical optimality and structural correctness. At the time, however, reference databases were at a very preliminary development stage and CPU limitation was making it difficult to explore datasets informative enough. We decided to re-explore this issue by applying "MSA" (Lipman et al. 1989; Gupta et al. 1995) the only heuristic able to deliver optimal or near optimal MSAs onto BAliBASE 3. When doing so, using as an objective function unweighted SoP score based on a BLOSUM62 and gap penalties (open: -12, extend: -1), we only found "MSA" able to deal with 17 BAliBASE 3 datasets (6 sets in RV11 and 11 sets in RV12). For comparison, these same datasets were ran through PSI-Coffee (Chang et al. 2012c), a flavor of T-Coffee that combines consistency based alignments with homology extension and has been reported to produce highly accurate MSAs on that same dataset. Besides mathematical score, alignment in structural accuracy is measured as its similarity against the core region of the reference alignment reported by BAliScore program. The relation between SoP and BAliScore of PSI-Coffee and "MSA" is shown on Figure 1-1, where SoP score is normalized as divided by the SoP of the structural reference alignment. Alignment with higher SoP score is not necessary to have higher BAliScore. It suggests a poor correlation between SoP mathematical accuracy and structural correctness. This result is not surprising, as it merely reflects our limited understanding about relationship between sequence and structures.

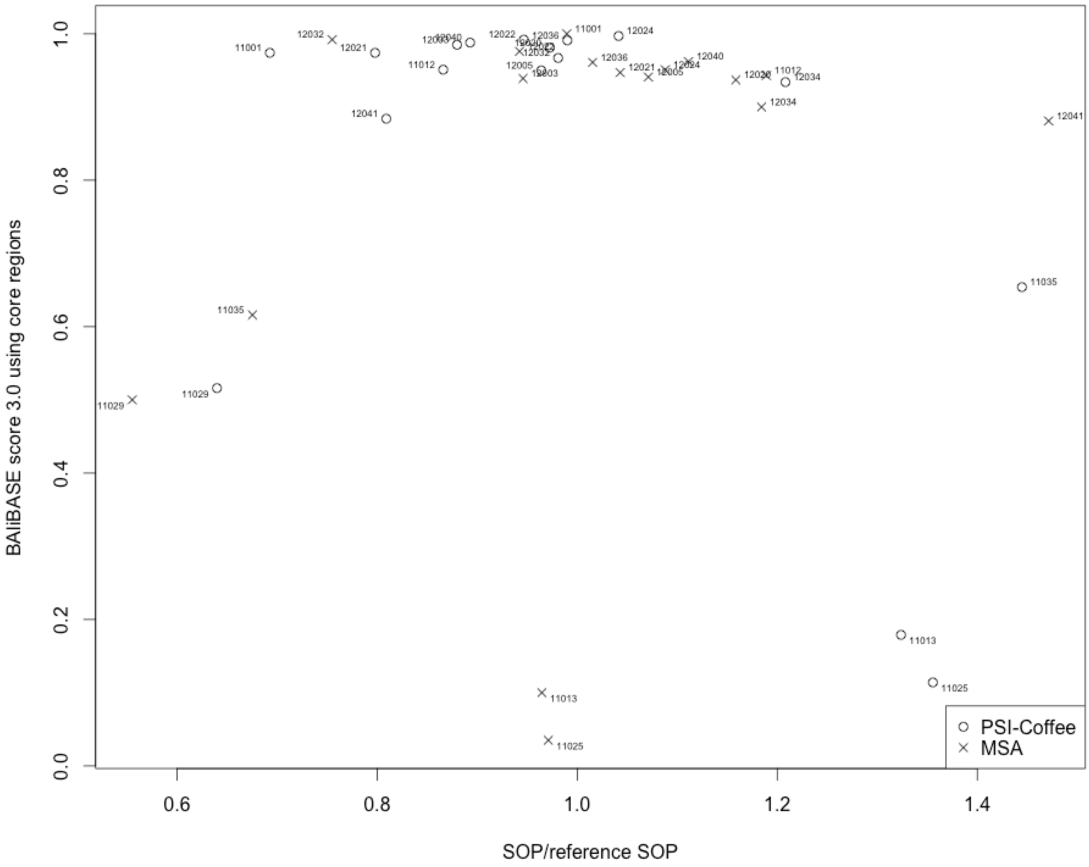

Figure 1-1 SOP against the SOP of reference alignment versus BAliScore of PSI-Coffee and “MSA” alignments. Labels represent corresponding BAliBASE sets.

Despite the inherent difficulties of modeling biological reality, many progresses have been made to improve alignment strategies by defining more realistic objective functions. The most notable progress has been the development of consistency based (Notredame et al. 2000) and probabilistic consistency based objective functions (Do et al. 2005; Roshan and Livesay 2006) that address both the optimization issue and the need to enrich similarity based objective functions with functional, structural and evolutionary information (Kemena and Notredame 2009). Consistency makes it possible to do a better use of data prior knowledge (i.e., like 3D structures, knowledge of active sites, a known pattern expected for a protein domain). This not only facilitates the computation of higher accuracy MSAs, but also makes it possible to systematically combine heterogeneous level of sequence analysis. As such, the systematic use of consistency may be seen as a useful step towards the establishment of a better balance between mathematical and biological optimality when computing MSAs, an issue often described as pressing (Anisimova et al. 2010).

Until recently, most methods were developed under the assumption that similarity was a reasonable indicator of homology, thus implying that similarity based models would be informative with respect to homology. This would imply that similarity based models should reflect the evolutionary process associated with the divergence of the considered sequences. In this context, one would expect aligners that do well at maximizing similarity to eventually support the most accurate phylogenetic reconstructions. This simple assumption has recently been questioned and there is indeed no direct evidence that a procedure based on maximizing similarity will result in an alignment reflecting accurately the positional homology between the residues (Morrison 2009). Indeed, Blackburne and Whelan found that “similarity-based” MSA methods (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 two alignments (Blackburne and Whelan 2012a). The class of an MSAM has a substantial impact on downstream analyses, phylogenetic inference (Blackburne and Whelan 2012a). They found tree topology estimates and their branch lengths show highly dependent on the class of MSAM used. The class of aligner used also affects the number of families, and the sites within those families, inferred to have undergone adaptive evolution. Similarity-based aligners favor to find more adaptive evolution (Blackburne and Whelan 2012a). David A. Morrison pointed out that phylogeneticists are usually dissatisfied with similarity-based alignment procedures and tend to manually edit alignment because they do recognize that similarity-based alignments are not likely to be homology alignments. The preponderance of manual alignments is simply a reflection of the above observation of two MSAMs groups (Morrison 2009). He concluded that there is currently no bioinformatics approach that is acceptable for phylogeneticists. This observation may explain why the results observed on simulated data significantly differ from those measured on empirical data (Kemena and Notredame 2009).

c) A new class of objective function based on functional data gathered through next generation sequencing

Whenever experiments reveal novel biological features, it is a common procedure to look for them across related species and to evaluate how these new features fit in the framework of comparative biology. If comparison is the engine of this framework, new data is its fuel. A major source of novel data is the emergence of novel technologies. I will argue that the emergence of Next Generation Sequencing (NGS) with its dropping sequencing cost defines a new era in biology, and that the change of scale has already started initiating one of the most important paradigmatic shifts since discovering DNA structure. Advances in sequencing technologies have allowed the rapid sequencing of full genomes, which in turn is driving advances in methodology for aligning and assembling short, reads and for multiple whole genome alignment. As recently pointed out by Anisimova *et al.* (Anisimova et al. 2010), “Despite a number of recent algorithmic advances, the genomics alignment field is still in its infancy, presenting succulent challenges, yet to be solved”.

Next generation sequencing may be seen as a functional wrapping for genomic data. For instance, transcriptomic analysis indicates the precise location of transcription activity while recordings made across alternative cell lines reveal cellular functions and their immediate low level phenotypic outcome. Likewise, ChIP-Seq data provides a direct evidence of binding activity. In ENCODE, this data has been systematically recorded across 18 human cell lines (http://genome.ucsc.edu/ENCODE/cellTypes.html). Even though some controversy has arose as weather binding and function may be considered equivalent (Graur et al. 2013), it is a fact that signals thus recorded are fit for cross species comparison and they are therefore opening the way of a new comparative genomics era.

It is therefore reasonable to believe that ENCODE-type NGS data will gradually become a key component of comparative genomics in the near future, playing on non-translated RNA a role similar to that of structures when dealing with proteins. It may first be used for benchmarking purposes, by providing the aligners with enriched DNA sequences in a controlled manner and by extrapolating the readouts on enriched sequences to the total dataset, including sequence for which no RNA-Seq or ChIP-Seq data was available. Recently, Erb *et al.* proposed a benchmark procedure based on

ChIP-Seq data for promoter alignment (Erb et al. 2012) gathered across several species: human, mouse, dog, and chicken. Methods were evaluated for their capacity to correctly align Transcription Factors Binding Sites (TFBSs) identified by using ChIP-Seq. The accuracy of a sequence alignment would be to simply count the numbers of TFBSs effectively matched across species. It is the first time that such a data set is used to determine the relative accuracy of multiple promoter alignment procedures.

By extending Erb's idea, I have evaluated the usefulness of RNA-Seq data for the comparison of alternative multiple genome alignments. My approach relies on the idea that given RNA-Seq expression data gathered in similar tissues, a correct alignment of the corresponding genomes should result in a strong overlap of the mapped data, both in terms of position and coverage, under the general assumption that orthologous genes tend to have the similar expression pattern across species (Liao and Zhang 2006; Zheng-Bradley et al. 2010). In order to avoid saturation, the measurements were limited to boundaries between high and low read coverage, which is most likely to represent intron/exon boundaries. Nucleotides corresponding to such boundaries were marked and used to compare alternative multiple genome alignments of the same sequences (Figure 1-2). For each pair of sequences, the values are summed up to produce the final score of this alignment named RNA-Seq score.

RNA-Seq data was obtained previously from six *Drosophila* species: *D. Ananassae*, *D. Erecta*, *D. Melanogaster*, *D. Virillis, D. Willistoni* and *D. Yakuba*. The reads from those data were mapped by using the segemehl program (Hoffmann et al. 2009) with default settings. Files containing the per-base read coverage for each of the six species were used for the subsequent analysis. The alignment framework is a special flavor of T-Coffee, named Robusta, which makes it possible to combine alternative multiple genome aligners using the T-Coffee consistency algorithm. We used it to test and compare various combinations. In this context, we tested 4 different packages: Pecan, Mavid, Mauve (progressive), and Lastz. Robusta is a meta-aligner and an extension of the M-Coffee package that combines the output of several alternative aligners into one unique final model. We considered a total of 14 combinations (Table 1-2). Besides Robusta, four aligners, Lastz, Mavid, Pecan, and Pro-Coffee, are also involved into benchmark. Those aligners are applied on two data sets, simulated Mammals and biology Flies, from Alignathon (Earl 2012). For simulated Mammals set, since we already knew the truth due to simulation, we therefore used mafComparator (Earl et al. 2012) to compare the true Multiple Alignment Format (MAF) file against a predicted MAF file by aligner. For the Flies set, each method is evaluated based on RNA-Seq score (Table 1-2).

It is interesting to note the lack of a clear agreement between the RNA-Seq evaluation of the Drosophila genome alignments, and a similar evaluation on the simulated data sets. For instance, mafComparator, suggests Pecan to the best method on simulated Mammalian data, while on Drosophila data, RNA-Seq suggest Pecan to be an average method, with the Robusta combination of pecan, mavid, lastz giving the best readout. Overall, however, it is worth pointing out a reasonable agreement between the both the specificity and sensitivity of the mafComparator score with our RNA-Seq readout (Figure 1-3).

Table 1-2 Sensitivity and specificity of the Mammels set and RNA-Seq score of the Flies set for all methods

	Mammals				Flies	
Method	Sen.(%)	rank	Spe.(%)	rank	RNA-Seq	rank
Robusta						
pecan	36.01%	3	79.89%	2	4369400	10
mavid	30.84%	15	66.79%	14	4361020	14
pmauve	26.42%	16	59.69%	16	4327600	17
pecan, mavid	35.70%	7	75.19%	6	4380320	4
pecan, pmauve	35.88%	4	77.61%	4	4367800	12
pecan, lastz	36.06%	2	78.76%	3	4383280	2
mavid, pmauve	33.29%	12	69.08%	13	4370700	9
mavid, lastz	34.00%	11	70.83%	10	4372020	7
pmauve, lastz	33.06%	13	69.31%	12	4357930	15
pecan, mavid, pmauve	35.71%	6	74.08%	8	4381400	3
pecan, mavid, lastz	35.68%	8	74.38%	7	4384130	1
pecan, pmauve, lastz	35.77%	5	76.81%	5	4368650	11
mavid, pmauve, lastz	34.34%	10	70.39%	11	4370830	8
pecan, mavid, pmauve, lastz	35.64%	9	73.67%	9	4379560	5
Lastz	32.99%	14	61.85%	15	4338160	16
Mavid	16.49%	17	56.26%	17	4206520	18
Pecan	36.88%	1	87.86%	1	4362240	13
Pro-Coffee	N.A.		N.A.		4376850	6

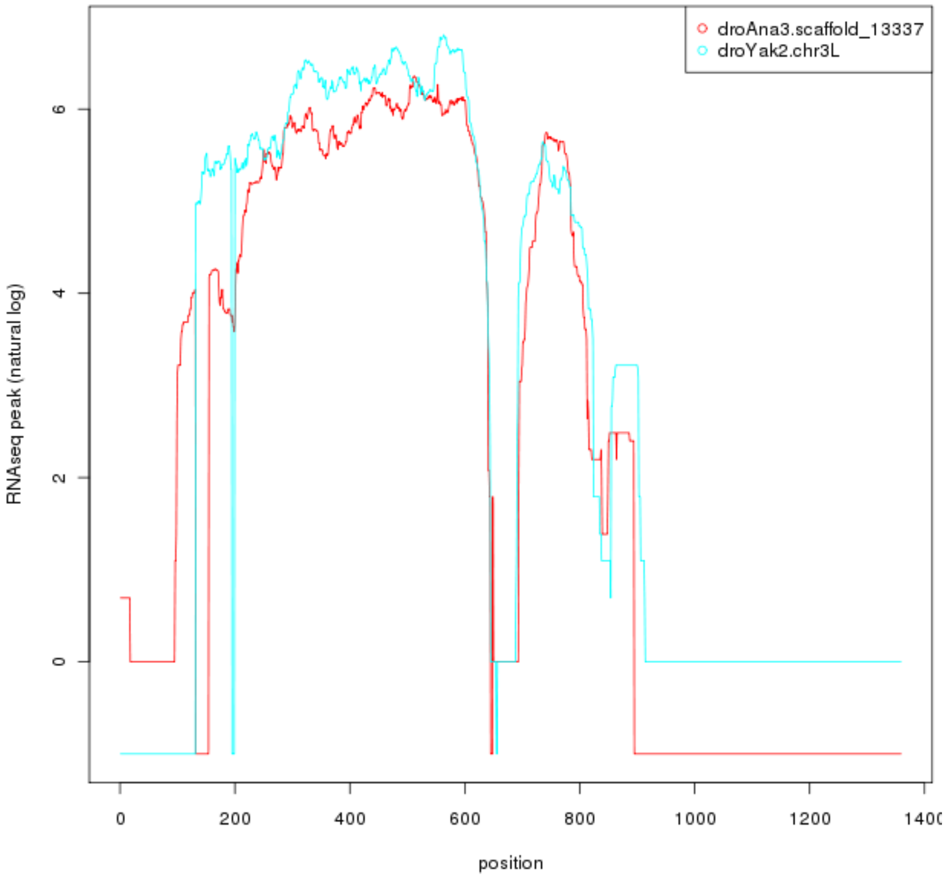

Figure 1-2 Correlation of peaks / peak boundaries for an example alignment (r = 0.89). Gap values are set to -1.

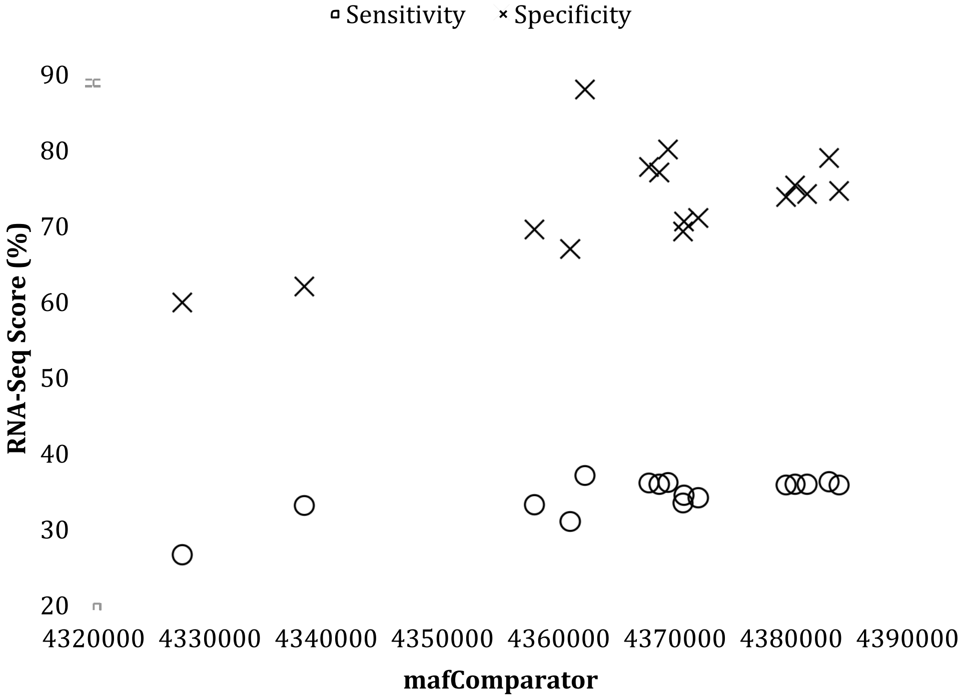

Figure 1-3 Correlation of RNA-Seq score (Flies) vs. Sensitivity/Specificity by mafComparator (mammals) (sensitivity r = 0.95/ specificity r = 0.67)

1.2 Alignment uncertainty: cause and measurement

Section 1.1 has reviewed the difficulty of objectively selecting a criterion for the computation of biologically meaningful MSAs. This problem, purely biological, leaves intact the issue of optimizing the chosen objective function. Indeed the computation of an MSA being NP-Complete under all the above formulation, one must use approximate heuristics that do not guaranty optimality. As one would expect, the wealth of available solutions means than many alternative MSAs may be derived from a single dataset. Furthermore, since all these packages do not always rely on an explicit objective function, it can be hard to estimate the level of optimality of a given model. To make things worse, objective function is defined in such a way that more than one MSA model may have the same optimal score. This issue is especially severe when dealing with distantly related sequence of low complexity like nucleic acids. In that case, the tiniest perturbation (change of the input sequence order, different substitution matrices) can have a dramatic impact on the final model, especially when dealing with very large datasets (Breen et al. 2012). In this section we discuss the main algorithmic reasons of this instability and how one can improve modeling robustness by quantifying it so as to identify the most trustworthy MSA features.

Interest for this long known issue (Notredame 2002; Wallace et al. 2006) is now rapidly growing, as illustrated by an entire session of the last international conference of the Society for Molecular Biology & Evolution (SMBE 2012), entirely dedicated to this specific topic under the heading “Multiple sequence alignment, alignment confidence, and impact on downstream analyses”. This renewed attention is also well illustrated by the recent publication of several high impact papers dealing with this precise aspect of MSA modeling (Talavera and Castresana 2007; Wong et al. 2008) and an attempt to quantify its effect on downstream modeling (Jordan and Goldman 2012). It turns out that being able to discriminate confidently aligned from problematically aligned parts within an alignment is more important than its overall accuracy (Wu et al. 2012). In fact, in the last few years special efforts have been made focusing on the development of alignment confidence measures and led to many new approaches (Landan and Graur 2007; Talavera and Castresana 2007; Capella-Gutierrez et al. 2009; Penn et al. 2010a; Penn et al. 2010b). One may distinguish two sources of instability when doing MSA modeling: internal and external. External uncertainty results from arbitrary algorithmic choices across various algorithms and their many implementations. The reason why this is a cause of uncertainty is the difficulty we have to discriminate a priori between two alternative alignments of the same sequences produced by two different packages. Without adequate structural information, one can only guess on the basis of the relative benchmark performances, which amounts to betting on the horse with the best odds. Internal uncertainty relates to the algorithm itself and the effect of any arbitrary tie breaking process taking place in the course of the optimization process. Internal uncertainty often results in degraded robustness and sensitivity to a priory neutral manipulation, like sequence input order. Recent packages developed to quantify alignment uncertainty are summarized in Table 1-3. In this section we review the main causes for internal and external instability.

a) Differences in aligners algorithm

It is well known that algorithmic variations in multiple aligners result in MSA variations. A given dataset will often lead to significantly different MSAs when processed with different methods. The level of similarity between various methods can be visually displayed as a "method tree" shown in Figure 1- 4 (Wallace et al. 2006).

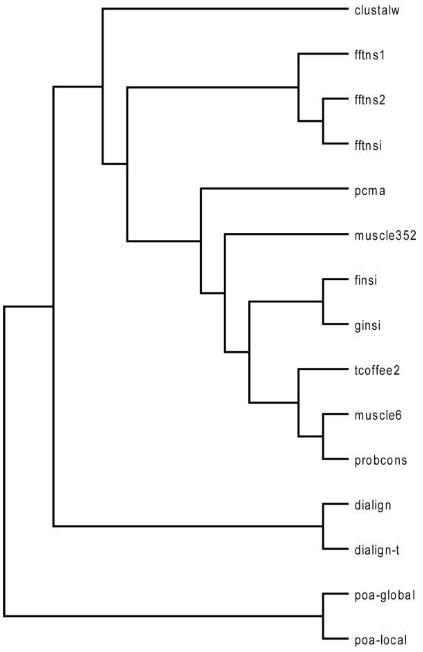

Figure 1-4 Methods Tree - A UPGMA (Sokal and Michener 1958) tree which shows the clustering of different MSA methods. Pairwise distances are calculated on the HOMSTRAD benchmark by computing the SoP differences of the alignments produced by individual methods (adapted from Figure 1 (Wallace et al. 2006)).

Yet, it is only recently that the community has started evaluating systematically the effect of these variations onto subsequent modeling, including phylogenetic reconstruction, by far the most common application of MSA reconstruction. Hickson *et al*. were the first to show how different alignments generated using various programs or even different alignment parameters can yield very diverse phylogenetic trees (Hickson et al. 2000). More recently, Wong *et al.* have pointed out the practical consequences related to this issue when reconstructing phylogenetic trees at genome scale (Wong et al. 2008). Given a collection of 1502 dataset of homologous sequences, the seven most widely used aligners only agree on less than 55% of the dataset (ClustalW (Thompson et al. 1994), Dca (Stoye 1998), Dialign2 (Morgenstern 1999), Mafft (Katoh et al. 2002), Muscle (Edgar 2004b), ProbCons (Do et al. 2005) and T-Coffee (Notredame et al. 2000)). In other words, for any given dataset there is less that 55% chance to obtain a unique tree topology. Even considering individual alignment software, different parameters usually result in different alignment and therefore different downstream analyses, i.e., phylogenetic conclusion (Kjer et al. 2007). As discussed in previous section, we refer to this uncertainty, which results from algorithmic variation, as “external” uncertainty as opposed to the internal uncertainty developed in the next section. In phylogeny, it is common practice to estimate the robustness of each node in a tree using a process known as “bootstrap”. In this context, the bootstrap is a re-sampling procedure used identifies any bias in the data that would undermine the full support of the considered trees. It may be described as an attempt to estimate the fraction of columns in the MSA model supporting every split of the sequences. In Chapter 3, I describe a novel bootstrap procedure, Weighted Partial Bootstrap, that makes it possible to simultaneously take into account the sampling biases (like regular bootstrap) and the algorithmic fluctuation, eventually combining them into a unique bootstrap support value for each node.

b) Input sequence orientation

A vast majority of all available aligners, and for sure the five most widely used, all rely on a similar algorithm known as the progressive alignment. Under that scheme, sequences are incorporated one by one in the MSA following an order defined by a binary guide tree. At every node the child sequences are merged into a sub-MSA using a pairwise sequence alignment algorithm able to align sequences, profiles or a combination. This algorithm is always based on Needlman and Wunch (Needleman and

Wunsch 1970), a variation of the original Bellman dynamic programming designed to solve a sink to source problem (Bellman 1958). NW is meant to estimate the best possible score for matching two sequences, given a position specific scoring scheme and a gap penalty:

$$score_{i,j}=\mathbf{max}(score_{i-1,j-1}+match_{i,j}\,,\ score_{i-1,j}+gap\ ,\ score_{i,j-1}+gap) \qquad (1)$$

Interestingly, the estimate of this optimal quantity can be obtained using at least two different iterations, outlined below. Overall, the score is the same but the difference between these two implementations is a break of ties in a different order that results in different final optimal alignments. These two alternative ways of resolving a tie are often referred to as high-road and low-road.

Code 1

```
if ($\text{score}_{i-1,j-1}+\text{match}_{i,j} \geq \text{score}_{i-1,j}+\text{gap}$) & ($\text{score}_{i-1,j-1}+\text{match}_{i,j} \geq \text{score}_{i,j-1}+\text{gap}$)
    $\text{score}_{i,j} = \text{score}_{i-1,j-1}+\text{match}_{i,j}$
else if ($\text{score}_{i-1,j}+\text{gap} \geq \text{score}_{i,j-1}+\text{gap}$)
    $\text{score}_{i,j} = \text{score}_{i-1,j}+\text{gap}$
else
    $\text{score}_{i,j} = \text{score}_{i,j-1}+\text{gap}$
```

Code 2

```
if ($\text{score}_{i-1,j-1}+\text{match}_{i,j} > \text{score}_{i-1,j}+\text{gap}$) & ($\text{score}_{i-1,j-1}+\text{match}_{i,j} > \text{score}_{i,j-1}+\text{gap}$)
    $\text{score}_{i,j} = \text{score}_{i-1,j-1}+\text{match}_{i,j}$
else if ($\text{score}_{i-1,j}+\text{gap} > \text{score}_{i,j-1}+\text{gap}$)
    $\text{score}_{i,j} = \text{score}_{i-1,j}+\text{gap}$
else
    $\text{score}_{i,j} = \text{score}_{i,j-1}+\text{gap}$
```

For instance, when aligning OPOSSUM (seq i) and BLOSUM62 (seq j), there are two optimal alignment results (Figure 1-5b), Aln1 from Code 1 (Figure 1-5a, organ path) and Aln2 from Code 2 (Figure 1-5a, blue path), respectively.

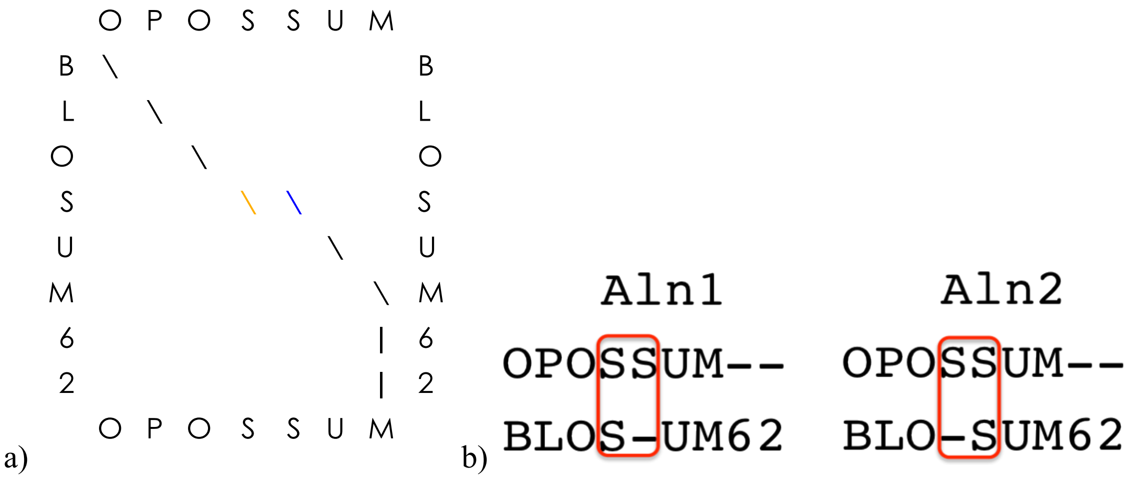

Figure 1-5 Alignment uncertainty when aligning OPOSSUM and BLOSUM62.

Given a pairwise alignment of reasonably related sequences, these arbitrary decisions often have little impact and usually result in shifting some gaps (those having similar residues on both edge). It is worth pointing out that changing the input order of the sequences is a strict equivalent of a change of formulation. Under this scheme the optimal alignment (not its score) will therefore depend on the sequences input order. With two sequences, this has little consequences but with more than two sequences, a combinatorial problem occurs that can significantly affect the reconstruction process. For instance, if we include the third sequence and estimate the MSA through the combination of two pairwise MSAs (1 vs. 2 and 2 vs. 3), we will have 4 possible MSAs. It is easy to anticipate that the problem will become increasingly severe with growing numbers of sequences (Figure 1-6).

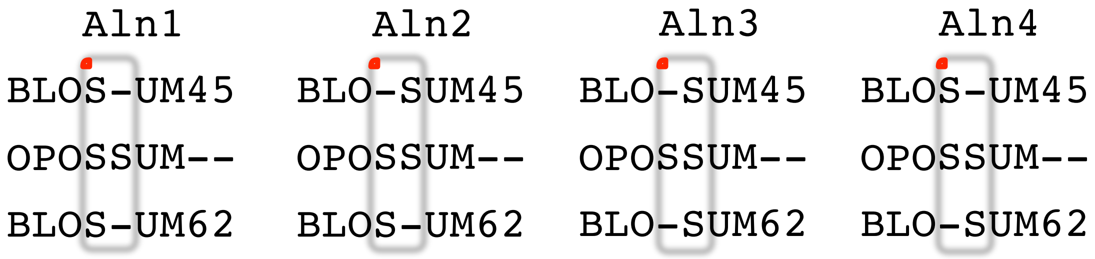

Figure 1-6 Alignment uncertainty when aligning OPOSSUM, BLOSUM62, and BLOSUM45.

This very precise phenomenon was recently used to develop the so Heads-or-Tails methodology (HoT), showing that substantial variations can occur when constructing an MSA on a set of sequences and subsequently on the same sequences reversed from left to right (Landan and Graur 2007). Observed variations merely results from the arbitrary tie breaking between high-road (i.e., Figure 1-5 a, blue path) and low-road (i.e., Figure 1-5 a, organ path) during dynamic programming (Landan and Graur 2008). It must be stressed, however, that reversing sequence input direction amounts to systematically changing all the ties. It is therefore a very limited exploration of the MSA space. Other packages, like PRANK do a more thorough exploration by systematically breaking ties in a random fashion thus making each run an independent sampling. Several variations have been described around the HoT algorithm, one of which involved applying the HoT procedure on the partition of the sequence defined by internal nodes of the guide tree. The updated HoT increases the alternative sample size to eight (Landan and Graur 2008). This allows a better exploration but looses one of the main advantages of the HoT algorithm, which is the possibility to apply it on any third party algorithm. It is also worth noting that in the original publication, the Head or Tail process was applied onto the ClustalW algorithm, whose gap penalties are position specific and were estimated using structural information. Under such a scheme, the non symmetry of dynamic programming is not a consequence of the high-road/low road issue, but merely the result of a different scoring scheme, resulting from the new amino-acid transitions imposed by the reverse order.

c) Input sequence order

We showed that the shuffle of the order in the sequences are provided to the aligner, which can estimate a similar result approximately (Chang et al. 2012b). In most dynamic programming implementation, a swap of the order in which two sequences are aligned effectively amounts to inverting the tiebreak priority of each tie (low-road becomes high-road and inversely).

We applied this strategy on two popular aligners: ClustalW2.1 (Larkin et al. 2007) and MAFFTv6.815b (Katoh and Toh 2008), which use the BAliBASE3.0 reference dataset (Thompson et al. 2005b). Our approach contained shuffling the sequences and realigning them 100 times. The resulting MSAs are then combined by using M-Coffee; a flavor of the T-Coffee package enables the combination of several alternative models into a consensus model (Wallace et al. 2006). The M-Coffee MSA model comes along with an estimation of the consistency between each position and the combined alignments, which is like a consensus tree when replicates are combined. Their results indicated that variations among alignments are like a consequence of the order shuffling. For instance, MAFFT replicates are on 92.1% consistent in average; while ClustalW alignments are 90.2% consistent of BAliBASE 3. Interestingly, our analysis also suggested that a correlation between alignment accuracy and overall consistency (ClustalW $r = 0.49$, MAFFT $r = 0.52$). This result also held locally. For instance, pairs of residues having a score of 5 or higher are 96.9% likely to be correctly aligned. Likewise, pairs of residues having a score of 4 or lower are 27.6% likely to be accurately aligned.

d) Guide tree topology in the progressive alignment

In theory, the computation of an optimal progressive alignment should not depend on the guide tree since, unless one uses a weighting scheme based on this same guide tree, the tree itself is not part of the scoring scheme. In practice, however, the guide tree is a very important component of the MSA reconstruction and has been shown (Wheeler and Kececioglu 2007) to significantly impact the reconstruction capacity of various methods (Edgar 2004b). GUIDANCE was recently developed in order to evaluate the impact of the guide tree onto the subsequent MSA robustness (Penn et al. 2010b). Guidance measures the robustness of MSA by estimating a set of perturbed trees obtained using bootstrap replicates drawn from an input MSA. These replicate trees are then used to generate an equal number of alternative MSAs using any progressive algorithm accepting a pre-computed guide tree along with the sequences. The final evaluation is obtained by measuring the agreement between the various MSA replicates and the target MSA. GUIDANCE has been reported to be more accurate than HoT in predicting regions most likely to affect phylogenetic reconstruction. This approach, however, is invasive and requires a substantial code modification in most MSA packages. Moreover, it is only suitable for progressive approach.

e) Alignment robustness

It has long been known that the correctness of optimal alignment may be estimated by comparing it to alternative sub-optimal alignments of the same sequences. This notion has considerably gained momentum when the notion of posterior decoding was introduced in the field of sequence alignment, along with pair Hidden Markov Models (pair-HMM). In this context, one can simultaneously estimate the combined posterior probability of all sub-optimal alignments. Posterior decoding then makes it possible to measure the score of any alignment in the light of its sub-optimal background. This idea was originally developed in the probabilistic Smith and Waterman algorithm (Bucher and Hofmann 1996) and later used in the ProbCons algorithm where alignment library is estimated using the posterior decoding of a pair-HMM. More recently, the PSAR algorithm was reported as a way to quantify an alignment reliability by comparison with alternative alignments of the same sequences (Kim and Ma 2011). In that case, the strategy relies on a leave one out approach where the MSA is re-estimated several time by removing all sequences in turn and then estimating the consistency between this collection of partial alignments and the full model. Kim and Ma reported a positive relation between the support by suboptimal alignments and the correctness of alignment, at least on DNA sequences.

The consistency framework of T-Coffee is ideally suited for such analysis. Consistency, as mean to estimate MSA accuracy was first introduced by Gotoh in an iterative MSA strategy (Gotoh 1990). The notion was then further refined within the Dialign algorithm that uses the so-called overlapping weights to drive an agglomerative assembly process, (Morgenstern et al. 1996) in a way reminiscent of Vingron's dot matrix multiplication (Vingron and Argos 1991). This concept was eventually combined with the progressive alignment framework in the T-Coffee package, and later further refined using in ProbCons probabilistic consistency algorithm. The main specificity of T-Coffee is to use a library of pairwise comparison in order to estimate the cost for aligning every pair of residues in the dataset. In this context, the cost for matching two residues is not anymore estimated by a substitution matrix, but simply by estimating how many times these two residues were found aligned in the library, either directly or through a third sequence. The support of all these combined alternative alignments (suboptimal or not) can then be used to estimate the reliability of every pair of aligned residues. This score, named Transitive Consistency Score (TCS) is a slight modification of the CORE index originally reported for the same purpose.

In Chapter 4, I report a comparison between GUIDANCE, HoT and TCS, as an indicator of alignment reliability, as estimated on BAliBASE 3 and PREFAB4 (Edgar 2004a). We found that the performance of TCS was superior to the GUIDANCE score in two discriminating tests: correct/incorrect regions of alignments and accurate/inaccurate aligners. On the former test, the Area Under the Curve (AUC) of

TCS is not only the best, ranging from 94.44% to 98.80%, but also the most stable among different properties of testing sets and aligners. On the later analysis, the accuracy of determining more accurate alignment of two methods by TCS is from 80% to 87%, which is highly significant from a statistical point of view. Detail description can be found in Chapter 4.

1.3 Using MSA scoring schemes for post processing purposes

Alignment internal uncertainty can be used to reveal poorly supported portions of an MSA model, while the external uncertainty may allow an estimate of the MSA modeling global robustness. This information is arguably more useful than any global estimate of the full MSA accuracy since it can allow a systematic use of the most trustworthy regions for modeling purpose. Those two issues are related to each other. For the first issue, how to summary the information among alternative alignments is to make the consensus alignment of those alternative alignments, i.e., a consensus tree from bootstrap or supertree, a consensus tree from different genes (Delsuc et al. 2005). Furthermore, these alternative sample alignments reflect uncertainty and come by different sources. Therefore, we can evaluate the confidence region of a given alignment by checking the proportion of aligned pairs in a given alignment existing among alternative alignments. The low confident region of the alignment might be filtered. Methods related to those two issues will be reviewed in the following sections.

a) Consensus alignment

The alignment pattern might be quite common among or various among alternative alignments. A consensus alignment is one way to summary those information from alternatives. Regions of high agreement have been shown usually well aligned (Wallace et al. 2006). Although there are many packages for MSA, there are few packages available for making a consensus alignment from input alternative alignments. For the combination of alternative DNA alignments, ComAlign was first described by Bucka-Lassen *et al.* (Bucka-Lassen et al. 1999) and then M-Coffee was proposed for protein MSAs (Wallace et al. 2006). ComAlign extracts qualitatively good sub-alignments from a set of multiple alignments and merges them into a new alignment which is showed often improved. Their algorithm is implemented as a variant of the traditional dynamic programming strategies. M-Coffee is a special mode of T-Coffee and uses consistency technique to estimate a consensus alignment. It does not explicitly align sequences but compiles externally produced alignments as consistency libraries. Then, those libraries are combined into a final MSA during the aligning procedure. Their benchmarks suggest that the resulting alignments tend to be more accurate than the individual methods (Wallace et al. 2006). Another framework, MergeAlign, represents constituent MSAs as a weighted directed acyclic (Collingridge and Kelly 2012). By using dynamic programming, the path, which maximizes the weights of edges, corresponds to a consensus MSA. Interestingly, both studies found that there is a positive correlation between the consistency of the part of the consensus alignment and its accuracy.

b) Trimming alignment

Filtering alignment ambiguity region is supposed to increase the single-versus-noise ratio of phylogenetic information but this assumption is still controversial. For phylogenetic construction, it is not clear that the single-versus-noise ratio is guaranteed to increase after filtering.

On one side, it has been shown that trimming alignment by removing non-reliable regions usually lead to an improvement on the overall accuracy of downstream phylogenetic analysis (Castresana 2000; Talavera and Castresana 2007; Capella-Gutierrez et al. 2009). In 2000, Jose Castresana proposed a method, Gblock. It selects the blocks of alignment according to the number of contiguous conserved positions, less gappy and high conservation of flanking positions. Many possible nonhomologous positions will be eliminated after trimming (Castresana 2000). Then, they showed that cleaned alignments produce better tree topologies by using simulated protein sequences aligned by ClustalW, MAFFT, and ProbCons (Talavera and Castresana 2007). trimAl determines the confidence of alignment column based on gap score, residue similarity score, and identity score. It will automatically select the corresponding thresholds to be used in each specific alignment so that the signal-to-noise ratio is optimized. Its performance is better than Gblock in most scenarios (Capella-Gutierrez et al. 2009).

In contrast, using simulated data (Liu et al. 2009; Wang et al. 2011) or biological data (Aagesen 2004; Simmons et al. 2008; Dessimoz and Gil 2010; Saurabh et al. 2012b) as a way to evaluate the final alignment, trimming does not always give more informative results. Dessimoz and Gil concluded gaps carry substantial phylogenetic signals; therefore, excluding gaps and variable regions is harmful (Dessimoz and Gil 2010). Saurabh *et al*. reported that indel (insertion/deletion) process might create considerable information that is potentially benefic for phylogenetic inference. However, a better model of indel process is required before the information in gaps can be fully exploited for phylogenetic inference (Saurabh et al. 2012b).

To overcome the trade-off between filtering problematic alignment regions and keeping evolutionary signals, we proposed a weighted replication strategy. It uses replicating instead of filtering, such that the information of a given alignment will not only be kept but also be enriched for phylogenetic analysis according to its confidence. The detail is described in the Section 4.

Table 1-3 Alignment uncertainty software currently available to users (incomplete list).

Software	Characteristic description	Reference
Confidence		
HoT	Sampling by reversing input sequence	(Landan and Graur 2007)
COS	Sampling by reversing input sequence according to the partition of the internode of guide tree	(Landan and Graur 2008)
Guidance	Sampling by perturbed guide trees by using the bootstrap method	(Penn et al. 2010b)
PASR	Sampling from pairwise comparisons between each single sequence and the sub-alignment. Only for DNA	(Kim and Ma 2011)
ZORRO	Probabilistic masking program based on pair Hidden Markov Model	(Wu et al. 2012)
TCS	Quantify the transitive consistency of the aligned pair of MSA with pairwise alignment	(Chang et al. 2012a)
Consensus		
ComAlign	Combing good sub-alignments	(Bucka-Lassen et al. 1999)
M-Coffee	Consistency framework	(Wallace et al. 2006)
MergeAlign	Dynamic programming in weighted directed acyclic graph	(Collingridge and Kelly 2012)
Filter		
Gblock	Two models: strict, relaxed. It is suitable for close related sequences	(Talavera and Castresana 2007)
TrimAl	Three models: gappyout, strictplus, automated1	(Capella-Gutierrez et al. 2009)
BMGE	Conservation approach based on entropy value weighted with BLOSUM or PAM similarity matrices.	(Criscuolo and Gribaldo 2010)

2. IMPROVING MULTIPLE SEQUENCE ALIGNMENT

2.1 Accurate multiple sequence alignment of transmembrane protein

Jia-Ming Chang, Paolo Di Tommaso, Jean-François Taly, Cedric Notredame

BMC Bioinformatics. 2012; 13 (Suppl 4): S1

2.2 Improving the alignment quality of consistency based aligners with an evaluation function using synonymous protein words

Hsin-Nan Lin, Cédric Notredame, Jia-Ming Chang, Ting-Yi Sung, Wen-Lian Hsu

PLoS ONE. 2011; 6 (12): e27872

3. IMPROVING SUBSEQUENT PHYLOGENETIC ANALYSIS

Title: Phylogenetic tree reliability can be improved using alignment uncertainty

3.1 Abstract

Motivation: Most evolutionary analyses are based upon pre-estimated multiple sequence alignment models. From a computational point of view, it is as complex to estimate a correct alignment, as it is to derive a correct tree from that alignment. Wong et al. established the uncertainty that multiple alignment procedures induce when reconstructing phylogenies. They were able to show that in many cases different aligners produce different phylogenies, with no simple objective criterion sufficient to distinguish among these alter-natives.

Result: We show that it is possible to significantly increase the dis-criminative capacities of bootstrap measures used to estimate phylogenetic trees reliability. Our procedure involves concatenating several alternative multiple sequence alignments of the same sequences, produced using different commonly used aligners. Concatenated alignments are used to draw bootstrap replicates. We named this method Weighted Partial Super Multiple Sequence Alignment (wpSMSA). On a collection of 853 datasets made of 7 one-to-one yeast orthologues, wpSMSA significantly improves the capacity to discriminate between topologically correct and incorrect trees. Over-all, we show that the combined use of wpSMSA trees along with single aligners bootstrap value makes it possible to identify 68% of the correct trees with a confidence of 92%. In contrast, a single method can only identify 14% of the correct trees at a similar confidence level. Bootstrap values were estimated for entire trees and are therefore suitable for large scale filtering. The values themselves are comparable to similar readouts estimated using a single method.

Availability: The automated generation of replicates has been implemented in the T-Coffee package, which is available as open source freeware from www.tcoffee.org

3.2 Introduction

Phylogenetic reconstruction tools are among the most widely used modeling methods in biology (Stamatakis 2006; Guindon et al. 2010; Tamura et al. 2011). Phylogenetic reconstruction has now become a method of choice for a wide range of applications that range from regulatory network evolution analysis (Brawand et al. 2011) to protein structure comparison (Magis et al. 2010). The availability of an increasing amount of sequence data, obtained by high throughput sequencing, is rapidly amplifying this trend (Rokas et al. 2003). Nonetheless, correctly estimating phylogenetic trees remains a challenging task from both computational and biological standpoints.

A phylogenetic tree is a binary representation of an evolutionary scenario where each node represents either duplication or a speciation event. Phylogeny aims at reconstructing the most likely scenario, that is to say the one best supported by the variations measured when comparing the sequences. Assuming one knows perfectly how mutations separate each pair of sequences, the task of turning this estimate into the most likely tree is NP-Complete under its most common formulations.

The problem of reconstructing a tree usually starts with precisely estimating the number of mutations that have occurred in each sequence. Two confounding effects hamper this quantification. The first one, known as multiple sampling, is a consequence of successive mutations altering the same site and occasionally reverting it. Ignoring multiple sampling results in an under-estimation of evolutionary distances. This problem is well known and can be addressed using a wide range of corrections involving more or less sophisticated evolutionary models (Kimura 2 parameters, etc.). The second confounding factor is the uncertainty inherent to any sequence alignment (Rost 1999; Capriotti and Marti-Renom 2010). This phenomenon, often referred to as twilight zone, results from protein alignments tendency to decrease in accuracy when dealing with sequences less than 25% identical (60% for RNA). Below these ranges, similarity measures based on pairwise alignments are usually over-estimated. This issue can be addressed by estimating distances from a MSA rather than through pairwise comparisons. Computing accurate MSAs, however, comes with issues of its own.

Estimating pairwise distances on an MSA yields at least three immediate advantages. Firstly, pairwise distances measured on an MSA tend to be more accurate because they result from the entire dataset. Secondly, the intrinsic nature of the MSA model means that pairwise projections are all constrained to remain consistent with one another, thus resulting in pair-wise distances more likely to be ultra metric, a very useful property when using distance based methods like neighbor joining. Thirdly, the MSA materializes columns of homology that can span the whole dataset hence making it possible to estimate local evolutionary models, for in-stance when using Maximum

Likelihood (ML) methods. These benefits come at a cost and MSA estimation is also an NP-Complete problem under its most common formulations (Wang and Jiang 1994), including Sankoff's that requires the simultaneous computation of a phylogenetic tree along with its underlying MSA (Sankoff et al. 1973). This problem has been addressed using a wide range of iterative heuristics such as PRRN (Gotoh 1996), MUSCLE (Edgar 2004b) and SATé (Liu et al. 2009). All the methods can be loosely described as alternative optimization protocols. None of them can guarantee optimality, even within some limits. As a consequence they tend to produce MSAs of comparable accuracy but often significantly different from one another.

Not being able to define and estimate unambiguously a correct MSA is a major problem when doing phylogenetic reconstruction. Most available aligners have been optimized for their capacity to reconstruct structure based sequence alignments while using sequence information only. They rely on a variety of scoring schemes (objective function) that, given the same dataset, explicitly define different optimal alignments. Wong et al. have recently exposed practical consequences of this problem when reconstructing phylogenetic trees (Wong et al. 2008). They have shown that given a collection of 1502 dataset of homologous sequences, the seven most widely used aligners only agree on less than 50 % of the dataset. In other words, for any given dataset there is less than 1 chance out of 2 to obtain a unique tree topology when applying the seven most widely used aligners.

Until recently, this problem had been all but ignored by the community, with the vast majority of published trees relying most of the time on a single ClustalW MSA. When doing so, the issue of MSA reliability is usually addressed using a post processing method, like trimming, in order to systematically remove the portions of an MSA unlikely to be correct (Castresana 2000; Capella-Gutierrez et al. 2009) even though recent results suggest trimming to have only a limited impact on phylogenetic estimation (Liu et al. 2009; Dessimoz and Gil 2010; Saurabh et al. 2012b). One may also argue that by removing some of the uncertainty inherent to a dataset, trimming results in an alteration of the signal to noise ratio, a process that can hamper robustness estimates. New protocols are now addressing this problem through systematic MSA sampling. For instance, the Heads-or-Tails (HoT) procedure involves aligning the reversed sequences and comparing the direct and re-verse versions of the MSA (Landan and Graur 2007).

Algorithmically speaking, HoT amounts to systematically reverse the order in which ties are broken when processing the dynamic programming matrix. Therefore, it only allows for two replicates. PRANK uses a more sophisticated sampling strategy, where each MSA replicate is estimated by randomly breaking all dynamic programming ties (Loytynoja and Goldman 2008). The most elaborate (and time intensive) protocol is probably GUIDANCE where MSA replicates are obtained by re-estimating progressive MSAs using guide trees obtained from the bootstrap replicates of a seed MSA (Penn et al. 2010a; Penn et al. 2010b). All these sampling strategies produce a similar output, in the form of an index summarizing the alignment robustness each residue across the MSA sampling process.

Various benchmarks, by others and us have shown these indexes to be very informative as accuracy estimators, and comparable in their specificity. They address the uncertainty issue raised by Wong et al. but they stop short of providing a definite answer to the effect of MSA uncertainty onto phylogenetic reconstruction, especially when dealing with non-controlled cases. In this work we are proposing an alternative method that precisely addresses this issue. We show that rather than combining the alternative MSAs into a unique consensus model or to locally trim them, one can concatenate the alternative MSAs and use them to draw bootstrap replicates. Doing so results in a bootstrap value that simultaneously reflects evaluative sampling (as does regular bootstrap) along with the uncertainty induced by the MSA procedure. While this procedure does not improve the tree accuracy, it makes the global bootstrap index more informative and therefore more useful when doing large high-throughput automated analysis.

3.3 Methods

a) Reference dataset

Validation was made using a reference collection of orthologous datasets adapted from Wong et al. Wong's collection consists of 1502 one-to-one orthologous datasets estimated using 7 yeast complete genomes with the phylogeny extending back >100 Ma. Each dataset comes along with seven alternative MSAs: ClustalW (Thompson et al. 1994), Dca (Stoye 1998), Dialign2 (Morgenstern 1999), MAFFT (Katoh et al. 2002), Muscle (Edgar 2004b), ProbCons (Do et al. 2005) and T-Coffee (Notredame et al. 2000) and their corresponding PAUP Maximum Likelihood (PAUP ML) tree. In order to compile a more phylogenetically homogenous set and avoid false orthologs, we selected the 853 datasets for which at least one of the 7 aligners yields the established yeast Tree of Life (ToL) topology which was shown in Figure 4 of Rokas's paper (Rokas et al. 2003), as estimated using the Robinson and Foulds topological comparison implemented in treedist (Felsenstein 1989).

b) Tree computation

For each dataset we used the seven alternative MSAs to estimate PAUP ML and its associated bootstrap support, as defined by Wong et al. Under this formulation, bootstrap is estimated for an entire tree and defined as the fraction of bootstrap replicates for which PAUP ML recovers the topology of the original tree estimated on the complete MSA. In addition to the original seven MSAs, we used the T-Coffee package to produce another two MSA models: a consensus alignment, estimated with the M-Coffee mode, and a super multiple sequence alignment (SMSA) containing a con-catenation of the seven individual MSAs. The term Super-MSA refers to the super-matrix procedure (Delsuc et al. 2005) where several genes are combined in order to estimate a tree. SMSAs were used to draw bootstrap replicates by either drawing a number of columns identical to the full SMSA length, or by drawing a number of columns equal to the average length of the concatenated MSAs. The last procedure is referred to as partial SMSA bootstrap (pSMSA). Individual MSAs, consensus and SMSA were all estimated in one single operation using the T-Coffee package:

t_coffee –seq <dataset> -method msa_method1, msa_method2, ... -log <concatenate> -outfile <conscensus>

Alignment procedure is done followed by Wong et al. so alignments are done through backtranslating from protein to DNA.

c) Weighted sampling scheme

Different alignment programs may use similar strategies such that columns of SMSA are not generated independently. In order to reflect the similarity across concatenated MSAs, a weighting scheme was designed based on the column similarity across the alignments. For each MSA_x, the weight is defined as:

$$W\left(MSA_x\right) = 100 - \frac{\sum_{i \neq x}^{N} ColumnSim\left(MSA_x, MSA_i\right)}{N-1}$$

where N is the number of concatenated MSAs and *ColumnSim* is the percentage of columns in MSA_x that are found identically aligned within MSA_i, as returned by the aln_compare algorithm:

```
t_coffee -other_pg aln_compare -al1 <aln1> -al2 <aln2> -compare_mode column
```

The resulting weight is then used when drawing replicates from the concatenated dataset, with each column having a probability to be selected proportional to the weight of the MSA it comes from. According to this weighting scheme, an alignment program similar with others will contribute less during bootstrap procedure.

d) Evaluation

A strategy was designed to compare the relative merits of all the bootstrap strategies described here. Given any MSA procedure and the associated collection of PAUP ML trees estimated on the 853 reference datasets, we proceeded as follows. Each tree topologically identical to the ToL was labeled as Proven Positive and the remaining trees were labeled as Proven Negatives. A bootstrap value was then estimated for every tree and the collection was sorted in reverse bootstrap order. This ordered list was then scanned in order to plot a true positive versus false positive graph. This same sorted set was also used to obtain Receiving Operator Characteristic curves and their associated Area Under the Curve (AUC) values with the ROCR R package (Sing et al. 2005).

3.4 Results

Our goal was to determine whether the uncertainty associated with MSA estimation could be incorporated within phylogenetic tree bootstrapping processes in order to refine tree robustness estimates. We started by analyzing the relationship between bootstrap support and alignment uncertainty. To that effect, we used the 1502 Wong's datasets made of 7 one-to-one yeast orthologues and binned them by the number of different tree topologies recovered when aligning each dataset with seven popular aligners and turning the resulting MSAs into PAUP ML trees. We then used Wong's methodology to estimate a global bootstrap for each tree and boxed-plotted the distribution by topological bin (Figure 3-1). Results clearly show that bootstrap values decrease when the number of distinct topologies increases. This rather intuitive finding indicates the existence of a relationship between MSA instability (i.e. alternative MSAs yielding different tree topologies) and low bootstrap support.

High bootstrap values are not direct indicators of phylogenetic accuracy. They merely reflect the homogeneity of the evolutionary sampling revealed by the MSA. Drawing an absolute correlation between phylogenetic correctness and bootstrap robustness is difficult because it would require access to validated reference phylogenetic trees, a scarce commodity. In theory, one could work with Wong's collection using the yeast ToL as a gold standard, since each dataset is supposed to be made of one-to-one orthologues. In practice, things are less straightforward, owing to the complex yeast genome history, with the suspicion of several duplication rounds (Wolfe and Shields 1997) and potential linage sorting effect. It is therefore unclear for which fraction of Wong's datasets the ToL can be considered the true history. We addressed this problem by identifying within Wong's collection a subset of sequences likely to be enriched in ToL-like evolutionary relation-ships. We did so by selecting the 853 families for which at least one of the 7 aligners yields an MSA supporting the ToL. Such datasets are not guaranteed to be ToL compliant, but it is reasonable to expect some enrichment. Indeed, if one assumes the 7 MSAs in each dataset to be independent (which they are not) and any MSA with PAUP_ML leads to a purely random tree topology (which is not totally appropriate as tree reconstruction), the probability of obtaining by chance the unrooted ToL topology is, which is about 1% (or even less if one considers the MSAs non independent). Out of 853 datasets, one would therefore expect in the order of 6 mistakenly selected datasets. This figure is approximately 10 times smaller than most differences reported in subsequent benchmarks and can therefore be considered an acceptable error margin.

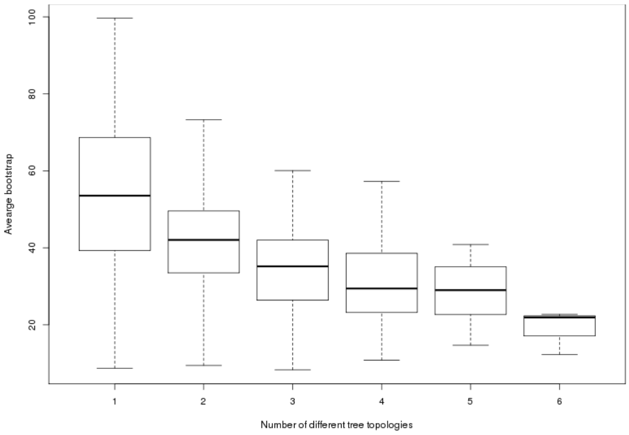

Figure 3-1 Number of different tree topologies versus average bootstrap values measured across the seven individual aligners.

a) MAFFT provides the most informative bootstrap among individual aligners

We used the 853 family collections to compare alternative bootstrap strategies on alternative MSA methods (Table 3-1, Figure 3-2a). Our intention was not to compare the tree reconstruction accuracy, but rather to quantify the discriminative power of various bootstrap procedures. We estimated a PAUP ML tree on each MSA produced with each of the seven aligners and labeled the resulting trees as Proven Positives when their topology recapitulates the ToL or as Proven Negatives otherwise. We then estimated the bootstrap support for each tree and sorted, for each aligner, the trees in decreasing bootstrap order to plot True Positives (TPs) versus False Positives (FPs) curves (Figure 3-2a, Table 3-1). We found the individual aligners to be very comparable in terms of overall accuracy. They all manage to reconstruct a similar number of trees having the ToL topology, with MAFFT being the most effective method (665 ToL topologies) and T-Coffee the less accurate (620 ToL topologies). The overall topological correctness is a poor indicator of suitability for an aligner, because it does not give any indication on a method discriminative capacity and leaves users unable to select the trees most likely to be correct. To complement this figure, we measured the number of reported TPs for a certain number of accepted FPs (10 and 25). For both these values, MAFFT performed best with ClustalW and DCA was the less accurate. When measuring the Person Correlation Coefficient between the number of reported TPs at a given FP threshold and the average bootstrap values (Table 3-1), we found the 25 FPs limit to be significantly more correlated than the 10 FPs limit (0.77 and 0.44 respectively).

Table 3-1 Average bootstrap, AUC values and the number of TPs for 10 and 25 accepted FPs of each method.

			TPs		
method	ave. bootstrap	AUC	for 10 FPs	for 25 FPs	total
ClustalW	51.31	0.7521	185	274	643
DCA	50.62	0.7694	194	284	624
DIALIGN	51.94	0.7618	253	340	659
MAFFT	52.82	0.7750	253	359	665
Muscle	52.35	0.7771	224	315	639
Probnt	50.96	0.7790	256	312	642
T-Coffee	51.21	0.7889	234	311	620
M-Coffee	51.41	0.7688	193	325	646

SMSA	77.31	0.8301	329	425	661
pSMSA	50.96	0.8140	342	385	661
wpSMSA	50.86	0.8215	353	423	661

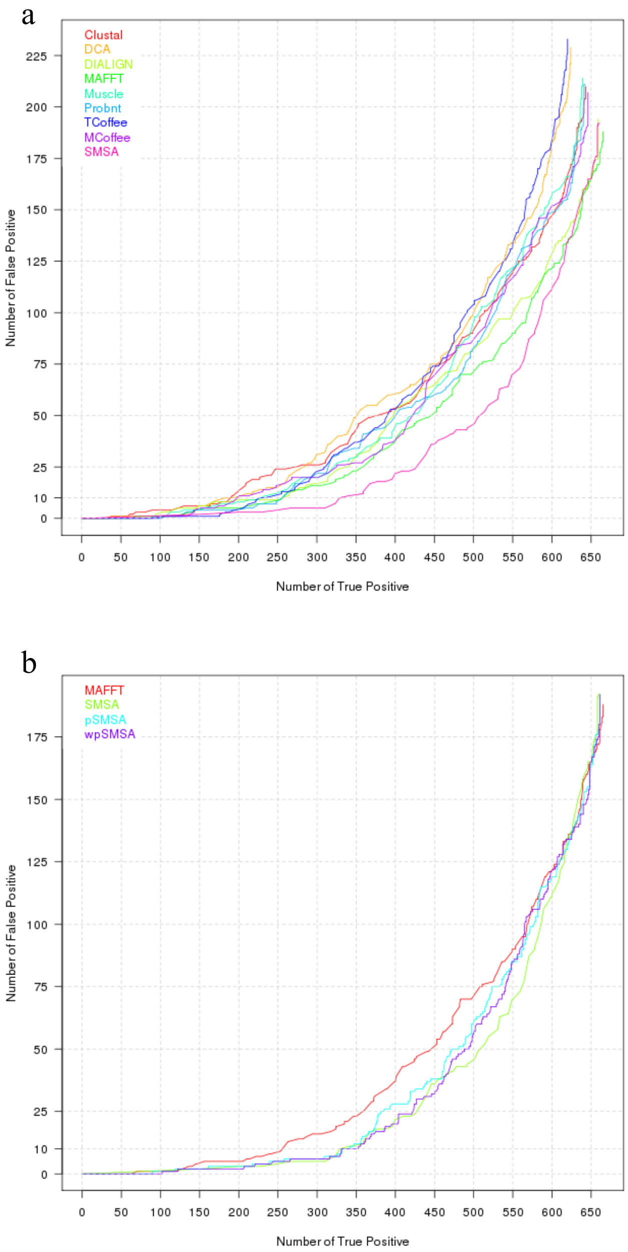

Figure 3-2 The number of FPs (non-ToL) versus the number of TPs (ToL) analysis on a) individual aligners and SMSA b) alternative sampling procedures.

b) SMSA performs better than MAFFT but artificially increases bootstrap scale

We then used this same dataset and approach to estimate the effect of combining individual MSAs into either a consensus MSA (build with M-Coffee) or into a concatenated set of MSAs (Super MSA, SMSA). Results show that M-Coffee, the consensus, is a poor performer delivering only slightly more than the average of individual aligners on most readout considered here. For instance, when considering the number of TPs for 25 FPs, M-Coffee ranks 5th out of 8 considered aligners. SMSA, the concatenated alignments, yields very different results. If one excludes the average bootstrap that will be discussed in the next paragraph, the SMSA procedure outperforms all the single aligners and appears to be dramatically more discriminative. For instance, at the 25 FP threshold, SMSA is 18% (435 vs. 359) more discriminative than MAFFT and on average 35% more than individual aligners. Similar levels of improvement can be measured when considering the 10 FPs threshold or the AUC that summarizes the sensitivity/specificity trade-off of a method across all possible thresholds (Table 3-1). This improvement does not come at a significant cost in overall accuracy and SMSA only delivers a few ToL topologies less than MAFFT (661 vs. 665).

Bootstrap values on SMSA come in a range that does not make them comparable to bootstrap values on standard alignments. For practical usage, these values would need to be re-calibrated each time one changes the concatenated methods or the number of methods. The reason why bootstrap values increase so much when drawing replicates from the SMSA has to do with the non-independence of the concatenated MSAs that induce a multiple re-sampling effect. In practical terms, concatenating non-independent MSAs amounts to artificially increasing self-agreement within the dataset, hence the inflated bootstrap values. This increase does not, however, correspond to any improved discriminative capacity. One can easily show this by self-concatenating all the individual MSAs seven times in a row. The result is an increase of average bootstrap value from 51.6 up to 78.8 with no significant variation in discriminative capacity. The increased bootstrap values and increased discrimination capacity we report on Table 3-1 are there-ore disconnected observations.

c) wpSMSA as good as SMSA and its bootstrap compatible with individual aligner

We hypothesized the possibility to counterbalance the multiple-sampling effect by generating shorter replicates. To that effect we re-ran a bootstrap procedure where each replicate only contains a number of columns equal to the average size of the concatenated MSAs, i.e. 1/7 of the total SMSA length (Zharkikh and Li 1995). We compensated the potential information loss by drawing 7 times more replicates (700 rather than a 100). This protocol, named Partial SMSA (pSMSA), behaved exactly as anticipated. Its various readouts suggest accuracy comparable to SMSA and bootstrap values in the same range as those measured on individual aligners. Partial bootstrap is not, however, an entirely satisfying solution to the multiple sampling issue. It does address the mechanical bootstrap inflation induced by concatenation, but ignores the fact that all aligner methods are not equally different and do not contribute equal amounts of information. Some are more likely to produce similar MSAs, as a consequence of arbitrary algorithmic implementation similarities. Such methods will be effectively over-weighted when doing the sampling. One can easily address this problem by adding a corrective weighting scheme to the sampling. We selected a scheme ("Weighted Sampling Scheme" section, Methods) that up-weights MSAs in which sequences are aligned differently from the rest of the concatenated MSAs. Columns from such MSAs are more likely to contribute a column when drawing replicates. Results (Table 3-1, Figure 3-2b and 3-3) show a net improvement of the Weighted Partial SMSA (wpSMSA) over most alternative procedures with no significant variation of the bootstrap value range. Outputting compatible bootstrap value to individual aligner makes wpSMSA practical because a usual bootstrap threshold can be also applied to wpSMSA.

In order to establish the relationship between the wpSMSA bootstrap value, tree topological correctness and similar readouts on the individual aligners, we plotted the respective bootstrap values on the 853 families (Figure 3-3). On this plot, the most striking feature is the high level of topological correctness for trees having a bootstrap value higher than 60 (Table 3-2). Nearly 98% of the 248 trees in this range are topologically correct as opposed to a mere 67% (=382/565) below the 60% bootstrap limit. In the lower range, one can clearly see that correct topologies (in blue) are more often above the main diagonal than below. This observation can be quantified (Table 3-2) and it appears that whenever the wpSMSA bootstrap is higher than the average bootstrap value measured on single aligners, the resulting tree is 86% (=213/247) more likely to be topologically correct. This observation suggests that there exists a very informative relationship between the individual MSA bootstrap readouts and their concatenated counterpart, with the confrontation of these two quantities likely to improve interpretation. Decreases in bootstrap values within the wpSMSA are by contrast less informative.

Table 3-2 The average bootstrap comparison between wpSMSA and single aligner respect to whether the tree topology of wpSMSA is identical to ToL or not.

	wpBS>single BS	wpBS=single BS	wpBS<single BS
Overall			
wpSMSA == ToL	383	2	276
wpSMSA != ToL	38	1	153
wpBS and single BS > 60			
wpSMSA == ToL	146	1	96
wpSMSA != ToL	3	0	2
wpBS and single BS < 60			
wpSMSA == ToL	213	1	168
wpSMSA != ToL	34	1	148

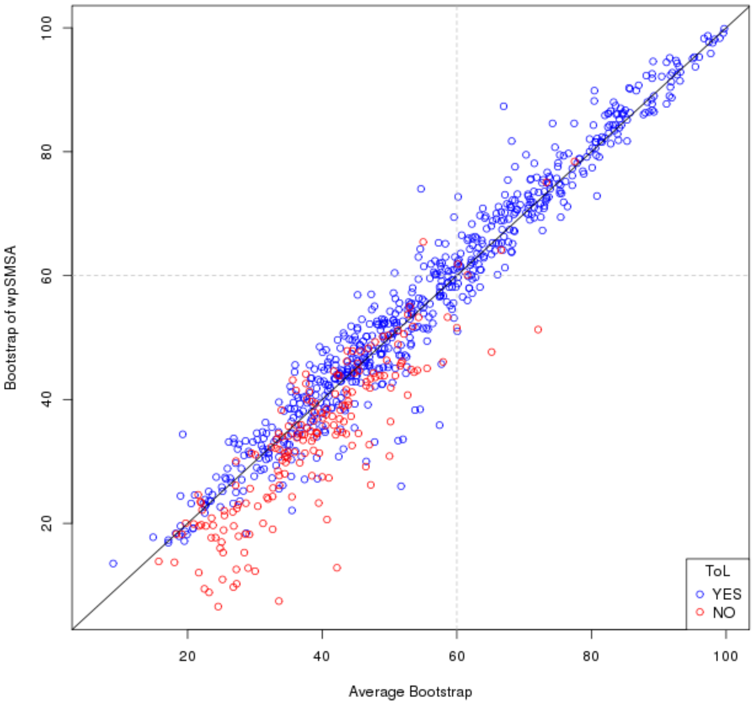

Figure 3-3 The bootstrap of wpSMSA versus the average bootstrap of individual aligners. Blue points indicate datasets for which the wpSMSA topology is identical to the ToL, red points otherwise.

3.5 Discussion

One of the most complex aspect of the work reported here is the notion of an “aligners collection”. Such collections exist, merely reflecting the difficulty of properly aligning multiple sequences. More than 50 MSA methods (Kemena and Notredame 2009) have been reported over the last 20 years, and choosing the right number and combination of aligners is important. As for the collection used here, no strong rationale exists except that these aligners have been shown to perform well and have been evaluated for their tendency to deliver different alignments (Wallace et al. 2006). The number of concatenated MSA is an issue that can more easily be addressed. We did so by systematically considering all possible combinations of an increasing number of methods (Figure 3-4). We found the combination of at least three aligners to be a critical point. With seven aligners, bootstrap increase appears to flatten; yet, the sharp increase of AUC suggests that there might be some merit in combining a few more methods.

We report a variation of standard bootstrap procedures and its validation on a reference dataset. While bootstrap is usually carried out in order to quantify the effect of uneven evaluative sampling across homologous genomic sites, we show here how minor adaptation of standard protocols makes it possible to integrate within this process MSA uncertainty and its effect on phylogenetic reconstruction. Our procedure is relatively simple as it only involves computing alternative MSAs using different methods, comparing the MSAs in order to weight them, concatenating them and drawing partial replicates from the weighted columns. All other aspects of the tree reconstruction are left to external third party methods (PAUP ML in the context of this work). This approach is named wpSMSA. Its main merit is to combine within a single numerical value the combined effect of evolutionary sampling and MSA uncertainty. This combination is a very desirable property, considering that evolutionary sampling and MSA uncertainty are strongly correlated and virtually impossible to disentangle from one another. Once combined, they become less of confounding factor. The effect of our procedure is not to derive more accurate phylogenetic trees, but rather more informative bootstrap values. For instance, we show that when ranking trees by their bootstrap values, wpSMSA is much more accurate than alternative method at separating correct and incorrect trees on the basis of their bootstraps.

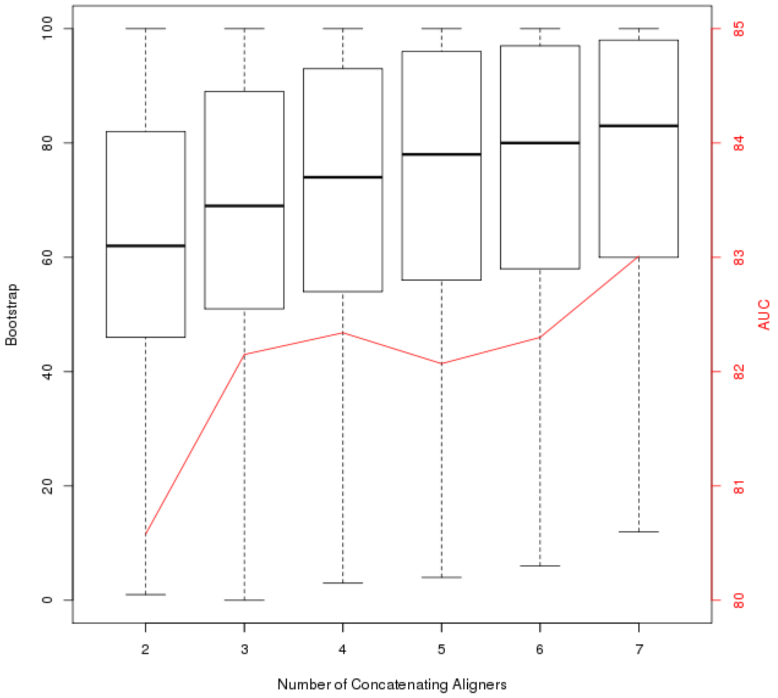

Figure 3-4 The average bootstrap and AUC (red line, %) of concatenating different number of aligners.

The main issue with the concatenation of non-independent MSAs is bootstrap values inflation; a phenomenon that would require sophisticated calibration if one is to use our method in a standard set up. We have addressed this issue by designing and validating a partial bootstrap procedure that makes the wpSMSA bootstrap support directly comparable with that of single aligners. When using these values and comparing them with single aligners, we could show that above 60 units of bootstrap, about 97% (=146/149) of the reported topologies are accurate, regardless of the method. Yet, wpSMSA reports 52% (146 vs. 96) more correct topologies than single aligners do. This effect is especially significant when considering that wpSMSA bootstrap is slightly lower than that measured on individual aligners. Below 60 bootstrap units, the combined use of our method also makes it possible to identify correct topologies. We show that whenever the wpSMSA bootstrap is higher than the average value estimated on individual MSAs, the resulting trees are 86% likely to be topologically correct. On a dataset like the one analyzed here, such a filtering would have revealed 213 correct trees (out of a total of 661 ToLs). Overall, if one had been using a 60 units threshold to decide between correct and inaccurate trees, one would have identified 22% (=146/661) of the correct trees with our method (with 3 FPs). If one had also been using the bootstrap shift effect below 60 bootstrap units, this figure would have increased to

68.8% (= 243+213 / 661) with a confidence level of about 92%.

One class of MSA programs are designed by focusing on phylogenetic constrain. Blackburne and Whelan pointed out this class usually cluster together compared to another class, similarity based MSA methods in terms of similarity of alignments (Blackburne and Whelan 2012a). Two evolutionary based methods, PRANK v.121018 (Loytynoja and Goldman 2008) and SATé v2.2.5 (Liu et al. 2012), are selected as representative for comparison with wpSMSA. PRANK is run under default parameters (-protein) and SATé is run under auto mode ("--auto"). Figure 3-5 shows the TP versus FP analysis. SATé performs better than the best individual aligner, MAFFT, but still worst than wpSMSA, that is, wpSMSA produces more TPs than SATé under the same FP threshold (353 versus 313 at 10 FPs, 423 versus 395 at 25 FPs).

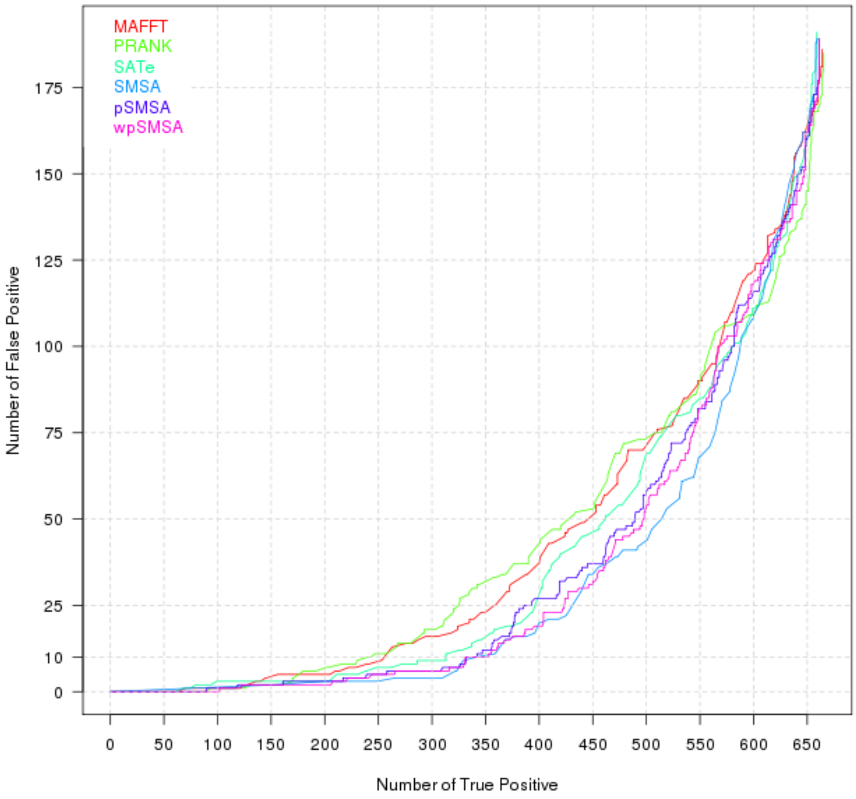

Figure 3-5 TP versus FP analysis of PRANK (green line) and SATé (cyan line).

3.6 Conclusion

The approach developed here leaves untouched the validation of many key properties of the methodology. For instance, one would want to know how informative is the node support delivered by wpSMSA. With only 7 species, our reference dataset is probably not the best to ask such detailed questions, and one would probably need a larger set of species. The other critical component of wpSMSA is the number of combined methods. When doing the first validation of a consensus MSA approach, Wallace et al. did show the various levels of correlation that exist among methods (Wallace et al. 2006). We also show how wpSMSA is sensitive to the number of combined methods but stop short of proposing a methodology suitable for the assembly of an optimal aligner cocktail. It is nonetheless obvious that a change in the combined meth-od will most likely have an effect on the reported trees. One may therefore argue about our aligner choice being too arbitrary. This is certainly true; on the other hand, the fact that alternative aligners fail to agree merely results from our incapacity to estimate optimal MSAs. In this context, each aligner can be seen as arbitrary heuristics in its own right. Their combination is therefore another heuristic aligner, better suited for estimating phylogenetic trees. Because our approach has been designed to handle both the use of non-independent models and the effect of uneven correlations, it should remain robust, regardless of the number and the nature of combined methods and one may even consider combining alternative alignments generated while exploring some alignment parameter, like gap penalties. Another important question would be to deter-mine how sensitive the wpSMSA approach would be to dataset difficulty. The reference benchmark assembled here is not very challenging and one may therefore consider asking if the accuracy figures we report here will hold when dealing with larger datasets or sequences harder to align. It is reasonable to expect that concatenation will remain more informative than single aligners, if only because it incorporates in the global readout an element about MSA overall accuracy.

The approach is very generic and makes it easy to systematically integrate in a phylogenetic analysis the potential biases that may require systematic sampling, including dynamic programming artifacts or guide tree effects as pointed out by Löytynoja and Goldman (Loytynoja and Goldman 2009). However, inferring posterior probabilities across a set of MSAs from different aligners is challenging (Blackburne and Whelan 2012a) such that how to model those uncertainty into BAliPhy (Redelings and Suchard 2005) framework remains open.

4. IMPROVING MODEL USAGE

Title: Transitive consistency provides a Unified Framework For Homology and Evolutionary Modeling

4.1 Abstract

Motivation: Multiple sequence alignment (MSA) is a key modeling procedure when analyzing biological sequences. MSAs are often used as primary models when estimating phylogenetic, functional or structural relationships. It has long been known that the accuracy of MSAs can affect the accuracy and the reliability of these inferences. This effect results from an uncertainty inherent to the computation of the MSA models. Recently, new methods have been developed to show how these phenomena can be quantified in order to increase the trustworthiness of MSA models and subsequent analysis based upon them. Such methods include Heads-or-Tails (HoT) and GUIDANCE. The methods are powerful but often slow, not generic enough and of limited accuracy in some cases. Furthermore, no validation exists that would clarify if the same approach can be used to estimate the reliability of an MSA from different points of view such as a structural, a functional and an evolutionary one.

Results: Our approach is named TCS (T-Coffee Consistency Score). We show here that TCS can be used to identify reliable positions of BAliBASE structure-based sequence alignments, and it does so in a way superior to both HoT and GUIDANCE. Using the same approach, we show that this measure can be used to do a weighted replication that results in a more accurate tree topology, both on simulated and empirical reference datasets. A main strength of TCS is the possibility to apply it to any third party MSA model generated by any available method. Because it is based on an estimate of the consistency between an MSA and a library of pairwise alignment, our method can be used along with different types of such collections. We show here that the best results are obtained using an all against all collection, but we also found that good quality results can be obtained when populating the library with pairwise projections extracted from fast approximate MSAs.

Availability: TCS is part of the T-Coffee package, a web server is also available from http://tcoffee.crg.cat/core and a freeware open source code can be downloaded from http://www.tcoffee.org/Packages/Stable/Latest.

4.2 Introduction

Multiple sequence alignment (MSA) is an important initial step for many applications in biology, the main applications being phylogenetic reconstruction, structural homology modeling and functional inference through domain profile comparisons. More than 100 publications describing novel MSA methods have been published over the last 30 years (Kemena and Notredame 2009), and the MSA Wikipedia page lists 47 available MSA packages (http://en.wikipedia.org/wiki/Sequence_alignment_software). The accuracy of MSA is limited by both the complexity of the problem (known to be NP-Complete in all its useful formulations), and the difficulty to describe an MSA in mathematical terms to accurately reflect biological relationships. This explains why the problem is such an active field of research. Over the last years, new fronts have started emerging in the MSA research. Departing from the canonical attempt at generating more accurate algorithms and aligners, several groups have started exploring the issue of reliability and the feasibility of using approximate models along with some index indicating the trustworthy parts of the MSA. The rationale of this approach could be described as "A bird in the hand is worth two in the bush", which, in MSA words, can be translated into the higher biological relevance of a model that accurately describes a small fraction of the data compared with one describing the entire data with lower accuracy.

The main reason why MSA reliability fluctuates lies in our limited capacity to describe sequence homology, especially when dealing with distantly related sequences having less than 20% similarity. At this level of identity, the homology signal is nearly saturated and lower than background noise. Aligners that are meant to maximize similarity often over-estimate the level identity and yield inaccurate sequence alignment with an identity level often higher than the correct one. This global phenomenon can be heterogeneous across the sequences with heterotopia, the difference in evolutionary rates across sites, which is a common feature in protein sequences. In this case it is reasonable to expect significant variation in alignment accuracy, especially between the slowly evolving internal part of the proteins and the fast evolving exposed loops. This problem, which mostly results from our limited capacity to trace evolutionary relationships across long distances, is worsened by most alignment methods, which almost always rely on pairwise comparisons carried out by dynamic programming. When doing so, one uses the Needleman and Wunsch algorithm in order to estimate the relationship between two sequences. NW estimates the optimal edit score of two sequences and delivers a pairwise alignment having an optimal score. Under most formulations, there often exist more than one optimal alignment. In most implementations, the algorithm arbitrarily resolves the ties that may arise and always returns the same alignment. The order in which ties are resolved is sometimes referred to as “low-road/high-road”. Given two sequences, these arbitrary tiebreaks have little consequence. It is worth mentioning that the order in which the ties are resolved depends on the order of the sequences themselves. By swapping them, one may get a different alignment with the same optimal score. This issue is important when dealing with multiple sequence datasets.

On a pairwise alignment, these arbitrary decisions have little consequence. By contrast, they can be major shaping forces when computing a multiple sequence alignment. Indeed, each tiebreak may be seen as a flip of a coin bringing one bit of uncertainty. With most MSAs being resolved in a progressive way, and sequences aligned along the phylogenetic tree, the dynamic programming uncertainty would require all tiebreaks to be somehow resolved in a coordinated fashion, so that the resulting patterns remain compatible while the algorithm works its way along the guide tree and coalesces the sequences. Such coordination, however, would dramatically increase the algorithm complexity.

Many alternative solutions have been proposed to either deal with this issue or to take it into account when estimating MSA accuracy. The most complete solution is probably the MSA algorithm that estimates an optimal MSA in multi-dimensional space. The cost of MSA, however, becomes prohibitive when dealing with even moderately distant sequences. Another alternative, implemented in the PRANK algorithm, is to turn each MSA computation into a sampling across the tiebreak space (Loytynoja and Goldman 2008). In PRANK all ties are broken randomly and the resulting MSAs may therefore be seen as replicates whose robustness yields an indication of the model robustness. Consistency based progressive alignments constitute a powerful alternative to this expansive sampling strategy. In a progressive consistency framework, like T-Coffee, MSAs are estimated by maximizing their consistency with a set of pre-computed pairwise alignments. The main strength of this approach relies in its capacity to re-interpret all possible pairwise alignments. In the original implementation, the algorithm was computing for each pair of sequences two global pairwise alignments by feeding in the sequences to a pairwise aligner in the two possible orders. By doing so, the algorithm was populating a pairwise library, which was then used to estimate the propensity of every pair of residue to be aligned, given its compatibility with the rest of the library. This process dramatically decreases the number of ties and though it remains to be established, one may suggest that a large part of the reported improvement resulted from the synchronization of tiebreaks. In their implementation, the ProbCons authors went a bit further and populated the library with all sub-optimal pairwise alignments above some threshold. As noted by both groups, the resulting score for aligning two symbols reflects the support of the whole sequence dataset, and as such, it may be used as a reliability indicator. More recently, a method, named Heads-or-Tails (HoT) (Landan and Graur 2007; Landan and Graur 2008) was reported, based on the observation that MSAs may vary when aligning a set of sequences after flipping them from left to right. As discussed in subsequent publications this effect is due to a systematic inversion of the tiebreak order resulting from the inversion of the sequences. At the pairwise level, this only affects the alignment and not its score, but when dealing with an MSA in a progressive alignment framework, these effects usually add up and may result in significant differences across replicates.

The main motivation of HoT is not so much to reveal MSA instability, but rather to determine to which extent this instability can be used to estimate model reliability. In this case the authors used the estimate in order to show that phylogenetic reconstruction can be significantly increased when filtering out unstable positions. This concept was recently taken a bit further by the GUIDANCE approach (Penn et al. 2010b). In GUIDANCE, the authors showed how random guide trees could help identify the less trustworthy positions in an alignment, thereby increasing its phylogenetic reconstruction potential. These approaches depart significantly from the most common filtering procedures like Gblocks and trimAl that rely on a more static interpretation of the MSA and eliminate positions on the basis of their indel propensities. More dynamic filterings have, however, often been used when doing homology modeling. For instance, the CASPER server (Claude et al. 2004) uses the T-Coffee CORE index, originally described as a means to identify unreliable positions in an MSA in order to filter out them when doing homology modeling. In practice, any approach introducing an element of instability in an MSA can be used to estimate its robustness. As an alternative to random guide trees, the authors of PARS (Kim and Ma 2011) have recently shown that accuracy can be estimated by comparing alternative projections of the same sequences in an MSA when removing in turn every sequence and realigning the remaining set. These perturbation methods are in many ways comparable to a phylogenetic bootstrap, where a model is re-estimated through data re-sampling in order to estimate how well supported the final model is by the data. By contrast, consistency based methods are much less intensive, since they rely on a finite and limited sampling (typically quadratic with the number of sequences) and do not require any replication stage, as the model is generated along with the reliability estimates.

It is rather surprising to see that so far no filtering method has yet been analyzed for its combined capacity to reflect both structural and evolutionary correctness. This question is rather important as it amounts to asking whether correct structural alignments are a good substrate when growing phylogenetic trees. This issue is becoming highly relevant in a context where novel phylogeny aware methods like PRANK or SATe (Liu et al. 2009) that are meant to estimate MSAs suitable for phylogenetic reconstruction have recently been reported. In most cases, these methods tend to fare poorly on structure based reference datasets, while they perform with superior accuracy when reconstructing phylogeny on simulated datasets. The most common explanation is that evolutionary and structural reconstructions consider extreme processes in the evolutionary spectrum. We show in this work that this apparent discontinuity between structural and phylogenetic aligners may not exist. Our accuracy evaluation method uses consistency in order to estimate the reliability of every pair of aligned residue in an MSA. We show that this score correlates better than HoT or GUIDANCE with structural correctness on BAliBASE3 (Thompson et al. 2005a) or PREFAB4 (Edgar 2004b) reference MSAs. We also show that this accuracy estimation can be used to weight a standard bootstrap procedure in order to significantly increase the accuracy of the estimated trees. The result is that using that same methods we find the TCS score able to outperform all alternative filtering methods for the reconstruction of accurate phylogenetic trees, either on simulated or empirical datasets. We find this effect to be significant on simulated data and even more pronounced on real empirical datasets.

4.3 Methods

a) Transitive consistency measure

The transitive consistency measure presented here is an extended version of the T-Coffee scoring scheme. Given a library of pairwise alignments, this score is used to estimate the score of aligning two residues A_x and B_y from two sequences A and B of the MSA, by identifying all intermediate residues I_z from a third sequence I that may be part of two pairs A_xI_z and I_zB_y. Given the entire pairwise library, the reliability score is then calculated as a ratio between the sum of the weight of all A_xB_y pairs linked through an I_z residue defined as $TCS(A_x,B_y|I_z)$, divided by the sum of the score of all possible pair combinations involving A_x or/and B_y through an intermediate I_z. This formulation, shown below, amounts to estimating the fraction of all compatible pairs that support the alignment of A_x and B_y.

$$TCS(A_x, B_y) = 2\frac{\sum_I^S TCS(A_x, B_y|I_z)}{\sum_I^S TCS(A_x, B_*|I_z) + \sum_I^S TCS(A_*, B_y|I_z)}$$

Libraries were estimated using the three following protocols:

(1) T-Coffee original, in which the library is made by combining ClustalW (Thompson et al. 1994) and Lalign (Huang and Miller 1991) pairwise alignments. In this library, named TCS_original, pairs of residues are weighted by the average identity measured on the pairwise alignment they were extracted from.

```
t_coffee –seq=<seq_file> -method clustalw_pair, lalign_id_pair –out_lib
<lib.TCS_original>
```

(2) Libraries populated with pairwise alignments produced by the ProbCons pair-HMM (Do et al. 2005). In these libraries (TCS), weights are defined as the posterior probability of the two considered residues being aligned.

```
t_coffee –seq=<seq_file> -proba_pair –out_lib <lib.TCS>
```

(3) Libraries built using the pairwise projections extracted from several fast multiple aligners (TCS_FM). In the context of this work, we used the protocol developed for the Ensembl Compara pipeline that combines MAFFT (Katoh et al. 2002), MUSCLE (Edgar 2004b) and Kalign (Lassmann and Sonnhammer 2005).

```
t_coffee –seq=<seq_file> -method kafft_msa,kalign_msa,muscle_msa –out_lib
<lib.TCS_FM>
```

These three kind of libraries (TCS_original, TCS, TCS_FM) were then used to evaluate MSAs produced by three alternative methods: ClustalW 2.1 (Larkin et al. 2007), MAFFT 6.711 (Katoh and Toh 2008), and MUSCLE 3.8.31(Edgar 2004a). T-Coffee was voluntarily excluded in order to rule out any interference between the actual computation and the subsequent evaluation. MSAs were evaluated using the following command:

t_coffee –infile=<target MSA> –evaluate –lib <library> [-output score_ascii,sp_ascii]

, where *sp_ascii* is a format reporting the score of every aligned pair in the target and *score_ascii* is a command reporting the average score of every individual residue along with the average score of every column.

b) Structural reference datasets

Two datasets were used in order to estimate structural correctness: BAliBASE3 (Thompson et al. 2005b) that contains 218 sets classified in 5 categories. BAliBASE datasets contain several sequences having a known structure and have annotated blocks in which the structural superposition is considered reliable and fit for benchmarking. We also used PREFAB4 (Edgar 2004a), a much more extensive collection where each set is made of about 50 sequences embedding 2 sequences with a known structure. The reference alignments come along with block indication suggesting the reliable positions for benchmark. PREFAB4 is classified into four groups: 0~20, 20~40, 40~70 and 70~100 according to the pairwise identity of reference sequence. RV11 of BAliBASE3 and 0~20 of PREFAB4 are the most challenging sets because their sequence identity falls in the Twilight Zone (Rost 1999). RV11 has been shown to the most informative subset across all these categories (Kemena, 2009).

c) Structural evaluation procedure

In order to compare alternative evaluation methods, like HoT and GUIDANCE, the score of every aligned pair was estimated using TCS, HoT or GUIDANCE. Pairs containing residues that are part of the reference block were then extracted, labeled as either Proven Positives (when they corresponded to the reference) or Proven Negatives otherwise. The list of ordered pairs was then used to do a Receiver Operator Curve (ROC) and the Area Under Curve (AUC) was estimated in order to compare performances with the ROCR R package (Sing et al. 2005). Subgroups of BAliBASE3 and PREFAB4 reflect different protein properties. Average AUC was computed for each BAliBASE and PREFAB subgroup. We also used the provided packages to estimate the BAliScore and the PREFAB score on all the considered MSAs

d) Phylogenetic validation

Reference datasets. Validation was done using the Gblocks (Talavera and Castresana 2007) simulated dataset (16 tips), trimAl (Capella-Gutierrez et al. 2009) (32 and 64 tips) and an empirical dataset. Both simulated datasets were generated by their respective authors using Rose (Stoye et al. 1998). We only used the asymmetric mode reported to be the most challenging. Alignments were constructed on the 16 tips dataset using ClustalW, MAFFT and ProbCons. On the 32 and 64 tips, we only used MAFFT. Maximum Likelihood (ML), Neighbor Joining (NJ) and Parsimony trees were estimated on each MSA and the benchmark was carried out using the Gblocks protocol. The empirical dataset was extracted from Wong *et al.* (Wong et al. 2008) in which the authors assembled 1502 clusters of 7 orthologues that they aligned with 7 aligners (DCA, ClustalW, Dalign, MAFFT, Muscle, ProbCons and T-Coffee) in order to estimate as many phylogenies using Maximum likelihood (PAUP). Out of this set we selected the 853 datasets in which at least one aligner yields a phylogeny identical to the canonical yeast ToL (Rokas et al. 2003).

Filtering procedures. In order to evaluate and compare filtering procedures, MSAs were filtered with Gblocks using the stringent and the relaxed procedure that keeps all positions containing less than 50% of gapped positions. trimAl was benchmarked in two modes: *gappyout,* which automatically selects gap cut-off score depending on MSA's gap distribution and *strictplus,* which automatically selects block size.

TCS replicates. Rather than being used as a filtering score, the TCS scoring scheme was used as a weighting scheme when building replicates for the benchmarking. Under this scheme, the number of times a column is chosen for replication is proportional to its average score.

The MSAs and the trees were estimated using computation on the Amazon elastic cloud. Tree accuracy was estimated using the Robinson-Foulds distance measure implemented in *treedist* to compare the target trees with their reference (Felsenstein 1989).

4.4 Results

a) Prediction of structural accuracy

We first computed the BAliBASE MSAs using ClustalW, MAFFT and Muscle (Table 4-1). We found the Sum-of-Pairs (SPs) accuracy to be in broad agreement with reported figures in the literature. We then used the ROC approach described in the methods section to test the capacity of our scoring schemes (HoT, GUIDANCE and TCS) to separate between accurate and inaccurate pairs of aligned residues (as judged from comparison). By this criterion, we found the TCS to outperform both GUIDANCE and HoT on BAliBASE. We also found the TCS to be much more robust across aligners, and being little affected by the overall method accuracy. We then refined the analysis by only considering the behavior of the best method (MAFFT) on the extreme datasets, 'easy' and 'difficult', of BAliBASE and PREFAB (Table 4-2). This analysis confirmed the superiority of the TCS scoring scheme, which is much less affected than its counterparts by variations in accuracy.

Table 4-1 AUC/average AUC analysis of different confidence schemes for different alignments on BAliBASE 3 set.

	ClustalW	MAFFT	Muscle
SPs	0.714	0.807	0.793
TCS	96.46/98.80	94.44/95.81	94.51/96.37
HoT	90.95/96.72	82.66/89.87	-*
GUIDANCE	87.69/95.11	90.28/93.95	94.51/95.16

*HoT does not support the Muscle aligner.

Table 4-2 The average AUC of easy and difficult protein families from BAliBASE and PREFAB by MAFFT.

	difficult		easy	
	RV11	0~20	RV12	70~100
SPs	0.536	0.465	0.888	0.942
TCS	91.11	87.16	96.83	78.98
HoT	72.63	81.35	78.79	57.96
GUIDANCE	83.51	86.03	92.64	62.01

We then compared the effectiveness of alternative TCS protocols (Table 4-3). Our result indicates the superior discriminative capacity of ProbCons-based libraries. This protocol is about 3 times faster than GUIDANCE and significantly more informative. It is, however, interesting to note that the fast protocol is only modestly less accurate than GUIDANCE but nearly 10 times faster, which may be convenient when running high throughput pipelines.

Table 4-3 The average AUC (%) of different library protocols and their running time analysis

	BAliBASE	PREFAB	Time (s)
TCS	94.44	89.24	17,244
TCS_original	91.20	83.83	113,624
TCS_FM	87.28	80.03	6,957
GUIDANCE	90.28	85.74	66,368

b) Overall alignment correctness

Establishing the relative accuracy of individual pairs of residues within an MSA has limited practical applications. In reality, one is often more interested in deciding objectively between 2 or more alternative MSAs. We therefore asked if TCS is a suitable method to compare alternative MSAs of the same sequences. In order to estimate this capacity, we did a non-parametric analysis by estimating how often the relative accuracy of two alternative MSAs could be inferred from the relative TCS (or GUIDANCE) score of these same sequences. Such analyses typically yield plots like the ones in Figure 4-1. Given an ideal method, such plots should only contain points in the top right and the bottom left quadrant, which correspond to situations where the two differences have the same sign. We used the three alignments (ClustalW, MAFFT and Muscle) of each dataset as well as the reference, which was treated as a fourth method. For each alignment, we estimated the BAliBASE and PREFAB score on the one hand, and the TCS (or GUIDANCE) score on the other hand. We then estimated for each combination dataset/evaluation method the proportion of points for which the relation of order between the structural evaluation and the sequence evaluation were in agreement. Results are reported in Table 4-4. In this table, an ideal method would get a score of 100, and TCS dominates by far GUIDANCE, both on BAliBASE and on PREFAB. These performances are comparable to those reported when using structural information for similar analysis (Kemena et al. 2011).

Table 4-4 TCS, TCS_original, TCS_FM and GUDIANCE applied to BAliBASE 3 and PREFAB 4.

	BAliBASE3							PREFAB4				
	RV11	RV12	RV20	RV30	RV40	RV50	all	0~20	20~40	40~70	70~100	all
# comp.	228	264	246	180	294	96	1,308	2,391	1,962	345	294	4,992
TCS_original	69.2	84.0	**87.4**	**91.7**	**82.0**	**90.6**	83.1	62.8	71.2	**68.3**	73.6	66.8
TCS	**82.4**	**87.0**	81.7	85.6	80.6	86.5	**83.5**	**71.7**	**74.8**	67.9	62.7	**72.5**
TCS_FM	67.4	70.6	70.3	70.0	70.7	69.8	69.9	65.2	70.7	62.6	**85.5**	67.8
GUIDANCE	68.3	73.7	64.2	77.8	72.1	72.9	71.1	59.9	61.7	56.1	62.7	60.5

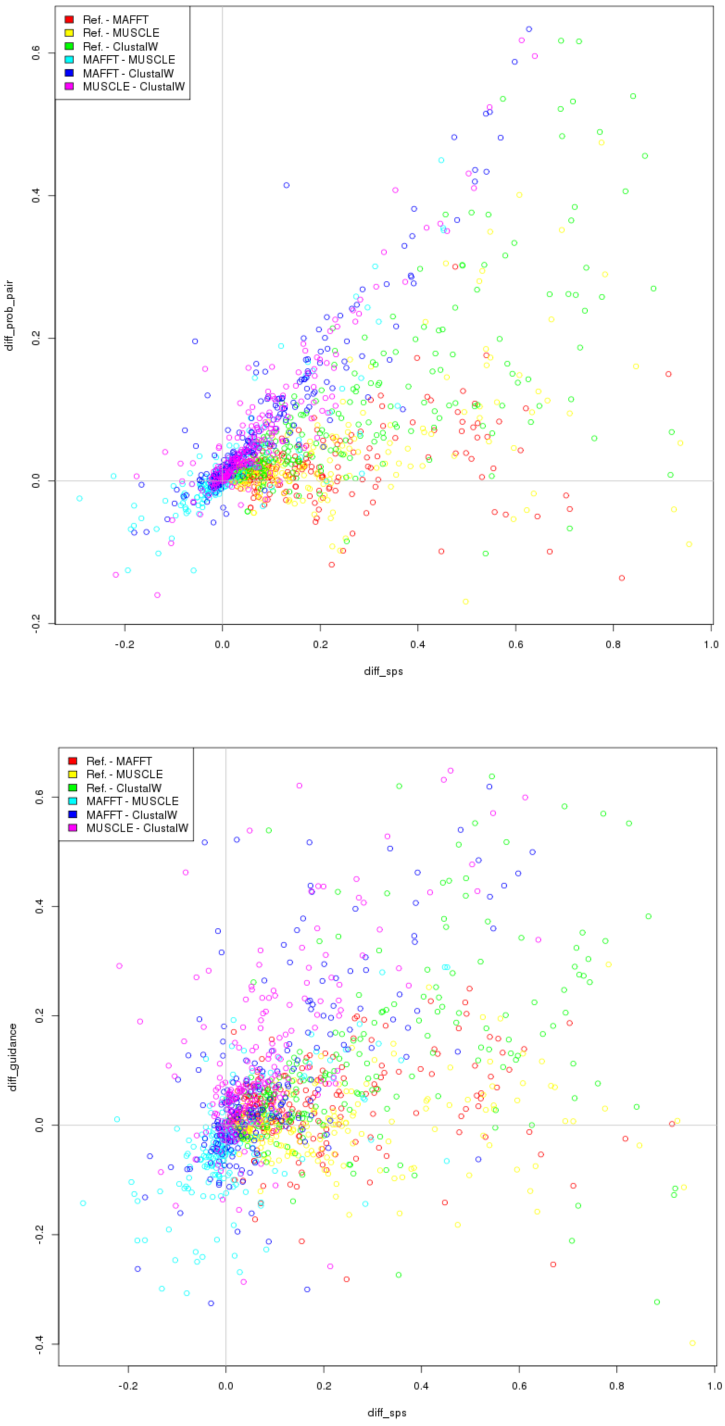

Figure 4-1 Comparison of Δ SPS and Δ confidences by GUIDANCE (upper) and TCS (bottom) on BAliBASE3 using alignments produced by MAFFT, MUSCLE and ClustalW as well as the reference alignment. All points that have the same algebraic sign are correctly classified.

c) Usefulness of the TCS score for phylogenetic reconstruction

In order to estimate the usefulness of the TCS when building phylogenetic trees, we ran this analysis on three simulated datasets (Figure 4-2). We used the approach to compare filtering methods that remove columns, like trimAl and Gblocks, with our much less invasive approach that merely uses MSA stability to weight the contribution of each position. On almost every configuration, when using ML, we found our weighted replication to yield more accurate trees than any of the alternative protocols. We also found the benefits of weighted replication to increase with the number of tips. Interestingly we found the benefits of TCS to be much less significant when using lower-accuracy tree reconstruction methods like Neighbor Joining and Maximum Parsimony.

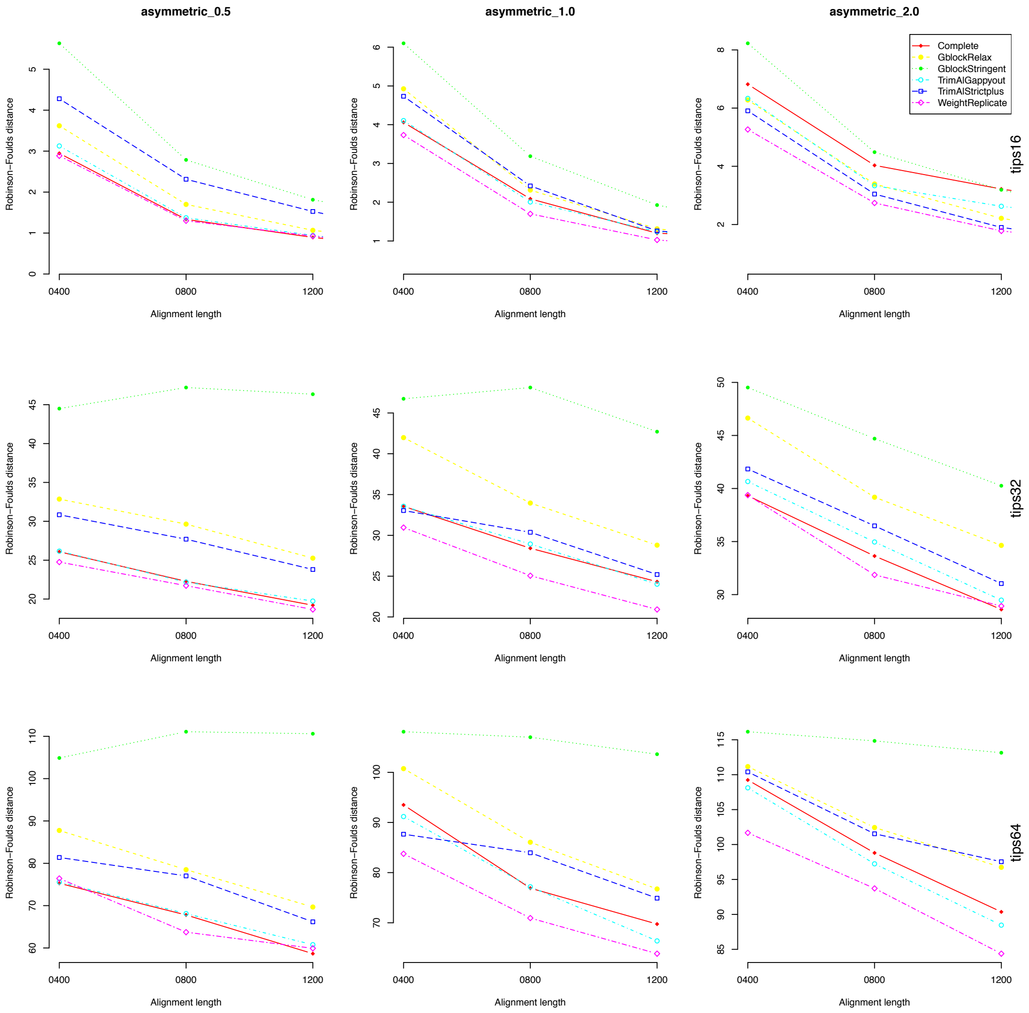

Figure 4-2 Average Robinson-Foulds distances of ML tree to reference tree with 16, 32 and 64 tips from the MAFFT complete alignments, the same alignments after treatment with Gblock relaxed, Gblock stringent, trimAl gappyout, trimAl strictplus and TCS replicated. The asymmetric tree with three different divergence levels was used for the simulations with different alignment lengths.

We did run a similar analysis on the empirical yeast datasets and found the results to be in excellent agreement with those observed on the simulated dataset. On these empirical datasets, the use of the TCS replication results in more correct topologies on all methods. The overall correctness of the topologies is also significantly higher (lower RF). Overall, the TCS approach recovers a significantly larger number of correct topologies than most trimming methods, which seem to have a tendency to degrade accuracy on this specific dataset. This interesting observation raises the issue of the usefulness of simulated datasets, whose value depends entirely on the expectation that these simulations effectively manage to recapitulate biological processes.

The yeast data shows that the stricter the mode is, the less well it performs. For example, Gblocks relaxed works better than Gblocks stringent and trimAl gappyout works better than trimAl strictplus (Table 4-5). Filtering procedures will damage the evolutionary signal in the yeast data set. By replicating instead of filtering, TCS is the only strategy that can recover more yeast ToLs than using the original alignments, improving the number of true topologies from 641.71 to 656.14 (averaging over the seven aligners)

Table 4-5 853 genes respect to the Yeast ToL.

	Original		Gblocks relaxed		Gblocks stringent		trimAl gappyout		trimAl strictplus		TCS replicate	
	RF	TPs	RF	TPs	RF	TPs	RF	TPs	RF	TPs	RF	TPs
ClustalW	0.90	643	0.99	629	1.24	584	0.95	628	1.31	561	0.91	649
DCA	1.07	624	1.01	626	1.32	552	1.13	606	1.31	569	0.93	647
Dalign	0.84	659	0.95	638	1.36	561	0.85	655	1.31	563	0.80	668
MAFFT	0.80	665	0.83	653	1.26	573	0.83	657	1.28	562	0.76	669
Muscle	0.95	639	0.91	646	1.26	578	0.96	633	1.29	559	0.84	662
ProbCons	0.94	642	0.96	632	1.31	579	1.02	624	1.25	566	0.87	657
T-Coffee	1.02	620	1.08	612	1.30	570	1.06	612	1.30	565	0.91	641
AVE	0.93	641.71	0.96	633.71	1.29	571.00	0.97	630.71	1.29	563.57	0.86	656.14

4.5 Discussion and Conclusion

In this work, the TCS, a score that is based on a library of pairwise alignments is shown to have a high discriminative power regarding alignment accuracy. An additional advantage is that the library can be constructed in many ways (e.g. from structural alignments), so that the transitive consistency of the input alignment can be judged according to the criteria of interest. The library platform is so flexible to integrate different information that it can meet a variety of different needs.

One simple way of automatically curating alignments to improve quality is to remove ambiguously aligned regions from subsequent phylogenetic analysis, which can be done with the popular program Gblocks. However, the effect of applying Gblocks on downstream tree accuracy is controversial due to the fact that the filtered part of the alignment might contain an important phylogenetic signal (Anisimova et al. 2010). Some researches have shown that a gappy region may indeed contain an evolutionary signal (Liu et al. 2009; Dessimoz and Gil 2010; Saurabh et al. 2012a). In this paper, we propose a replication scheme based on the TCS score. Our replication strategy is new in two points: First, it keeps all columns of the alignment, and the confidence of the columns is reflected in the number of their replications. It does not filter out any position. Second, TCS goes beyond the popular approach based on sequence conservation in that it captures the alignment uncertainty in terms of pairwise alignments. It is shown that this new strategy helps the downstream phylogenetic analysis. It will be interesting to incorporate the TCS score with the gappy score or the conserved sequence score used in other tools like Gblocks and trimAl.

5. DISCUSSION

David Morrison reviewed 1,280 articles in 2007 from 26 scientific journals where an original MSA, based on empirical data, was used in phylogenetic context (Morrison 2009). He studied separately generalist journals and journals specialized in systematic. He found the majority of the practitioners (78% and 76% for general and systematics, respectively) manually intervene alignment processes, which includes modifying the result from computer software or by manually constructing the alignment from the beginning. He concluded that, at that time, no bioinformatics approach was acceptable for phylogeneticists thus leaving a gap that needs to be filled (Morrison 2009).

In this manuscript, I detailed the implementation of new methods aim to contribute filling this lack of accurate alignments mentioned by Morrison in 2009. While I worked on the improvement of the T-Coffee alignment package (TM-Coffee and SymAlign), I also developed the TCS (section 4) that uses the consistency approach to evaluate the accuracy of an alignment position per position. Therefore, a manual alignment edition process can be speed up under the TCS score guidance. Besides TCS for internal alignment uncertainty, we found that phylogenetic tree reliability can be improved by using a weighted sampling strategy among external alignment uncertainty. Further extensions are discussed in the following sections.

5.1 Future work

a) Alignment space

One sequence set might come out of many alternative alignments due to uncertainty effects, which we have discussed in Section 1.2. All alternative alignments may appear equally good by visual inspection; then users will be confused to choose a final suitable one. One solution proposed by current tools is to summarize these alternative alignments into a consensus one (section 1.3.a). However, it is also interesting to investigate the distribution of these alignments, the so-called alignment space. For that matter, Blackburne and Whelan proposed a strategy to project alignments into a space in which they found there exist two MSAs clusters: similarity-based and evolution-based approaches (Blackburne and Whelan 2012a). Their mapping procedure consists of measuring pairwise distances among alignments using their own distance metric (Blackburne and Whelan 2012b). Then, the pairwise distance matrix is processed through the multidimensional scaling. Finally, each alignment is plotted as a point in the same *N*-dimensional space. Even though the projection strategy manages to capture two MSA classes, drive forces for two separated groups are unknown. This problem can be answered by using a MSA feature matrix instead of the pairwise distance matrix.

Feature representation of proteins is a usual initial step for predicting protein functional classes (Chang et al. 2008; Su et al. 2012). Then, a protein and its features can be projected into the same space trough Correspondence Analysis (CA) (Chang et al. 2012d). CA has been shown able to identify functional residues inside a MSA based on a projection of the columns and sequences (Casari et al. 1995; Rausell et al. 2010). Going a bit further, it will be interesting to investigate the "features" of a MSA, i.e., a gap distribution (Dessimoz and Gil 2010) and information bit, such that a MSA will be quantified in term of features. Taking the "information" feature as an example, we may consider an alignment as the process of sequences permutation from chaos to order, so different alignments represent different permutations with various information bits. Suddenly, the uni-probability model of "alignment space" will be built based on the distribution of the bits. Therefore, each alternative alignment is a point inside the space according to its containing information. Finally, MSA points can be clustered into sub-groups thanks to typical clustering algorithms like *k*-means. This platform will allow us to investigate how alignment features drive the distribution of alignments such that a cluster of MSAs might be explained due to its specific properties, i.e., compact alignment size and gappy columns.

Our goal is to provide a unified solution for handling alignment uncertainty effect through a visual platform of all alternative alignments. A downstream application of an alignment can be attached on its corresponding alignment point. Consequently, we are able to globally observe the effect of alignment accuracy on downstream analysis (Jordan and Goldman 2012). On other hand, a biological conclusion could be certainly conducted when the overall view of those different inferences is available.

b) Alignment confidence format

Although there are many programs measuring the uncertainty of MSAs, there is no standard alignment output format integrating uncertainty information. A popular format, FASTQ, developed to deal with data produced by Next Generation Sequencing, is a text-based format storing both a biological sequence and a character string giving the quality score position per position (Cock et al. 2010). We believe this format is a good starting point for deriving a format including alignment uncertainty. It should be adapted to alignment uncertainty information because it gives information for individual sequence while alignment offer also other type of information besides the confidence of a single character, i.e., the confidence of the aligned residue pair or the confidence of the aligned column. As said before MSAs are often the staring point of a protocol using the multiple comparisons of sequences to infer biological conclusions. Among following analyses we often cite phylogeny, as they are very dependent on the MSAs accuracy. A tree reconstruction method taking into account the reliance information into the calculating process should improve the confidence of the resulting phylogeny. In other words, the more convinced a region is, the more it should contribute in the phylogenetic inference. Thus, the bias of phylogenetic inference can be minimized by alignment uncertainty. As a conclusion, it appears to us that a standard format as described here would be of a great benefit for the scientific community.

c) Phy-Coffee server

The T-Coffee web server paper we published in 2011 is highly cited. Indeed, with 76 citations according to Google Scholar in June 2013, it is the highest citation among all issue according to the Bioinformatics Links Directory (Brazas et al. 2012). It indicates that T-Coffee is useful for the community. We believe that our new methodology for the detection of uncertain region in MSA will improve the interest for our tools. Therefore, we aim to construct a new public server, Phy-Coffee, including partial weighted super MSA for uncertainty issue from alternative aligners (Section 3) and the TCS evaluation tool detecting low confidence regions of input alignments (Section 4). With Phy-Coffee, the sequence alignment service by T-Coffee can be further extended to downstream homology and phylogenetic modeling. Phy-Coffee plus T-Coffee will provide a complete sequence analysis for biologists. We hope the web servers can reduce the gap between computational tools and phylogenetics' need, which is pointed out by Morrison (Morrison 2009).

6. CONCLUSION

The following points give a summary of the presented projects:

1. TM-Coffee addressed the problem of aligning transmembrane proteins. Beside the algorithm that incorporates homology information to derive accurate alignments, we also estimated the influence of the reference database content used for homology extension. We showed that highly non-redundant UniRef databases could be used to obtain similar results at a significantly reduced computational cost over full protein databases.

2. The SymAlign project proposed not only a novel local similarity measure based on conserved words but also better differentiates between the similar and the non-similar protein structures.

3. Weighted Partial Super MSA significantly contributed in the improvement of the discriminative capacities of bootstrap measures used to estimate phylogenetic trees reliability. Furthermore, the values themselves were comparable to similar readouts estimated by using a single method.

4. The TCS score developed in this thesis allows us to estimate the structural accuracy of an alignment by using pairwise alignment information. This method permits to use different aligners to align a set of sequences and choose the best alignment according to the TCS score. In addition to estimate the structural accuracy of an alignment, TCS weighted replicate scheme can enrich evolutionary signals for downstream phylogenetic analyses.

BIBLIOGRAPHY

Aagesen, L. 2004. The information content of an ambiguously alignable region, a case study of the trnL intron from the Rhamnaceae. Org Divers Evol 4:35 - 49.

Altschul, SF, RJ Carroll, DJ Lipman. 1989. Weights for data related by a tree. J Mol Biol 207:647-653.

Anisimova, M, G Cannarozzi, DA Liberles. 2010. Finding the balance between the mathematical and biological optima in multiple sequence alignment.

Bellman, RE. 1953. An introduction to the theory of dynamic programming. Santa Monica, CA: RAND Corporation.

Bellman, RE. 1954. The theory of dynamic programming. Bulletin of the American Mathematical Society 60:503-515.

Bellman, RE. 1958. On a routing problem. Quarterly of Applied Mathematics 16:87-90.

Blackburne, BP, S Whelan. 2012a. Class of multiple sequence alignment algorithm affects genomic analysis. Mol Biol Evol.

Blackburne, BP, S Whelan. 2012b. Measuring the distance between multiple sequence alignments. BIOINFORMATICS 28:495-502.

Bonizzoni, P, GD Vedova. 2001. The complexity of multiple sequence alignment with SP-score that is a metric. Theor. Comput. Sci. 259:63-79.

Brawand, D, M Soumillon, A Necsulea, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343-348.

Brazas, MD, D Yim, W Yeung, BF Ouellette. 2012. A decade of Web Server updates at the Bioinformatics Links Directory: 2003-2012. Nucleic Acids Res 40:W3-W12.

Breen, MS, C Kemena, PK Vlasov, C Notredame, FA Kondrashov. 2012. Epistasis as the primary factor in molecular evolution. Nature 490:535-538.

Bucher, P, K Hofmann. 1996. A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. Proc Int Conf Intell Syst Mol Biol 4:44-51.

Bucka-Lassen, K, O Caprani, J Hein. 1999. Combining many multiple alignments in one improved alignment. BIOINFORMATICS 15:122-130.

Capella-Gutierrez, S, JM Silla-Martinez, T Gabaldon. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. BIOINFORMATICS 25:1972-1973.

Capriotti, E, MA Marti-Renom. 2010. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. BMC Bioinformatics 11:322.

Casari, G, C Sander, A Valencia. 1995. A method to predict functional residues in proteins. Nat Struct Biol 2:171-178.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-552.

Chang, J-M, PD Tommaso, C Notredame. 2012a. Transitive consistency provides a unified framework for measuring homology and phylogenetic modeling. in preparation.

Chang, J-M, S Capella, PD Tommaso, J-F Taly, T Gabaldon, C Notredame. 2012b. Uncertainty on multiple sequence alignment: an approach using the bootstrap. Annual Meeting of the Society for Molecular Biology and Evolution. Dublin, Ireland.

Chang, J-M, P Di Tommaso, J-Fß Taly, C Notredame. 2012c. Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinformatics 13.

Chang, J-M, J-F Taly, I Erb, T-Y Sung, W-L Hsu, CY Tang, C Notredame, ECY Su. 2012d. Efficient and interpretable prediction of protein functional classes by correspondence analysis and compact set relations. submitted.

Chang, JM, EC Su, A Lo, HS Chiu, TY Sung, WL Hsu. 2008. PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. Proteins 72:693-710.

Chothia, C, AM Lesk. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J 5:823-826.

Claude, JB, K Suhre, C Notredame, JM Claverie, C Abergel. 2004. CaspR: a web server for automated molecular replacement using homology modelling. Nucleic Acids Res 32:W606-609.

Claverie, J-M, C Notredame. 2003. Bioinformatics for Dummies: Wiley Publishing Inc.

Cock, PJ, CJ Fields, N Goto, ML Heuer, PM Rice. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767-1771.

Collingridge, PW, S Kelly. 2012. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. BMC Bioinformatics 13:117.

Criscuolo, A, S Gribaldo. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10:210.

Delsuc, F, H Brinkmann, H Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361-375.

Dessimoz, C, M Gil. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol 11:R37.

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

Earl, D. 2012. Alignathon. p. http://compbio.soe.ucsc.edu/alignathon/.

Earl, D, B Paten, M Diekhans. 2012. mafTools.

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

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

Elias, I. 2006. Settling the intractability of multiple alignment. J Comput Biol 13:1323-1339.

Erb, I, JR Gonzalez-Vallinas, G Bussotti, E Blanco, E Eyras, C Notredame. 2012. Use of ChIP-Seq data for the design of a multiple promoter-alignment method. Nucleic Acids Res 40:e52.

Felsenstein, J. 1989. PHYLIP-Phylogeny Inference Package(Version 3.2). Cladistics. p. 164-166.

Fleissner, R, D Metzler, A von Haeseler. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. Syst Biol 54:548-561.

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 refinement as assessed by reference to structural alignments. J Mol Biol 264:823-838.

Graur, D, Y Zheng, N Price, RB Azevedo, RA Zufall, E Elhaik. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 5:578-590.

Guindon, S, JF Dufayard, V Lefort, M Anisimova, W Hordijk, O Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307-321.

Gupta, SK, JD Kececioglu, AA Schaffer. 1995. Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. J Comput Biol 2:459-472.

Hall, BG. 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol Biol Evol 22:792-802.

Hickson, RE, C Simon, SW Perrey. 2000. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. Mol Biol Evol 17:530-539.

Hoffmann, S, C Otto, S Kurtz, CM Sharma, P Khaitovich, J Vogel, PF Stadler, J Hackermuller. 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5:e1000502.

Huang, X, W Miller. 1991. A time-efficient, linear-space local similarity algorithm. Advances in Applied Mathematics 12:337-357.

Jordan, G, N Goldman. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol 29:1125-1139.

Just, W. 2001. Computational complexity of multiple sequence alignment with SP-score. J Comput Biol 8:615-623.

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

Katoh, K, H Toh. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286 - 298.

Kemena, C, G Bussotti, E Capriotti, MA Marti-Renom, C Notredame. 2013. Using tertiary structure for the computation of highly accurate multiple RNA alignments with the SARA-Coffee package. BIOINFORMATICS 29:1112-1119.

Kemena, C, C Notredame. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. BIOINFORMATICS 25:2455-2465.

Kemena, C, JF Taly, J Kleinjung, C Notredame. 2011. STRIKE: evaluation of protein MSAs using a single 3D structure. BIOINFORMATICS 27:3385-3391.

Kim, J, J Ma. 2011. PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. Nucleic Acids Res 39:6359-6368.

Kjer, KM, JJ Gillespie, KA Ober. 2007. Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between POY and structural alignment. Syst Biol 56:133-146.

Kumar, S, A Filipski. 2007. Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res 17:127-135.

Landan, G, D Graur. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol 24:1380-1383.

Landan, G, D Graur. 2008. Local reliability measures from sets of co-optimal multiple sequence alignments. Pac Symp Biocomput:15-24.

Landan, G, D Graur. 2009. Characterization of pairwise and multiple sequence alignment errors. Gene 441:141-147.

Larkin, M, G Blackshields, N Brown, et al. 2007. Clustal W and Clustal X version 2.0. BIOINFORMATICS 23:2947 - 2948.

Lassmann, T, EL Sonnhammer. 2005. Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 6:298.

Liao, BY, J Zhang. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol 23:530-540.

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

Liu, K, S Raghavan, S Nelesen, CR Linder, T Warnow. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 324:1561-1564.

Liu, K, TJ Warnow, MT Holder, SM Nelesen, J Yu, AP Stamatakis, CR Linder. 2012. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst Biol 61:90-106.

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

Loytynoja, A, N Goldman. 2009. Evolution. Uniting alignments and trees. Science 324:1528-1529.

Magis, C, F Stricher, AM van der Sloot, L Serrano, C Notredame. 2010. T-RMSD: a fine-grained, structure-based classification method and its application to the functional characterization of TNF receptors. J Mol Biol 400:605-617.

Magis, C, AM van der Sloot, L Serrano, C Notredame. 2012. An improved understanding of TNFL/TNFR interactions using structure-based classifications. Trends Biochem Sci 37:353-363.

Markova-Raina, P, D Petrov. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res 21:863-874.

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

Morgenstern, B, A Dress, T Werner. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc Natl Acad Sci U S A 93:12098-12103.

Morrison, DA. 2009. Why would phylogeneticists ignore computerized sequence alignment? Syst Biol 58:150-158.

Needleman, SB, CD Wunsch. 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. 2002. Recent progress in multiple sequence alignment: a survey. Pharmacogenomics 3:131-144.

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

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

Novak, A, I Miklos, R Lyngso, J Hein. 2008. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. BIOINFORMATICS 24:2403-2404.

Nuin, PA, Z Wang, ER Tillier. 2006. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7:471.

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

Ogdenw, TH, MS Rosenberg. 2006. Multiple sequence alignment accuracy and phylogenetic inference. Syst Biol 55:314-328.

Penn, O, E Privman, H Ashkenazy, G Landan, D Graur, T Pupko. 2010a. GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Res 38:W23-28.

Penn, O, E Privman, G Landan, D Graur, T Pupko. 2010b. An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol 27:1759-1767.

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

Rausell, A, D Juan, F Pazos, A Valencia. 2010. Protein interactions and ligand binding: from protein subfamilies to functional specificity. Proc Natl Acad Sci U S A 107:1995-2000.

Redelings, BD, MA Suchard. 2005. Joint Bayesian estimation of alignment and phylogeny. Syst Biol 54:401-418.

Rokas, A, BL Williams, N King, SB Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Rosenberg, MS. 2005. Evolutionary distance estimation and fidelity of pair wise sequence alignment. BMC Bioinformatics 6:102.

Roshan, U, DR Livesay. 2006. Probalign: multiple sequence alignment using partition function posterior probabilities. BIOINFORMATICS 22:2715-2721.

Rost, B. 1999. Twilight zone of protein sequence alignments. Protein Eng 12:85-94.

Sankoff, D, C Morel, RJ Cedergren. 1973. Evolution of 5S RNA and the non-randomness of base replacement. Nat New Biol 245:232-234.

Saurabh, K, BR Holland, GC Gibb, D Penny. 2012a. Gaps: an elusive source of phylogenetic information. Syst Biol 61:1075-1082.

Saurabh, K, BR Holland, GC Gibb, D Penny. 2012b. Gaps: An Elusive Source of Phylogenetic Information. Syst Biol.

Sibbald, PR, P Argos. 1990. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J Mol Biol 216:813-818.

Simmons, M, D Richardson, A Reddy. 2008. Incorporation of gap characters and lineage-specific regions into phylogenetic analyses of gene families from divergent clades: an example from the kinesin superfamily across eukaryotes. Cladistics 24:372 - 384.

Sing, T, O Sander, N Beerenwinkel, T Lengauer. 2005. ROCR: visualizing classifier performance in R. BIOINFORMATICS 21:3940-3941.

Sokal, RR, CD Michener. 1958. A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin 28:1409-1438.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. BIOINFORMATICS 22:2688-2690.

Stebbings, LA, K Mizuguchi. 2004. HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. Nucleic Acids Res 32:D203-207.

Stoye, J. 1998. Multiple sequence alignment with the Divide-and-Conquer method. Gene 211:GC45-56.

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

Su, EC, JM Chang, CW Cheng, TY Sung, WL Hsu. 2012. Prediction of nuclear proteins using nuclear translocation signals proposed by probabilistic latent semantic indexing. BMC Bioinformatics 13 Suppl 17:S13.

Suchard, MA, BD Redelings. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. BIOINFORMATICS 22:2047-2048.

Talavera, G, J Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564-577.

Tamura, K, D Peterson, N Peterson, G Stecher, M Nei, S Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739.

Taylor, WR. 2000. Protein structure comparison using SAP. Methods Mol Biol 143:19-32.

Thompson, J, P Koehl, R Ripp, O Poch. 2005a. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61:127 - 136.

Thompson, JD, DG Higgins, TJ Gibson. 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 22:4673-4680.

Thompson, JD, P Koehl, R Ripp, O Poch. 2005b. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61:127-136.

Thompson, JD, O Poch. 2006. Multiple Sequence Alignment as a Workbench for Molecular Systems Biology. Current Bioinformatics 1:95-104.

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

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

Varón, A, LS Vinh, WC Wheeler. 2010. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics 26:72-85.

Vingron, M, P Argos. 1991. Motif recognition and alignment for many sequences by comparison of dot-matrices. J Mol Biol 218:33-43.

Wallace, IM, O O'Sullivan, DG Higgins, C Notredame. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34:1692-1699.

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

Wang, LS, J Leebens-Mack, P Kerr Wall, K Beckmann, CW dePamphilis, T Warnow. 2011. The impact of multiple protein sequence alignment on phylogenetic estimation. IEEE/ACM Trans Comput Biol Bioinform 8:1108-1119.

Wheeler, TJ, JD Kececioglu. 2007. Multiple alignment by aligning alignments. BIOINFORMATICS 23:i559-568.

Wolfe, KH, DC Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708-713.

Wong, KM, MA Suchard, JP Huelsenbeck. 2008. Alignment uncertainty and genomic analysis. Science 319:473-476.

Wu, M, S Chatterji, JA Eisen. 2012. Accounting for alignment uncertainty in phylogenomics. PLoS One 7:e30288.

Zhang, Y, J Skolnick. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302-2309.

Zharkikh, A, WH Li. 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. Mol Phylogenet Evol 4:44-63.

Zheng-Bradley, X, J Rung, H Parkinson, A Brazma. 2010. Large scale comparison of global gene expression patterns in human and mouse. Genome Biol 11:R124.