Influence of alignment uncertainty on homology modeling

Jia-Ming Chang[1], Cedric Notredame[2]
[1]Department of Computer Science
National Chengchi University, Taipei, Taiwan
jmchang@cs.nccu.edu.tw
[2]Centre for Genomic Regulation (CRG)
The Barcelona Institute of Science and Technology, Barcelona, Spain
Cedric.Notredame@crg.eu

Abstract

Most evolutionary analyses or structure modeling are based upon pre-estimated multiple sequence alignment (MSA) models. From a computational point of view, it is too complex to estimate a correct alignment. Hence, increasing or identifying signal inside sequence alignment has intensified over the last few years. In this work, we show how this problem can be partly overcome using the transitive consistency score (TCS), an extended version of the T-Coffee scoring scheme. Using this local evaluation function, we show that one can identify the most reliable portions of an MSA, as judged from BAliBASE and PREFAB structure-based reference alignments. We compared TCS with Heads-or-Tails, GUIDANCE, Gblocks, and trimAl and found it to lead to significantly better estimates of structural accuracy and more accurate phylogenetic trees.

1 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 [1], and the MSA Wikipedia page lists 47 available MSA packages (http://en.wikipedia.org/wiki/Sequence_alignment_software). 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 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. 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.

More recently, a method, named Heads-or-Tails (HoT) [2] 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 [7]. 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.

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 [3] or PREFAB4 [4] 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.

2 Method

The transitive consistency score (*TCS*) 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.

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

Two datasets were used in order to estimate structural correctness: BAliBASE3 [3] 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 [4], 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 [5]. RV11 has been shown to the most informative subset across all these categories [1].

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 [6]. 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.

3 Results

We first computed the BAliBASE MSAs using ClustalW, MAFFT and Muscle (Table 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 2). This analysis confirmed the superiority of the TCS scoring scheme, which is much less affected than its counterparts by variations in accuracy.

Table 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 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

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 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.

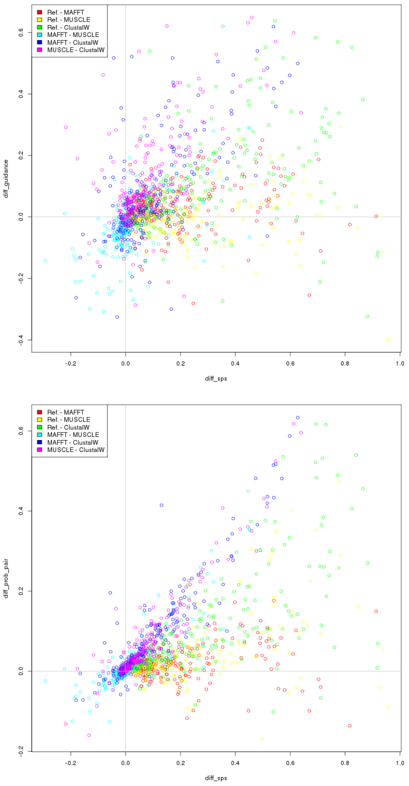

Figure 1. Comparison of Δ SPS and Δ confidences by GUIDANCE (upper) and TCS (bottom) on BAliBASE3 using alignments produced by MAFFT, MUSCLE and ClustalW as well as the reference alignment. All points that have the same algebraic sign are correctly classified.

4 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.

The web server is available at http://tcoffee.crg.cat/tcs.

Acknowledgments

This manuscript is adapted from [8], [9]. We would also like to thank two anonymous reviewers for their comments.

References

[1] C. Kemena and C. Notredame, “Upcoming challenges for multiple sequence alignment methods in the high-throughput era.,” *Bioinformatics*, vol. 25, no. 19, pp. 2455–65, Oct. 2009.

[2] G. Landan and D. Graur, “Heads or tails: a simple reliability check for multiple sequence alignments.,” *Molecular biology and evolution*, vol. 24, no. 6, pp. 1380–3, 2007.

[3] O. Penn, E. Privman, G. Landan, D. Graur, and T. Pupko, “An alignment confidence score capturing robustness to guide tree uncertainty.,” *Molecular biology and evolution*, vol. 27, no. 8, pp. 1759–67, 2010.

[4] J. Thompson, P. Koehl, R. Ripp, and O. Poch, “BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark,” *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. 1, pp. 127–36, 2005.

[5] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput.,” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–7, Jan. 2004.

[6] B. Rost, “Twilight zone of protein sequence alignments.,” *Protein Eng.*, vol. 12, no. 2, pp. 85–94, Feb. 1999.

[7] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, “ROCR: visualizing classifier performance in R.,” *Bioinformatics (Oxford, England)*, vol. 21, no. 20, pp. 3940–1, 2005.

[8] J.-M. M. Chang, P. Di Tommaso, and C. Notredame, “TCS: a new multiple sequence

alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction.," *Mol. Biol. Evol.*, vol. 31, no. 6, pp. 1625–37, Jun. 2014.

[9] J.-M. Chang, P. Di Tommaso, V. Lefort, O. Gascuel, and C. Notredame, "TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction.," *Nucleic acids research*, vol. 43, no. W1, pp. W3–6, 2015.