IMPROVING ACCURACY AND SCALABILITY FOR CORE BIOINFORMATICS ANALYSES

Allocation: Illinois/125 Knh
PI: Tandy Warnow[1]
Collaborators: Erin Molloy[1], Michael Nute[1], Ehsan Saleh[1], Kodi Collins[2]

[1]University of Illinois at Urbana-Champaign
[2]University of California, Los Angeles

EXECUTIVE SUMMARY

This project developed new methods for large-scale evolutionary tree construction and multiple sequence alignment that can be used to address fundamental science problems such as "How did life evolve on Earth?" and "What function does this protein have?" The most important outcome is the discovery we made regarding protein sequence alignment methods: Our research suggests that BAli-Phy, the leading statistical method for multiple sequence alignment, has the best accuracy of all methods tested on simulated data sets, but is less accurate than standard multiple sequence alignment methods when evaluated on protein benchmark data sets. While the cause for this difference in performance between biological and simulated data is not yet known, each of the most likely explanations (i.e., either model misspecification or errors in the protein benchmark data sets) presents troubling ramifications for other problems in biology, including molecular systematics and protein structure and function prediction.

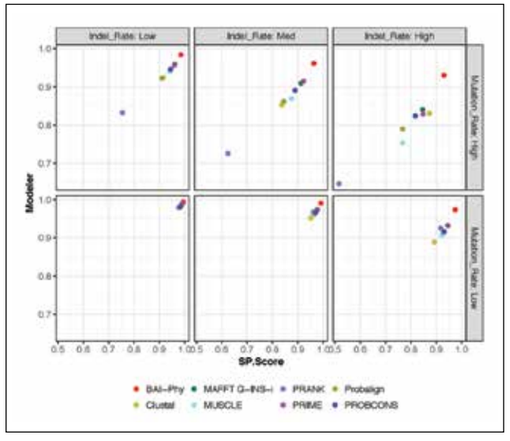

Figure 1: Accuracy on simulated data sets (each with 27 sequences) for different multiple sequence alignment methods, showing averages (over 20 replicates) for modeler score (precision) and SP score (recall). Note that BAli-Phy has the highest accuracy of all methods under all conditions. (Figure taken from [5].)

RESEARCH CHALLENGE

Much biological research—including the estimation of evolutionary histories, the prediction of protein structure and function, and the detection of positive selection—requires that a set of molecular sequences first be "aligned" with each other. Furthermore, multiple sequence alignment of large data sets is necessary for many biological studies. Most obviously, the construction of the tree of life will require millions of sequences, spanning large evolutionary distances. Less obviously, protein structure prediction also benefits from large data sets: the most accurate protocols for predicting the structure of an unknown protein from its sequence of amino acids begins by collecting a very large number of related sequences and then computing a multiple sequence alignment on that set.

Unfortunately, large-scale multiple sequence alignment is enormously difficult to perform with high accuracy. The only methods that have been able to run on ultra-large data sets (with up to one million sequences) are PASTA [1] and UPP [2], which use a divide-and-conquer approach to scale other alignment methods to large data sets. Both PASTA and UPP have excellent accuracy on simulated data sets, but less is known about their accuracy on biological data sets, and especially on protein data sets, where alignment estimation may be enabled through the use of inferred or known structural elements. BAli-Phy [3], one of the most promising approaches, infers an alignment under a statistical model of sequence evolution and is expected to have the best accuracy of all methods. Yet, BAli-Phy is too computationally intensive to use directly on large data sets. Alternatively, BAli-Phy can be incorporated into PASTA and UPP so that it scales to large data sets [4]. These "boosted" versions of BAli-Phy have outstanding accuracy on simulated nucleotide data sets [4]; however, when this project began, it was not known if BAli-Phy or the "boosted" version of BAli-Phy would provide improved accuracy on biological data sets (nucleotides or amino acids) in comparison to standard multiple sequence alignment methods. Efficient and scalable multiple sequence alignments that have improved accuracy on ultra-large biological data sets, and especially for protein sequences, would therefore provide major benefits for many downstream analyses.

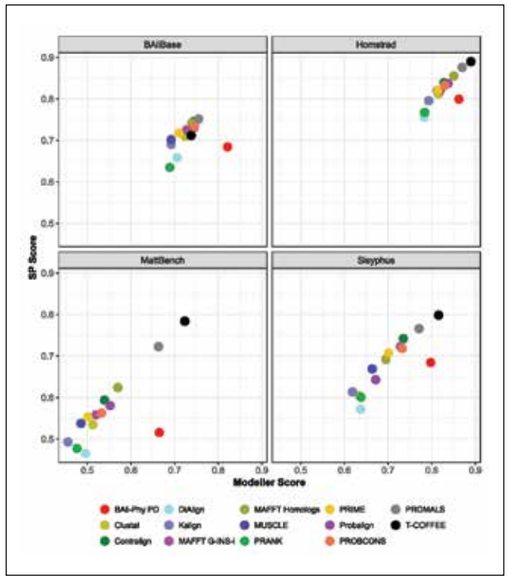

Figure 2: Average accuracy on four biological benchmark collections, each with at most 25 amino acid sequences (1,192 sequence data sets in all) for different multiple sequence alignment methods, showing modeler score (precision) and SP score (recall). Note that BAli-Phy tends to have the best average recall of all methods for all benchmarks. (Figure taken from [5].)

METHODS & CODES

We performed two major studies regarding protein sequence alignment. In the first study [5], we compared BAli-Phy to leading protein sequence alignment methods on data sets from four established benchmark collections of protein sequences. Since BAli-Phy is computationally intensive, we limited the study to small data sets. The second study evaluated the impact of integrating the best-performing methods from this first study into PASTA.

RESULTS & IMPACT

BAli-Phy has outstanding accuracy on the simulated data sets (Fig. 1), clearly dominating all the other methods with respect to both recall (*sum of pairs* or SP score) and precision (modeler score). Yet, on the biological data sets (Fig. 2), BAli-Phy has much poorer precision and recall. In fact, BAli-Phy typically has only average recall and often is among the poorest of the top alignment methods. The best-performing multiple sequence alignment methods in this study have been integrated into PASTA (thus attaining improved scalability and reduced running time, while maintaining accuracy), and the new version of PASTA is available at [6] in open-source form.

The distinction in accuracy on biological data sets and simulated data sets is troubling and requires further investigation. BAli-Phy's excellent accuracy on simulated data is expected since the simulated data sets are generated under statistical models of sequence evolution that are close, even if not identical, to the statistical models under which BAli-Phy performs its inference. However, since BAli-Phy is so much less accurate on biological data sets, this suggests that the protein data sets have evolved under processes that are quite different from the ones that are well modelled by BAli-Phy. While it has always been expected that there would be some level of model misspecification (as no model is perfect), for there to be a substantial difference in relative accuracy between simulated and biological data sets suggests that the level of model misspecification must be quite large. This would be a troubling conclusion, since many bioinformatics analyses are performed under statistical models similar to the one assumed in BAli-Phy. However, there are other potential explanations, one of which is that the reference alignments in the biological benchmarks may themselves not be highly accurate (i.e., they may be inferred through a combination of information about structural features in the proteins and then interpolation among the structurally derived parts of the alignment using software tools). If this is a reason for the discordance, then BAli-Phy may still be useful, but the benchmarks must be questioned. Future research is needed to explore these possible explanations, as well as the others that we posit, as discussed in [5].

WHY BLUE WATERS

This study used 230 CPU years for the BAli-Phy analyses alone and would not have been feasible on other computational systems available to the project team.

PUBLICATIONS & DATA SETS

Nute, M., E. Saleh, and T. Warnow, Benchmarking statistical alignment. *bioRxiv* (2018), DOI:10.1101/304659.

Nute, M., et al., The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics*, 19: Supplement 5 (2018), DOI:10.1186/s12864-018-4619-8.

Christensen, S., et al., OCTAL: Optimal completion of gene trees in polynomial time. *Algorithms for Molecular Biology*, 13:6 (2018), DOI:10.1186/s13015-018-0124-5.

Mirarab, S., et al., PASTA github site (https://github.com/smirarab/pasta), accessed May 13, 2018.

Collins, K., and T. Warnow, PASTA for Proteins Data (BALiBASE). University of Illinois at Urbana-Champaign (2018), DOI:10.13012/B2IDB-4074787_V1.

Christensen, S., et al., Datasets from the study "OCTAL: Optimal Completion of Gene Trees in Polynomial Time." University of Illinois at Urbana-Champaign (2018), DOI:10.13012/B2IDB-1616387_V1.

Nute, M., et al., Data from: The Performance of Coalescent-Based Species Tree Estimation Methods under Models of Missing Data. University of Illinois at Urbana-Champaign (2017), DOI:10.13012/B2IDB-7735354_V1.