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 BAi-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.

METHODS & CODES

We performed two major studies regarding protein sequence alignment. In the first study [5], we compared BAi-Phy to leading protein sequence alignment methods on data sets from four established benchmark collections of protein sequences. Since BAi-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

BAi-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 (modeller score). Yet, on the biological data sets (Fig. 2), BAi-Phy has much poorer precision and recall. In fact, BAi-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 retaining 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. BAi-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 BAi-Phy performs its inference. However, since BAi-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 BAi-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 BAi-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 BAi-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 BAi-Phy analyses alone and would not have been feasible on other computational systems available to the project team.

PUBLICATIONS & DATA SETS


