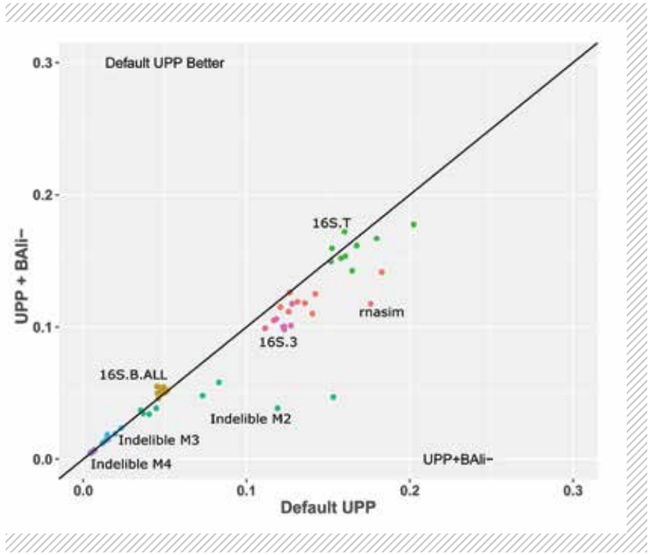
ADVANCING GENOME-SCALE PHYLOGENOMIC ANALYSIS

Allocation: Illinois/158 Knh
PI: Tandy Warnow[1]

[1]University of Illinois at Urbana-Champaign

EXECUTIVE SUMMARY

The project has three main goals, each geared towards advancing the accuracy of large-scale estimation of evolutionary history. The first year has focused on the first goal of method development for multiple sequence alignment, maximum likelihood phylogeny estimation, and species tree estimation. Highlights include (a) scalable versions of BAli-Phy, a Bayesian method for statistical co-estimation of multiple sequence alignments and trees, (b) a new method (HIPPI) for classifying sequences into gene families, and (c) new supertree methods with improved accuracy and scalability. One journal paper has been published, two others have been submitted, three others are in preparation, and six projects are underway. Three graduate students and two postdoctoral fellows at University of Illinois at Urbana-Champaign participated in this project.



INTRODUCTION

Because phylogenies and multiple sequence alignments are the basis of many biological discoveries, improving the accuracy of methods that estimate these alignments and phylogenies will improve downstream biological inferences. These inferences range from the timing of different evolutionary events, to understanding which genes are involved in trait evolution, to how species adapt to their environments, to analyzing the human gut microbiome. All of these inferences depend on accurate gene trees (i.e., how genes evolve within the species tree) and species trees (how organisms diversified from a common ancestor), which in turn depend on accurate multiple sequence alignments. The estimation of alignments and trees is computationally difficult, as the best methods are attempts to solve NP-hard optimization problems or use Markov Chain Monte Carlo techniques. This project developed methods and software to enable highly accurate estimations of large-scale multiple sequence alignments and phylogenetic trees and thus advance scientific discovery.

METHODS & RESULTS

Our research focused on the development of methods with improved accuracy for multiple sequence alignment, large-scale maximum likelihood, species tree estimation from multiple genes, remote homology detection and protein family classification, and supertree estimation. We also used Blue Waters to perform large-scale phylogenomic and metagenomic analyses of biological datasets, in collaborations with research groups here at Illinois and around the country. These analyses gave us insight into how we can improve the methods regarding accuracy and computational performance. Many of these biological analyses involved estimating multiple sequence alignments using our new methods (PASTA [1] and UPP [2]) and estimating species trees from multi-locus datasets using maximum likelihood heuristics such as RAxML [3] or FastTree [4]. Some species tree analyses have also been based on ASTRAL-2 [5], a method we developed for estimating species trees in the presence of gene tree heterogeneity, and that is statistically consistent in the presence of incomplete lineage sorting. One of the major outcomes of this year was a study performed in collaboration with Illinois Professor Bryan White (Carle Woese Institute for Genomic Biology) and Mihai Pop (University of Maryland at College Park). Our recently published study [6] compared computational methods for defining operational taxonomic units. To perform this study, we computed a multiple sequence alignment of approximately 40,000,000 16S sequences, using UPP; this may be the largest multiple sequence alignment ever computed, and Blue Waters was essential for this study. Two Ph.D. students and one postdoctoral fellow worked on research supported by this Blue Waters allocation; the Ph.D. dissertations of the students will include research reported here.

The highlights of the method development were:

- Scalable versions of BAli-Phy [7], a Bayesian method for statistical co-estimation of multiple sequence alignments and phylogenetic trees, so that it can analyze datasets with 10,000 sequences. The improvement in scalability was obtained by integrating BAli-Phy into PASTA and UPP, two of our divide-and-conquer methods for estimating multiple sequence alignments. The existing BAli-Phy code is otherwise limited to, at most, 100 or so sequences.
- A new method (HIPPI) for classifying sequences (including short reads generated by sequencing technologies) into gene families. HIPPI is based on a new machine learning approach we have developed called "ensembles of Hidden Markov Models," which we have also used for metagenomic taxon identification and abundance profiling [8]. HIPPI improved accuracy compared to existing gene family classification methods based on BLAST [9] or single Hidden Markov Models [10].
- A new supertree method, FastRFS, which finds an optimal solution to an NP-hard optimization problem (Robinson-Foulds Supertrees) within a constrained search space. FastRFS provides improved accuracy and greatly reduced running times compared to other supertree methods.

WHY BLUE WATERS

Blue Waters was necessary for the development and testing of computationally intensive methods, which is not feasible on other platforms. Also, we were able to analyze several large-scale biological datasets using Blue Waters, which would not have been feasible using the other computational platforms that were available. Finally, the Blue Waters staff helped us port our codes and get them running, which was very helpful.

NEXT GENERATION WORK

The significant advances we are making are obtained by enabling statistical estimation methods, some employing computationally intensive MCMC techniques, to scale to large datasets. Should a next generation Track-1 system become available for us to use, we would be able to analyze the largest biological datasets coming online, including whole genome datasets with tens of thousands of species and metagenomic datasets with up to billions of reads. These analyses would transform biological and biomedical research.

PUBLICATIONS AND DATA SETS

Nguyen N., T. Warnow, M. Pop, B. White. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *Nature Biofilms and Microbiome Analysis*, 2, article number 16004 (2016), doi:10.1038/npjbiofilms.2016.4

FIGURE 1 (LEFT): Scatterplot of multiple sequence alignment error rates of UPP-default compared to UPP+Bali-Phy on datasets with up to 10,000 sequences. We show alignment error of UPP run in default mode to UPP using BAli-Phy (written as UPP+BAli-Phy) on biological and simulated datasets, ranging in size from about 5000 to 10,000 sequences. On nearly all these datasets, using BAli-Phy within UPP resulted in reduced alignment error, showing that integrating BAli-Phy within UPP improved UPP's accuracy. Since Bali-Phy cannot run on these datasets due to computational limitations, this analysis also shows that integrating BAli-Phy within UPP improves BAli-Phy's scalability.