Final Report

1 Project Information

- Project Title: High Performance Methods for Big Data Phylogenomics, Proteomics, and Metagenomics
- PI: Tandy Warnow, Founder Professor of Engineering, The University of Illinois at Urbana-Champaign
- Names and affiliations of co-PIs and collaborators: N.A.
- Corresponding author name and contact information: Tandy Warnow, warnow@illinois.edu

2 Executive Summary

This project addressed three inter-related problems in computational molecular biology, where large datasets present substantial and difficult computational and statistical challenges: *phylogenomics* (species tree estimation from multiple loci), *protein sequence analysis* (inference of structure and function from amino acid sequences), and *metagenomics* (analysis of environmental samples from shotgun sequence datasets). Highlights of this project's activity include several advances in coalescent-based species tree estimation methods (students Pranjal Vachaspati and Erin Molloy), improved multiple sequence alignment methods for use in all these research areas (students Mike Nute, Ehsan Saleh, and Kodi Collins), and taxon identification methods for microbiome analysis (students Mike Nute and Erin Molloy). Five journal papers were published this year supported by this allocation, another two were submitted, and two studies are underway.

3 Description of research activities and results

3.1 Key Challenges

Each of the three problem areas we identified involve statistical estimation on large datasets where standard methods either do not run or provide poor accuracy. For example, Markov Chain Monte Carlo (MCMC) approaches are standard statistical techniques for estimating multiple sequence alignments and phylogenetic trees, and when they can be run sufficiently well (i.e., to stationarity), they provide outstanding accuracy. However, the sizes of the biological datasets we focus on in this project go well beyond what MCMC methods can analyze. Conversely, fast methods have been developed to compute multiple sequence alignments and trees, and to analyze microbiome datasets, but these do not provide anywhere near the level of accuracy that sophisticated statistical methods can provide, and the accuracy of downstream analyses is compromised. Thus, providing high accuracy in multiple sequence alignments, genome-scale phylogenies, or analysis of microbiomes, requires new approaches that can scale to large datasets and not degrade in accuracy with increasing dataset size and complexity.

3.2 Why it Matters

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).

Multiple sequence alignments are used to infer phylogenies, and so are important to compute with high accuracy for that reason. In addition, the inference of selection depends on multiple sequence alignments, as does the inference of protein structure and function. Hence, multiple sequence alignment estimation is a fundamental bioinformatics step that has many downstream uses and consequences beyond phylogenetic estimation.

There are very good statistical methods for multiple sequence alignment and phylogeny estimation (both gene trees and species trees), but they are generally limited to very small datasets. For example, BAli-Phy (*11*) is the most scalable of the Bayesian methods for co-estimation of multiple sequence alignments and phylogenetic trees, and it is limited to about 50 sequences. Similarly, *BEAST (*4*) is a Bayesian method for co-estimation of gene trees and species trees in the presence of gene tree heterogeneity due to incomplete lineage sorting, and it is limited to about 20 species and 50 genes. Yet even datasets of these sizes can take several weeks or months for the Bayesian MCMC analysis to converge - a necessary condition for this type of method. In contrast, the methods that can run on realistic phylogenomic datasets with 50 or more species and many thousands or tens of thousands of species do not use Bayesian MCMC techniques, and tend to have reduced accuracy in comparison to these statistical methods.

The need for new methods is particularly urgent as more and more studies attempt to analyze phylogenomic datasets with many thousands or tens of thousands of genes, and hence encounter massive gene tree heterogeneity, which can be due to multiple biological processes (incomplete lineage sorting, gene duplication and loss, horizontal gene transfer, etc.). The Genome 10K group (https://www.genome10k.soe.ucsc.edu) is encountering these challenges in its plans to assemble phylogenies of the major groups of life on earth. This project addresses multiple computational needs of phylogenomics projects through the development of methods with strong statistical guarantees, excellent accuracy on biological and simulated datasets, and excellent scalability to large datasets.

3.3 Why Blue Waters

Blue Waters is necessary for at least two reasons. First, the development of these methods requires extensive testing, which is not feasible on other platforms. Second, the analysis of large biological datasets (and even of moderate-sized datasets) often requires years of CPU time (e.g., the avian phylogenomics project spent 450 CPU years to analyze approximately 50 whole genomes); Blue Waters makes this feasible, and enables biological discovery.

3.4 Accomplishments

Highlights of the research activity (overview) A brief description of the main contributions is given first, with some elaboration on selected topics below.

Phylogenomics

- *FastRFS: a new supertree method with improved accuracy compared to prior methods.* Journal publication (*14*), work with PhD student Pranjal Vachaspati. (This project was initiated in the previous year, and then completed and published in this year.)

- *Evaluation of strategies for screening genes in multi-locus phylogenomic analyses on species tree estimation methods.* In revision after encouraging reviews from Systematic Biology, and is joint work with PhD student Erin Molloy. Description: As different loci (i.e., regions of the genome) can have different evolutionary histories, multi-locus phylogenomic analysis is the standard protocol in species tree estimation. However, many loci have substantial amounts of missing data (i.e., sequence data may not be present for all species of interest) and gene trees estimated on these loci may have poor accuracy due to insufficient ”phylogenetic signal” in highly conserved regions of the genome. There is substantial concern in the evolutionary biology community that species tree estimation will not be accurate when these poor quality loci are included, and many studies filter loci prior to species tree estimation. In this project, we evaluate the impact of current gene filtering practices on species tree accuracy, using a combination of simulated and biological datasets. We observed that the best methods for species tree estimation are generally highly robust, and that in most conditions the most accurate species trees are obtained by including all the loci – even those with large amounts of missing data and high gene tree estimation error.

- *SVDquest: new coalescent-based species tree method based on an improved optimization technique.* Submitted for publication, joint work with PhD student Pranjal Vachaspati. Description: Species tree estimation from multiple loci is often estimated by combining estimated gene trees, but when gene trees are poorly estimated then the estimated species trees that are produced can also have high error. Alternative species tree estimation approaches have been developed that bypass gene tree estimation and that have the potential to provide higher accuracy. One such approach is to use SVDquartets (*2*) to estimate four-leaf species trees for every four species, and then combine the quartet trees into a species tree using a quartet amalgamation method. SVDquartets is designed to compute quartet trees when gene trees differ from the species tree due to incomplete lineage sorting (ILS). In this paper, we presented an effective method for combining quartet trees that is guaranteed to find an optimal species tree (one that agrees with the maximum number of quartet trees) within a constrained search space. Our method, which we call SVDquest, is competitive with leading species tree methods in the presence of ILS in terms of accuracy, and is fast enough to run on datasets with 100 species and many thousands of genes.

Protein Sequence Analysis

- *Scaling statistical multiple sequence alignment to large datasets.* Journal publication (*9*), work with PhD student Mike Nute. (This project was initiated in the previous year, and then completed and published in this year. The previous year's report provides a description of this research.)

- *HIPPI: a new method for classifying protein sequences into existing protein families.* Journal publication (*8*), work with PhD student Mike Nute, former postdoctoral fellow Nam-phuong Nguyen, and former PhD student Siavash Mirarab. See below for a detailed description of HIPPI and some results from the publication. (This project was initiated in the previous year, and then completed and published in this year.)

- *Evaluation of leading multiple sequence alignment methods on biological datasets.* In preparation, with PhD students Mike Nute and Ehsan Saleh; we plan to complete the study and submit it for publication in 2017. Description: Our earlier work (*9*) established that statistical methods such as BAli-Phy (*11*) for co-estimating multiple sequence alignments and trees can produce much more accurate alignments than standard methods on simulated datasets. In this study, we are benchmarking BAli-Phy in comparison to standard methods on biological benchmarks with structurally-inferred alignments. Our study (not yet complete) suggests that BAli-Phy does provide excellent accuracy on biological benchmarks, but that comparable (and arguably higher) accuracy can be obtained using standard alignment methods that are much more efficient. This study therefore suggests the possibility that model misspecification is an obstacle to the applicability of BAli-Phy for aligning protein datasets.

Metagenomics

- *Taxon identification and abundance profiling of microbiome datasets.* TIPP (*7*) is a method for abundance profiling of metagenomic datasets that employs an ensemble of HMMs technique. The first step in TIPP uses BLAST to assign reads to marker genes, so that the reads mapping to the marker genes can then be taxonomically identified using statistical methods to predict its species membership. Because HIPPI (described below) improves on BLAST for precision and recall when applied to protein sequences, we expect that replacing BLAST by HIPPI within TIPP will also improve TIPP.

TIPP was optimized for abundance profiling, which provides an estimate of the relative frequency of the different species in a microbiome sample. A different but related problem is "taxon identification", which seeks to identify the species for each of the reads in the metagenomic sample. The challenge in this problem is that each read is short and has multiple errors produced by sequencing technologies. Our ongoing work (with PhD student Mike Nute, Erin Molloy, and collaborator Mihai Pop, Prof at University of Maryland) is examining how to modify TIPP for taxon identification. In this study, we are varying the TIPP algorithm design to find the settings of its algorithmic parameters that produce the most accurate taxonomic classifications of reads sampled from a microbiome. Our preliminary study on biological data suggests that very simple changes to TIPP's design result in very substantial changes to its classification rate. Our ongoing work will evaluate TIPP and its variants on simulated data, so that accuracy can be assessed.

Selected research result: HIPPI. Gene family identification is a basic step in many bioinformatics pipelines, such as metagenomic taxon identification and abundance profiling (first steps in microbiome analysis) and is closely related to remote homology detection, which is a basic step in protein function and structure prediction. BLAST (*1*) is the most well known method for this problem, but other approaches based on profile Hidden Markov Models have been used as well. In this paper, we developed a novel machine learning technique to detect membership in existing protein families, where we construct an ensemble of profile Hidden Markov models (HMMs) to represent each protein family, and then compares each sequence (which can be short reads or full length sequences) to each HMM in each ensemble to find the best fitting protein family. We provided an extensive study based on the PFAM database of protein families and their associated profile HMMs from HMMER (*3*) to compare our method to the previous best methods. This study showed that the technique outperformed all the current methods (including BLAST, HMMER, and HHsearch (*12*)) in terms of both precision and recall, especially when analyzing short sequences; see Figure 1.

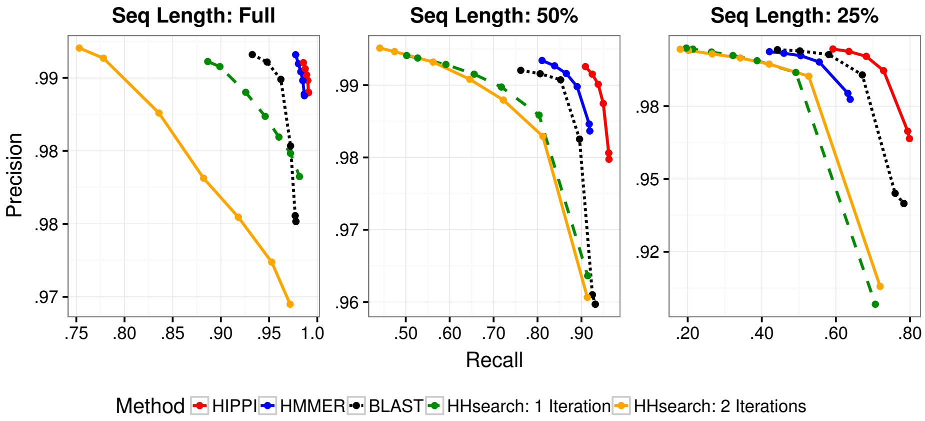

Figure 1: **Precision-recall curves for methods for protein family classification, evaluated on one cross-fold subset of the PFAM seed sequence data.** Our new method - HIPPI (shown in red) – strictly dominates all the other methods with respect to precision and recall, with the largest improvement on the most fragmentary datasets (25% sequence length). Figure taken from N. Nguyen et al., BMC Bioinformatics 17 (Suppl. 10): 765, 2016. The scales for both axes vary between panels due to the significant impact of sequence fragmentation.

4 List of publications and products associated with this work (and supported by the Blue Waters allocation)

I include the journal publications that acknowledge this Blue Waters allocation. Many of these include my current UIUC graduate students or postdoc. All of the papers listed here advance the research into method development specified in the proposal for the Blue Waters allocation, either through explicit algorithm design exploration and testing, or through the use of the methods on biological datasets. Note that the research in these papers will be, in nearly every case, in the PhD dissertation of one of my students. Thus, this allocation is also providing support to the educational mission of UIUC.

Journal Publications

1. P. Vachaspati and T. Warnow (2016). “FastRFS: Fast and accurate Robinson-Foulds Supertrees using constrained exact optimization.” *Bioinformatics* 2016; doi: 10.1093/bioinformatics/btw600. (Special issue for selected papers from RECOMB-CG). Description: Large-scale phylogeny estimation is likely to depend on the development of supertree methods, in which trees on small datasets can be combined into a tree on the full dataset. Yet supertree estimation is NP-hard, and to date the heuristics that have been developed have been limited to approaches that combine hill-climbing with randomness to explore

an exponentially-sized space. In this paper, we present a very different kind of algorithmic approach for the NP-hard Robinson-Foulds supertree problem, where the objective is a supertree that minimizes the topological distance to the input trees. Instead of using heuristic search strategies, we provably solve the problem optimally within a constrained search space that we construct from the input trees, and we do so in polynomial time using dynamic programming. Our study on both biological and simulated datasets, shows that our new method, FastRFS, provides better accuracy and in less time than all the current methods for this problem, and can analyze the largest supertree dataset (2228 species) where other methods fail to be able to run.

2. M. Nute and T. Warnow (2016). "Scaling statistical multiple sequence alignment to large datasets." *BMC Genomics* 17(Suppl 10): 764, special issue for RECOMB-CG. Description: Multiple sequence alignment is generally a required step in phylogeny estimation. The gold standard has been statistical co-estimation of alignments and trees under stochastic models of sequence evolution, yet the best of the methods for such estimation, BAli-Phy (*13*) uses an MCMC approach and for convergence reasons is generally restricted to very small datasets. In fact, the user guide at (*10*) states "BAli-Phy is quite CPU intensive, and so we recommend using 50 or fewer taxa in order to limit the time required to accumulate enough MCMC samples. (Despite this recommendation, data sets with more than 100 taxa have occasionally been known to converge.)" The Warnow Lab adapted PASTA (*5*) and UPP (*6*), two divide-and-conquer approaches to multiple sequence alignment, to enable BAli-Phy to be used on large sequence datasets. As shown in (*9*), this technique allowed BAli-Phy to run on very large datasets (10,000 sequences), and produced more accurate trees and alignments than the previous most accurate methods.

3. N. Nguyen, M. Nute, S. Mirarab, and T. Warnow (2016). "HIPPI: Highly accurate protein family classification with ensembles of HMMs." *BMC Genomics* 17 (Suppl 10):765, special issue for RECOMB-CG. (See description above.)

4. B.M. Boyd, J.M. Allen, N. Nguyen, A.D. Sweet, T. Warnow, M.D. Shapiro, S.M. Villa, S.E. Bush, D.H. Clayton, and K.P. Johnson. "Phylogenomics using Target-restricted Assembly Resolves Intra-generic Relationships of Parasitic Lice (Phthiraptera: Columbicola)." Systematic Biology 2017, doi: 10.1093/sysbio/syx027. Methods developed and improved during the first Blue Waters allocation were used to analyze this biological dataset. Kevin Johnson (senior author on the paper) is an Assistant Research Scientist at the Illinois Natural History Museum.

5. J.M. Allen, B. Boyd, N. Nguyen, P. Vachaspati, T. Warnow, D.I. Huang, P.G. Grady, K.C. Bell, Q.C. Cronk, L. Mugisha, B.R. Pittendrigh, L.M. Soledad, D.L. Reed, and K.P. Johnson. "Phylogenomics from Whole Genome Sequences Using aTRAM". Systematic Biology 2017, doi:10.1093/sysbio/syw105. Methods developed and improved during the

first Blue Waters allocation were used to analyze this biological dataset. Kevin Johnson (senior author on the paper) is an Assistant Research Scientist at the Illinois Natural History Museum.

Software All the methods developed in this project are made available in open source form, on github.

- HIPPI: https://github.com/smirarab/sepp, a github site maintained by Siavash Mirarab (former student).
- FastRFS: https://github.com/pranjalv123/FastRFS, a github site maintained by Pranjal Vachaspati (current PhD student).
- PASTA+BAli-Phy: https://github.com/MGNute/pasta, a github site maintained by Michael Nute (current PhD student).

References and Notes

1. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–10, 1990.
2. J. Chifman and L. Kubatko. Quartet inference from SNP data under the coalescent. *Bioinformatics*, 30(23):3317–3324, 2014.
3. S.R. Eddy. A new generation of homology search tools based on probabilistic inference. *Genome Inform*, 23:205–211, 2009.
4. Joseph Heled and Alexei J Drummond. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580, 2010.
5. S. Mirarab, N. Nguyen, L.-S. Wang, S. Guo, J. Kim, and T. Warnow. PASTA: ultra-large multiple sequence alignment of nucleotide and amino acid sequences. *J. Computational Biology*, 22:377–386, 2015.
6. N. Nguyen, S. Mirarab, K. Kumar, and T. Warnow. Ultra-large alignments using phylogeny aware profiles. *Genome Biology*, 16(124), 2015. A preliminary version appeared in the Proceedings RECOMB 2015.
7. N. Nguyen, S. Mirarab, B. Liu, M. Pop, and T. Warnow. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics*, 30(24):3548–3555, 2014.
8. N. Nguyen, M. Nute, S. Mirarab, and T. Warnow. HIPPI: highly accurate protein family classification with ensembles of hidden Markov models. *BMC Bioinformatics*, 17 (Suppl 10):765, 2016. Special issue for RECOMB-CG 2016, to appear.

9. M. Nute and T. Warnow. Scaling statistical multiple sequence alignment to large datasets. *BMC Bioinformatics*, 2016. Special issue for RECOMB-CG 2016, to appear.

10. B. Redelings. Bali-phy user guide, 2017. Available at `http://www.bali-phy.org/README.html#mixing_and_convergence`.

11. B. Redelings and M. Suchard. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.*, 54(3):401–418, 2005.

12. J. Söding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–960, 2005.

13. M.A. Suchard and B.D. Redelings. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, 22:2047–2048, 2006.

14. P. Vachaspati and T. Warnow. FastRFS: fast and accurate Robinson-Foulds Supertrees using constrained exact optimization. *Bioinformatics*, 2016. doi: 10.1093/bioinformatics/btw600.