

Final Report

1 Project Information

- Project Title: Advancing Genome-scale Phylogenomic Analysis
- 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

The project has three main aims, each geared towards advancing the accuracy of large-scale estimation of evolutionary history. The first year has focused on the first aim, method development for multiple sequence alignment, maximum likelihood phylogeny estimation, and species tree estimation. Highlights include (a) scalable versions of BALi-Phy (*11*), 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 UIUC participated in this project.

3 Description of research activities and results

3.1 Key Challenges

The main goals of this project are all related to phylogenomics, i.e., the estimation of evolution for large datasets, up to hundreds of thousands of species and using whole genomes. In our proposal, we identified three specific scientific questions that arise in phylogenomic analyses – multiple sequence alignment, large-scale maximum likelihood, and species tree estimation from multiple loci – none of which is adequately solved using existing methods when large datasets are considered. We have expanded this set of scientific problems to include remote homology detection and gene family assignment, since biological dataset analysis begins with data collection where answering these two questions is necessary.

These scientific challenges are also computationally very difficult, because nearly all desirable approaches are computational methods that attempt to solve NP-hard optimization problems, or that seek to find optimal statistical models in a high-dimensional parameter space.

Many biological datasets require months of analysis, and some recent biological studies (e.g., the Avian phylogenomics project (5)) have spent hundreds of CPU years in computing phylogenies for their phylogenomic datasets.

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 inferences 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, about 50 single gene sequences for 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

In our proposal, we stated three main aims, all geared at improving large-scale phylogenomic analysis:

- Aim 1: New methods with improved accuracy for multiple sequence alignment, large-scale maximum likelihood, and species tree estimation from multiple genes.
- Aim 2: Parallel implementations of these methods that can take advantage of the Blue Waters architecture and provide excellent scalability.
- Aim 3: An automated pipeline for use by biologists to analyze their genome-scale datasets.

As planned, the first year's activity focused on Aim 1, and produced advances in each target problem. Overall this work has lead to 12 papers that are in various stages of preparation. Of these, 1 paper has been published, 2 others have been submitted, and 3 other papers are being prepared for submission and should be submitted before the end of summer.

Highlights in method development:

- Scalable versions of BALi-Phy (*11*), a Bayesian method for statistical co-estimation of multiple sequence alignments and phylogenetic trees, so that it can analyze datasets with thousands of sequences. The previous version was limited to at most 50 or so sequences. This research has been presented as a poster at the Pacific Symposium on Biocomputing (PSB) 2016, and will be submitted to a journal before the end of Summer 2016.
- A new method (HIPPI) for classifying sequences (including short reads generated by sequencing technologies) into gene families and improves the ability to detect remote homology, and can be used to improve analyses of metagenomic samples and phylogenetic dataset assembly. A paper on this has been submitted to a journal.
- A new supertree method with improved accuracy and greatly reduced running times based on dynamic programming. A paper on this has been submitted to a journal.

We are also using Blue Waters to perform large-scale phylogenomic and metagenomic analyses of biological datasets, in collaborations with research groups here at UIUC and around the country. These analyses are helping to publicize the methods, but also give us insight into how we can improve the methods in terms of accuracy and computational performance. Many of these analyses involve estimating multiple sequence alignments using our new methods (PASTA (6) and UPP (9)) and estimating species trees from multi-locus datasets using maximum likelihood heuristics such as RAxML (14) or FastTree (10). Some species tree analyses have also been based on ASTRAL-2 (7), 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. In the rest of this section, we go into greater detail about the major activities this year.

HIPPI: Hierarchical Profiles for Homology Detection. The detection of homology and gene family identification is a basic step in many bioinformatics analyses, with applications to protein structure and function prediction, metagenomic taxon identification and abundance profiling, and systems biology, among others. For decades, BLAST (1) was the leading method for homology detection, but the use of profile Hidden Markov Models (HMMs) (2) in various forms has become increasingly popular, especially for homology detection and gene family classification of amino acid sequences. Currently, the most accurate methods for homology detection include BLAST, HMMER (3), and HHSearch (13). Yet even these methods have substantial difficulties in correctly classifying protein sequences in the presence of low sequence similarity (the “twilight zone” (12)), or for fragmentary sequences produced by Illumina and other short read sequencing technologies. Indeed, remote homology detection remains one of the more difficult analytical problems in bioinformatics.

My postdoctoral fellow Nam-phuong Nguyen, my students Mike Nute and Siavash Mirarab (now my former student), and I have developed a new technique for homology detection and gene family identification called “HIPPI” (Hierarchical Profiles for Homology Prediction). HIPPI decomposes a representative set of sequences for a protein family or superfamily into hierarchical subsets, and represents the family by an ensemble of such HMMs, one on each subset, rather than by a single HMM for the whole family as previous methods had done. Then, given a query sequence (i.e., one for which the correct protein family is not known), it uses these ensembles of HMMs to find the family with the best fit to the query sequence. Thus, HIPPI can be used to perform gene family identification for unclassified sequences, and also to quantify the fit between a query sequence and a given protein family. To evaluate HIPPI, we downloaded 11,157 protein families from the Pfam database, and used Blue Waters to run the entire experimental pipeline. This included estimating thousands of maximum likelihood trees and assigning more than one million (1,000,000) protein sequences to Pfam families.

Our study (submitted for publication) show that HIPPI has better overall precision and recall than any of the leading methods for gene family classification. In addition, HIPPI is robust to fragmentary data, and is able to maintain high precision and recall while other methods (and in particular single profile HMMs) have greatly reduced recall.

Scalable statistical co-estimation of multiple sequence alignments and phylogenetic trees.

BALi-Phy (11) is a Bayesian statistical method that co-estimates multiple sequence alignments and trees, and is the leading available software for statistical co-estimation of alignments and trees. However, because of the MCMC strategy, BALi-Phy is limited to very small numbers of sequences, and can take weeks to converge on these datasets (as we and others have noted). We proposed to improve the scalability of BALi-Phy by using divide-and-conquer. Specifically, we proposed to divide sequence datasets into small subsets (using tree-based decomposition strategies), align small subsets of sequences using BALi-Phy, and then combine the alignments together into an alignment on the full taxon set; we would then use maximum likelihood to estimate trees from these merged alignments.

Initial results on this research has been performed by my PhD student Mike Nute. Mike integrated BALi-Phy into two of our earlier methods – PASTA and UPP. Mike has tested these methods, and demonstrated that the integration of PASTA with BALi-Phy allows BALi-Phy to scale to at least 100 sequences, and that the integration of UPP with BALi-Phy allows BALi-Phy to scale to 10,000 sequences. Given that the published version of BALi-Phy cannot realistically be run on more than 25 sequences, this is very satisfying. Also, the integration of BALi-Phy into PASTA and UPP showed improvements over their default versions (which use MAFFT to estimate alignments on subsets of sequences), thus confirming our hypothesis that this integration would achieve improved accuracy compared to prior methods.

As an example, Figure 1 shows a scatterplot of the difference in alignment error (measured as the average of the false positive and false negative rates) of UPP in default mode (where it computes a backbone alignment on 100 randomly selected sequences using PASTA, and then aligns the remaining sequences to the backbone alignment) and UPP using a BALi-Phy backbone alignment on the same randomly selected sequences. We explored performance on a range of datasets, including some biological RNA sequence datasets with structural alignments (the datasets with “gutell” in their names) and simulated datasets; these datasets ranged in size from above 5000 sequences (gutell_16S.3) to 10,000 sequences (the indelible_10000M2 through indelible_10000M4 datasets). Points *above* the $x = y$ diagonal refer to datasets on which Default UPP had lower alignment error than UPP+BALi-Phy, while points below the $x = y$ diagonal refer to datasets on which Default UPP had higher alignment error than UPP+BALi-Phy. On nearly all these datasets, using BALi-Phy instead of PASTA to produce the backbone alignment resulted in reduced alignment error, showing that integrating BALi-Phy within UPP improved UPP’s accuracy.

To analyze large datasets, we decomposed each dataset into subsets with at most 25 sequences so that BALi-Phy could converge efficiently on each subset. Hence the analysis of a dataset with 500 sequences requires running BALi-Phy on 40 datasets with 25 sequences each; hence analyzing a dataset with 1,000,000 sequences using the combination of PASTA and BALi-Phy would involve 40,000 analyses using BALi-Phy. Since these can be run in parallel, Blue Waters is an ideal platform for this type of method. We plan to submit a manuscript on this research by the end of the summer.

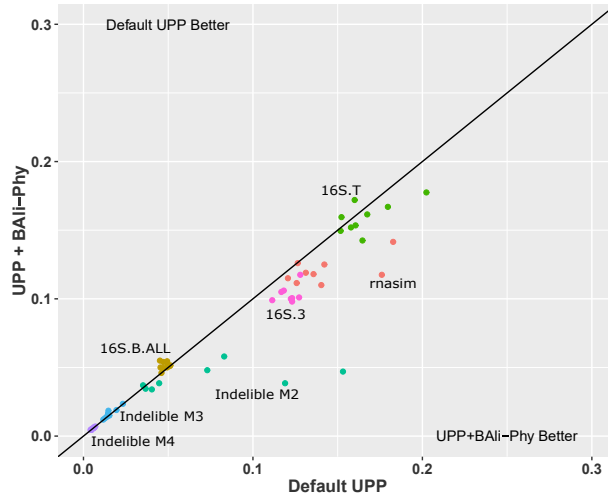


Figure 1: **Scatterplot of multiple sequence alignment error rates of UPP-default compared to UPP+Bali-Phy on datasets with up to 10,000 sequences.** We show alignment error, using the average of false positive and false negative rates, of UPP in default mode to UPP using Bali-Phy (written as UPP+Bali-Phy), in both cases using a backbone of size 100. UPP in default mode uses PASTA to compute the backbone alignment and tree, while UPP+Bali-Phy uses Bali-Phy to compute the backbone alignment and tree. We explore performance on biological and simulated datasets, ranging in size from a bit above 5000 sequences (gutell_16S.3) to 10,000 sequences (the indelible_10000M2 through indelible_10000M4 datasets). Points *above* the $x = y$ diagonal refer to datasets on which Default UPP had lower alignment error than UPP+Bali-Phy, while points *below* the $x = y$ diagonal refer to datasets on which Default UPP had higher alignment error than UPP+Bali-Phy. On nearly all these datasets, using Bali-Phy instead of PASTA to produce the backbone alignment resulted in reduced alignment error, showing that integrating Bali-Phy within UPP improved UPP's accuracy.

Fast and accurate supertree estimation using constrained optimization. The construction of large evolutionary trees is computationally challenging since nearly all good approaches involve attempts to solve NP-hard optimization problems, and rely on search heuristics to explore an exponentially sized treespace. In addition, large evolutionary trees are statistically challenging, because different parts of the Tree of Life evolve under different statistical models, thus requiring the application of methods for estimating trees under highly complex statistical models with large numbers of parameters. Divide-and-conquer strategies that divide the species set into local overlapping subsets, construct trees on each subset, and then combine the subset trees into a tree on the full set of species, can be highly accurate and efficient techniques for analyzing large datasets.

Supertree methods are methods that combine overlapping subset trees into trees on the full set of species, and are key parts of any divide-and-conquer strategy. My student Pranjali Vachaspati and I have developed a new approach to supertree estimation that has shown excellent performance on preliminary data. Specifically, we have developed FastRFS, a method for supertree estimation that uses a dynamic programming algorithm to solve a constrained version of a standard optimization problem for supertree construction, Robinson-Foulds Supertrees.

Our new method, FastRFS, has a basic version and an enhanced version, depending on how the constraint space is selected. We compared these two variants of FastRFS to the best performing methods for this optimization problem (MulRF and PluMiST) and also to two other supertree methods (ASTRAL, ASTRID, and MRL). As shown in Figure 2, FastRFS-enhanced clearly dominated all the other methods we tested in terms of criterion scores: on every model condition of the simulated datasets, FastRFS-enhanced found trees with scores that were at least as good as the other methods. In second place was FastRFS-basic, which also found the best score in three conditions, and came in second in all but two of the remaining model conditions. The other methods each did well in some cases. These analyses showed that FastRFS's dynamic programming algorithm produced better solutions to the Robinson-Foulds Supertree optimization problem than MulRF and PluMiST, so that the exact solution to the constrained optimization problem was more accurate than the heuristic search used by either of these methods. Also, and even more important, FastRFS produced supertrees that matched or exceeded the topological accuracy of these other methods. For example, on 500-species simulated datasets with small (20%) scaffolds, MulRF had 50% tree error, MRL had 16% tree error, and our DP algorithm had 15% tree error. (Results on other model conditions with large numbers of species where all methods could be run showed similar trends.)

Furthermore, FastRFS was much faster than both of the alternative methods, and could run on much larger datasets than the other methods. For example, on the benchmark simulated datasets with 500 species, our algorithm completed in just over one minute (61-75 seconds per dataset), MRL completed in 34 minutes, and MulRF completed in 424 minutes (just over 7 hours). On the 1000-species datasets, only our algorithm was able to complete in under 24 hours; all other methods failed to complete within 24 hours. On these 1000-species datasets, FastRFS had average running times on these datasets of under 5 minutes. Running time results on biological supertree datasets show similar advantages of FastRFS; see Figure 3.

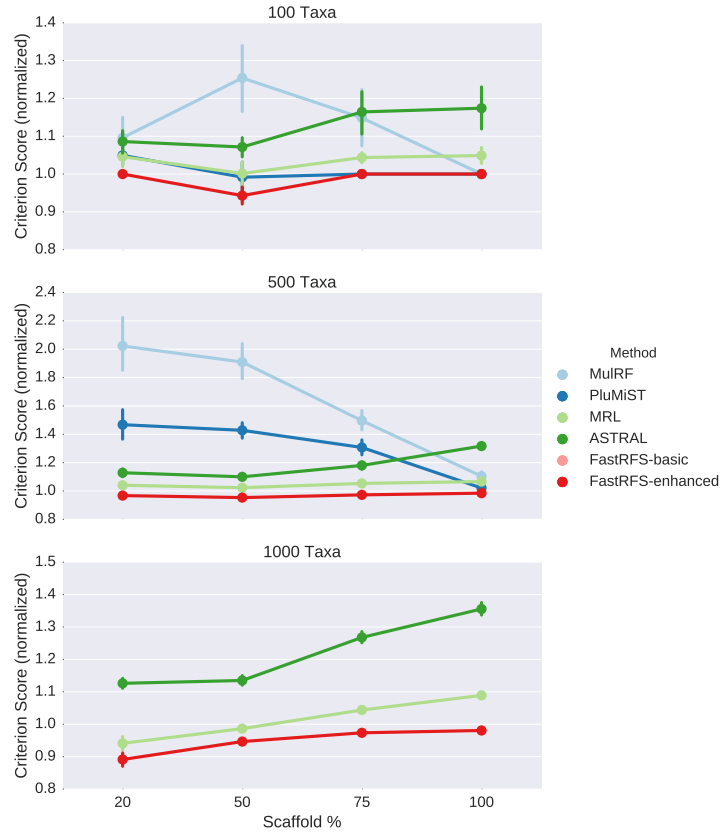


Figure 2: Normalized Robinson-Foulds Supertree criterion scores on the simulated datasets; lower is better. Normalized scores are obtained by taking the raw score and then dividing by the score for FastRFS-basic; hence, FastRFS-basic always has a score of 1.0. Results for FastRFS-basic are not shown for this reason. Results for ASTRID are not shown because (with the exception of the 100%-scaffold datasets) they were exceedingly poor. Results are not shown for MulRF and PluMiST on the 1000-taxon datasets because they were too computationally intensive to run on those data. Values below 1.0 indicate that the method returned a better score than FastRFS-basic; values above 1.0 indicate a worse score.

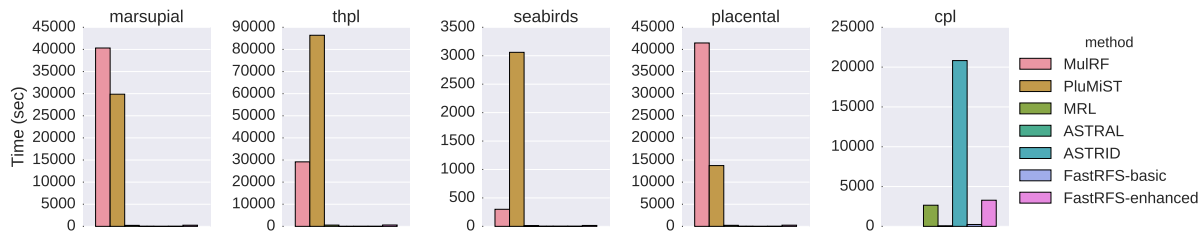


Figure 3: Sequential running times (in seconds) on biological data of supertree methods. MulRF and PluMiST could not be run on the CPL dataset, due to its large size; hence no values are shown for those methods on that dataset.

Biological dataset analyses: We have collaborated with several biologists in analyzing their datasets, and used Blue Waters for those large-scale analyses that could not be completed using other resources. These studies provide valuable information to us in terms of our method development, and also help to promote the methods for use in the biological and biomedical research community. Furthermore, because our methods are able to provide improved accuracy compared to other methods, they also lead to improved biological discoveries.

For example, in collaboration with UIUC Professor Bryan White (IGB), we performed a study (in revision) comparing methods for defining OTUs (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. In another collaboration (with UIUC professor Derek Wildman (IGB)), we have used our methods for multiple sequence alignment and phylogeny estimation (PASTA and UPP) and species tree estimation from multi-locus data (ASTRAL-2) to estimate the evolution of 48 highly diverged nuclear receptor proteins using more than 2000 sequences, in order to investigate the origins of “orphan receptors”. We have also worked with Sarah Tishkoff (Professor at the University of Pennsylvania) in an analysis of human gut microbiomes from African populations using TIPP (8), a method we have developed for taxonomic identification and abundance profiling of shotgun metagenomic sequence data; this study shows how dietary lifestyles influence gut microbiomes. Two other studies using Blue Waters to perform large-scale analyses of biological datasets have also been performed, and are mentioned in the list of publications and products below.

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

I include all papers that were used the Blue Waters allocation to me. Many of these include my graduate students (Siavash Mirarab, Mike Nute, Pranjal Vachaspati, and Ashu Gupta) or postdocs (Nam-phuong Nguyen and Ruth Davidson). These names (and my own) are boldfaced in the papers below.

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. *These papers acknowledge (or will acknowledge) Blue Waters support.*

I note also that the research in these papers will be, in nearly every case, in the PhD dissertation of one of my students (Mike Nute and Pranjal Vachaspati) or has appeared in the MS thesis of my student Ashu Gupta. Thus, this allocation is also providing support to the educational mission of UIUC.

Publications

1. **Nguyen N., Warnow T.,** 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.

Submitted

1. **Nguyen N., Nute M., Mirarab S., and Warnow T.** “HIPPI: Highly accurate protein family classification with Ensembles of HMMs.”
2. **Vachaspati P. and Warnow T.** “FastRFS: Fast and accurate Robinson-Foulds Supertrees using constrained exact optimization.”

In preparation, Blue Waters analyses completed

1. Allen J., Boyd B., **Nguyen N., Vachaspati P.,** Huang D., Gero P., Bell K., Cronk Q. , **Warnow T.,** and Johnson K. “Phylogenomics using Genome Sequencing and aTRAM”. (Blue Waters analyses completed.)
2. Boyd B., Allen J., **Nguyen N., Vachaspati P.,** Quicksall Z., **Warnow T.,** Johnson K., and Reed D. “Lousey trees obscure the origins of human primate louse symbionts.” (Blue Waters analyses completed.)
3. Ortiz X., **Nguyen N., Warnow T.,** and Wildman D. “Vertebrate nuclear receptor phylogenetics: a tool for understanding the evolution of orphan nuclear receptors. (Blue Waters analyses completed.)

Studies requiring further analyses on Blue Waters

1. **Nute M., Nguyen N., Peng J., and Warnow T.** “Improved protein alignment in the twilight zone.”
2. **Davidson, R., Vachaspati P., and Warnow T.** “Statistically consistent estimation of rooted species trees from unrooted gene trees under the multi-species coalescent using algebraic statistics.”
3. **Gupta A., Bayzid Md. S., Vachaspati P., and Warnow T.** “Scalable and accurate co-estimation of gene trees and species trees.” (This is part of the UIUC Computer Science MS thesis for Ashu Gupta, which was submitted in April 2016.)
4. Hansen M., Rubel M., Bailey A., Bittinger K., Laughlin A., **Nguyen N.,** Beggs W., Ranciaro A., Thompson S., **Warnow T.,** Bushman F., and Tishkoff S. “Diet, Environment, and Parasites: Factors Shaping Rural African Gut Microbiomes.”
5. **Nute M., Nguyen N., and Warnow T.** “Scaling Bayesian co-estimation of multiple sequence alignments and trees to thousands of sequences.”
6. **Vachaspati P., Davidson R., and Warnow T.** “A new implementation of SVDquartets, a site-based species tree method, with improved accuracy and speed”.

References and Notes

1. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–10, 1990.
2. Sean R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.
3. Sean 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. Erich D. Jarvis, Siavash Mirarab, et al. Whole genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, 2014.
6. Siavash Mirarab, Nam-phuong Nguyen, Sheng Guo, Li-San Wang, Junhyong Kim, and Tandy Warnow. PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology*, 2014.

7. Siavash Mirarab and Tandy Warnow. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, 2015. Preliminary version appeared in the Proceedings of Intelligent Systems for Molecular Biology (ISMB) 2015.
8. N.-p. Nguyen, S. Mirarab, B. Liu, M. Pop, and T. Warnow. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics*, 30(24):3548–3555, 2014.
9. Nam-phuong Nguyen, Siavash Mirarab, Keerthana Kumar, and Tandy Warnow. Ultra-large alignments using phylogeny-aware profiles. *Genome Biol.*, 16(1):124, jun 2015.
10. Morgan N Price, Paramvir S Dehal, and Adam P Arkin. FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS One*, 5(3):e9490, 2010.
11. Ben Redelings and Marc Suchard. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.*, 54(3):401–418, 2005.
12. B Rost. Twilight zone of protein sequence alignments. *Protein engineering*, 12(2):85–94, 1999.
13. Johannes Söding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–960, 2005.
14. Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.