PARALLEL ALGORITHMS FOR BIG DATA PHYLOGENOMICS, PROTEOMICS, AND METAGENOMICS

Allocation: Illinois/125 Knh
PI: Tandy Warnow[1]

[1]University of Illinois at Urbana-Champaign

EXECUTIVE SUMMARY

This project addressed three interrelated problems in computational molecular biology, where large data sets present substantial computational and statistical challenges: phylogenomics (genome-scale phylogeny estimation), proteomics (protein structure and function prediction), and metagenomics (analysis of environmental samples from shotgun sequence data sets). Highlights of this project's activity include: (1) SVDquest, a method for species tree estimation from multilocus data sets that bypasses gene tree estimation (Tandy Warnow, with Ph.D. student Pranjal Vachaspati); (2) HIPPI: a method for protein family classification (Tandy Warnow, with Ph.D. student Mike Nute and two others); (3) an evaluation of the impact of screening genes in multilocus phylogenomic analyses (Tandy Warnow, with Ph.D. student Erin Molloy); and (4) an evaluation of statistical methods for multiple sequence alignment on protein benchmark data sets (Tandy Warnow, with Ph.D. students Ehsan Saleh and Mike Nute, and undergraduate researcher Kodi Collins). Five journal papers based on this work were published this year and another two were submitted.

RESEARCH CHALLENGE

This project aimed to develop methods for large-scale statistical estimation problems of phylogenies, multiple sequence alignments, and analyses of metagenomic data sets, where standard methods either do not run or provide poor accuracy. The need for new methods is particularly urgent as more and more studies attempt to analyze phylogenomic data sets 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 is encountering these challenges in its plans to assemble phylogenies of the major groups of life on earth.

METHODS & CODES

We made progress on each problem using a combination of algorithmic approaches. In many cases we used divide-and-conquer, which allows powerful statistical off-the-shelf techniques to be applied to small subsets of a large data set, followed by innovative approaches to combine results from the small data sets.

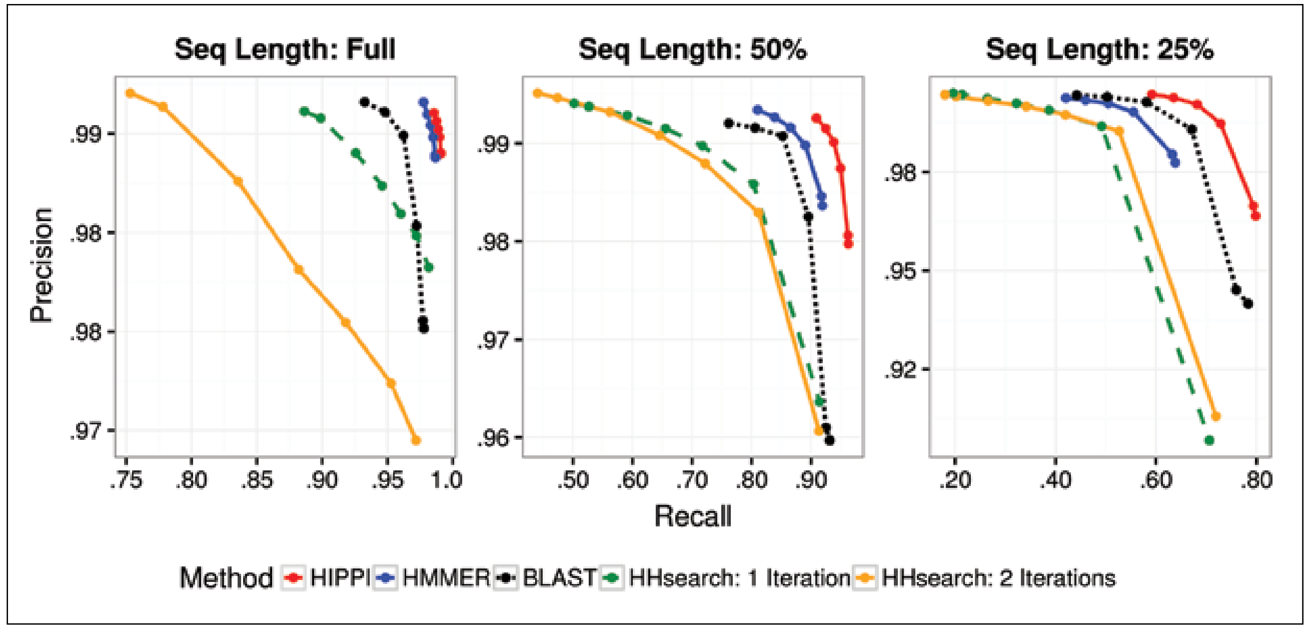

Figure 1: Precision-recall curves for methods for protein family classification, evaluated on one cross-fold subset of the PFAM seed sequence data. Our new method—HIPPI (shown in red)—strictly dominates all the other methods. Figure taken from [3].

RESULTS & IMPACT

One of the main outcomes of this project is HIPPI [1], a method for protein family classification. 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 [2] is the most well-known method for this problem, but other approaches based on profile Hidden Markov Models (HMMs) have been used as well. In this work, we developed a novel machine-learning technique to detect membership in existing protein families, where we construct an ensemble of profile HMMs to represent each protein family, and then compare 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 [3] database of protein families and their associated profile HMMs from HMMER [4] 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 [5]) in terms of both precision and recall, especially when analyzing short sequences (Fig. 1).

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 data sets (and even of moderate-sized data sets) 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.

PUBLICATIONS AND DATA SETS

Vachaspati, P., and T. Warnow, FastRFS: Fast and accurate Robinson-Foulds Supertrees using constrained exact optimization. *Bioinformatics*, (2016), DOI: 10.1093/bioinformatics/btw600.

Nute, M. and T. Warnow, Scaling statistical multiple sequence alignment to large datasets. *BMC Genomics*, 17 (Supplement 10):764 (2016), DOI: 10.1186/s12864-016-3101-8.

Nguyen, N., M. Nute, S. Mirarab, and T. Warnow, HIPPI: Highly accurate protein family classification with ensembles of HMMs. *BMC Genomics 17* (Supplement 10), 765 (2016), DOI: 10.1186/s12864-016-3097-0.

Boyd, B. M., et al., Phylogenomics using Target-restricted Assembly Resolves Intra-generic Relationships of Parasitic Lice (Phthiraptera: Columbicola). *Systematic Biology*, (2017), DOI: 10.1093/sysbio/syx027.

Allen, J.M., et al., Phylogenomics from Whole Genome Sequences Using aTRAM. *Systematic Biology* (2017), DOI: 10.1093/sysbio/syw105.

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 Ph.D. student).

PASTA+BAli-Phy: https://github.com/MGNute/pasta, a github site maintained by Michael Nute (current Ph.D. student).