Optimization of Microbial Community Proteomics

W.R. Cannon[1]*, A. Hugo[1], D. J. Baxter[2], A. Kalyanaraman[3] and [1]B.-J. Webb-Robertson

[1]Computational Biology and Bioinformatics Group, [2]Molecular Sciences Computing Facility, Pacific Northwest National Laboratory, Richland, WA 99352; [3]School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, 99164.

Goals: Analyzing proteomics data from environmental samples is extremely challenging. Generally, the analyses fall into one of two categories, either the complete lack of metagenome sequence information to match proteome data against, or the incomplete nature of metagenome sequence information. The goals of this 3-year SciDAC/GSP project were to alleviate this situation by leveraging high performance computing, advanced modeling and algorithms to develop more accurate and more complete analyses. Achievements 1-3 below address the accuracy issue, and achievements 4-6 address issues regarding complete analysis of metaproteome/metagenome data.

1. **Developed advanced models based on the mathematical overlap between statistical data analysis and statistical thermodynamics [1].**
2. **Implemented the advanced models to bear on problems relevant to DOE's mission in Bioenergy [2].**
3. **Released code for these tools that can be used on multiple computing platforms including workstations, HPC clusters and cloud resources [2-4].**
4. **Developed HPC solutions to the challenge of detecting protein sequence homology for expanding genome and metagenome datasets [5, 6].**
5. **Developed methods to analyze samples when no reference genomes or metagenomes exist [7].**
6. **Developed methods to maximally use metagenome sequences when uncertainty about the gene calls is not small [8].**

These last four achievements are discussed briefly below.

Abstract

Proteotyping Environmental Samples without a Metagenome. When a metagenome sequence is not available, proteomic analysis is often not possible. This year we reported the development of a novel high performance computing method for proteotyping environmental samples. The method uses computational optimization to provide an effective way to control the false discovery rate. The method provides phylum/species information based on the expressed proteins in a microbial community, and thus complements DNA-based methods. Testing on blind samples demonstrates that the method provides 79-95% overlap with analogous results from searches involving only the correct genomes. Scaling and performance evaluations for the software demonstrates the ability to carry out large-scale optimizations on 1258 genomes containing 4.2M proteins [7].

Proteome Matching against a Metagenome: Bayesian Integration of Evidence. When a metagenome sequence is available, specific identifications of proteins are possible. However, due to the imperfect nature of metagenome sequence information, the protein identification rate is usually quite a bit lower than what is typical for single organism laboratory studies. This need not be the case if multiple levels of evidence are assessed when identifying proteins from an environmental sample. We report the development of a Bayesian framework that incorporates evidence based on peptide detectibilities, significance of MSMS matches, and prior probabilities of protein occurrence. The method more than doubles the number of spectra that are identified with proteins, and increases the number of proteins that are identified by 40-50%.

Proteomics Analysis Code: MSPolygraph. The analyses above were carried out using high performance computing. MapReduce, MPI and serial versions of *MSPolygraph* for peptide identification from mass spectrometry data have been developed. The MPI version runs on any MPI-capable cluster [2]. The MapReduce implementation can run on any Hadoop cluster environment [3]. Availability: The source code along with user documentation is available at omics.pnl.gov/software/MSPolygraph.php.

Remote Homology Detection. Large-scale projects generate millions of *new* sequences that need to be matched against themselves and against already available sequences. For example, the ocean microbiota survey project in 2007 analyzed a total of 28.6M sequences. The most time consuming step during analysis was homology detection, which accounted for 10^6 CPU hours despite the use of fast approximation heuristics such as BLAST.

We developed a novel parallel algorithm, *pGraph,* to efficiently parallelize the construction of sequence homology graphs from large-scale protein sequence data sets based on dynamic programming alignment computation. The parallel design is a hybrid of multiple-master/worker and producer-consumer models, which effectively addresses the unique set of irregular computation issues and input data availability issues. The implementation scales linearly up to 2,048 processors using up to 2.56×10^6 metagenomic protein sequences [6].

Computations were performed in the National Energy Research Scientific Computing Center (NERSC) in Berkeley, CA, and in the Molecular Sciences Computing Facility at the Environmental Molecular Sciences Laboratory (EMSL).

Funding: This research was supported by the Genomic Science Program, Office of Biological and Environmental Research, and the Scientific Discovery through Advanced Computing program, Office of Advanced Scientific Computing Research, U.S. Department of Energy.

[1] W. R. Cannon and M. M. Rawlins, "Physicochemical/Thermodynamic Framework for the Interpretation of Peptide Tandem Mass Spectra," *Journal of Physical Chemistry C,* vol. 114, pp. 5360-5366, Apr 1 2010.

[2] W. R. Cannon, M. M. Rawlins, D. J. Baxter, S. J. Callister, M. S. Lipton, and D. A. Bryant, "Large Improvements in MS/MS-Based Peptide Identification Rates using a Hybrid Analysis," *J Proteome Res,* Mar 30 2011.

[3] A. Kalyanaraman, W. R. Cannon, B. Latt, and D. J. Baxter, "MapReduce implementation of a hybrid spectral library-database search method for large-scale peptide identification," *Bioinformatics,* vol. 27, pp. 3072-3, Nov 1 2011.

[4] G. Kulkarni, A. Kalyanaraman, W. R. Cannon, and D. Baxter, "A Scalable Parallel Approach for Peptide Identification from Large-Scale Mass Spectrometry Data," presented at the International Conference on Parallel Processing Workshops (ICPPW 09), 2009.

[5] C. Wu, A. Kalyanaraman, and W. R. Cannon, "A scalable parallel algorithm for large-scale protein sequence homology detection," in *39th International Conference on Parallel Processing (ICPP)*, 2010, pp. 333-342.

[6] C. Wu, A. Kalyanaraman, and W. R. Cannon, "pGraph: Efficient Parallel Construction of Large-Scale Protein Sequence Homology Graphs," *Revision submission- IEEE Transactions on Parallel and Distributed Systems,* 2012.

[7] A. Hugo, D. J. Baxter, W. R. Cannon, A. Kalyanaraman, G. Kulkarni, and S. J. Callister, "Proteotyping of Microbial Communities using High Performance Optimization of Proteome-Spectra Matches," in *Pacific Symposium BioComputing*, In Press, 2012.

[8] A. Hugo, B. J. Webb-Robertson, and W. R. Cannon, "Increased Proteome Coverage from Environmental Samples using Bayesian Analysis of Evidence," in *In Preparation*, 2012.