SEQUENCE SIMILARITY NETWORKS FOR THE PROTEIN "UNIVERSE"

Allocation: Illinois/0.625 Mnh
PI: John A. Gerlt[1,2]
Personnel: Katie L. Whalen[2], Daniel Davidson[1], Boris Sadkhin[1], David Slater[1]
Collaborator: Alex Bateman[3]

[1]Institute for Genomic Biology, University of Illinois at Urbana–Champaign
[2]Enzyme Function Initiative
[3]European Bioinformatics Institute, European Molecular Biology Laboratory

EXECUTIVE SUMMARY:

The Enzyme Function Initiative (EFI) is supported by the National Institutes of General Medical Sciences (U54GM093342-04). The EFI is devising strategies and tools to facilitate prediction of the *in vitro* activities and *in vivo* metabolic functions of uncharacterized enzymes discovered in genome projects. This project is enabling generation of a library of precomputed sequence similarity networks (SSNs) for all 14,831 Pfam sequence-based families, 515 Pfam sequence-based clans, and 2,737 Gene3D/CATH structure-based superfamilies in the UniProtKB protein database; the SSNs will be provided to the scientific community via a webserver [1]. At present, the UniProtKB database contains 92,672,207 nonredundant sequences (release 2015_3; 03-March-2015). We used Blue Waters to calculate the required all-by-all BLAST sequence comparisons as well as generate statistical analyses of the BLAST results. We plan to calculate the networks on a two-month refresh cycle so that the library will remain current.

INTRODUCTION

The Enzyme Function Initiative (EFI) is supported by the National Institutes of General Medical Sciences (U54GM093342) and is multi-institutional (nine academic institutions in the U.S. and the European Molecular Biology Laboratory/European Bioinformatics Institute in the U.K.) and multi-disciplinary (bioinformatics, protein production, structural biology, homology modeling, *in silico* ligand docking, experimental enzymology, microbiology, transcriptomics, and metabolomics) [2]. The goal of the EFI is to devise tools and strategies to enable prediction of the *in vitro* activities and *in vivo* metabolic functions of uncharacterized enzymes discovered in genome projects.

The UniProtKB database contains 92,672,207 sequences (release 2105_03; 04-March-2015). The functions for 547,964 entries (0.5%) have been manually curated [3]; the functions for the remaining have been assigned by automated procedures [4]. The majority of the entries were obtained from microbial genome sequencing projects, with the rationale that knowledge of the complete set of proteins/enzymes encoded by an organism will allow its biological/physiological capabilities to be understood. However, if many of the proteins/enzymes have uncertain or unknown functions, people cannot capitalize on the investments in genome projects. The EFI was conceived to meet this challenge.

Bioinformatic tools are integral to the EFI's strategies. Phylogenetic trees and dendrograms are the usual bioinformatic representations of relationships among homologous proteins. However, their construction requires structure-based sequence alignments and is computationally intensive. Recently, Babbitt described the use of sequence similarity networks (SSNs) to visualize relationships in families of homologous proteins [5]. Sequence similarities are quantitated by the BLAST bit-scores between pairs of sequences.

The EFI's goal is to provide to the biological community an on-demand library of SSNs for all 14,831 Pfam sequence-based families, 515 Pfam sequence-based clans, and 2,737 Gene3D/CATH structure-based superfamilies in the UniProtKB protein database and to update this library on a minimum two-month refresh cycle.

METHODS & RESULTS

In the first year of our Blue Waters allocation, we optimized two pieces of codes as well as the Perl scripts that control the flow of data and collect the results.

BLAST v 2.x (*blastall*) is a widely used program developed by the National Center for Biotechnology Information (NCBI). *Blastall* is not efficiently multi-threaded, so we ran as many single-threaded processes per node as there are integer cores available. The bulk of the Blue Waters CPU time is used by *blastall*.

CD-Hit is a sequence clustering algorithm [6] that we used to both generate merged datasets of input sequences and/or post-process sequences flagged as being similar to each other by *blastall*.

In the second year, we initially generated SSNs for virtually all of the Pfam families in Release 49.0 of the InterPro database/Release 2014_10 of the UniProt database. We also developed a webtool that allows members of the scientific community to download the precomputed SSNs [7].

This experience involved improvements in our data generation algorithms that reduced the computational complexity, optimized the performance, and dramatically decreased the node-hours required for the BLAST. In particular, we improved the data generation pipeline by increasing automation, clustering highly similar protein sequences, binning the sequences, and improving our data plot generation algorithms.

With these improvements, we are generating SSNs for virtually all of the Pfam families in Release 50.0 of the InterPro database/Release 2015_02 of the UniProt database. When complete, these also will be made available to the community with our webtool. As new releases of InterPro/UniProt databases are made available, we will continue to generate the library of Pfam SSNs and make these available to the scientific community.

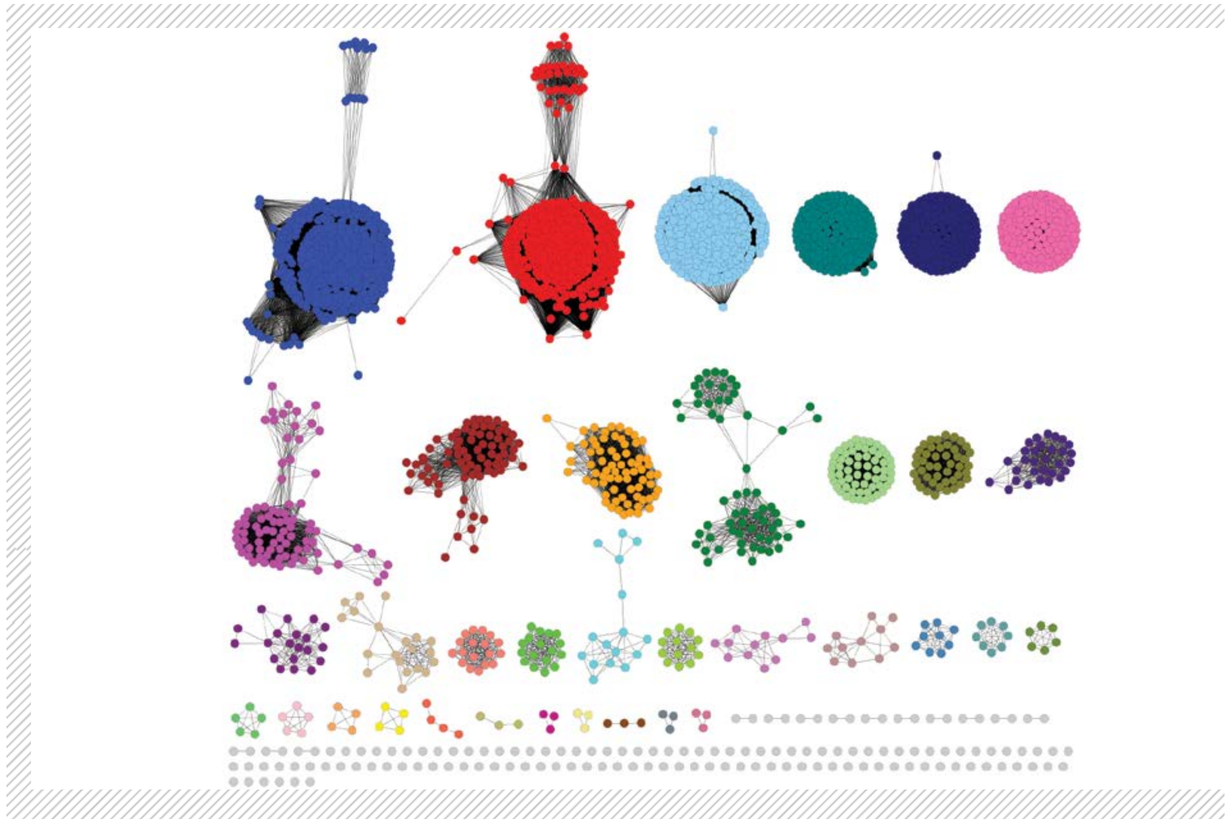

FIGURE 1: Sequence similarity network (SSN) for PF08794, the proline racemase family, displayed with a BLAST e-value threshold of 10^{-110} (50% sequence identity). The colors are used to distinguish predicted isofunctional clusters.

WHY BLUE WATERS?

The project uses an embarrassingly parallel computing model to perform the BLAST analyses and, in principle, could be run on any cluster of sufficient size. However, because of the scales of the computations used (number and sizes of Pfam families and clans and the number of sequences in each family and clan) and the time sensitivity of the production of the output relative to database updates, only a resource of the scale of Blue Waters can perform the job in a reasonable time frame.

Only Blue Waters is able to provide the ultra-large-scale computational power required to generate SSNs for Pfam-defined families and clans as the databases are updated. The sequence database is expected to continue to grow exponentially for the foreseeable future, and only Blue Waters has the necessary capacity to support these increasing needs.