DOE final report

Division: Biological and Environmental Research
Title: Phylogenomic tools and web resources for the Systems Biology Knowledgebase
Grant number: DE-SC0004916

Report date: December 8, 2014

Kimmen Sjolander, PI
University of California Berkeley
kimmen@berkeley.edu

Overview

This project as a whole aimed to develop high-throughput functional annotation systems that exploit information from protein 3D structure and evolution to provide highly precise inferences of various aspects of gene function, including molecular function, biological process, pathway association, Pfam domains, cellular localization and so on. We accomplished these aims by developing and testing different systems on a database of protein family trees: the PhyloFacts Phylogenomic Encyclopedia (at http://phylogenomics.berkeley.edu/phylofacts/).

Major developments in 2011:

In 2011, we worked primarily on revising our database schema, on creating software tools for clustering protein families for the PhyloFacts resource, and on improving the graphical user interfaces for external users to explore PhyloFacts data.

Software engineering efforts have focused on improving the robustness and extensibility of the code base, rewriting code, debugging and refactoring, and developing new software tools. Many of these improvements are not visible on our public-facing PhyloFacts database and webservers (which we call PhyloFacts 2.0). The new PhyloFacts webservers (PhyloFacts 3.0) will be launched later this summer and are currently being extended and tested with Quality Assurance tests.

Specific details of software engineering progress include:
- We have made tangible progress in moving our code and data to new webservers and phasing out dependency on the old servers
- We have substantially improved the UI for family data display
- We have concentrated on database optimization; as a result, several large datasets are loaded and displayed quickly (where previously it might take several minutes for a webpage for a very large protein family to load).
- We have imported data from 3rd party sources : iRefIndex, MetaCyc, KEGG, etc.
- We have written several parsers for bioinformatics data
- We have instituted systematic testing of programs to improve code robustness
- We have started making changes to our database schema to allow for proper querying of BPG data

Quality assurance: Yaoqing Shen (postdoctoral scholar) is playing a lead role in quality assurance. Dr Shen has a PhD in bioinformatics and is responsible for identifying sources of data for inclusion in PhyloFacts, for testing software tools developed by programmers on the project and for assisting in the development of novel software tools (e.g., to reduce redundancy in PhyloFacts protein families).

We have targeted key microbial genomes for coverage in PhyloFacts by building gene family trees for PFAM domains found in microbial genomes and for whole domain architectures (that is, the full-length proteins, clustered on the basis of global similarity). As a result, our coverage of microbial genomes has increased significantly.

Major developments in 2012: In 2012, we worked on increasing our coverage of microbial genomes in PhyloFacts. We have significantly expanded our coverage of microbial gene families. More than 7.3M proteins are included in PhyloFacts families, representing >99K unique taxa (including strains) across >92K families (>25K grouped by PFAM domain and >67K grouped by multi-domain architecture agreement). For many species our coverage is almost complete. For instance, within Archaea, >90% of *Halobacterium salinarum* and >87% of *Sulfolobus solfataricus* are represented by at least one PhyloFacts family. Within Bacteria, 100% of *Escherichia coli K12*, >94% of *Bacillus subtilis*, >91% of *Thermotoga maritime*, >90% of *Geobacter sulfurreducens*, >87% of *Sulfolobus solfataricus* and >79% of *Deinococcus radiodurans* are represented. Within Eukarya, >90% of *Saccharomyces cerevisiae* and >86% of *Arabidopsis thaliana* genes are included. Detailed coverage of representative species with whole genomes is presented at http://phylogenomics.berkeley.edu/phylofacts/coverage/.

We also worked on building phylogenetic trees for Pfam domains in **The PhyloFacts-Pfam project.** This arose out of our work on ortholog identification. In "Ortholog identification in the presence of domain architecture rearrangement," Briefings in Bioinformatics 2011, I showed that phylogenetic tree accuracy and ortholog identification were often superior when based on individual conserved regions and domains rather than when sequences were clustered based on global sequence similarity (i.e., a conserved multi-domain architecture). This surprising finding can be attributed to the robustness of clustering based on individual domains to errors in gene models; the more inclusive clustering criteria (domain-based matches) increases taxon sampling significantly, improving phylogenetic tree accuracy, and, by extension, orthology prediction based on these trees. Based on this finding, we have started emphasizing trees for Pfam domains. PhyloFacts-Pfam is designed to provide biologists with a mechanism to find all PhyloFacts families matching specific Pfam domains. Data can be downloaded from individual PhyloFacts family pages and can also be downloaded in bulk from http://phylogenomics.berkeley.edu/phylofacts/downloads/ .

Inclusion of additional data from 3rd party resources. From discussions with numerous people at the DOE Grantees meeting (February 2012), it became clear that many biologists depend on the BioCyc database for pathway inference. We have developed parsers for BioCyc data and extended the PhyloFacts PostgreSQL schema to include these data, and to display these data on sequence, family and PHOG pages. We are now working on parsers for Gene Ontology data (previously, we retrieved GO data from UniProt; it appears that UniProt GO annotations are not quite current).

Continued development and testing of the FAT-CAT (Fast-Approximate Tree Classification) algorithm. The PhyloFacts library construction pipeline is computationally expensive; both CPU and disk space constraints prevent the use of this pipeline every time a new genome is sequenced. The FAT-CAT system is designed to provide rapid and highly specific functional sub-classification of novel sequences without the computational burden of library construction. FAT-CAT uses HMMs at internal nodes of PhyloFacts trees, which are annotated with the observed functions (e.g., EC numbers, GO annotations, etc.) of sequences within the trees. Classification of sequences to these HMMs allows us to predict function (and possibly taxonomic origin). We are exploring different techniques to improve FAT-CAT scalability to large datasets while maintaining high precision. First, we use the HMMER 3.0 suite which has been optimized for speed. We then reduce the number of HMM scores required using a two-step protocol. First, we select families for sub-classification by scoring query sequences against family HMMs (these correspond to HMMs located at the root nodes of PhyloFacts trees). Even with the almost 100,000 family HMMs in PhyloFacts, this initial step takes under a minute on average. Families with significant scores are then selected for phylogenetic placement using FAT-CAT. HMMER's super-fast hmmscan software makes a brute-force approach feasible (i.e., scoring the query sequence against all HMMs in the tree). We are also developing a tree-traversal approach. Tree traversal recursively traces a path from the root to a leaf, starting at the root node and scoring the query against the HMMs located at child nodes. We then follow the edge to the child node corresponding to the HMM giving the strongest score. The process is repeated until a leaf node is reached. The HMM giving the query the most significant score on that path is identified, and the corresponding subtree node is used to derive a functional (and perhaps taxonomic) annotation. Using tree traversal reduces the number of HMM scores required to place a sequence in a balanced binary tree of K sequences to only O(log K) making phylogenetic placement efficient. Our preliminary data shows FAT-CAT is competitive with the top-ranked methods in phylogenetic placement (e.g. EPA). We are also testing FAT-CAT accuracy at functional classification using the Structure Function Linkage Database produced by Dr. Patsy Babbitt and colleagues at UCSF.

Lastly, we worked to improve the accessibility and interpretability of PhyloFacts family data. Users can access the data in PhyloFacts in several ways, including sequence accession (UniProt and GenBank accessions are both accepted) and text search. Very rapid text search is now provided on our website using Solr and Lucene.

Major developments in 2013: The major advance during this reporting period is our launch of the FAT-CAT (Fast Approximate Tree Classification) web server at http://phylogenomics.berkeley.edu/phylofacts/fatcat/, and a publication describing the webserver and validation experiments in Nucleic Acids Research. Afrasiabi, et al, "The PhyloFacts FAT-CAT Webserver: Ortholog Identification and Function Prediction using Fast Approximate Tree Classification," *Nucleic Acids Research 2013; doi: 10.1093/nar/gkt399*

FAT-CAT: Fast Approximate Tree Classification Web Server.
The PhyloFacts FAT-CAT web server provides ortholog identification and functional annotation based on phylogenetic placement of protein sequences to pre-calculated gene trees in the PhyloFacts database (2). FAT-CAT uses a novel subtree-HMM-based classification protocol to allow flexible phylogenetic classification and highly precise ortholog identification. PhyloFacts trees are overlaid with functional and annotation data from numerous resources, including Gene Ontology, UniProt (SwissProt and TrEMBL), Pfam, BioCyc/MetaCyc and EC, so that subtree-HMM-based classifications to pre-defined orthology groups can be used to derive a precise functional sub-classification. PhyloFacts has broad taxonomic and functional coverage, with >7.3M proteins from 99K unique taxa across the Tree of Life, allowing FAT-CAT to predict orthologs and assign function for most user queries. Sequences from metagenome projects, such as the human microbiome or environmental samples from soil and marine environments, can be submitted for simultaneous functional annotation and prediction of taxonomic origin. Benchmarking experiments comparing FAT-CAT against the major orthology web servers – eggNOG, KEGG, OrthoMCL, InParanoid, PhylomeDB and OrthoDB – demonstrate FAT-CAT's high precision and robustness to both promiscuous domains and recent duplication events. FAT-CAT was the only webserver to have no errors on test proteins, with OMA and PhylomeDB having a very small number of errors. By contrast, other orthology web servers mix paralogs with predicted orthologs and include proteins with only partial homology to query sequences. The FAT-CAT webserver is available at http://phylogenomics.berkeley.edu/phylofacts/fatcat/. Details on these and other experiments are available online at http://phylogenomics.berkeley.edu/phylofacts/fatcat/supplementary/.

Major developments in 2014: The major advance during this year is our release of data and software tools produced by this project on the PhyloFacts website.

Major data, including phylogenetic trees, multiple sequence alignments and other data for protein families are now available for download from http://phylogenomics.berkeley.edu/data/.

Software tools are available for download from http://phylogenomics.berkeley.edu/software/.

Project Member Activities

The DOE-PhyloFacts project team included myself, three postdoctoral scholars (Ruchira Datta, Yaoqing Shen, David Dineen), three programmers (Grant Shoffner, Cyrus Afrasiabi, Shailen Tuli and Jonathan Dobbie) and a graduate student (Bushra Samad).

Kimmen Sjölander (PI). My activities have focused on supervising the overall project, training team members in bioinformatics, identifying bugs in our system, and directing algorithm development. Highlights of the specific contributions of funded project members are listed below.

Ruchira Datta (Postdoctoral scholar and Assistant Specialist).

Yaoqing Shen (Postdoctoral scholar and Assistant Specialist).

David Dineen, Ph.D. (Assistant Specialist). Dave worked on PF3.0 Webpage development and maintenance, extensions to the PF3.0 PostgreSQL database, Biocyc data integration, PhyloFacts-Pfam: database development and webpage design, PhyloFacts Genome pages. He also contributed to the FAT-CAT webserver.

Grant Shoffner (applications programmer). Grant worked on various webservers associated with the PhyloFacts resource.

Cyrus Afrasiabi (Applications Programmer). Cyrus worked on PF3.0 webpage development and maintenance**,** bioinformatics software development for PF 3.0 family clustering, quality assurance of PF3.0 webpages**,** extensions to the PF3.0 PostgreSQL database.

Shailen Tuli (Applications Programmer). Shailen worked on the PF3.0 webpage development and maintenance and to modifications to the database schema to improve performance.

Jonathan Dobbie (Applications Programmer). Jonathan worked on major revisions to the PhyloFacts PostgreSQL database schema, including several required to enable rapid database queries, rapid text search enabled using Solr and Lucene, RESTFUL API for PhyloFacts searching and external job management (FatCat), revisions to the PHOG algorithm for orthology identification, bioinformatics software development for PF3.0 family tree analysis and interpretation, data integration from UniProt (revisions to scripts to retrieve and load UniProt data).

Bushra Samad (Ph.D. student). Bushra worked on algorithm development and benchmarking experiments for the FAT-CAT phylogenetic placement and orthology identification.

Publications supported by DOE funding

1. Erik Sonnhammer, Toni Gabaldón, Alan Wilter Sousa da Silva, Maria Martin, Marc Robinson-Rechavi, Brigitte Boeckmann, Paul Thomas, Christophe Dessimoz, and the Quest for Orthologs consortium. "Big Data and Other Challenges in the Quest for Orthologs," *Bioinformatics (2014) doi: 10.1093/bioinformatics/btu492*
2. Brian P. Anton, Yi-Chien Chang, Peter Brown, Han-Pil Choi, Lina L. Faller, Jyotsna Guleria, Zhenjun Hu, Niels Klitgord, Ami Levy-Moonshine, Almaz Maksad, Varun Mazumdar, Mark McGettrick, Lais Osmani, Revonda Pokrzywa, John Rachlin, Rajeswari Swaminathan, Benjamin Allen, Genevieve Housman, Caitlin Monahan, Krista Rochussen, Kevin Tao, Ashok S. Bhagwat, Steven E. Brenner, Linda Columbus, Valérie de Crécy-Lagard, Donald Ferguson, Alexey Fomenkov, Giovanni Gadda, Richard D. Morgan, Andrei L. Osterman, Dmitry A. Rodionov, Irina A. Rodionova, Kenneth E. Rudd, Dieter Söll, James Spain, Shuang-yong Xu, Alex Bateman, Robert M. Blumenthal, J. Martin Bollinger, Woo-Suk Chang, Manuel Ferrer, Iddo Friedberg, Michael Y. Galperin, Julien Gobeill, Daniel Haft, John Hunt, Peter Karp, William Klimke, Carsten Krebs, Dana Macelis, Ramana Madupu, Maria J. Martin, Jeffrey H. Miller, Claire O'Donovan, Bernhard Palsson, Patrick Ruch, Aaron Setterdahl, Granger Sutton, John Tate, Alexander Yakunin, Dmitri Tchigvintsev, Germán Plata, Jie Hu, Russell Greiner, David Horn, Kimmen Sjölander, Steven L. Salzberg, Dennis Vitkup, Stanley Letovsky, Daniel Segrè, Charles DeLisi, Richard J. Roberts, Martin Steffen, Simon Kasif. "The COMBREX Project: Design, Methodology, and Initial Results," *PLoS Biol 11(8): e1001638. doi:10.1371/journal.pbio.1001638*
3. Afrasiabi, C., Samad, B., Dineen, D., Meacham, C. **Sjölander, K.**, "The PhyloFacts FAT-CAT Webserver: Ortholog Identification and Function Prediction using Fast Approximate Tree Classification," *Nucleic Acids Research 2013; doi: 10.1093/nar/gkt399* PDF
4. Liberles, D., Teichmann, S., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L., de Koning, J., Dokholyan, N., Echave, J., Elofsson, A., Gerloff, D., Goldstein, R., Grahnen, J., Holder, M., Lakner, C., Lartillot, N., Lovell, S., Naylor, G., Perica, T., Pollock, D., Pupko, T., Regan, L., Roger, A., Rubinstein, N., Shakhnovich, E., Sjölander, K., Sunyaev, S., Teufel, A., Thorne, J., Thornton, J., Weinreich, D., Whelan, S., "The interface of protein structure, protein biophysics, and molecular evolution," *Protein Science 2012; doi: 10.1002/pro.2071*
5. Dessimoz, C., Gabaldon, T., Roos, D., Sonnhammer, E., Herrero, J., and the Quest for Orthologs Consortium, "Toward community standards in the quest for orthologs," *Bioinformatics 2012; doi: 10.1093/bioinformatics/bts050*
   (Members of the Quest for Orthologs Consortium: Adrian Altenhoff, Rolf Apweiler, Michael Ashburner, Judith Blake, Brigitte Boeckmann, Alan Bridge, Elspeth Bruford, Mike Cherry, Matthieu Conte, Durand Dannie, Ruchira Datta, Christophe Dessimmoz, Jean-Baka Domelevo Entfellner, Ingo Ebersberger, Toni Gabaldon, Michael Galperin, Javier Herrero, Jacob Joseph, Tina Koestler, Evgenia Kriventseva, Odile Lecompte, Jack Leunissen, Suzanna Lewis, Benjamin Linard, Michael S. Livstone, Hui-Chun Lu, Maria Martin, Raja Mazumder, David Messina, Vincent Miele, Matthieu Muffato, Guy Perriere, Marco Punta, David Roos, Mathieu Rouard, Thomas Schmitt, Fabian Schreiber, Alan Silva, Kimmen Sjölander, Nives Skunca, Erik Sonnhammer, Eleanor Stanley, Radek Szklarczyk, Paul Thomas, Ikuo Uchiyama, Michiel Van Bel, Klaas Vandepoele, Albert J. Vilella, Andrew Yates and Evgeny Zdobnov).
6. Shen, Y., Bonnot, F., Imsand, E., Rosefigure, J., Sjölander, K., Kilnman, J., "Distribution and Properties of the Genes Encoding the Biosynthesis of the Bacterial Cofactor, Pyrroloquinoline Quinone," *Biochemistry 2012; doi: 10.1021/bi201763d*
7. Sjölander, K., Datta, R., Shen, Y., Shoffner, G., "Ortholog identification in the presence of domain architecture rearrangement," *Briefings in Bioinformatics 2011; doi: 10.1093/bib/bbr036*