

Evolutionary and Experimental Evidence-based Functional Annotation of Genes

Morgan N. Price¹, **Adam P. Arkin**^{1,2}, Ben Bowen³, Barbara E. Engelhardt^{4,5}, Ameet Talkwalkar⁴, Jeffrey M. Yunes⁶, Michael L. Souza⁷, Gaurav Pandey^{8,9}, Susanna Repo⁸, Michael I. Jordan^{4,10}, **Steven E. Brenner**^{1,2,6,7,8*} (brenner@compbio.berkeley.edu), Nitin Baliga¹¹, and Paul D. Adams¹

¹Physical Biosciences Division, Lawrence Berkeley National Lab, Berkeley, California; ²Department of Bioengineering, University of California, Berkeley; ³Life Sciences Division, Lawrence Berkeley National Lab, Berkeley, California; ⁴Department of Electrical Engineering and Computer Science, University of California, Berkeley; ⁵Currently at Biostatistics and Bioinformatics Department, Duke University, Durham, North Carolina; ⁶Joint Graduate Group in Bioengineering, University of California, Berkeley and San Francisco; ⁷Biophysics Graduate Group, University of California, Berkeley; ⁸Department of Plant and Microbial Biology, University of California, Berkeley; ⁹Currently at Mount Sinai School of Medicine, New York, New York; ¹⁰Department of Statistics, University of California, Berkeley; ¹¹Institute for Systems Biology, Seattle, Washington

<http://enigma.lbl.gov>
<http://compbio.berkeley.edu>

Project Goals

The ENIGMA SFA aims to understand the architecture of microbial communities from a molecular level, which requires understanding in detail the molecular biology of key organisms. Although sequencing the genomes of these organisms is now straightforward, determining the molecular function of genes remains a challenge. However, many genome sequences are now available, and rich genome-wide functional data is becoming available outside of traditional model organisms. Therefore, we are developing improved tools for using evolutionary comparisons and functional-genomic data to predict the molecular function of proteins.

Abstract

Phylogenetic analysis has been employed to infer the molecular function of a target gene by finding a function that is consistent with the evolutionary history of the gene. Over the past decade, this has been recognized as a highly accurate approach, but its manual application requires laborious effort by a domain expert. We have developed the SIFTER method, which automates phylogenetic-based function annotation by finding the most likely assignments of functions to proteins given a phylogenetic tree, model of evolution, and known functions. SIFTER uses a Bayesian graphical model framework to propagate molecular functions across the tree in a way that is statistically rigorous and robust. SIFTER explicitly takes account of evidence quality, to account for the variable quality of annotations from different sources.

Benchmarking studies of SIFTER show that it outperforms other widely-used homology-based approaches. Recently, we improved the core SIFTER algorithm, enabling it to run on large and diverse protein families, to work on a genome-scale, and to participate in the Critical Assessment of Function Annotation in 2011. We are extending it to share information between protein families based on gene-gene “association” relationships such

as protein-protein interactions, co-expression, co-fitness, genome proximity, or genetic interactions. In doing so, we will be able to incorporate a larger variety of experimental data developed by and applicable to the ENIGMA project than other prediction approaches. We hope that with these enhancements, SIFTER will be the first successful method to statistically incorporate both homology and association data.

Another challenge is to interpret large-scale “fitness” data or knockout mutant phenotypes that are becoming available for diverse microbes due to approaches such as tagged transposon mutagenesis or TnSeq. In a pilot study in *Shewanella oneidensis* MR-1, we were able to confirm many annotations and to revise the annotations of 40 genes or operons, but this required extensive manual curation. To streamline the analysis, we are developing heuristics to find “re-annotatable” proteins to focus the manual curation. We are also automating the comparison of fitness data to metabolic models; in principle, it should be possible to automate much of the manual curation that now goes into producing a high-quality metabolic model.

This work partially conducted by ENIGMA (Ecosystems and Networks Integrated with Genes and Molecular Assemblies) was partially supported by the Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Other support for SIFTER comes from NIH K22 HG00056, NIH R01 GM071749, DOE SciDAC BER KP110201.