182. PhyloFacts FAT-CAT Ortholog Identification and Functional Annotation

Kimmen Sjölander[1]* (kimmen@berkeley.edu) and Cyrus Afrasiabi[1] [1]University of California, Berkeley
http://phylogenomics.berkeley.edu/phylofacts/

Project goals: The PhyloFacts project aims to develop novel methods and web servers to integrate multiple data types and to infer and curate (meta)genomic functional annotations. We address these aims by using evolutionary reconstruction to integrate and organize heterogeneous data from homologous genes across thousands of species, to simultaneously derive functional and taxonomic annotations for environmental sample sequences, and for functional annotation of whole genomes.

The PhyloFacts FAT-CAT web server is designed to improve the accuracy of genome functional annotation. Standard pipelines use a BLAST-based annotation-transfer protocol to assign predicted functions to genes. This approach is now known to have high error rates due to hidden paralogy and promiscuous domains, and errors can be propagated through sequence databases by annotation transfer protocols. Although the actual fraction of gene annotation errors is unknown (and errors can be difficult to detect), numerous studies suggest that as many as 25% of genes have errors in their functional annotations. If we add to this number the ~30% of genes labeled simply as "hypothetical", it becomes clear that significant work remains to improve on this status quo. We address these challenges in the PhyloFacts resource through the use of phylogenomic analysis, protein structure information and integration of heterogeneous experimental and annotation data.

The fundamental assumption underlying phylogenomic analysis is that accurate prediction of protein function depends on the phylogenetic identification of orthologs. Orthology is a phylogenetic term: two proteins are each other's orthologs if they are related by speciation from a common ancestor. We combine phylogenomic ortholog identification with protein structure analysis, for instance, building trees for Pfam domains as well as for multi-domain architectures, an approach we call *structural phylogenomics*. Protein function is multi-faceted, and informed by experimental data of different types; we therefore retrieve data from numerous resources, including GO annotations for molecular function, biological process, and cellular location, pathway role, protein-protein interaction, Pfam multi-domain architecture, Enzyme Commission numbers, and other types of experimental and annotation data. We overlay these data on protein family trees, allowing the function(s) of individual proteins to be inferred based on annotations retrieved for their homologs in the tree, weighting sequences nearby in the tree more than those that are distant, and weighting sequence annotations with experimental support more than those that have been derived using noisy annotation transfer protocols.

Because building phylogenetic trees is computationally expensive, we have developed a system to enable phylogenomic classification for novel sequences (e.g., newly sequenced genomes) using hidden Markov models placed at all nodes (or vertices) in PhyloFacts protein family trees. Sequences can be assigned to positions in trees based on the top-scoring HMM in the tree. From this phylogenetic placement, we identify orthologs and derive a functional annotation.

PhyloFacts' broad taxonomic and functional coverage, with >7.3 M proteins from across the Tree of Life, enables FAT-CAT to predict orthologs and assign function for most sequence inputs. 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. On this dataset, FAT-CAT was the only webserver with perfect precision, with OMA and PhylomeDB making 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 web server is available at http://phylogenomics.berkeley.edu/phylofacts/fatcat/. Details on the method and validation experiments are presented in Afrasiabi *et al*, "The PhyloFacts FAT-CAT Webserver: Ortholog Identification and Function Prediction using Fast Approximate Tree Classification," *Nucleic Acids Research* 2013. Supplementary data are available online at http://phylogenomics.berkeley.edu/phylofacts/fatcat/supplementary/.

The PhyloFacts FAT-CAT webserver is supported by the Office of Biological and Environmental Research in the DOE Office of Science.