The PhyloFacts Phylogenomic Encyclopedia of Microbial Gene Families

Yaoqing Shen[1], Bushra Samad[2], Ajithkumar Warrier[1], Ruchira Datta[1] and **Kimmen Sjölander**[1,2,3*] (kimmen@berkeley.edu)

1. QB3 Institute, University of California Berkeley, Berkeley, CA; 2. Department of Bioengineering, University of California Berkeley, Berkeley, CA; and 3. Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA.

http://makana.berkeley.edu/phylofacts/

Project Goals: To improve the precision of microbial (meta)genome functional annotation by providing phylogenomic analyses of microbial gene families, providing access to these analyses to the scientific community in the PhyloFacts Database. This pipeline includes: clustering all microbial genes from whole genomes into gene families including homologs from other species; constructing multiple sequence alignments and estimating protein family trees; ortholog identification; integrating experimental and annotation data; computationally scalable methods for HMM classification of (meta)genome sequences to PhyloFacts families and orthology groups.

PhyloFacts is a phylogenomic encyclopedia of gene family trees across the Tree of Life (1). Gene families are defined based on (1) agreement at the multi-domain architecture, and (2) on containing a single Pfam domain in common. The FlowerPower algorithm (2) is used to retrieve homologs from the UniProt database, parameterized separately for the two homology clustering criteria. For each gene family, we construct a multiple sequence alignment and phylogenetic tree; phylogenetic trees are used to identify orthologs in different species using the PHOG algorithm (3). This combination of single domain and domain-architecture clustering enhances the recall and precision of functional classification and orthology identification (4). We construct a hidden Markov model (HMM) for the family and use it to identify homologous protein structures by scoring the Protein Data Bank. Finally, we retrieve experimental and annotation data from various external resources, including UniProt, Gene Ontology, Pfam and KEGG and use these to provide informative descriptions of each family and orthology group, from which the functions of family and orthology-group members can be inferred.

We have focused during the last year on expanding our coverage of microbial gene families. More than 7M proteins are included in PhyloFacts families, representing a broad swath of species. 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://makana.berkeley.edu/phylofacts/coverage/.

Users can access the data in PhyloFacts in several ways, including sequence accession and inputting protein sequences in FASTA format for HMM classification. Data can be downloaded from individual PhyloFacts family pages and can also be downloaded in bulk from http://makana.berkeley.edu/phylofacts/downloads/.

We have also developed a prototype phylogenomic HMM classification system we call FAT-CAT (for Fast Approximate Tree Classification), to allow the functional classification of novel proteins and the simultaneous taxonomic and functional classification of metagenome reads using HMMs placed at internal nodes of gene trees. We will present the results of this analysis at the meeting.

References

1. Krishnamurthy, N., Brown, D.P., Kirshner, D. and Sjölander, K. (2006) PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome biology*, **7**, R83.
2. Krishnamurthy, N., Brown, D. and Sjölander, K. (2007) FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. *BMC evolutionary biology*, **7 Suppl 1**, S12.
3. Datta, R.S., Meacham, C., Samad, B., Neyer, C. and Sjölander, K. (2009) Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic acids research*, **37(Web Server issue):** , W84-W89.
4. Sjölander, K., Datta, R.S., Shen, Y. and Shoffner, G.M. (2011) Ortholog identification in the presence of domain architecture rearrangement. *Briefings in bioinformatics*. 12 (5): 413-422.

The PhyloFacts Microbial Encyclopedia is supported by the Office of Biological and Environmental Research in the DOE Office of Science.