BERKELEY PHOG: PHYLOGENOMIC ORTHOLOGY PREDICTION, WITH APPLICATIONS TO INFERRING INTERACTIONS

Ruchira S. Datta*

QB3 Institute, University of California, Berkeley
Berkeley, CA 94720-3220, USA
**Email: ruchira@berkeley.edu*

Kimmen Sjölander

QB3 Institute, Department of Bioengineering, and Department of Plant and Microbial Biology, University of California, Berkeley
Berkeley, CA 94720-3220, USA
Email: kimmen@berkeley.edu

We present the Berkeley PHOG (PhyloFacts Orthology Group) phylogenomic orthology prediction algorithm. Ortholog detection is essential in functional annotation of genomes, with applications to predicting protein-protein interaction and pathway participation as well as other essential tasks for systems biology. PHOG employs a novel algorithm to identify orthologs using phylogenetic analysis. PHOGs integrate information from various sources such as literature references, Gene Ontology (GO) annotations, and Enzyme Classification (EC) numbers. The Phyloscope tree viewer presents these data in their evolutionary context. The PHOG interactome Viewer presents protein-protein interactions predicted by interolog analysis. The PHOG Pathway Viewer (forthcoming) presents metabolic pathways inferred from orthology. Predictions are available from the webserver http://phylofacts.berkeley.edu/orthologs/

1. OVERVIEW

Phylogenomic analysis uses evolutionary history, differentiation between orthologs and paralogs, and existing functional information to predict the functions of novel proteins, greatly reducing the error rates relative to simpler methods of function prediction.[12] The PhyloFacts database and associated webservers provide an online platform for robust phylogenomic analysis, freely available to the research community.[34] PhyloFacts includes over 60,000 protein families and domains with predicted phylogenies, structures, and functions for millions of proteins, and over 1.5 million hidden Markov models for classification of user-submitted sequences to families and subfamilies.

Berkeley PHOG (PhyloFacts Orthology Group) is a novel scalable algorithm for identifying orthologs using phylogenetic analysis.[5] This tree-distance based method does not require a trusted species tree, and can be tuned to targeted precision levels and taxonomic distances. PHOG uses trees based on both global domain architecture and individual domains, and is not limited to sequences from fully sequenced genomes. Each PHOG has a specific taxonomic distribution and is linked to GO annotations, EC numbers, KEGG pathway information, Pfam domains, and biological literature. Classifying an input sequence to a PHOG enables propagating this information, resulting in functional annotation of the unknown sequence.

Results on a benchmark dataset from the TreeFam-A manually curated orthology database show that PHOG provides a combination of high recall and precision competitive with both InParanoid and OrthoMCL, and allows users to target different taxonomic distances and precision levels through the use of tree-distance thresholds. For instance, OrthoMCL-DB achieved 76% recall and 66% precision on this dataset; at a slightly higher precision (68%) PHOG achieves 10% higher recall (86%). InParanoid achieved 87% recall at 24% precision on this dataset, while a PHOG variant designed for high recall achieves 88% recall at 61% precision, increasing precision by 37% over InParanoid.

The PHOG Interactome Viewer uses interolog analysis to predict protein-protein interactions. Given a protein sequence of interest from one species, the PHOG Interactome Viewer checks its orthologs in other species for experimentally characterized protein-protein interactions, and predicts that the ortholog of the interacting partner in the original species interacts with the sequence of interest. The observed and predicted

[*] Corresponding author.

interacting partners are presented in a graphical display with links to supporting evidence.

The PHOG Pathway Viewer (forthcoming) will similarly predict the participants in a metabolic pathway in a given species based on their experimentally characterized orthologs in other species. In this way, PHOG aligns the metabolic networks of different organisms.

The user can change the threshold targeting a given precision level interactively on the webserver. This capability extends from PHOG to the PHOG Interactome Viewer and (in the future) the PHOG Pathway Viewer.

Because PHOG is a tree-based orthology prediction method, it naturally handles many species at once. A PhyloFacts multi-gene family tree may include hundreds of species. PhyloFacts is expanding coverage to target a diverse set of microbial species, enabling high-throughput functional annotation of microbial sequences using Berkeley PHOG. This makes PHOG particularly well suited for metagenomics, the "systems biology of the biosphere": PHOG can infer metabolic pathways carried out cooperatively by members of a microbial community. Future work includes placement of fragmentary, noisy sequences such as those produced by environmental sampling into trees, so that PHOG can be used for simultaneous taxonomic and functional annotation of metagenomic sequences.

2. METHODS

The foundation of the PHOG algorithm is the Reciprocal Nearest Neighbor (RNN) condition. Two sequences in a gene family tree satisfy RNN if each sequence is the nearest neighbor from its species to the other sequence, by tree distance. Two sequences in a gene family tree are inparalogs if they are from the same species, and the smallest subtree containing them contains only sequences from that species. The PHOG-S (for Superorthology) variant of the algorithm finds maximal subtrees such that every sequence in the subtree satisfies RNN (or would do so after collapsing inparalagous subtrees). PHOG-S computes these subtrees efficiently using dynamic programming. The parent of a PHOG-S node is a putative duplication node, since it contains two distinct, noninparalogous sequences from the same species, each of which is a nearest neighbor to some of the sequences descending from that node. The PHOG-T (for Thresholded) variant computes the "duplication distance" of each of these putative duplication nodes: half the distance between these two sequences. If the duplication distance does not exceed the threshold (supplied as a user-specified paramter), PHOG-T proceeds up the tree to the next putative duplication node, until it finds one greater than the threshold (or reaches the root). Whereas PHOG-S partitions the gene family tree into subtrees, PHOG-T may partition it into subtrees or subtrees minus other subtrees.

Acknowledgments

RSD and KS acknowledge funding from the National Science Foundation (The PhyloFacts Phylogenomics Encyclopedia of Microbial Protein Families) and the Department of Energy (Phylogenomic tools and web resources for the Systems Biology Knowledgebase).

References

1. K. Sjölander, *Bioinformatics*, 2004, **20**, 170-179.
2. D. Brown and K. Sjölander, *PLoS Comput Biol*, 2006, **2**, e77.
3. N. Krishnamurthy, D. Brown, D. Kirshner, and K. Sjölander, *Genome Biology*, 2006, **7**, R83.
4. J. G. Glanville, D. Kirshner, N. Krishnamurthy, and K. Sjölander, *Nucl. Acids Res.*, 2007, **35**, W27-32.
5. R. S. Datta, C. Meacham, B. Samad, C. Neyer, and K. Sjölander, *Nucl. Acids Res.*, 2009, gkp373.