

THE PHYLOFACTS PHYLOGENOMIC ENCYCLOPEDIAS: STRUCTURAL PHYLOGENOMIC ANALYSIS ACROSS THE TREE OF LIFE

Kimmen Sjölander

Berkeley Phylogenomics Group
University of California, Berkeley
<http://phylogenomics.berkeley.edu>

Protein families evolve a multiplicity of functions and structures through gene duplication, domain shuffling, speciation and other processes. Phylogenomic analysis, combining phylogenetic tree construction, integration of experimental data, and differentiation of orthologs and paralogs, has been shown to address the systematic errors associated with standard protocols of protein function prediction. The explicit integration of structure prediction and analysis in this framework, which we call *structural phylogenomics*, provides additional insights into protein superfamily evolution, and improves function prediction accuracy.

The Berkeley Phylogenomics Group has developed the PhyloFacts Phylogenomic Encyclopedia for protein

families across the Tree of Life. At present (April 17, 2007), PhyloFacts contains over 27,000 “books” for protein families and domains and over 988,000 hidden Markov models (HMMs) enabling classification of proteins to functional families and subfamilies. Other functionality provided by PhyloFacts includes prediction of protein structure, active site residues, and cellular localization.

In this talk, I will present new methods developed by my group for key tasks in a phylogenomic pipeline, including multiple sequence alignment, phylogenetic tree construction, subfamily identification and critical residue prediction.