Protein Function Prediction by COFACTOR in CAFA3

Chengxin Zhang, Peter L. Freddolino*, Yang Zhang*
Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI, 48109, USA
*Correspondence should be addressed to: PLF: petefred@umich.edu, YZ: zhng@umich.edu

1. INTRODUCTION

The Zhang-Freddolino lab protein function prediction pipeline in CAFA3 is built on an updated version of COFACTOR [1], which was originally designed for inferring protein function from known function analogies detected on local and global structure comparisons. In the new version, two sequence and protein-protein interaction (PPI) based pipelines are developed to improve the accuracy and coverage of functional template identification. In the sequence-based pipeline, the target sequence is searched against the UniProt-GOA database using BLAST and PSI-BLAST to detect sequence homologs from which Gene Ontology (GO) terms are extracted. In the PPI-based pipeline, the GO terms of the targets are inferred from the known function of the interaction partners as annotated by the STRING database [2], under the assumption that the interacting partners tend to participate in the same biological pathway at the same sub-cellular location and therefore share similar functional terms. Finally, a consensus of function terms from the three pipelines is collected as the final model of the function prediction.

Protein targets are categorized into two groups: a target is considered “Easy” if there is one or more close sequence homologues identified, or “Hard” otherwise. Given the huge number of testing targets (130,827 sequences in CAFA3), the structure-based pipeline, starting from structural assembly simulation by I-TASSER [3] followed by global and local structure alignment search against the BioLiP structure-function database [4], is too expensive to apply to all targets. Therefore, for Easy targets, GO terms are generated only using the sequence- and PPI-based pipelines, which are less time-consuming. For Hard targets, however, structure- and PPI-based pipelines are used for predicting the GO terms. For both categories of targets, the first model is from the consensus predictions from all available pipelines (sequence + PPI for Easy targets, structure + PPI for Hard targets), and the second model selected from the sequence and PPI-based pipelines (submitted for Easy targets only). The third model, generated by the structure-based pipeline, is submitted only for Hard targets. The prediction process is fully automated.

2. AVAILABILITY

The on-line COFACTOR server is available at
http://zhanglab.ccmb.med.umich.edu/COFACTOR/.

3. REFERENCES

[1] Roy A, Yang JY, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res. 2012;40:W471-W7.
[2] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43:D447-D52.
[3] Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5:725-38.
[4] Yang JY, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res. 2013;41:D1096-D103.