Application Note

SUNPRO: Structure and function predictions of proteins from representative organisms

Hongyi Zhou, Mu Gao, Narendra Kumar and Jeffrey Skolnick*

Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, GA, 30318

This manuscript is unpublished

ABSTRACT

Summary: Protein structures and functions are essential for understanding all biological processes in living systems. Three-dimensional protein structures, which provide the basis for protein functions, are determined by protein amino acid sequences. With the rapid progress of proteome sequencing, the experimental determination of protein structure and function lags well behind the demands of the biological and drug discovery communities. To fill this gap, SUNPRO provides structure and function predictions of tens of thousands of proteins from representative organisms that are commonly studied by researchers. It employs the state-of-the-art structure prediction algorithm TASSER^{VMT}-lite to predict the structures of all sequences ≤ 600 amino acids in size from 10 representative proteomes; and the accurate sequence or threading/structure based function prediction approaches EFICAz, DBD-Threader and FINDSITE for functional annotation and ligand virtual screening.

Availability: http://cssb.biology.gatech.edu/genome/index.html

Contact: skolnick@gatech.edu

1 INTRODUCTION

Advances in gene sequencing technology have provided the sequences of hundreds of organisms (Adams, et al., 2000; Mukhopadhyay, 2009; Venter, et al., 2001). These sequences have made possible the large-scale studies of all genes and gene products in an organism(Stein, 2001). One common goal of these studies is to understand the function of all molecules in a cell using a systems biology approach. Computational approaches that predict the structure and functions of proteins can contribute significantly towards this goal. For instance, sequence-based methods using evolutionary relationships can provide insights into the enzymatic function of about 50% of the ORFs in a given proteome(Tian and Skolnick, 2003; Tian, et al., 2004). Since the three dimensional structures of proteins are the basis for their functions and are also useful for structure based drug discovery, protein design and engineering, the emerging new field of Structural Genomics seeks to determine the three dimensional structure of every protein in a given genome (Baker and Sali, 2001; Chandonia and Brenner, 2006; Skolnick, et al., 2000). This can only be accomplished by a combination of experimental and modeling approaches(Baker and Sali, 2001; Skolnick, et al., 2000). Our recently developed fast version of protein structure prediction method, TASSER^{VMT}-lite (Zhou and Skolnick, 2012; Zhou and Skolnick, 2012) is such a modeling approach applicable for proteome scale structure modeling, with an accuracy comparable to the top methods in CASP9 (http://predictioncenter.org) (Zhou and Skolnick, 2012).

Given experimental or modeled structures, the next task is to utilize them for function inference and drug discovery. Recently, effective approaches that can use modeled low-resolution structures for function inference and ligand virtual screening have been developed, such as the threading/structure based FINDSITE(Brylinski and Skolnick, 2008), DBD-Hunter(Gao and Skolnick, 2008) and DBD-Threader(Gao and Skolnick, 2009). FINDSITE can not only predict function, but also predict binding sites and perform ligand virtual screening. DBD-Hunter and DBD-Threader predict DNA-binding domains and associated binding-sites if they exist. In addition to these structure based methods, EFICAz is an accurate sequence-based enzyme function inference method that provides good accuracy and precision in large scale benchmarking(Arakaki, et al., 2009; Tian, et al., 2004).

In this note, we present the SUNPRO database that is result of the application of TASSER^{VMT}-lite, DBD-Threader, FINDSITE, and EFICAz, to the **S**tructure and f**UN**ction predictions of **P**roteins from 10 **R**epresentative **O**rganisms. There are databases that provide a repository of modeled structures for many of annotated protein sequences and domains (Kiefer, et al., 2009; Ursula Pieper, et al., 2004). However, SUNPRO distinguishes itself using one of the most accurate structural modeling approaches and integrating comprehensive functional annotation and virtual screening for all targets whose structures are modeled. In our earlier work, we have implemented an interactive server PSiFR that uses the same methodology and produces the same data and gives the user flexibility of selecting methods(Pandit, et al., 2010). In this note, we have updated the structure modeling approach from the previously used METATASSER(Zhou, et al., 2007), that is relatively computationally expensive and not as accurate as TASSER^{VMT}-lite and EFICAz to its latest version EFICAz$^{2.5}$. For DNA-binding prediction, a more general threading/structure based method DBD-Threader is implemented. Most importantly, the pre-computed database saves users the time of selecting methods and waiting for results. Users can get all structure/function/virtual screening data for a given target by a single search in seconds. For those targets absent in the SUNPRO database, PSiFR can be used.

2 METHODS

At present, included organisms in SUNPRO are: human, *E. coli,* budding yeast, *C. elegans*, fruit fly, social amoebae, zebrafish,

*To whom correspondence should be addressed.

mouse, rat, and *A. thaliana*. The first two are commonly used organisms and the latter 8 organisms are NIH model organisms for biomedical research (http://www.nih.gov/science/models/). All proteome sequences, except for the social amoebae, are downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/. Social amoebae sequences are from http://dictybase.org/. Currently, only sequences whose size is ≤ 600 amino acids (AA) are predicted. Under this length cutoff, about 70% of the sequences are covered for a typical proteome. SUNPRO uses the recently developed TASSER^{VMT}-lite (Zhou and Skolnick, 2012) for structure modeling of Easy targets, defined as those having a SP^3 threading Z-score≥ 6.0 (Zhou and Zhou, 2005). In a typical proteome, Easy targets constitute around 75% of the total number of proteins. For the other 25% of targets, chunk-TASSER is applied(Zhou and Skolnick, 2007). Five models are predicted for each target. For function annotations, we use the latest update of sequence-based enzyme function assignment approach EFICAz version 2.5(Arakaki, et al., 2009)(trained with new experimental data) to assign three field or four field EC numbers to the target. DBD-Threader(Gao and Skolnick, 2009) is used to predict if a target is likely a DNA-binding protein and if so, then the DNA-binding domains and sites are predicted. Finally, FINDSITE is employed for ligand binding site prediction, Gene Ontology (GO) term(Ashburner, et al., 2000) assignment and virtual screening against KEGG compound & drug libraries(Kanehisa, et al., 2012) and the open compound database from the National Cancer Institute. Predictions for all 10 organisms can be searched at http://cssb.biology.gatech.edu/genome/index.html. Users can input the NCBI gi ID or RefSeq ID, a keyword or a FASTA formatted sequence to search results of the desired target. The results of the top first hit will be returned as a link to retrieve the data. Available data are the predicted first model quality in terms of the TM-score(Zhang and Skolnick, 2004) and RMSD; five structure models; list of top SP^3 threading templates; target-template threading alignments; DNA-binding data; EC number assignments; FINDSITE binding pockets, GO term assignments and virtual screening data. The predicted binding pockets can also be used by users to search against their own compound libraries for drug discovery.

3 DISCUSSION

SUNPRO provides proteomic scale structure and function annotation of commonly studied organisms. The structure models are generated by a variant of an accurate comparative modeling approach TASSER(Zhang and Skolnick, 2004; Zhou, et al., 2007). Functions are predicted by three state-of-the-art methods EFICAz$^{2.5}$, DBD-Threader and FINDSITE. Furthermore, FINDSITE also provides virtual screening data. These predictions should not only be useful for biomedical studies but also facilitate drug target and drug discovery. Possible updates of SUNPRO in the future include modeling of the 30% larger (> 600AA) proteins, virtual screening using the extended version of FINDSITE(Zhou and Skolnick, 2012) and the addition of the proteomes of more organisms such as frog and chicken in the NIH model organism database.

ACKNOWLEDGEMENTS

Funding: This work was supported in part by grant Nos. GM-37408 and GM-48835 of the Division of General Medical Sciences of the National Institutes of Health.

REFERENCES

Adams, M., et al. (2000) The genome sequence of Drosophila melanogaster, Science, 287, 2185-2195.

Arakaki, A., et al. (2009) EFICAz2: enzyme function inference by a combined approach enhanced by machine learning., BMC bioinformatics., 10:107.

Ashburner, M., et al. (2000) Gene ontology: tool for the unification of biology., Nat Genet, , 25, 25-29.

Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics, Science 294 93--96.

Brylinski, M. and Skolnick, J. (2008) FINDSITE: A threading-based method for ligand-binding site prediction and functional annotation, Proc Natl Acad Science, 105, 129-134.

Chandonia, J. and Brenner, S. (2006) The Impact of Structural Genomics: Expectations and Outcomes., Science, 311, 347-351.

Gao, M. and Skolnick, J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions., Nucleic Acids Research, 36, 3978-3992.

Gao, M. and Skolnick, J. (2009) A threading-based method for the prediction of DNA-binding proteins with application to the human genome., PLoS computational biology, 5, e1000567.

Kanehisa, M., et al. (2012) KEGG for integration and interpretation of large-scale molecular datasets., Nucl. Aci. Res., 40, D109-D114.

Kiefer, F., et al. (2009) The SWISS-MODEL Repository and associated resources., Nucleic Acids Research., 37, D387-D392.

Mukhopadhyay, R. (2009) DNA sequencers: the next generation, Anal. Chem., 81, 1736-1740.

Pandit, S., et al. (2010) PSiFR: an integrated resource for prediction of protein structure and function., Bioinformatics 26, 687-688.

Skolnick, J., et al. (2000) Structural genomics and its importance for gene function analysis, Nat. Biotechnol, 18, 283--287.

Stein, L. (2001) Genome annotation: from sequence to biology, Nature Reviews Genetics 2, 493-503.

Tian, W. and Skolnick, J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity?, J. Mol. Biol., 333, 863-882.

Tian, W.D., et al. (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference Nucleic Acids Research, 32, 6226-6239.

Ursula Pieper, et al. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources, Nucleic Acids Res., 32, D217–D222.

Venter, J., et al. (2001) The Sequence of the Human Genome, Science, 291, 1304-1351.

Zhang, Y. and Skolnick, J. (2004) Automated structure prediction of weakly homologous proteins on genomic scale, Proc. Natl. Acad. Sci. (USA), 101, 7594--7599.

Zhang, Y. and Skolnick, J. (2004) A scoring function for the automated assessment of protein structure template quality, Proteins, 57, 702--710.

Zhou, H., et al. (2007) Analysis of TASSER based CASP7 protein structure prediction results, Proteins, 69(S8), 90--97.

Zhou, H. and Skolnick, J. (2007) Ab initio protein structure prediction using chunk-TASSER, Biophys. J., 93, 1510--1518.

Zhou, H. and Skolnick, J. (2012) FINDSITE^X: A structure based, small molecule virtual screening approach with application to all identified human GPCRs., PLoS Computational Biology, submitted.

Zhou, H. and Skolnick, J. (2012) FINDSITE^X: A structure based, small molecule virtual screening approach with application to all identified human GPCRs., submitted.

Zhou, H. and Skolnick, J. (2012) Template-based protein structure modeling using TASSER^{VMT}, Proteins, 80, 352-361.

Zhou, H. and Zhou, Y. (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments, Proteins, 58 321--328.