BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 5 2010, pages 687–688
doi:10.1093/bioinformatics/btq006

Structural bioinformatics

Advance Access publication January 14, 2010

PSiFR: an integrated resource for prediction of protein structure and function

Shashi B. Pandit, Michal Brylinski, Hongyi Zhou, Mu Gao, Adrian K. Arakaki and Jeffrey Skolnick*

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

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: In the post-genomic era, the annotation of protein function facilitates the understanding of various biological processes. To extend the range of function annotation methods to the twilight zone of sequence identity, we have developed approaches that exploit both protein tertiary structure and/or protein sequence evolutionary relationships. To serve the scientific community, we have integrated the structure prediction tools, TASSER, TASSER-Lite and METATASSER, and the functional inference tools, FINDSITE, a structure-based algorithm for binding site prediction, Gene Ontology molecular function inference and ligand screening, EFICAz^2, a sequence-based approach to enzyme function inference and DBD-hunter, an algorithm for predicting DNA-binding proteins and associated DNA-binding residues, into a unified web resource, Protein Structure and Function prediction Resource (PSiFR).

Availability and implementation: PSiFR is freely available for use on the web at http://psifr.cssb.biology.gatech.edu/

Contact: skolnick@gatech.edu

Received on September 4, 2009; revised on November 10, 2009; accepted on January 5, 2010

1 INTRODUCTION

Over the past decade, the success of genome sequencing has produced a large number of gene products with unknown structure and function (Benson *et al.*, 2009). The description of protein function ranges from its biochemical role to its role in determining phenotypical response (Skolnick and Fetrow, 2000). Thus, the complete functional annotation of a protein is a time consuming process that involves data curation from both computational and experimental studies (Consortium, 2009). Most computational function annotation tools use protein sequences to detect evolutionary relationships between proteins of unknown and known function that provide functional clues for ∼40–60% of the gene products in a given proteome (Gerstein, 1998; Muller *et al.*, 1999). However, these methods begin to fail as the sequence becomes more distant from proteins of known function (Gerstein, 1998; Tian and Skolnick, 2003). Since protein tertiary structure is more conserved than sequence, structure can play an important role in functional annotation (Baker and Sali, 2001; Skolnick and Fetrow, 2000). However, experimental determination of protein structure is time consuming, expensive and not always feasible (Slabinski *et al.*, 2007). This motivated the development of proteome-scale automated methods for protein tertiary structure and molecular function prediction.

In recent years, we developed the protein structure prediction algorithm TASSER that employs a hierarchical approach consisting of template identification by threading, followed by tertiary structure assembly from continuous template fragments (Zhang and Skolnick, 2004). TASSER can often refine the threading templates generating final models closer to their native tertiary structure than the input templates (Zhang and Skolnick, 2004). Recently, TASSER was improved by incorporating information from consensus threading templates (METATASSER) (Zhou *et al.*, 2007). The assessment of TASSER's performance in CASP 6-8 shows that it is among the best structure prediction algorithms (Zhang *et al.*, 2005; Zhou *et al.*, 2007). Furthermore, to provide rapid results, while retaining TASSER's ability to improve structure quality, we developed TASSER-Lite, a version applicable when the sequence identity between target and template is $\geq$25% (Pandit *et al.*, 2006).

For function annotation, we have implemented FINDSITE (Brylinski and Skolnick, 2008) for ligand-binding site prediction, DBD-Hunter (Gao and Skolnick, 2008) for DNA-binding prediction and EFICAz^2 (Enzyme Function Inference by a Combined Approach) (Arakaki *et al.*, 2009) for enzyme function inference. FINDSITE predicts ligand-binding pockets based on the binding site similarity among superimposed groups of template structures identified from threading. Here, threading acts as a filter to establish that the set of identified template structures are evolutionary related. FINDSITE also specifies the chemical properties of ligands that are likely to occupy the binding site and provides a collection of ligand templates for use in fingerprint-based virtual ligand screening. Furthermore, FINDSITE assigns Gene Ontology (GO) terms (Ashburner *et al.*, 2000) with a probability that corresponds to the fraction of threading templates annotated with that molecular function (Brylinski and Skolnick, 2008). DBD-Hunter is a recently developed structure-based method for identifying DNA-binding proteins and associated binding sites (Gao and Skolnick, 2008). The method first selects potential DNA-binding proteins through structural comparison to known DNA-binding proteins, and further assesses DNA-binding propensity with a DNA–protein interfacial potential. Finally, EFICAz^2 is an enzyme function inference approach that employs machine learning techniques to combine predictions from six different methods developed and optimized to achieve high prediction accuracy (Arakaki *et al.*, 2009).

*To whom correspondence should be addressed.

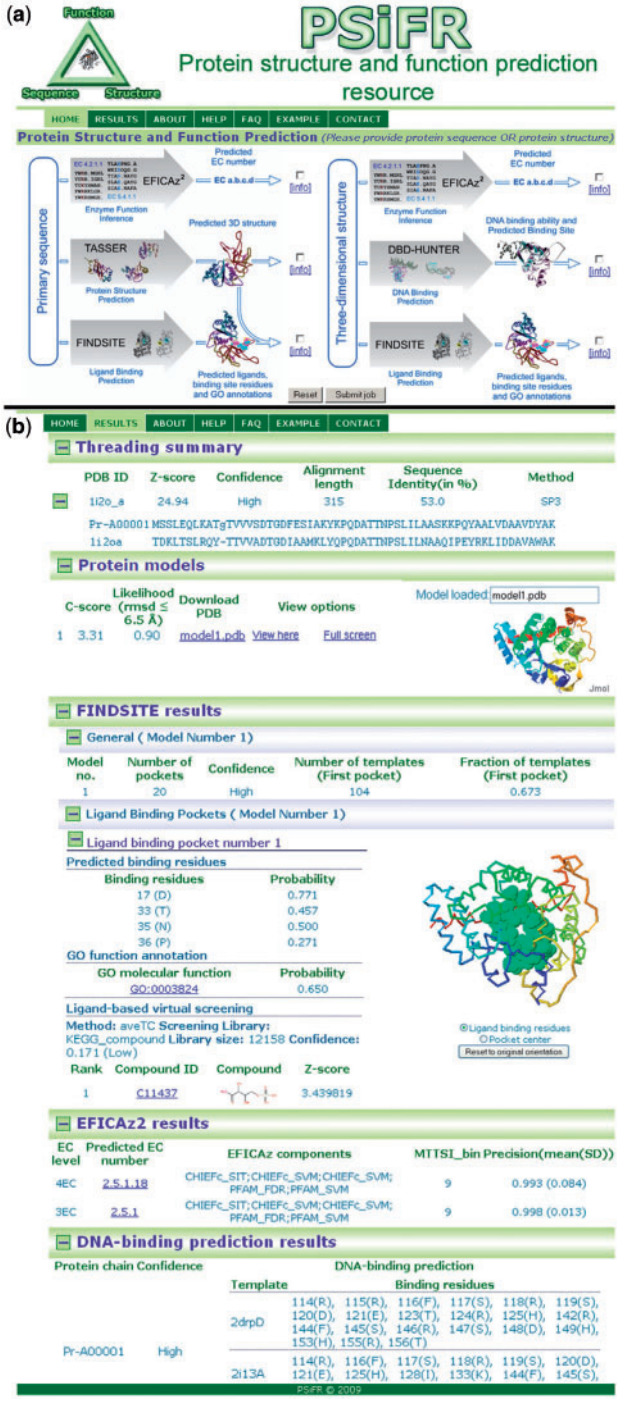

Fig. 1. Snapshots of (**a**) submission and (**b**) results pages from PSiFR.

2 PSiFR SERVER OVERVIEW

The Protein Structure and Function prediction Resource (PSiFR) server provides integrated tools for protein tertiary structure prediction and structure and sequence-based function annotation. Users can submit the protein sequence or structure with options for structure/function prediction (Fig. 1a) and a user-friendly output for visualizing the results is provided (Fig. 1b).

The results from tertiary structure prediction (TASSER/METATASSER) include the display of the threading alignments and protein models, visualized using Jmol (Jmol: an open-source Java viewer for chemical structures in 3D) (Fig. 1b). FINDSITE uses the top predicted model or the user's submitted structure for function prediction. For each model, the results for the top five binding pockets are displayed. For each pocket, predicted ligand-binding residues and predicted GO terms are shown (Fig. 1b). In addition, results of ligand-based virtual screening against the KEGG compound library are also displayed. DBD-Hunter reports whether or not the target protein is likely to bind DNA and, for putative DNA-binding proteins, shows the predicted DNA-binding residues. Similarly, results from EFICAz^2 show whether the target protein is likely to be an enzyme, in which case the predicted three field or four field enzyme commission (EC) number is also reported. The important feature of PSiFR is that its component methods provide confidence measures of the prediction quality. This helps a user to assess the reliability of the predictions. The PSiFR web service is accessible via http://psifr.cssb.biology.gatech.edu/.

Funding: National Institutes of Health (GM-48835, GM-37408 and RR-12255).

Conflict of Interest: none declared.

REFERENCES

Arakaki,A.K. *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. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

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

Benson,D.A. *et al.* (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.

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

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

Gerstein,M. (1998) Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins*, **33**, 518–534.

Muller,A. *et al.* (1999) Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.*, **293**, 1257–1271.

Pandit,S.B. *et al.* (2006) TASSER-Lite: an automated tool for protein comparative modeling. *Biophys. J.*, **91**, 4180–4190.

Skolnick,J. and Fetrow,J.S. (2000) From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol.*, **18**, 34–39.

Slabinski,L. *et al.* (2007) The challenge of protein structure determination–lessons from structural genomics. *Protein Sci.*, **16**, 2472–2482.

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.

Uniprot Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.

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

Zhang,Y. *et al.* (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins*, **61**(Suppl. 7), 91–98.

Zhou,H. *et al.* (2007) Analysis of TASSER-based CASP7 protein structure prediction results. *Proteins*, **69**(Suppl. 8), 90–97.