Demo: I-TASSER Gateway for Protein Structure Prediction and Structure-based Function Annotation

Chengxin Zhang, S. M. Mortuza, Yang Zhang*
*Corresponding author address: Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI, 48109, USA;
email: zhng@umich.edu

Abstract: *I-TASSER (Iterative Threading ASSembly Refinement) is a composite pipeline for protein structure prediction and structure-based protein function annotation. Starting from sequence of a target protein, structure templates are identified by threading from the PDB. Full-length target structure is then constructed by fragment re-assembly simulation. The final structure model is further compared to entries in BioLiP structure-function database for biological function interference. Recently, I-TASSER is implemented as an XSEDE science gateway, which helped >14,000 users to decipher structure and function of >38,000 proteins in the last 12 months. The I-TASSER gateway is available at http://zhanglab.ccmb.med.umich.edu/I-TASSER/.*

1. Introduction

The gap between the overwhelming number of protein sequences and the slow accumulation of experimentally characterized protein structures is increasing. As of May 2017, for example, there are >85 million protein sequences deposited in the UniProt database, while >41 thousands of them have experimentally characterized structures in the PDB. The lack of structures for the vast majority of protein sequences significantly hinders our understanding of their biological functions. To address this issue, I-TASSER (Iterative Threading ASSembly Refinement) [1] has been developed for automated protein structure prediction. It has been consistently ranked as one of the best servers in the most recently CASP community-wide protein structure prediction experiments [2].

In order to understand the biological function of proteins, including their Gene Ontology (GO) terms, Enzyme Commission (EC) numbers and Ligand Binding Sites, two algorithms, COFACTOR [3] and COACH [4], are developed and integrated into the I-TASSER protocol for function annotations. Both algorithms are ranked as the top webservers in CASP9 and the CAMEO community-wide protein function prediction experiments [5], respectively.

To make these state-of-the-art structure and function prediction algorithms accessible to the biological community, the I-TASSER science gateway is developed to provide an integrated platform for the biologists in the community. The I-TASSER gateway is unique compared to other webservers for protein structure prediction [6-8] and function annotation [9-11] for the tight integration of structure and function modeling, and feature-rich result webpage that facilitates biological interpretation. Since its integration with the XSEDE-Comet computer cluster in October 2016, the gateway has been used by 14,503 researchers around the world (Fig. 1.).

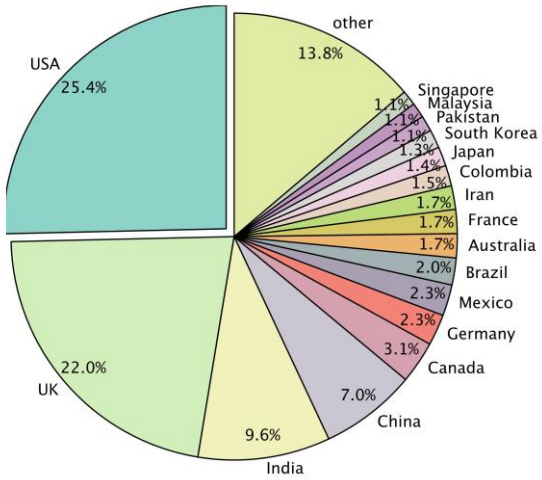

Fig. 1. Distribution of I-TASSER science gateway users among 132 countries.

Presented at Gateways 2017, University of Michigan, Ann Arbor, MI; October 23-25, 2017. https://gateways2017.figshare.com/.

2. Protein Structure and Function Prediction in I-TASSER

The I-TASSER protocol implemented by I-TASSER gateway consists of two major stages: protein structure prediction and structure-based protein function annotation.

In the structure prediction stage, the target sequence provided by a user is mapped to structure templates in the PDB database using LOMETS [12], a meta-server combining 10 different state-of-the-art threading programs [13-21]. Continuous fragments are excised from the template structures identified by each threading programs. These template fragments areto be assembled into full-length structural models by replica-exchange Monte Carlo simulation. The force field guiding the REMC simulations consists of inherent knowledge-based energy terms, contrained by the external contact and distance restraints calculated from the threading template structures. To select models, the "decoy" structures generated by the REMC simulations are clustered by SPICKER [22] according to their structural similarity, where the centroids of the top five biggest clusters, which correspond to near-native structures with low-free energy, are further refined by fragment re-assembly simulations. The re-assembled structure models are refined at atomic-level by FG-MD [23] to generate the final structure models.

The function annotation stage of the I-TASSER protocol relies on two algorithms: COFACTOR and COACH. In COFACTOR, TM-align [24] is used to compare the I-TASSER structure model with template structures in the BioLiP [25], a database of known protein structure and function associations. Biological insights, including GO terms, EC numbers and ligand binding sites are inferred from the template structures identified by local geometry matching and global structure similarity. In COACH, the ligand binding site prediction is further improved by combining the COFACTOR prediction with four other programs: TM-SITE [4], S-SITE [4], FINDSITE [26] and Concavity [27].

Upon job completion, a link of the result webpage is sent to the user through email. The on-line result page includes comprehensive report of structure template identification, structure modeling quality estimation, final structure models, and predicted biological functions. Each section of the report is organized in a detailed manner with interactive applet to facilitate visualization of the protein structures (Fig. 2).

3. Conclusion

I-TASSER science gateway is a feature-rich and unified platform for high-resolution modeling of protein structure and function. The webserver interface is designed with specific emphasis on biological interpretation of modeling results. The

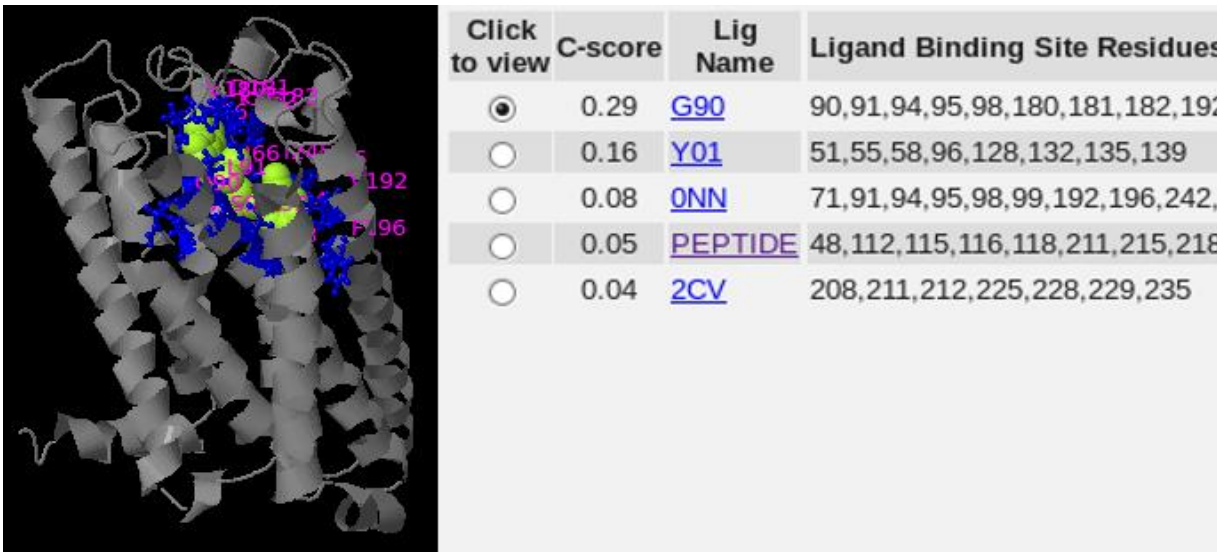

Fig. 2. An excerpt of a result from the I-TASSER webserver. The structure model (grey) in complex with the predicted ligand (light green) is shown in an interactive applet on the left. The predicted ligand binding residues are highlighted in blue on the left and listed in the table on the right.

computationally expensive simulations of I-TASSER are sped up by Comet cluster provided by XSEDE [28], which enables to serve a broader community of biological and medical researchers.

4. Acknowledgments

I-TASSER science gateway uses the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

5. References

[1] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nature Protocols,* vol. 5, pp. 725-738, 2010.

[2] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)--round x," *Proteins,* vol. 82 Suppl 2, pp. 1-6, Feb 2014.

[3] A. Roy, J. Y. Yang, and Y. Zhang, "COFACTOR: an accurate comparative algorithm for structure-based protein function annotation," *Nucleic Acids Research,* vol. 40, pp. W471-W477, Jul 2012.

[4] J. Y. Yang, A. Roy, and Y. Zhang, "Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment," *Bioinformatics,* vol. 29, pp. 2588-2595, Oct 15 2013.

[5] J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, *et al.*, "The Protein Model Portal--a comprehensive resource for protein structure and model information," *Database (Oxford),* vol. 2013, p. bat031, 2013.

[6] J. Soding, A. Biegert, and A. N. Lupas, "The HHpred interactive server for protein homology detection and structure prediction," *Nucleic Acids Research,* vol. 33, pp. W244-W248, Jul 1 2005.

[7] D. E. Kim, D. Chivian, and D. Baker, "Protein structure prediction and analysis using the Robetta server," *Nucleic Acids Research,* vol. 32, pp. W526-W531, Jul 1 2004.

[8] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, and M. J. Sternberg, "The Phyre2 web portal for protein modeling, prediction and analysis," *Nat Protoc,* vol. 10, pp. 845-58, Jun 2015.

[9] F. Minneci, D. Piovesan, D. Cozzetto, and D. T. Jones, "FFPred 2.0: Improved Homology-Independent Prediction of Gene Ontology Terms for Eukaryotic Protein Sequences," *Plos One,* vol. 8, May 22 2013.

[10] D. Piovesan, M. Giollo, E. Leonardi, C. Ferrari, and S. C. E. Tosatto, "INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity," *Nucleic Acids Research,* vol. 43, pp. W134-W140, Jul 1 2015.

[11] Q. T. Gong, W. Ning, and W. D. Tian, "GoFDR: A sequence alignment based method for predicting protein functions," *Methods,* vol. 93, pp. 3-14, Jan 15 2016.

[12] S. T. Wu and Y. Zhang, "LOMETS: A local meta-threading-server for protein structure prediction," *Nucleic Acids Research,* vol. 35, pp. 3375-3382, May 2007.

[13] S. T. Wu and Y. Zhang, "MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information," *Proteins-Structure Function and Bioinformatics,* vol. 72, pp. 547-556, Aug 1 2008.

[14] J. Soding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics,* vol. 21, pp. 951-960, Apr 1 2005.

[15] Y. Xu and D. Xu, "Protein threading using PROSPECT: Design and evaluation," *Proteins-Structure Function and Genetics,* vol. 40, pp. 343-354, Aug 15 2000.

[16] R. X. Yan, D. Xu, J. Y. Yang, S. Walker, and Y. Zhang, "A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction," *Scientific Reports,* vol. 3, Sep 10 2013.

[17] L. Jaroszewski, L. Rychlewski, Z. W. Li, W. Z. Li, and A. Godzik, "FFAS03: a server for profile-profile sequence

alignments," *Nucleic Acids Research,* vol. 33, pp. W284-W288, Jul 1 2005.

[18] M. Madera, "Profile Comparer: a program for scoring and aligning profile hidden Markov models," *Bioinformatics,* vol. 24, pp. 2630-2631, Nov 15 2008.

[19] A. Lobley, M. I. Sadowski, and D. T. Jones, "pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination," *Bioinformatics,* vol. 25, pp. 1761-1767, Jul 15 2009.

[20] D. Xu, L. Jaroszewski, Z. W. Li, and A. Godzik, "FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking," *Bioinformatics,* vol. 30, pp. 660-667, Mar 1 2014.

[21] H. Y. Zhou and Y. Q. Zhou, "Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition," *Proteins-Structure Function and Bioinformatics,* vol. 55, pp. 1005-1013, Jun 1 2004.

[22] Y. Zhang and J. Skolnick, "SPICKER: A clustering approach to identify near-native protein folds," *Journal of Computational Chemistry,* vol. 25, pp. 865-871, Apr 30 2004.

[23] J. Zhang, Y. Liang, and Y. Zhang, "Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling," *Structure,* vol. 19, pp. 1784-1795, Dec 7 2011.

[24] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Research,* vol. 33, pp. 2302-2309, 2005.

[25] J. Y. Yang, A. Roy, and Y. Zhang, "BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions," *Nucleic Acids Research,* vol. 41, pp. D1096-D1103, Jan 2013.

[26] M. Brylinski and J. Skolnick, "A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 105, pp. 129-134, Jan 8 2008.

[27] J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, and T. A. Funkhouser, "Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure," *Plos Computational Biology,* vol. 5, Dec 2009.

[28] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, *et al.*, "XSEDE: Accelerating Scientific Discovery," *Computing in Science & Engineering,* vol. 16, pp. 62-74, Sep-Oct 2014.