Future Generation Computer Systems 99 (2019) 73–85

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

I-TASSER gateway: A protein structure and function prediction server powered by XSEDE

Wei Zheng[1], Chengxin Zhang[1], Eric W. Bell, Yang Zhang *

Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109-2218, United States

HIGHLIGHTS

- The I-TASSER gateway is a server for protein structure and function prediction.
- The XSEDE-Comet supercomputer provides the computational backend for the gateway.
- The gateway has become a popular service for the biological community.

ARTICLE INFO

Article history:
Received 18 June 2018
Received in revised form 4 November 2018
Accepted 3 April 2019
Available online 9 April 2019

Keywords:
Protein structure prediction
Structure-based protein function annotation
I-TASSER web-server
XSEDE science gateway

ABSTRACT

There is an increasing gap between the number of known protein sequences and the number of proteins with experimentally characterized structure and function. To alleviate this issue, we have developed the I-TASSER gateway, an online server for automated and reliable protein structure and function prediction. For a given sequence, I-TASSER starts with template recognition from a known structure library, followed by full-length atomic model construction by iterative assembly simulations of the continuous structural fragments excised from the template alignments. Functional insights are then derived from comparative matching of the predicted model with a library of proteins with known function. The I-TASSER pipeline has been recently integrated with the XSEDE Gateway system to accommodate pressing demand from the user community and increasing computing costs. This report summarizes the configuration of the I-TASSER Gateway with the XSEDE-Comet supercomputer cluster, together with an overview of the I-TASSER method and milestones of its development.

1. Introduction

As the workhorses of life, proteins perform a wide range of biological functions in living organisms, including synthesis and breakdown of biomolecules, regulation of developmental processes and biological pathways, and constitution of protein complexes and subcellular structures. To carry out its functions, a protein usually needs to be folded into a specific tertiary structure from its amino acid sequence. Therefore, obtaining the structure of a protein is usually an important step towards elucidation of that protein's function. Unfortunately, experimental solution of protein structures is expensive in terms of time, labor, and cost. As a result, the gap between the overwhelmingly large number of protein sequences and the small number of experimentally solved protein structures has kept increasing over the years. As of June 2018, for example, there are more than 116 million proteins sequences deposited in the UniProt [1] database, and less than 45 thousand of them have an experimentally solved structure in the Protein Data Bank (PDB) [2] database.

To address this issue as well as to serve the high demand from the community for structure and function information of uncharacterized proteins, we have developed the I-TASSER (Iterative Threading ASSembly Refinement) [3–6] algorithm for automated protein structure prediction. I-TASSER structure prediction consists of threading based fold recognition, followed by fragment-based structure reassembly and refinement simulations. It has been consistently ranked as the top method in the community-wide CASP experiments [7–12] for accurate 3D structure prediction. To make full use of the I-TASSER models for understanding the biological function of proteins, two algorithms, COFACTOR [13–15] and COACH [16], have been developed and integrated into the I-TASSER protocol for structure-based function annotations, with focuses on Gene Ontology (GO) [17] terms, Enzyme Commission (EC) [18] numbers and ligand binding sites. Both COFACTOR and COACH have been examined in community-wide blind function prediction experiments [19,20], including CASP and CAMEO, and showed significantly higher performance than other state-of-the-art function prediction methods.

* Corresponding author.
E-mail address: zhng@umich.edu (Y. Zhang).
[1] These two authors contributed equally to this work.

https://doi.org/10.1016/j.future.2019.04.011
0167-739X/

Apart from the I-TASSER gateway, there are other on-line servers designed for protein structure and function prediction. For example, the HHpred [21], Phyre2 [22] and SWISS-MODEL [23] servers all implement template based protein structure prediction services by combining threading algorithms with fast homology modeling. However, these servers are usually not focused on structure quality improvement beyond template alignments, which may cause difficulty in modeling the hard targets that lack homologous structure templates, as evidenced by recent CASP experiments. Similarly, there are other protein function annotation portals, but they also have their own caveats. For example, the Bologna annotation resource [24] and INGA [25] make function predictions based on sequence alignment, protein family searching, or protein interaction networks, but these servers do not directly use protein structure information, which is generally believed to be more conserved than sequence in functional evolution. ProFunc [26] is one of the few examples of a server that generates functional annotations using structural information, but it requires experimental structure information of the query protein. Therefore, the advantage of the I-TASSER online webserver [3–5] is in the integration of structure and function modeling, feature-rich result display that facilitates biological interpretation, and high accuracy in structure and function modeling even for challenging protein targets. These features are essential to making the I-TASSER server a useful computational platform to the biological community.

To alleviate the unprecedented computational challenges brought about by the ever-growing user base, we have recently converted the I-TASSER protocol into an Extreme Science and Engineering Discovery Environment (XSEDE) [27] science gateway. By migrating the computational backend of the I-TASSER web-service to the XSEDE-Comet supercomputer cluster, we are able to continue serving a broader community with improved user experiences. Since its integration with the XSEDE-Comet computer cluster in October 2016, the I-TASSER gateway has created modeling results for 38,333 researchers around the world and has become one of the most heavily used gateways running on XSEDE-Comet [28]. In this paper, we report the web front end design and underlying implementation of this newly developed I-TASSER gateway.

2. Pipeline of the I-TASSER gateway

The I-TASSER (Iterative Threading ASSembly Refinement) protocol implemented by the I-TASSER gateway consists of four major steps: template threading, structural fragment assembly, model refinement and structure-based protein function annotation (Fig. 1).

Step 1: template threading. Threading refers to the process of identifying template proteins from the PDB database of experimentally solved structures that have a potentially similar structure to the query protein. In the I-TASSER gateway, we incorporated the LOcal MEta-Threading Server (LOMETS) [29] to identify homologous templates from a non-redundant structure database (https://zhanglab.ccmb.med.umich.edu/library/). LOMETS combines 14 state-of-the-art threading programs (FFAS03 [30], FFAS-3D [31], HHsearch [32], MUSTER [33], pGenThreader [34], PPA [35], PROSPECT2 [36], SP3 [37], SAM-T99 [38] and SPARKS-X [39]). It should be noted that we run two variants of MUSTER and three variants of HHsearch using different parameters, which is shown to improve threading result [6]. In each individual threading program, the top template hits are ranked by profile-based and local-structure-based scores, where the raw alignment score is converted to a Z-score [3–5] that signifies the statistical significance of the alignments. When combining templates from different programs, the Z-score and consensus score are used to define the target type (easy or hard) and to select the templates that are used for the next step of structure assembly simulations.

Step 2: structural fragment assembly. In the second step, continuous fragments from threading alignments are excised from the template structures, and the unaligned gap regions (mostly loops) will be built by I-TASSER *ab initio* modeling. Aligned fragments and unaligned loops are used for structural conformation assembly by Replica-Exchange Monte Carlo simulation (REMC) [40]. To improve the speed of conformational sampling, I-TASSER adopts a reduced structure model to represent the protein, where each residue is represented by its Cα atom and side-chain center of mass. The unaligned loop regions are confined to a lattice system of grid size 0.87 Å, which reduces the conformational search space. Since the threading-aligned regions usually have good modeling quality, conformations in these regions are off lattice and are kept rigid during the simulations. The I-TASSER REMC simulation implements 40 replicas in parallel with different temperatures. Within each replica, the acceptance of movements obeys the Metropolis–Hastings criterion. The force field guiding the REMC simulation consists of inherent knowledge-based energy terms, distance restraints calculated from the threading template structures, and optionally, external contact restraints or additional template restraints specified by server users. This REMC simulation generates tens of thousands of low energy conformations, which are commonly referred to as 'decoys'.

Step 3: model refinement. The decoys generated by the low-temperature replicas in the REMC simulation are clustered by SPICKER [41] in order to group decoys with similar tertiary structures. Cluster centroids are then obtained by averaging the 3D coordinates of all the clustered structural decoys. 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 in order to obtain re-assembled structure models with more physically realistic geometry. The re-assembled structure models are refined at an atomic level by FG-MD [42] in order to generate the final structure models. In FG-MD refinement, soft cutoffs for the Lennard-Jones and Coulombic terms are applied to the AMBER [43] force-field for guiding the MD simulation. After the generation of a refined final model, a confidence score (C-score) estimating the global quality of the models is computed based on the template quality and REMC simulation convergence [3]. Additionally, in order to provide residue level structure quality, we use our recently developed algorithm, ResQ [44], to estimate the residue displacement and provide them in the B-factor field in the PDB files of the top 5 models. This ResQ prediction is obtained through a Support Vector Regression predictor that combines information of structure template agreement, local structure consistency, and simulation convergence.

Step 4: structure-based protein functional annotation. The function annotation step of the I-TASSER protocol relies on two algorithms: COFACTOR and COACH. In COFACTOR, TM-align [45] is used to compare the I-TASSER structure model with function template structures in BioLiP [46], a weekly curated database of known protein structure and function associations (416,452 entries by May 2018). 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. The ligand binding site prediction from COFACTOR is further refined by COACH, which additionally combines four other programs, TM-SITE [16], S-SITE [16], FINDSITE [47] and ConCavity [48] to derive a consensus ligand binding site prediction.

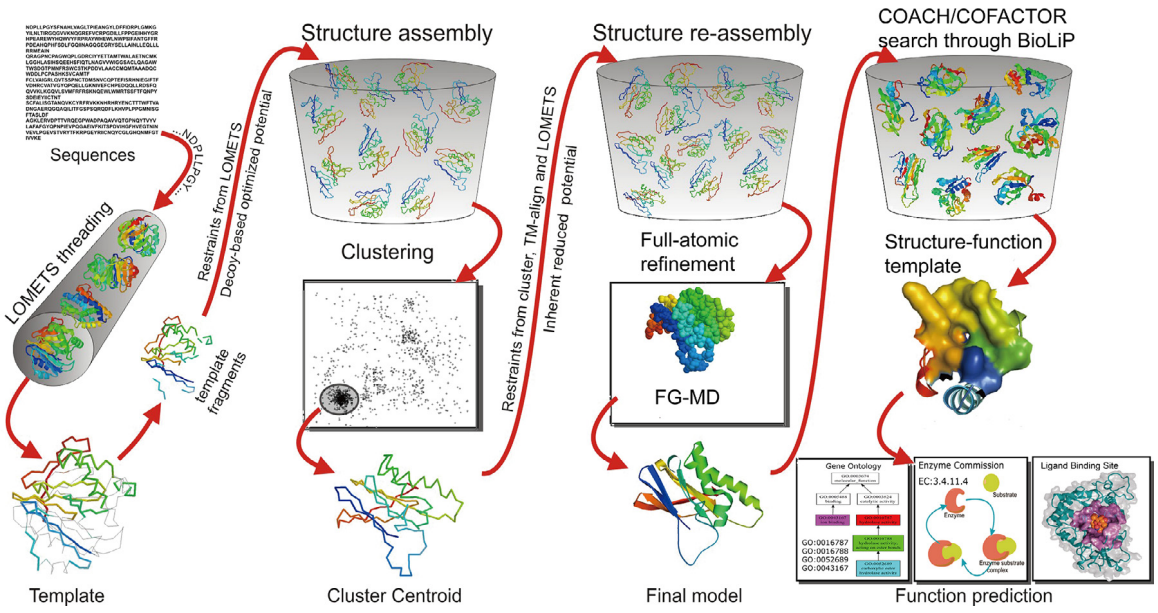

Fig. 1. Flowchart of the I-TASSER structure and function prediction protocol, which includes four stages: structural template threading (first column), structural fragment assembly (second column), model refinement (third column) and structure-based protein function annotation (fourth column).

3. Implementation of the I-TASSER gateway (hardware/background server)

The hardware implementation of the I-TASSER gateway is briefly shown by Fig. 2. When a user submits a protein sequence to the I-TASSER webserver at https://zhanglab.ccmb.med.umich.edu/I-TASSER/, the Common Gateway Interface (CGI) script run by the webserver checks the user input for illegal amino acids (such as B, Z, X or non-alphabetical characters), duplicated sequence submission, corrupted template structures, and mismatching user emails and passwords. If all the input information is correct, the user input is deposited to webserver by the CGI script, which also updates the list of pending server jobs. A crontab script run by the Zhang cluster, which periodically runs every 10 min, copies the user data from webserver to the Zhang cluster through Secure Copy Protocol (SCP). This crontab script also preprocesses the user input file by converting protein sequences and user provided template structures to a uniform format. After these preprocessing steps, this crontab script transfers the data files to the XSEDE-Comet cluster by rsync. At the Comet cluster, another crontab script running at the head node of the Comet cluster every 10 min checks the incoming new user input files and submits I-TASSER structure and function prediction jobs to the Comet computing nodes. After the structure and function predictions are finished, the crontab script running at the Zhang cluster will again use rsync to synchronize data from Comet cluster to the Zhang cluster, prepare HTML webpage output, and copy the data to the webserver front end for webpage display. Meanwhile, another crontab script running on the webserver updates our MySQL database for user job statistics and informs the user of job completion by email. The MySQL database of I-TASSER consumes 237 MB of disk space and holds around 500,000 records for user credentials and job submission history. On average, only 0.08 s is needed to query the MySQL database by the CGI script. Considering its relatively fast speed and small disk space consumption, as well as being actively maintained by upstream developers, MySQL is chosen as our main solution to relational database management.

In the I-TASSER gateway, user input has been mainly handled by low-level CGI scripts since the launch of the webserver. We had previously experimented with higher-level frameworks and client-side solutions (such as JavaScript). However, we found they are not easy to implement especially due to our need for complicated user input validation. Therefore, we have been adhering to low-level CGI mechanisms and accepting the responsibility of maintaining the security of our in-house code. For example, on an annual basis, we coordinate with the IT team of our institute to scan and fix webserver security vulnerabilities, including, but not limited to, MySQL injections, cross-site scripting, SMTP MX injections, and Unix file parameter alterations, as reported by Burp, IBM Security AppScan, and Open Bug Bounty.

Three additional notes should also be made regarding file handling in the I-TASSER gateway setup. Firstly, since usually a large number of files (typically a few thousand files per day) need to be transferred between the Comet cluster and the Zhang cluster, the file transfer protocol must cope with the relatively unstable network connection between the two clusters. Neither the commonly used SCP mechanism nor the more advanced Globus protocol is efficient enough for the transfer of such a large volume of data files. Rsync, another open source utility for file transfer across computer systems, typically provides more efficient data synchronization for our purpose. By checking the timestamp and size of local and remote files, rsync can automatically detect which files were not fully transferred in the previous file transfer attempts and avoid re-synchronization of files that are already correctly transferred. For example, on Oct 27, 2018, we synchronized the newly updated structure library from the Zhang cluster to the Comet cluster. The synchronization involved 2,618 new files. Rsync finished the synchronization within 2 h and 10 min, whereas both SCP and Globus died during the file transfer process. Therefore, rsync is used as the default file transfer engine between the Zhang cluster and Comet cluster.

Secondly, file I/O is intensive for multiple steps in the I-TASSER structure modeling pipeline, which creates unprecedented load on the 'Data Oasis' Lustre file system mounted since the early days of the I-TASSER gateway. To cope with this issue, we have partially rewritten the I-TASSER structure prediction pipeline to utilize the local SSD scratch space utilized by every Comet computing node. This local SSD scratch space on the Comet computing nodes is of critical importance for the I-TASSER gateway, and impacts every step of the protein structure prediction process. First, for LOMETS threading, template profiles created from the whole

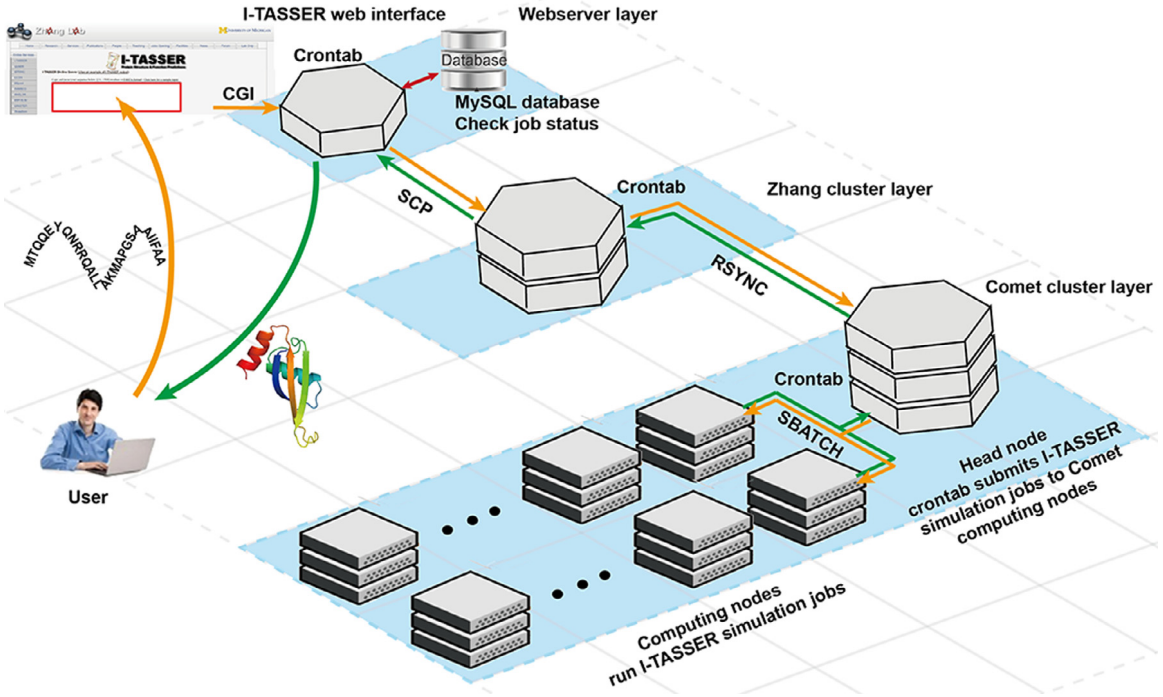

Fig. 2. Hardware implementation of the I-TASSER gateway. Users submit their jobs and visualize the prediction results at our webserver. Our internal computer cluster (Zhang cluster) preprocesses user input data, generates HTML webpage output and synchronizes data between the webserver front end and the XSEDE-Comet cluster backend. The Comet cluster performs the structure and function prediction computation.

template library need to be temporarily stored in the local SSD to avoid high I/O traffic due to the large size of the database; these profile files are later removed from the SSD after the threading process is completed. Second, all the decoy conformations created by the REMC simulation need to be stored on the SSD for the next step of clustering analysis. Since this step can result in heavy I/O traffic on the hard disk, all the simulation trajectories are cached at local SSD's before they are compressed and copied back to the shared file system on the Comet cluster at the end of simulation, thereby minimally impacting the shared file system. Finally, various intermediate output and temporary parameter files for clustering, structure re-assembly, and full-atomic refinement are also written to the local SSD to increase the speed of refinement simulation processes. We tested different I/O strategies for the I-TASSER pipeline on three proteins with 100 residues. The pipeline directly writing files to 'Data Oasis' shared file system spent 16.2% more running time compared to the current pipeline that utilizes local SSD for heavy I/O operations.

Thirdly, the I-TASSER system needs a large disk space for medium and long term storage, which we need to carefully administrate. The major storage needs of the I-TASSER system consist of two parts. The first part is for the structural and functional template libraries. These two libraries are essential for the template-based structure and function predictions, where target sequences and structure models are to be scanned through the libraries to identify structural and functional homologs to assist the corresponding structure assembly and function annotations. Both libraries are updated weekly according to the PDB and UniProt databases and a PubMed literature search. As of October 2018, the structure library takes up 372 GB and the functional library from BioLiP takes up 142 GB. The second part of the need for storage is for the results of the I-TASSER simulations. The I-TASSER pipeline consists of multiple steps of template recognition, structural assembly simulation, decoy clustering and model selection, structure refinement, and structure-based functional template searching and annotation. The modeling of each target generates on average ∼1 GB simulation data. We have adopted two strategies to save storage space. First, we compress all the structural files using the BZIP2 format. Second, we remove all the intermediate simulation decoy files after the final model generation. Nevertheless, each target still generates around 100 Mb of data, which includes mainly the modeling coordinates and image files. Currently, the I-TASSER server accepts ∼200–300 jobs per day from the user community, while the data files of each target are kept on the server system for 2 months before deletion, to give users a buffer time for data analysis and downloading. Thus, the simulation data occupies around 1.2 TB (=100 Mb × 200 users × 60 days). However, only 1 TB hard disk space is granted to the I-TASSER gateway by XSEDE, and since XSEDE gives supports to the I-TASSER gateway and 3,656 other science gateways and research groups, the average hard disk space allocated to each group is limited to around 700 GB (≈2.6PB/3656). To solve this dilemma, we run weekly crontab scripts on the Comet cluster to delete user data after job completion and successful transfer to the Zhang cluster, where we will keep the user data for up to 2 months.

4. Input and output of the I-TASSER gateway (user interface webpage)

4.1. Input of the I-TASSER gateway

To model a query protein of interest using the I-TASSER gateway, a user can paste a FASTA format amino acid sequence of the query to the input text box or upload a sequence file through the upload button (Fig. 3A). At present, the I-TASSER server accepts protein sequences of up to 1,500 amino acids in length. For sequences longer than 1,500 amino acids, as these are potentially multi-domain proteins, the user may split the sequence into domains by our in-house domain partition servers ThreaDom [49] or ThreaDomEx [50] and separately submit the sequences of the individual domains.

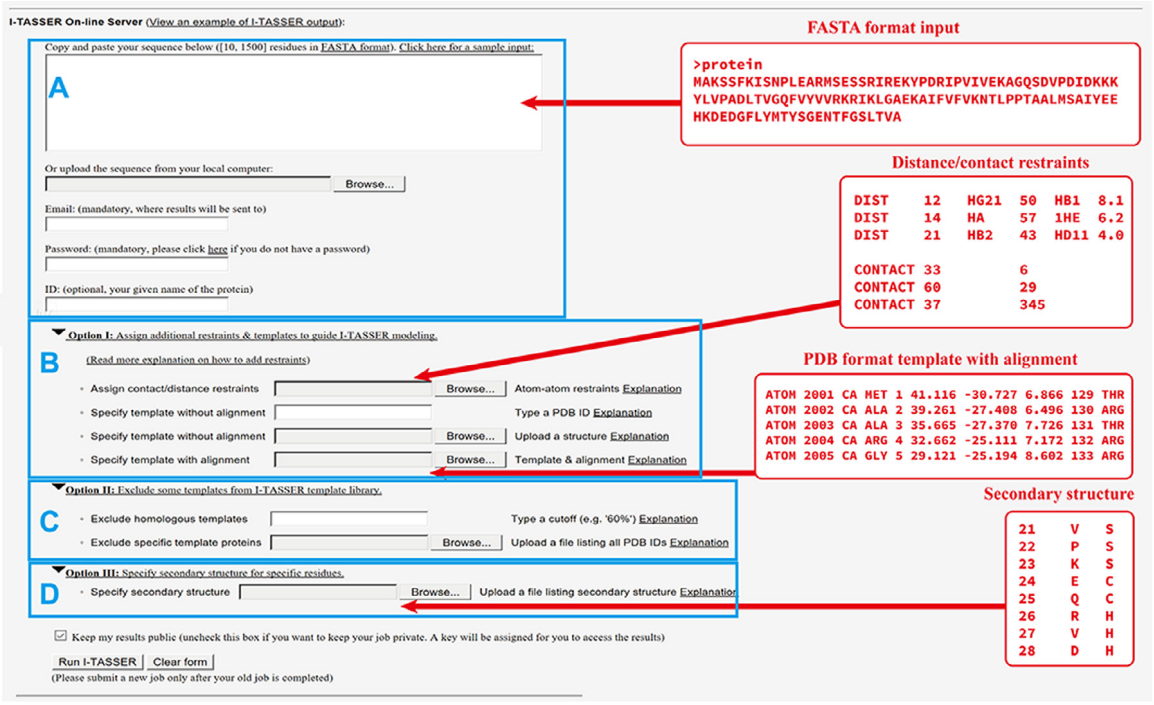

Fig. 3. Input interface of the I-TASSER gateway. Through this input webpage, a user can specify (A) the mandatory protein sequence and user email input; (B) additional distance/contact restraints and/or additional structure templates, both optionally provided by user; (C) exclusion of specific templates optionally specified by user; and (D) optional user-provided secondary structure restraints.

The I-TASSER server also provides three additional options for users to specify additional biological insights to guide modeling. These three options include: user specified distance/contact restraints, exclusion or inclusion of specific templates from modeling, and user specified secondary structures.

A user can specify external distance or contact restraints (Fig. 3B) to guide the I-TASSER structure assembly simulations. Normally, these restraints are collected from threading templates, but for some targets, especially for 'hard' targets, those restraints are often unreliable because of erroneous alignment between the query and distant-homology templates. As restraints from experiments often have a higher accuracy than those derived from threading alignments, user-specified restraints can be useful for improving the quality of the structure assembly simulation, especially for 'hard' targets. Restraint data from NMR or crosslinking experiments can be specified by uploading a restraint file. An example can be found in Fig. 3B, where column 1 specifies the type of restraint, i.e., 'DIST' or 'CONTACT'. For contact restraints, columns 2 and 3 mean positions (i, j) of the residues (the *i*th residue and *j*th residue in the sequence) that form a contact. I-TASSER will try to draw these two residues into contact during the structure assembly simulation. For distance restraints, residue positions (i, j) are shown in columns 2 and 4, and atom names specified by the PDB format are shown in column 3 and 5. Column 6 defines the distance between the two atoms. Similar to contact restraints, these atom pairs will be brought close to the specified distance by I-TASSER structure assembly simulations.

I-TASSER also allows users to specify their own templates. Normally, I-TASSER starts with protein templates identified by LOMETS threading programs from our PDB library (71,027 entries by June 2018). Our template library consists of representative protein structures at a pair-wise sequence identity cutoff of 70%. As the desired template may not be included in our library or may not be identified by LOMETS even if it is in our library, users can specify a solved PDB-formatted protein structure as the template (Fig. 3B). In this case, the I-TASSER structure simulation will start from the user-specified template, with restraints mainly collected from it; meanwhile, the templates from LOMETS threading will still be used for generating restraints in regions not covered by the user-specified template. To enable the I-TASSER server to use a template, the 3D structure and the alignment between the query and the template sequence are essential. If a user uploads a template structure without alignment, MUSTER will be utilized to align query with the template. More details can be found in https://zhanglab.ccmb.med.umich.edu/I-TASSER/restraint.html

As requested by many users, we have implemented a new feature in the I-TASSER server for excluding some templates from the I-TASSER template library, a feature designed for special purposes such as benchmark testing. Users can either set the upper limit of sequence identity between template and query, or specify the PDB ID to exclude the templates from the I-TASSER structure assembly simulation (Fig. 3C). Another new function we have implemented in I-TASSER server is allowing the user to specify secondary structure for specific residues (Fig. 3D). The first column shows the specified residue index of query protein sequence, the second column is residue name of the specified residue, and the last column is the secondary structure type of this residue ('H' for alpha helix, 'S' for beta strand and 'C' for coil). Finally, a user can keep their sequence private by unchecking the box above the 'Run I-TASSER' button.

4.2. Output of the I-TASSER gateway

I-TASSER takes around 10 h to generate results for a typical medium-size protein with 200 to 400 residues. However, when a user submits a sequence, the actual processing time also depends on the number of jobs in our queue. In reality, users typically receive results within 1–2 d. Users can check their job status and approximate completion time at https://zhanglab.ccmb.med.umich.edu/I-TASSER/queue.php. For each submission, a unique job ID corresponding to a unique URL will be generated for tracking modeling status and storing the results. When the

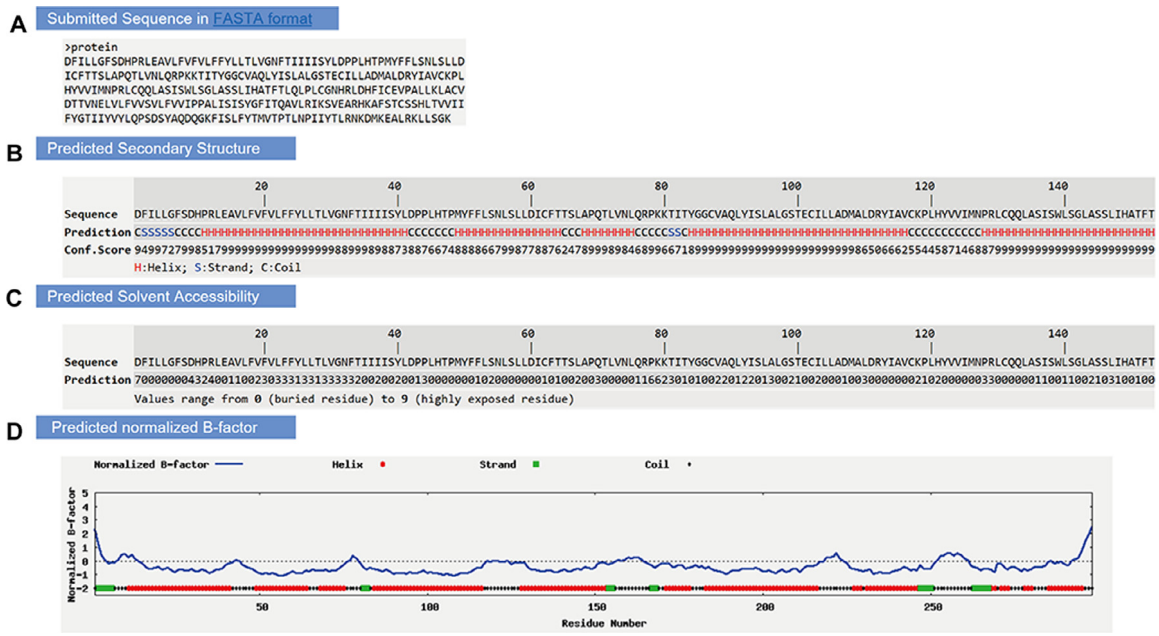

Fig. 4. The first four sections of the output result webpage, showing (A) the query protein sequence; (B) predicted secondary structure and the prediction confidence; (C) predicted solvent accessibility; and (D) ResQ predicted B-factor.

job is completed, users will receive a notification through their registered email. An example output page is available at https://zhanglab.ccmb.med.umich.edu/I-TASSER/example. The output data includes: (1) the submitted sequence (Fig. 4A) and predicted structural features including secondary structure information (Fig. 4B), predicted solvent accessibility (Fig. 4C), and normalized B-factor (Fig. 4D), (2) the top ten threading templates identified by LOMETS (Fig. 5A), (3) the top five final 3D models with measures for model quality estimation (Fig. 5B), (4) the top ten proteins found to be structurally closest to the first model of the query (Fig. 5C), (5) the results of structure-based function annotations, including EC numbers, GO terms, and ligand binding sites (Fig. 6). All results shown in the webpage are also available in a BZIP2 format compressed archive file (i.e. tarball), which users can download at the top of the same webpage. The output results are introduced briefly below. A detailed explanation of the results is provided at https://zhanglab.ccmb.med.umich.edu/I-TASSER/annotation.

The submitted sequence and predicted structural features are shown in the first four sections (Fig. 4) of the result page. A typical secondary structure prediction contains three states: alpha helix (H), beta strand (S) and coil (C), with confidence scores for each residue. The confidence score of predicted secondary structure and predicted solvent accessibility are also provided, with a score range [0, 9]. In general, positive B-factor values indicate that the residues are more flexible, while negative values suggest that the residues are more rigid. An additional link is provided to show the user-specified restraints, including distance/contact restraints, template alignments, and secondary structure restraints.

The information of the top 10 templates identified by LOMETS is listed in the fifth section of the results page (Fig. 5A), which includes: (1) the PDB IDs of templates, (2) the sequence identities between the template and the query, (3) the coverage of the alignments, (4) the normalized threading Z-score, (5) the alignments between the query and the templates. The normalized Z-score indicates the alignment quality, where a normalized Z-score >1 indicates a confident alignment. The query protein is defined as an 'easy' target if there is at least one template from each threading program having a normalized Z-score >1, otherwise, it is a 'hard' target. In the current version of LOMETS, 14 threading programs are employed to identify up to 280 templates (up to 20 templates per threading program) for distance restraint extraction. However, the top 10 templates are the most relevant because they are given a higher weight during restraint collection and are used as the starting conformation in the low-temperature replicas in the REMC simulation. By clicking on the PDB ID, the webpage will be re-directed to the RCSB Protein Data Bank website showing information about the template protein. Users can download the PDB-formatted threading alignment file by clicking on the 'Download' button. The alignment file can be viewed in any molecular visualization software (e.g. Jmol, PyMol, VMD, etc.).

Up to five full length atomic structural models with their C-scores are reported in the sixth section of the results page (Fig. 5B). The estimated TM-score and RMSD are also provided for the first model. All these scores are used for estimating the global accuracy of the models. The C-score is typically in the range $[-5, 2]$, where a higher score reflects a better quality. Normally, models with a C-score >-1.5 have the correct topology. More information about the C-score can be found at https://zhanglab.ccmb.med.umich.edu/I-TASSER/example/cscore.txt. Users can download the PDB-formatted structure model file by clicking the link under the figures of the models. The JSmol [51] applet is used for visualization of the model structure in the result pages. Users can drag, rotate or zoom the structure to check the model from different orientations.

The first I-TASSER model is searched against the PDB library by TM-align to find the top 10 analogs that are structurally similar to the query proteins (Fig. 5C). The results of this search are shown in the seventh section of the results page. The TM-scores, RMSDs, sequence identities, and alignment coverages between analogous structures and the query protein are also listed in this section. The alignment with the highest TM-score is visualized by JSmol by default. Users can click the 'Click to view' radio buttons to visualize alignments for other structure analogs. Here, users should note that the top 10 analogous structures (detected by TM-align using the I-TASSER first model) are usually not identical to the top 10 templates (detected by the threading programs from the query sequence) listed in fifth section. The alignment (in PDB format) can be downloaded through the 'download' link.

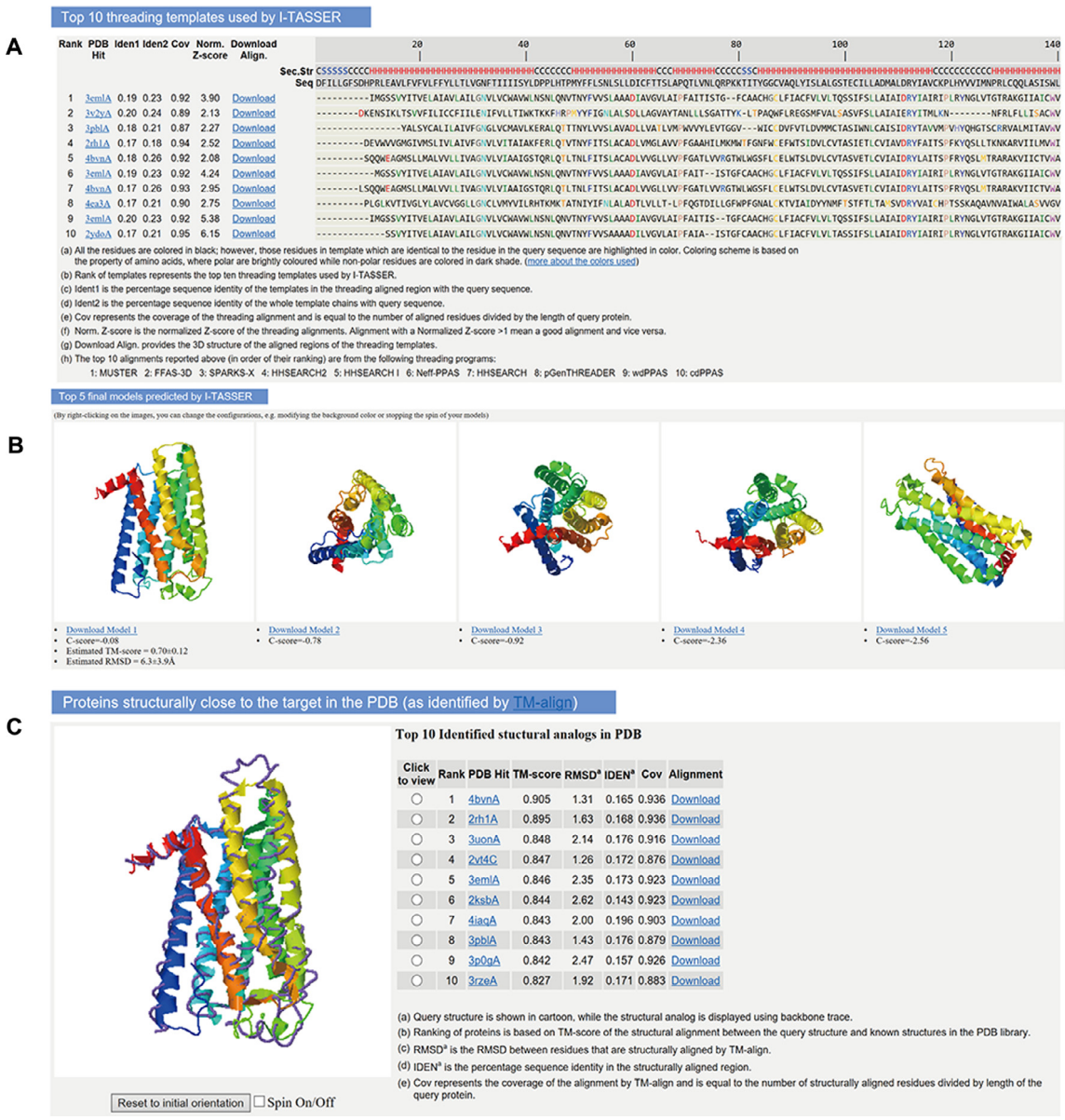

Fig. 5. Sections in the output webpage displaying the main structure modeling results, which include (A) the alignment between the query and the structure template identified from the PDB structure library by LOMETS; (B) up to five final structure models with estimated global structure quality metrics underneath the JSmol protein structure visualization applet; (C) structure analogs identified by a TM-align search using the first I-TASSER final model as the query structure.

The first I-TASSER model, which in general has the highest structure quality, is submitted to COFACTOR and COACH to predict its biological function, including ligand-binding site (Fig. 6A), EC number (Fig. 6B), and GO terms (Fig. 6C). Outputs of the structure-based function annotation by COACH and COFACTOR are listed in the last section of the results page. The predicted binding ligands, corresponding C-score, and binding site residues are shown in the sixth, third and eighth columns, respectively. By clicking on 'Lig Name', the browser will be redirected to the BioLiP entry for this ligand. Users can download the PDB format docked complex by clicking the 'Mult' link. The predicted binding site residues corresponding to each ligand are visualized in the left panel. Users can select different ligands by clicking the 'selection button'. The EC number prediction provides similar results for catalytic active site prediction, and the 'EC number' column is directed to the ExPASy enzyme [52] database, which provides a detailed description of the enzyme families. GO term predictions are available in the last part of the results page, which is further subdivided into two tables. The first table lists the top 10 ranked template proteins associated with their corresponding GO terms. The consensus GO terms for each of the three functional aspects (molecular function, biological process and cellular component) are presented in the second table.

5. Performance of the I-TASSER gateway and case studies

5.1. Performance of the I-TASSER gateway

The I-TASSER structure prediction server as well as its function annotation components COFACTOR and COACH have all been objectively evaluated in the CASP and CAMEO structure and function prediction challenges, in which these programs showed considerably higher performance when compared with other state-of-the-art approaches.

Critical Assessment of protein Structure Prediction (CASP) is a worldwide experiment for the objective benchmarking of protein structure prediction protocols. The CASP experiment is organized

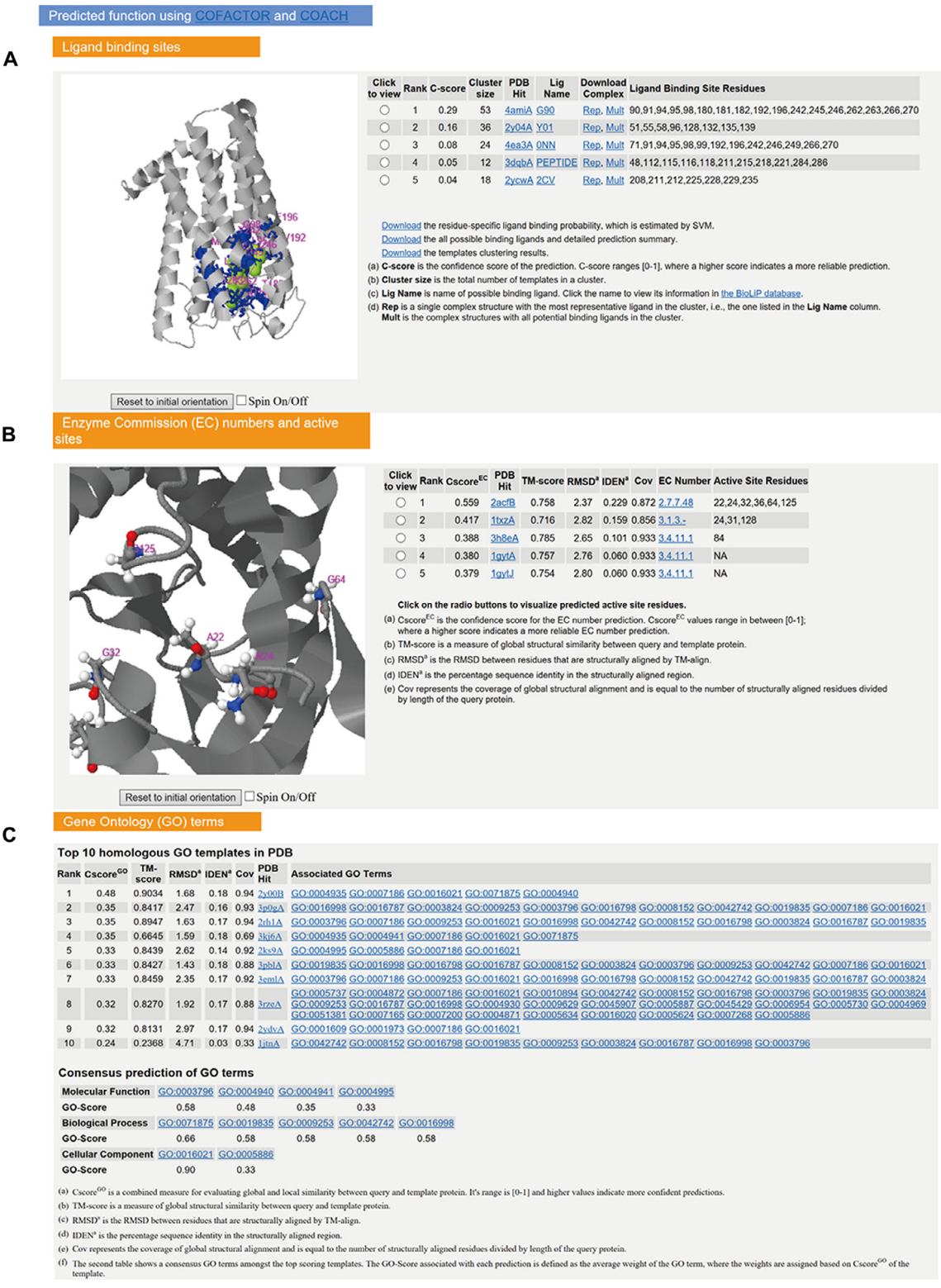

Fig. 6. Webpage output sections for the predicted protein functions, including (A) COACH predicted ligand binding residues in complex with putative ligands; (B) predicted EC numbers and the inferred active site residues by COFACTOR; (C) GO terms from the consensus COFACTOR prediction (bottom) and the structure templates from which the GO term annotations are inferred (top).

as blind modeling tasks, where participating servers are asked to predict protein structures that have been experimentally solved but not yet released. The I-TASSER server participated in the 7th CASP as 'Zhang-Server' in 2006 and was ranked as the best server among all 66 servers (Fig. 7). Starting from CASP9, an *ab initio* modeling tool, QUARK, from our lab has also participated in CASP. The I-TASSER server and QUARK server [53] have been ranked as the top 2 servers ever since.

Since 2012, I-TASSER has employed our in-house COFACTOR algorithm as one of its protein function annotation module to predict ligand binding sites, GO terms and EC numbers. COFACTOR

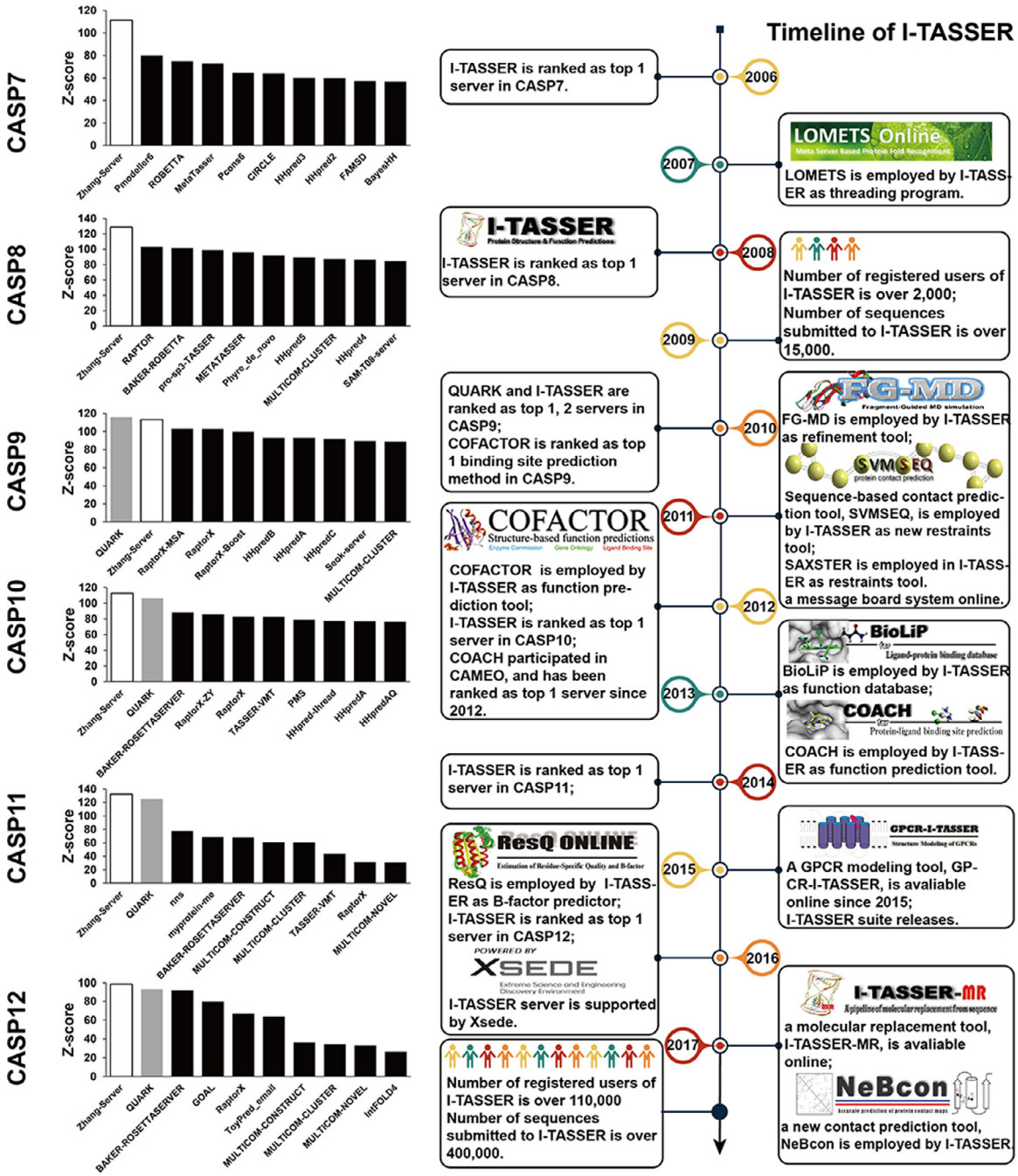

Fig. 7. The timeline of I-TASSER. The results of the CASP7 to CASP12 experiments are shown in the left panel. Each bar plot ranks the top 10 protein structure prediction servers in descending order of Z-score for GDT_TS, which is a metric quantifying overall structure quality. The white and gray bars are the I-TASSER server ('Zhang-Server' in CASP), and QUARK server, respectively, both from our group. It is shown that the I-TASSER is the top server from CASP7 to CASP12. The timeline for development of the I-TASSER server and its component algorithms is shown in the right panel. All the programs and databases illustrated by this panel are available online at https://zhanglab.ccmb.med.umich.edu/services/.

participated in the CASP9 ligand binding site prediction challenge, which was the last CASP to have any function prediction component. Achieving a precision of 72% and a Matthews Correlation Coefficient 0.69 for 31 blind test proteins [13], COFACTOR was ranked as the best server in the CASP function prediction challenge. Another of our protein function prediction modules, COACH, has participated in a similar community-wide challenge, Continuous Automated Model EvaluatiOn (CAMEO) [19] since 2012. In CAMEO, COACH generated ligand binding site predictions for 7,193 targets (between 2015-04-24 and 2016-04-16) with an average AUC score of 0.76, which is 0.08 higher than the second-best method in the experiment (https://www.cameo3d.org/lb/1-year/). These data demonstrate that the I-TASSER server represents one of the most reliable algorithms for automated protein structure and function prediction.

5.2. Evolution of the I-TASSER gateway

Since the I-TASSER server has been made publicly available online, we have made a continuous effort to improve the pipeline with the goal of providing the most accurate structural and function predictions. The timeline of the I-TASSER server (Fig. 7) briefly lists how the server pipeline has been updated with various new components in the last decade. Important events for

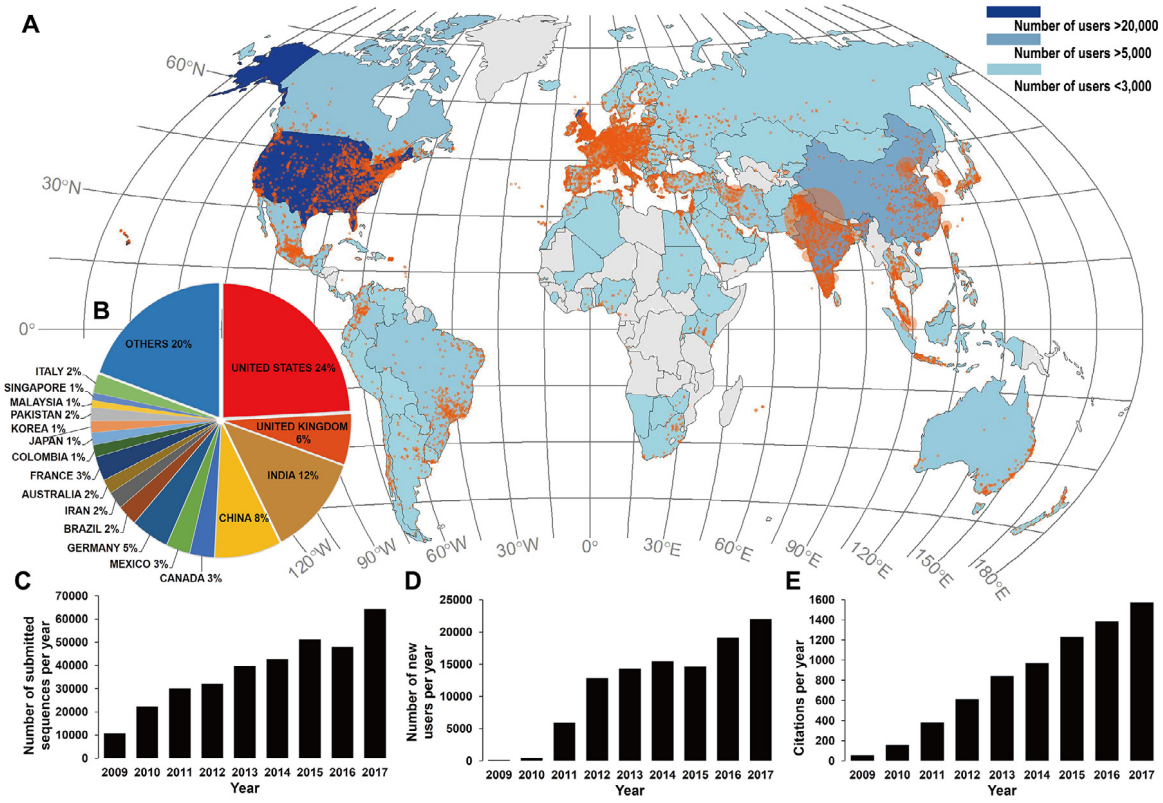

Fig. 8. Distribution of I-TASSER users. In the world map, different countries are colored from dark to light blue according to the descending number of registered I-TASSER users. Different cities are marked by orange points, whose size is proportional to the number of sequences submitted to I-TASSER server by these cities. The pie chart shows the percentage of the number of sequences submitted to I-TASSER by different countries among all submitted sequences. In the lower panel, from left to right, the three bar plots show the number of submitted sequences, the number of new users, and the number of citations per year. All three annual statistics represented by the three plots are steadily increasing year by year. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the improvement of our structure prediction protocol include (in chronological order), introduction of the LOMETS threading meta-server, addition of FG-MD for local structure refinement, inclusion of SVMSEQ [54] and NeBcon [55] for residue-residue contact prediction, and development of ResQ for local structure quality estimation. Taking advantage of these improvements in I-TASSER structure model quality, we have developed and incorporated three services for structure-based function prediction: COFACTOR for general function annotation in 2012, COACH for ligand binding site prediction in 2013, and a database, BioLiP, for compilation of protein-ligand binding structures in the same year. Recently, in response to user requests to develop a standalone version of the I-TASSER pipeline to run on their own computer, we have also packaged the I-TASSER structure and function prediction protocol into the standalone I-TASSER suite [6].

Due to the reliability and extensibility of the I-TASSER pipeline for generating high quality structure and function predictions, the pipeline has been extended to other more specific structure or function modeling tasks over the years. For instance, we developed in 2011 the SAXSTER [56] server for high-resolution protein structure determination by including small-angle X-ray scattering (SAXS) data as additional restraints in I-TASSER structure modeling. In 2015, GPCR-I-TASSER [57] was specifically developed for the prediction of G protein-coupled receptor (GPCR) structures. In 2017, I-TASSER-MR [58], a molecular replacement tool for solving the phase problem of distant-homology proteins in X-ray crystallography, was released.

Since all the tools mentioned in the previous two paragraphs are available online (https://zhanglab.ccmb.med.umich.edu/services/) with user bases whose needs diversify as they expand, we receive a great deal of questions and suggestions for the use and improvement of our services. As a forum for gathering all of these questions and responses, we have launched in 2011 a message board system (https://zhanglab.ccmb.med.umich.edu/bbs/) to facilitate discussions between our service developers and the outside users.

While the I-TASSER gateway provides a user-friendly interface to access I-TASSER, it might not satisfy all the special needs of power users, such as the need to process many protein targets using their own computing resources, check intermediate files, or use a specially built template database, partly due to the resource limit of the host machines. To address the diverse need of users, we have also prepared a standalone I-TASSER package [6] and made it freely available to academic community at https://zhanglab.ccmb.med.umich.edu/I-TASSER/download/. While the downloadable I-TASSER suite implements the same structure and function prediction methods as the webserver, it also preserves all intermediate files for users who are interested in the computational details. It also provides additional scripts for users to construct their own template databases so that their modeling results can be made for other research purposes.

5.3. Statistics of the I-TASSER gateway

By June 2018, over 110,000 users from 20,983 institutions have registered on the I-TASSER server, and these users come from 9,124 cities in 138 countries (Fig. 8A). More than 400,000

proteins have been successfully predicted by the I-TASSER gateway since it became available online. Researchers from the United States (24%), the United Kingdom (6%), India (12%) and China (8%) have submitted about 50% of the total number of sequences (Fig. 8B). Delhi (1,3677), Basel (8,838) and Beijing (4,763) are the top three cities, ranked by number of submitted sequences. Starting from 2012, we have over 10,000 new users registering for I-TASSER every year (Fig. 8C), and this number is still increasing by 11.2% per year. The number of submitted sequences has been increasing by 17.2% per year since 2010, the year that more than 20,000 protein models were generated by I-TASSER (Fig. 8D). Starting from October 2016, I-TASSER has been supported by the XSEDE-Comet cluster. The I-TASSER gateway consistently gains more computational resources from XSEDE for serving a high volume of users, and 38,333 new users have submitted 109,673 sequences since being powered by XSEDE. I-TASSER related papers [3–6] have been cited by more than 7,900 researches by 2017 (Fig. 8E), which indicates the I-TASSER gateway is also one of the most popular protein modeling tools.

5.4. Case studies

This section includes two successful examples of the I-TASSER server being used to elucidate essential biological insights into protein structure and function. In 2016, a new gene (MCR-1) [59] from an *E. coli* plasmid was found to be responsible for antibiotic resistance: an *E. coli* strain possessing the protein encoded by this gene could not be readily killed by polymyxin. Using I-TASSER, it was found that the MCR-1 protein structure resembles a topology typical of phosphoethanolamine transferase (Fig. 9A [59]). The I-TASSER structure model helped elucidate the molecular mechanism of antibiotic resistance induced by MCR-1, in which phosphoethanolamine is added to lipid A. Lipid A is an important component of endotoxin and is responsible for the toxicity of bacterial infection; the phosphoethanolamine modification reduces binding between lipid A and polymyxin and thus neutralizes polymyxin's power in curing *E. coli* infection.

The second case [60] concerns Mig-6 (Fig. 9B pink), which inhibits cell migration by binding to Cdc42 (Fig. 9B blue). In this study, the authors were interested in identifying 'hot spots', that is, the amino acids that are important for interaction between the two proteins. Since Mig-6 lacked an experimental structure, the interaction was modeled by predicting the Mig-6 structure with I-TASSER and docking this structure to the Cdc42 protein using Zdock [61]. From this Mig-6/Cdc42 structure model, four amino acids important for the interaction were identified. The importance of these four key residues, I11, R12, M26, R30, was confirmed by western blot experiments.

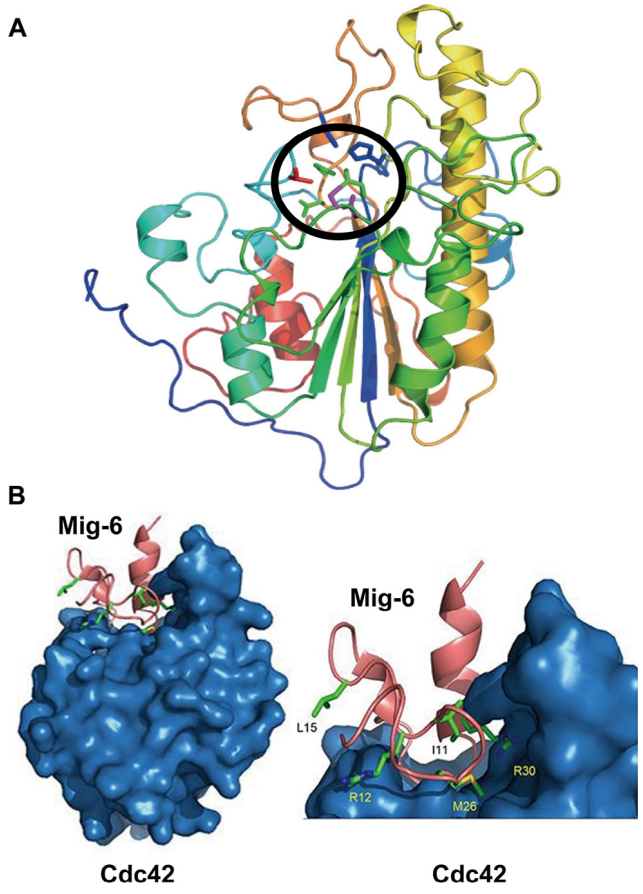

Fig. 9. Examples of the use of I-TASSER server by the user community for elucidating the molecular mechanism of important proteins. (A) The I-TASSER structure model of MCR-1 adopts a tertiary topology typical of phosphoethanolamine transferases, with putative catalytic residues shown as sticks (black circle). This figure is adapted from Liu et al. [59]. (B) I-TASSER predicted structure of Mig-6 (cartoon) in complex with the native protein structure of Cdc42 (surface). I11, R12, M26, and R30, four hot spot residues essential for interaction between Mig-6 and Cdc42, are shown as green sticks. An additional residue, L15, which is not essential for the interaction and was used as negative control in follow-up wet lab experiments, is also represented by green sticks. This figure is adapted from Jiang et al. [60] . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Conclusion

The I-TASSER gateway is an online server designed for automated prediction of protein structure and biological functions from protein sequence. The involved algorithms have been evaluated by community-wide blind tests (CASP and CAMEO) and demonstrated considerably higher prediction accuracy than other methods of protein structure and function prediction. The webserver interface is designed with specific emphasis on biological interpretation of modeling results. Attributed to good high-resolution structure models, accurately predicted protein functions and a user-friendly web interface, the I-TASSER server has become one of the most widely used gateways in the field of protein structure modeling. When generating a high-quality model, various computationally expensive simulations are needed, and internal computational resources from a single laboratory is not sufficient to provide a satisfying service for all users. To address these issues, the I-TASSER server has been powered by the XSEDE-Comet cluster since 2016, which enables the I-TASSER gateway to serve a broader community of biological and medical researchers. In 2016, as well as in 2017, the number of new registered users significantly increased. In 2017, the number of models generated by I-TASSER was around 70,000, which increased over 40% compared to the annual statistics from the last decade. All these data indicate the significant improvement of the I-TASSER server with support from XSEDE. We believe that with the integration of external resources such as XSEDE, the I-TASSER Gateway will have the potential and support to significantly enhance method development in protein structure and function modeling as well as make these tools accessible to an unprecedentedly broader range of users.

Acknowledgments

The I-TASSER gateway uses the Extreme Science and Engineering Discovery Environment (XSEDE) [27], which is supported

by the National Science Foundation, United States [grant number ACI-1053575]. The authors are grateful to Nancy Wilkins-Diehr and Mahidhar Tatineni, who provide outreach opportunities and technical support for the I-TASSER gateway's use of the XSEDE-Comet cluster. The authors also thank Dr. S.M. Mortuza and Qiqige Wuyun for critical review of the manuscript, and Dr. Xiaoqiong Wei for insightful discussions. This work was supported in part by the NIGMS, United States [grant numbers GM083107, GM116960], the NIAID, United States [grant number AI134678], the National Institutes of Health, United States Training Program in Bioinformatics [grant number T32 GM070449], and the National Science Foundation, United States [grant number DBI1564756].

References

[1] U. Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (2016) D158–D169.
[2] H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide protein data bank, Nature Struct. Mol. Biol. 10 (2003) 980.
[3] Y. Zhang, I-TASSER Server for protein 3D structure prediction, BMC Bioinform. 9 (2008) 40.
[4] A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction, Nat. Protoc. 5 (2010) 725.
[5] J. Yang, Y. Zhang, I-TASSER Server: new development for protein structure and function predictions, Nucleic Acids Res. 43 (2015) W174–W181.
[6] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER suite: protein structure and function prediction, Nat. Methods 12 (2015) 7.
[7] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, A. Tramontano, Critical assessment of methods of protein structure prediction—Round VII, Proteins: Struct. Funct. Bioinform. 69 (2007) 3–9.
[8] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, A. Tramontano, Critical assessment of methods of protein structure prediction—Round VIII, Proteins: Struct. Funct. Bioinform. 77 (2009) 1–4.
[9] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP)—Round XII, Proteins: Struct. Funct. Bioinform. 86 (2017) 7–15.
[10] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP) — round x, Proteins: Struct. Funct. Bioinform. 82 (2013) 1–6.
[11] J. Moult, K. Fidelis, A. Kryshtafovych, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP)—round IX, Proteins: Struct. Funct. Bioinform. 79 (2011) 1–5.
[12] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction: Progress and new directions in round XI, Proteins: Struct. Funct. Bioinform. 84 (2016) 4–14.
[13] A. Roy, Y. Zhang, Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement, Structure 20 (2012) 987–997.
[14] A. Roy, J. Yang, Y. Zhang, COFACTOR: an accurate comparative algorithm for structure-based protein function annotation, Nucleic Acids Res. 40 (2012) W471–W477.
[15] C. Zhang, P.L. Freddolino, Y. Zhang, COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information, Nucleic Acids Res. (2017) gkx366.
[16] J. Yang, A. Roy, Y. Zhang, Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment, Bioinformatics 29 (2013) 2588–2595.
[17] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, Gene ontology: tool for the unification of biology, Nat. Genetics 25 (2000) 25.
[18] E.C. Webb, Enzyme nomenclature 1992, in: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, Academic Press, 1992.
[19] J. Haas, A. Barbato, D. Behringer, G. Studer, S. Roth, M. Bertoni, K. Mostaguir, R. Gumienny, T. Schwede, Continuous automated model evaluation (CAMEO) complementing the critical assessment of structure prediction in CASP12, Proteins: Struct. Funct. Bioinform. 86 (2017) 387–398.
[20] J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, T. Schwede, The protein model portal—a comprehensive resource for protein structure and model information, Database 2013 (2013) bat031-bat031.
[21] L. Zimmermann, A. Stephens, S.-Z. Nam, D. Rau, J. Kübler, M. Lozajic, F. Gabler, J. Söding, A.N. Lupas, V. Alva, A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core, J. Mol. Biol. 430 (2018) 2237–2243.
[22] L.A. Kelley, M.J. Sternberg, Protein structure prediction on the web: a case study using the Phyre server, Nat. Protoc. 4 (2009) 363.
[23] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T.G. Cassarino, M. Bertoni, L. Bordoli, SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information, Nucleic Acids Res. 42 (2014) W252–W258.
[24] D. Piovesan, P. Luigi Martelli, P. Fariselli, A. Zauli, I. Rossi, R. Casadio, BAR-PLUS: the bologna annotation resource plus for functional and structural annotation of protein sequences, Nucleic Acids Res. 39 (2011) W197–W202.
[25] D. Piovesan, M. Giollo, E. Leonardi, C. Ferrari, S.C. Tosatto, INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity, Nucleic Acids Res. 43 (2015) W134–W140.
[26] R.A. Laskowski, J.D. Watson, J.M. Thornton, Profunc: a server for predicting protein function from 3D structure, Nucleic Acids Res. 33 (2005) W89–W93.
[27] CASP13, http://predictioncenter.org/casp13/doc/CASP13_Abstracts.pdf, (2018).
[28] http://ucsdnews.ucsd.edu/pressrelease/sdscs_comet_supercomputer_surpasses_10000_users_milestone.
[29] S. Wu, Y. Zhang, LOMETS: a local meta-threading-server for protein structure prediction, Nucleic Acids Res. 35 (2007) 3375–3382.
[30] L. Jaroszewski, L. Rychlewski, Z. Li, W. Li, A. Godzik, FFAS03: a server for profile–profile sequence alignments, Nucleic Acids Res. 33 (2005) W284–W288.
[31] D. Xu, L. Jaroszewski, Z. Li, A. Godzik, FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking, Bioinformatics 30 (2013) 660–667.
[32] J. Söding, Protein homology detection by HMM–HMM comparison, Bioinformatics 21 (2004) 951–960.
[33] S. Wu, Y. Zhang, MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information, Proteins: Struct. Funct. Bioinform. 72 (2008) 547–556.
[34] A. Lobley, M.I. Sadowski, D.T. Jones, pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination, Bioinformatics 25 (2009) 1761–1767.
[35] R. Yan, D. Xu, J. Yang, S. Walker, Y. Zhang, A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction, Sci. Rep. 3 (2013) 2619.
[36] Y. Xu, D. Xu, Protein threading using PROSPECT: design and evaluation, Proteins: Struct. Funct. Bioinform. 40 (2000) 343–354.
[37] H. Zhou, Y. Zhou, Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments, Proteins: Struct. Funct. Bioinform. 58 (2005) 321–328.
[38] K. Karplus, C. Barrett, R. Hughey, Hidden markov models for detecting remote protein homologies, Bioinform. (Oxford, England) 14 (1998) 846–856.
[39] H. Zhou, Y. Zhou, Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition, Proteins: Struct. Funct. Bioinform. 55 (2004) 1005–1013.
[40] Y. Zhang, J. Skolnick, Automated structure prediction of weakly homologous proteins on a genomic scale, Proc. Natl. Acad. Sci. USA 101 (2004) 7594–7599.
[41] Y. Zhang, J. Skolnick, SPICKER: a clustering approach to identify near-native protein folds, J. Comput. Chem. 25 (2004) 865–871.
[42] J. Zhang, Y. Liang, Y. Zhang, Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling, Structure 19 (2011) 1784–1795.
[43] D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, R.J. Woods, The Amber biomolecular simulation programs, J. Comput. Chem. 26 (2005) 1668–1688.
[44] J. Yang, Y. Wang, Y. Zhang, ResQ: an approach to unified estimation of B-factor and residue-specific error in protein structure prediction, J. Mol. Biol. 428 (2016) 693–701.
[45] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Res. 33 (2005) 2302–2309.
[46] J. Yang, A. Roy, Y. Zhang, BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions, Nucleic Acids Res. 41 (2012) D1096–D1103.
[47] M. Brylinski, J. Skolnick, A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation, Proc. Nat. Acad. Sci. 105 (2008) 129–134.
[48] J.A. Capra, R.A. Laskowski, J.M. Thornton, M. Singh, T.A. Funkhouser, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, PLoS Comput. Biol. 5 (2009) e1000585.
[49] Z. Xue, D. Xu, Y. Wang, Y. Zhang, Threadom: extracting protein domain boundary information from multiple threading alignments, Bioinformatics 29 (2013) i247–i256.

[50] Y. Wang, J. Wang, R. Li, Q. Shi, Z. Xue, Y. Zhang, Threadomex: a unified platform for predicting continuous and discontinuous protein domains by multiple-threading and segment assembly, Nucleic Acids Res. 45 (2017) W400–W407.
[51] M. Hanson Robert, J. Prilusky, Z. Renjian, T. Nakane, L. Sussman Joel, JSmol And the next-generation web-based representation of 3D molecular structure as applied to proteopedia, Isr. J. Chem. 53 (2013) 207–216.
[52] A. Bairoch, The ENZYME database in 2000, Nucleic Acids Res. 28 (2000) 304–305.
[53] D. Xu, Y. Zhang, Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field, Proteins: Struct. Funct. Bioinform. 80 (2012) 1715–1735.
[54] S. Wu, Y. Zhang, A comprehensive assessment of sequence-based and template-based methods for protein contact prediction, Bioinformatics 24 (2008) 924–931.
[55] B. He, S.M. Mortuza, Y. Wang, H.-B. Shen, Y. Zhang, NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers, Bioinformatics 33 (2017) 2296–2306.
[56] Marcelo A. dos Reis, R. Aparicio, Y. Zhang, Improving protein template recognition by using small-angle x-ray scattering profiles, Biophys. J. 101 (2011) 2770–2781.
[57] J. Zhang, J. Yang, R. Jang, Y. Zhang, GPCR-I-TASSER: a hybrid approach to g protein-coupled receptor structure modeling and the application to the human genome, Structure 23 (2015) 1538–1549.
[58] Y. Wang, J. Virtanen, Z. Xue, Y. Zhang, I-TASSER-MR: automated molecular replacement for distant-homology proteins using iterative fragment assembly and progressive sequence truncation, Nucleic Acids Res. (2017) gkx349.
[59] Y.-Y. Liu, Y. Wang, T.R. Walsh, L.-X. Yi, R. Zhang, J. Spencer, Y. Doi, G. Tian, B. Dong, X. Huang, Emergence of plasmid-mediated colistin resistance mechanism mcr-1 in animals and human beings in china: a microbiological and molecular biological study, Lancet Infectious Diseases 16 (2016) 161–168.
[60] X. Jiang, M. Niu, D. Chen, J. Chen, Y. Cao, X. Li, H. Ying, J. Bergholz, Y. Zhang, Z.-X. Xiao, Inhibition of Cdc42 is essential for Mig-6 suppression of cell migration induced by EGF, Oncotarget 7 (2016) 49180.
[61] R. Chen, L. Li, Z. Weng, ZDOCK: an initial-stage protein-docking algorithm, Proteins: Struct. Funct. Bioinform. 52 (2003) 80–87.

Wei Zheng is a postdoctoral research fellow at Department of Computational Medicine and Bioinformatics, University of Michigan. He obtained his Ph.D. in Bioinformatics from Nankai University, China, in 2017. His major research interest involves machine learning, drug screening, protein structure modeling and refinement, and structure-based protein function annotation.

Chengxin Zhang is a Ph.D. student in Bioinformatics at Department of Computational Medicine and Bioinformatics, University of Michigan. In 2015, he obtained his B.S. in Biological Sciences from Fudan University, China. His research focuses on protein structure prediction and structure-based protein function annotations.

Eric W. Bell is a Ph.D. student in Bioinformatics at Department of Computational Medicine and Bioinformatics, University of Michigan. He obtained his B.A. in Biochemistry from Oberlin College, Oberlin, in 2017. He is currently a part of the NIH Training Program in Bioinformatics (T32 GM070449) and his research focuses on drug discovery and inverse virtual screening.

Yang Zhang is a profess or at Department of Computational Medicine and Bioinformatics and Department of Biological Chemistry, University of Michigan. The major research interest of Dr. Zhang's lab is in protein folding and structure prediction, and protein design and engineering. Dr. Zhang is recipient of US NSF Career Award, Alfred P Sloan Award, and Dean's Basic Science Research Award, and was selected as the Thomson Reuters Highly Cited Researcher in 2015-2018.