CHAPTER 11

COMPOSITE APPROACHES TO PROTEIN TERTIARY STRUCTURE PREDICTION: A CASE-STUDY BY I-TASSER

AMBRISH ROY, SITAO WU, and YANG ZHANG
Center for Computational Medicine and Bioinformatics
University of Michigan
Ann Arbor, MI

11.1. INTRODUCTION

The post-genomic era is witnessing an upsurge of protein sequences in public databases. By the end of 2009, over 9 million protein sequences had been deposited in the Universal Protein Resource (UniProtKB)/TrEMBL [1]. However, this increase in the amount of sequence data does not necessarily reflect an increase in biological knowledge. One of the most challenging tasks that have emerged in recent years is to functionally characterize these sequences for better understanding of physiological processes and systems [2]. This has motivated computational biologists to develop a variety of fast and accurate methods for quickly characterizing these sequences.

One of the most significant efforts in this regard has been the development of powerful sequence alignment algorithms like Basic Local Alignment Search Tool **(**BLAST) [3], Position-Specific Iterative-BLAST (PSI-BLAST) [4], and hidden Markov model (HMM) techniques [5,6], which are frequently used for identifying evolutionary homologs and transferring functional annotations. The underlying assumption in these approaches is that evolutionarily related sequences fold similarly [7,8] and the functional similarity between these related proteins can be explored by detecting evolutionary relationship

Introduction to Protein Structure Prediction: Methods and Algorithms,
Edited by Huzefa Rangwala and George Karypis

between them [9,10]. It has been estimated that using these approaches, functional inference can be drawn for nearly 40–60% of the open reading frames (ORFs) in the genome [11]. However, there are numerous cases where functional conservation exists in evolutionarily diverged proteins but annotations cannot be transferred based on evolutionary-based approaches [12,13]. At this juncture, it is apparent that protein sequences are generally insufficient for determining protein functionality and providing support for functional genomics [14].

The three-dimensional (3D) structure of a protein is closely linked to its biological function [15]. As residues located far apart in the primary sequence may be very close in 3D space and only a few spatially conserved residues are generally responsible for a protein's function [16,17], the 3D structure of a protein provides useful insight into the key component(s) of its functionality. This awareness and the limited number of solved protein structures in Protein Data Bank (PDB) [18] have actuated the structural genomics (SG) project to increase the throughput of experimental structure elucidation [19–21] and provide a framework for inferring molecular function [22,23]. While the SG aims to structurally characterize the protein universe by an optimized combination of experimental structure determination and comparative modeling (CM) building, 3D structures of at least 16,000 optimally selected proteins would be required in order that the CM can cover 90% of protein domain families [24] and at the current rate it appears that this goal can be achieved only in the next 10 years [25]. This underscores the need for computational methods for protein structure prediction, so that 3D structural models can be built and can provide insight for functional analysis. Also, the development of better structural refinement and CM methods would dramatically enlarge the scope of structural genomics project.

Historically, protein structure prediction methods have been classified into three categories: CM [26,27], threading [28–33] and *ab initio* modeling [34–38]. In CM, the protein structure is constructed by matching the sequence of the protein of interest (query) to an evolutionarily related protein with a known structure (template) in the PDB [18], where the residue equivalency between query and the template is obtained by aligning sequences or sequence profiles. Threading-based methods match the query protein sequence directly to 3D structures of solved proteins with the goal of recognizing similar protein folds that may have no clear evidence of an evolutionary relationship with the query protein. The last resort for predicting the protein structure, when no good template is detected in the PDB library, is to predict the structure using *ab initio* modeling. Predictions based on this method assume that the native structure of a protein corresponds to its global free-energy minimum [39] and the conformational space is sampled to attain this state as guided by well-designed energy force fields. This is the most difficult category of protein structure prediction and if successful will provide the eventual solution to protein folding problem. However, the success of *ab initio* modeling is currently limited to small proteins with less than 100 amino acids [34–38].

As a general trend in the field of protein structure prediction, the borders between the conventional categories of methods have become blurred. For instance, both CM- and threading-based methods use sequence-profile and profile-profile alignments (PPA) for identifying templates. Similarly, most of the contemporary *ab initio*-based methods often use evolutionary information either for generating sparse spatial restraints or for identifying local structural building blocks. Recent community-wide blind tests have demonstrated significant advantages of the composite approaches in protein structure predictions [40–42], which combines the various techniques from threading, *ab initio* modeling, and atomic-level structure refinements [43,44].

In this chapter, we will focus on the methodology of I-TASSER [35,44,45], which serves as a case study of the composite approach for generating 3D structural models and predicting the function of a given query sequence. The performance of I-TASSER on benchmark tests and in the recent Critical Assessment of Protein Structure Prediction (CASP) experiments [44,46] will be discussed. Finally, in the concluding section, the current status and future perspective are summarized.

11.2. I-TASSER: A COMPOSITE METHOD FOR PROTEIN STRUCTURE PREDICTION

I-TASSER [35,44,45] is a hierarchical protein structure modeling approach based on the multiple threading alignments and an iterative implementation of the Threading ASSEmbly Refinement (TASSER) program [47]. Figure 11.1 represents the schematic diagram of I-TASSER methodology for protein structure and function prediction, which consist of four consecutive steps of threading, structure assembly, structure refinement, and function prediction.

11.2.1. Threading of Query Sequence

Given a query protein, the first step of I-TASSER is to thread the query sequence through a representative PDB structure library (sequence identity cutoff of 70%) with the objective of identifying the global or local threading alignments using either MUSTER [29] (single threading server) or LOMETS [33] (meta-threading server). In this section, we will first describe the methodology of MUSTER threading algorithm and then give an overview and advantage of using LOMETS, a meta-threading server.

11.2.1.1. MUSTER Threading Server. MUSTER is a sequence PPA method assisted by the predicted structural information like secondary structure, structure profiles, solvent accessibility, backbone dihedral torsion angles, and hydrophobic scoring matrix. The scoring function of MUSTER [29] for aligning the *i*th residue of the query and the *j*th residue of the template is defined as

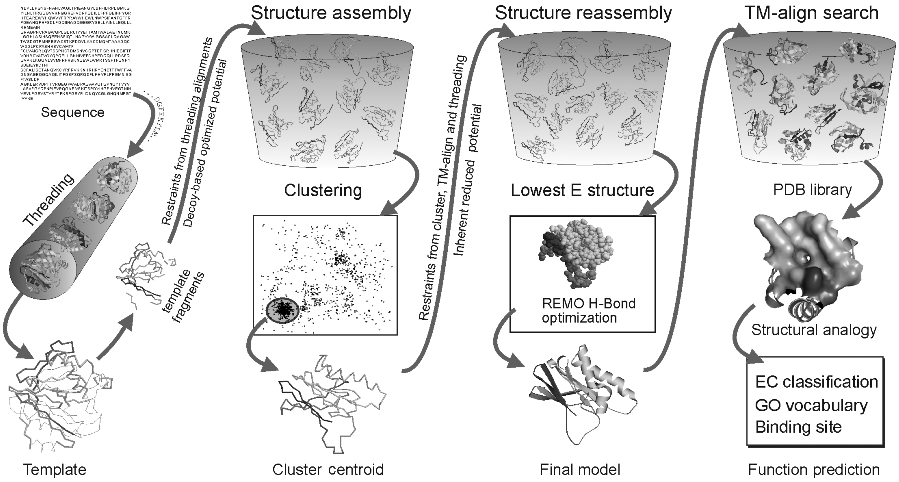

FIGURE 11.1 A schematic diagram of the I-TASSER [35,44,45] protein structure and function prediction protocol. Templates for the query protein are first identified by MUSTER [29] or LOMETS [33], which provide template fragments and spatial restraints. Template fragments are then assembled by modified replica-exchange Monte-Carlo simulations. The conformations generated during the simulation are clustered using SPICKER [48], in order to identify the structure with the lowest free energy. As an iterative refinement strategy, the cluster centroids are then subjected to the second round of simulation for refining the global topology and removing clashes. The final all-atom models are generated by REMO through the optimization of hydrogen-bonding networks [49]. Finally, functional homologs (protein structures with an associated EC number or GO terms) of final models are identified by using a sequence-independent structural alignment tool of TM-align [50] by ranking the hits based on their TM-score [51], RMSD and sequence identity in the structurally aligned region, coverage of the structural alignment, and confidence score (C-score [45]) of the model. (See color insert.)

$$\mathrm{Score}(i,j) = \mathrm{E}_{\mathrm{seq\_prof}} + \mathrm{E}_{\mathrm{sec}} + \mathrm{E}_{\mathrm{struc\_prof}} + \mathrm{E}_{\mathrm{sa}} + \mathrm{E}_{\mathrm{phi}} + \mathrm{E}_{\mathrm{psi}} + \mathrm{E}_{\mathrm{hydro}} + \mathrm{E}_{\mathrm{shift}}. \tag{1}$$

The first term, $\mathrm{E}_{\mathrm{seq\_prof}}$, is the alignment score of the sequence PPA. The second term, $\mathrm{E}_{\mathrm{sec}}$, computes the match between the predicted secondary structure of query and secondary structure of templates. The third term, $\mathrm{E}_{\mathrm{struc\_prof}}$, calculates the score of aligning the structure-derived profiles of templates to the sequence profile of query. The fourth term, E_{sa}, computes the difference between the predicted solvent accessibility of query and solvent accessibility of templates. The fifth and sixth terms ($\mathrm{E}_{\mathrm{phi}}$ and $\mathrm{E}_{\mathrm{psi}}$) calculate the difference between the predicted torsion angles (phi and psi) of query and those of templates. The experimental torsion angles for templates are calculated using STRIDE [52], while torsion angles of query are predicted by ANGLOR [53]. The seventh term, $\mathrm{E}_{\mathrm{hydro}}$, is an element of hydrophobic scoring matrix [54] that encourages the match of hydrophobic residue (V, I, L, F, Y, W, M) in the

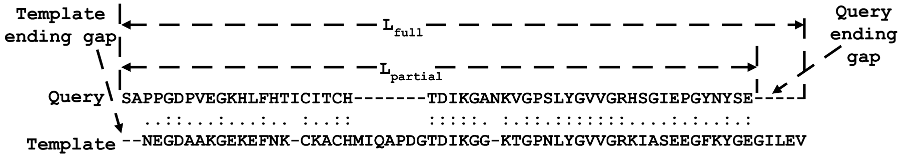
Template ending gap
L_full
Query ending gap
L_partial
Query SAPPGDPVEGKHLFHTICITCH-------TDIKGANKVGPSLYGVVGRHSGIEPGYNYSE----
..::...:...:.. :..::   :::::. :.::.::::::....:.:..:.:
Template --NEGDAAKGEKEFNK-CKACHMIQAPDGTDIKGG-KTGPNLYGVVGRKIASEEGFKYGEGILEV

FIGURE 11.2 Illustration of the full (L_{full}) and partial (L_{partial}) alignment lengths used to normalize the threading alignment score (R_{score}). Symbols "-", "." and ":" indicate an unaligned gap, an aligned nonidentical residue pair and an aligned identical residue pair, respectively. The query and template sequences are taken from 1hroA (first 53 residues) and 155c_ (first 61 residues), respectively, as an illustrative example. (From Wu and Zhang [29]).

query and the templates. Finally the last term, E_{shift}, is a constant, which is introduced to avoid alignment of unrelated residues in local regions. While the first term is sequence-based information, the second to seventh terms are related to structural information. If only the first two terms plus E_{shift} in Equation 1 are involved, the corresponding threading program is called PPA [33], which is the precursor of MUSTER.

The sequence and structural information are then combined into a single-body energy term, which can be conveniently used in the Needleman-Wunsch [55] dynamic programming algorithm for identifying the best match between the query and the templates. A position-dependent gap penalty in the dynamic programming is employed, i.e. no gap is allowed inside the secondary structure regions (helices and strands); gap opening (g_o) and gap extension (g_e) penalties apply to other regions; ending gap-penalty is neglected.

Following the dynamic programming alignments, the alignments on different structural templates are ranked based on their alignment score and the length of the alignment. In PPA [33] the templates are ranked based on a raw alignment score (R_{score}) divided by the full alignment length (L_{full}; including query and template ending gaps) as shown in Figure 11.2. In MUSTER, however, $R_{\text{score}}/L_{\text{partial}}$ is used as another possible ranking scheme, where L_{partial} is the partial alignment length excluding query ending gap as shown in Figure 11.2. A combined ranking is then taken as follows: If the sequence identity of the first template selected by $R_{\text{score}}/L_{\text{partial}}$ to the query is higher than that selected by $R_{\text{score}}/L_{\text{full}}$, then the template ranking is done by $R_{\text{score}}/L_{\text{partial}}$. Otherwise, the templates are ranked by $R_{\text{score}}/L_{\text{full}}$.

MUSTER was applied to a benchmark test of 500 non-homologous proteins (Fig. 11.3) and compared with PPA [29] at two different cutoffs: (i) all homologous templates with sequence identity >30% to the query were removed (cutoff 1); (ii) all homologous templates with sequence identity >20% or detectable by PSI-BLAST with an e-value <0.05 were removed (cutoff 2).

Here, the comparison between threading alignments and the native structures was done by evaluating the template-modeling score (TM-score), defined by Zhang and Skolnick [51], to assess the topological similarity of protein structure pairs with a value in the range of [0, 1]. Statistically, a TM-score <0.17 means a randomly selected protein pair with the gapless alignment taken from PDB; TM-score >0.5 corresponds to the protein pairs of similar folds. The statistical meaning of TM-score is independent of protein size [51].

At cutoff 1, the average TM-scores of the best threading alignment generated by PPA and MUSTER are 0.4285 and 0.4503, respectively, which shows that the additional structural information in MUSTER has improved the threading results by approximately 5%. Even at a more stringent cutoff (cutoff 2), MUSTER finds a better threading alignment with an average TM-score of 0.3638, while best threading alignment found by PPA has an average TM-score of 0.3423. Thus, the average TM-score of the first threading alignment by MUSTER at cutoff 2 is about 6% better than that of PPA alignment. The higher TM-score of MUSTER over PPA alignments is due to both the higher alignment coverage and the more accurate alignments as judged by the root-mean-square deviation (RMSD) within the aligned regions.

11.2.1.2. LOMETS: Meta-Threading Server. As shown in Figure 11.3, although MUSTER is better than PPA on average, it cannot outperform PPA on all protein targets. A similar trend has also been observed in CASP/Critical

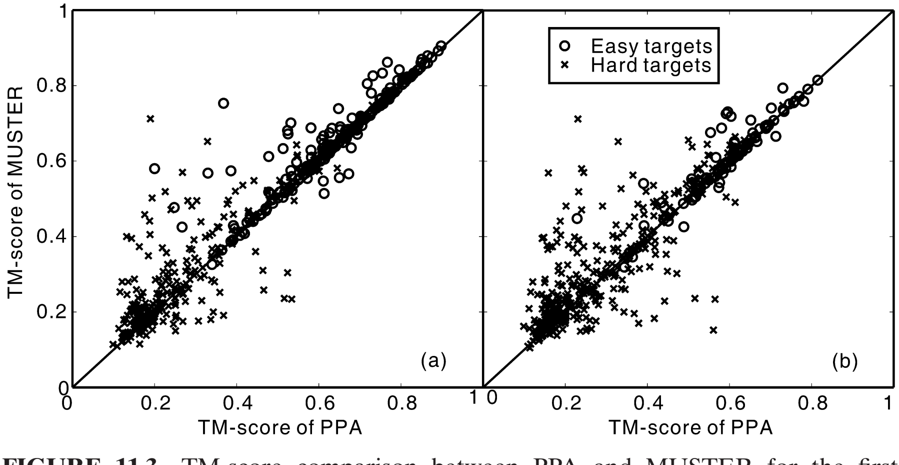

FIGURE 11.3 TM-score comparison between PPA and MUSTER for the first threading alignment of 500 non-homologous proteins. Circles represent the alignments from the "Easy" targets (z-score of alignment by MUSTER >7.5 and z-score of PPA alignment >7.0) and crosses indicate those from the "Hard" targets (z-score of alignment by MUSTER <7.5 and z-score of PPA <7.0). All homologous templates with sequence identity to targets (a) >30% (b) >20%, or detectable by PSI-BLAST with an e-value < 0.05 are excluded in this comparison. After removing homologous templates, the first template alignments by MUSTER for (a) 224 proteins and (b) 137 proteins have a correct fold (TM-score >0.5). (From Wu and Zhang [29]).

Assessment of Fully Automated Structure Prediction (CAFASP) experiments [42,56], where although the average Global Distance Test (GDT) or TM-score of some methods outperform others, there is no single method that can outperform others on all the targets. This inconsistency naturally leads to the prevalence of the metaserver [33,57], which is designed to collect and combine prediction results from a set of individual threading programs.

On the I-TASSER web server (http://zhanglab.ccmb.med.umich.edu/I-TASSER) this idea has been implemented using LOMETS [33], a locally installed meta-threading server. The threading programs in LOMETS represent a diverse set of the state-of-the-art algorithms using different approaches, namely, sequence profile alignments (PPA-I [33], PPA-II [33], SPARKS2 [33], SP^3 [58]), structural profile alignments (FUGUE [59]), pairwise potentials [PROSPECT2] [31], PAINT [33]), and the HMM (HHSearch [5], SAM-T02 [60]).

For each target, LOMETS first threads the query sequence through the PDB library to identify template threading alignments by each threading program and then ranks them purely based on consensus. The idea behind the consensus approach is simple: there are more ways for a threading program to select a wrong template than that to select a right one. Therefore, the chance for multiple threading programs working collectively to make a common wrong selection is lower than the chance to make a common correct selection.

Table 11.1 shows the improvement of LOMETS over individual threading programs. For the purpose of eliminating the dependence on the alignment coverage, the full-length models have been built here by MODELLER [26], using the templates from each threading program. Based on 620 non-homologous testing proteins, the models generated by LOMETS threading

TABLE 11.1 Performance Comparison of Component Threading Programs and LOMETS Metaserver on 620 Non-Homologous Testing Proteins

Threading Servers or Metaservers	TM-score (MODELLER models)		RMSD (Å) (MODELLER models)	
	First Model	Best in Top Five Models	First Model	Best in Top Five Models
PPA-I	0.4117	0.4531	16.66	14.02
SP3	0.4138	0.4551	13.86	12.83
PPA-II	0.4076	0.4512	14.89	13.02
SPARKS2	0.3973	0.4441	13.60	12.23
PROSPECT2	0.3914	0.4384	13.01	12.02
FUGUE	0.3721	0.4173	19.26	15.82
HHSEARCH	0.3827	0.4224	22.38	19.04
PAINT	0.3758	0.4210	15.74	14.21
SAM-T02	0.3575	0.3971	21.75	17.53
LOMETS	0.4434	0.4669	10.99	10.61

alignments achieve an average TM-score of 0.4434, which is at least 8% higher than that by any individual threading program.

11.2.2. Structure Assembly and Refinement

Following the threading procedure, the next step of I-TASSER is to generate the full-length model of the query protein and to refine the structure so that threading-aligned regions move closer to native structure. To achieve this, the complete protein chain in I-TASSER is divided into threading-aligned and unaligned regions, where the continuous fragments are excised from threading alignments, while the threading unaligned regions are build by *ab initio* modeling. The protein chain here is described by a reduced model, that is, a trace of alpha-carbon atoms and side chain center (SC) of mass to reduce the number of explicitly treated freedom and the intra-molecular interactions in the polypeptide chain. We will elaborately describe the whole procedure now.

For a given threading alignment, I-TASSER builds an initial full-length model by connecting the continuous secondary structure fragments (≥five residues) through a random walk of C_α–C_α bond vectors of variable lengths from 3.26 Å to 4.35 Å. To guarantee that the last step of this random walk can quickly arrive at the first C_α of the next template fragment, the distance l between the current C_α and the first C_α of the next template fragment is checked at each step of the random walk, and only walks with $l < 3.54n$ are allowed, where n is the number of remaining C_α–C_α bonds in the walk. If the template gap is too big to be spanned by a specified number of unaligned residues, a big C_α–C_α bond is kept at the end of the random walk and a spring-like force that acts to draw sequential fragments close will be applied in subsequent Monte-Carlo simulations, until a physically reasonable bond length is achieved.

The initial full-length models are then refined by the parallel replica-exchange Monte-Carlo sampling technique [61]. Two kinds of conformational updates (off-lattice and on-lattice) are implemented here: (i) Off-lattice movements of the aligned regions involve rigid fragment translations and rotations that are controlled by the three Euler angles. The fragment length normalizes the movement amplitude so that the acceptance rate is approximately constant for fragments of different sizes. (ii) The lattice-confined residues are subjected to 2–6 bond movements and multi-bond sequence shifts. Overall, the tertiary topology varies by the rearrangement of the continuously aligned substructures, where the local conformation of the off-lattice substructures remains unchanged during the assembly.

The movements in the structure assembly and refinement procedure are guided by an optimized force field that is described in the next section.

11.2.2.1. Force Field. The inherent I-TASSER assembly force field is similar to TASSER [47], which includes a variety of knowledge-based energy terms describing the predicted secondary structure propensities from PSIPRED [62], secondary structure-specific backbone hydrogen bonding, and a variety of statistical short-range and long-range correlation terms that are extracted from multiple threading alignments. Readers are recommended to read Zhang and Skolnick [47,63,64] for further details about these energy terms.

The new potentials terms that have been incorporated in I-TASSER include the predicted accessible surface area (ASA) [35] and sequence-based contact predictions [65]. Both the energy terms have been derived and optimized using machine learning methods.

For the purpose of fast calculations of the ASA effect, the hydrophobic energy in I-TASSER is defined by

$$E_{ASA} = -\sum\left(\frac{x_i^2}{x_0^2}+\frac{y_i^2}{y_0^2}+\frac{z_i^2}{z_0^2}-2.5\right)\times P(i), \tag{2}$$

where (x_i, y_i, z_i) is the coordinate of ith residue at the ellipsoid Cartesian system of the given protein conformation and (x_0, y_0, z_0) is the principal axes length. The constant parameter used for tuning the average depth of the exposed residues is 2.5, while $P(i)$ is the residue exposure index and is defined as $P(i)=\sum_{j=1}^{12} a(j)$, where a_j is the two-state neural network (NN) prediction of exposure ($a_j = 1$) or burial ($a_j = -1$) with the jth ASA fraction cutoff; $P(i)$ has a strong correlation with the real value of ASA. The overall correlation coefficient between the predicted $P(i)$ and the actual exposed area as calculated by STRIDE [52] on a test set of 2234 non-homologous proteins is 0.71. The same correlation for the widely used Hopp-Woods [66] and Kyte-Doolittle hydrophobicity indices [67] are 0.42 and 0.39, respectively. One of the probable reasons for the higher correlation by the NN prediction is because it explores the sequence-profile information, whereas the later methods are sequence-independent.

In the latest version of I-TASSER, sequence-based pairwise residue contact information from SVM-SEQ [65], SVMCON [68], and BETACON [69] are used to constrain the simulation search to a smaller conformational space and improve the minima of the landscape funnel of the overall energy function. Wu and Zhang recently showed that this additional information from SVM-SEQ can significantly increase the contact prediction accuracy in hard targets (when no good template is identified) by about 12–25%, compared with SVM-LOMETS [65], a template-based contact prediction method.

The predicted contacts from SVM-SEQ, SVMCON, and BETACON include contacts for C_α, C_β, and SC at distance cutoffs of 6Å, 7Å, and 8Å. These predicted contacts are implemented as restraints in the I-TASSER simulation in the following way: if two residues i and j are predicted to be in contact by sequence-based methods and they come in contact in the decoys during the course of I-TASSER simulation, then the residue pairs are preferred to keep in contact by giving an energy bonus, which is defined as

$$E_{contact} = -1-(\mathrm{conf}(i,j)-a), \tag{3}$$

where conf(i, j) is the confidence score of the predicted contact pair (i, j) and a ($\in [0,1]$) is an empirically determined score cutoff for each distance cutoff.

11.2.2.2. Iterative Strategy. The trajectories of the low-temperature replicas of the first-round I-TASSER simulations are clustered by SPICKER [48]. The cluster centroids are obtained by averaging all the clustered structures after superposition and are ranked based on the structure density of the clusters. However, the cluster centroids generally have a number of nonphysical steric clashes between C_α atoms and can be overcompressed. Starting from the selected SPICKER cluster centroids, the TASSER Monte-Carlo simulation [61] is performed again (see Fig. 11.1). While the inherent I-TASSER potential remains unchanged in the second run, external constraints are added, which are derived by pooling the initial high-confident restraints from threading alignments, the distance and contact restraints from the combination of the centroid structures, and the PDB structures identified by the structure alignment program TM-align [50] using the cluster centroids as query structures. The conformation with the lowest energy in the second round is selected as the final model.

The main purpose of this iterative strategy is to remove the steric clashes of the cluster centroids. On a benchmark test set of 200 proteins with <300 residues it was found that the average number of steric clashes (residue pairs with C_α distance <3.6 Å) for the cluster centroids of the first cluster dramatically reduces from 79 to 0.8. As strong distance map and contact restraints are implemented in this step, the topology of the models also improves. In these test cases, the average TM-score increased from 0.5734 to 0.5801 (1.2%) and the C_α-RMSD to native decreased from 6.67 Å to 6.52 Å compared with the cluster centroid of the first round.

11.2.3. Reconstruction of Atomic Model

The models generated after I-TASSER Monte-Carlo simulations [61] and SPICKER clustering [48] are reduced models, where each residue is represented by the C_α atom and the SC of mass. To increase the biological usefulness of protein models, all atom models are generated by REMO [49] simulations, which include three general steps: (i) removing steric clashes by moving around each of the C_α atoms that clash with other residues; (ii) backbone reconstruction by scanning a backbone isomer library collected from the solved high-resolution structures in the PDB library; and (iii) hydrogen-bonding network optimization based on predicted secondary structure from PSIPRED [62]. Finally, Scwrl3.0 [70] is used to add the side chain rotamers.

Figure 11.4 shows the performance of REMO [49] on 230 non-homologous test proteins. Figure 11.4a shows the number of the steric clashes (residue pairs with C_α distance <3.6 Å) in the cluster centroids of the test proteins after the first round of I-TASSER simulations (average clash = 119). After the REMO procedure (Fig. 11.4b), only 15 proteins had 2–6 clashes, 44 proteins had one clash, and in the remaining all clashes had been effectively removed.

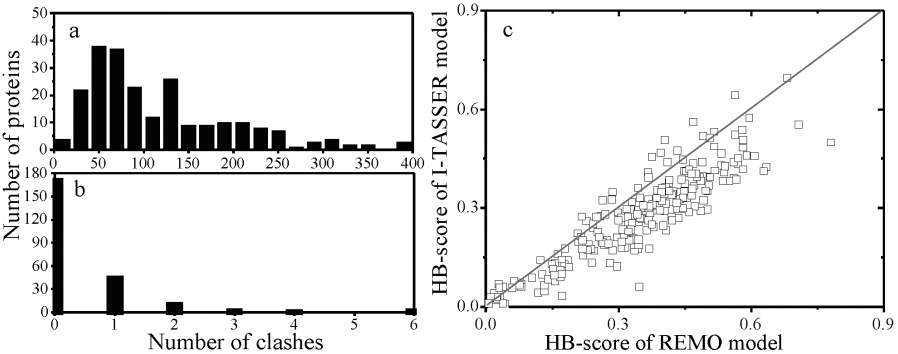

FIGURE 11.4 Histogram of steric clashes in (a) cluster centroids and (b) REMO models of 230 test proteins. (c) Comparison of HB-score of REMO models and I-TASSER models (From Li and Zhang [49]).

Remarkably, although the steric clashes had been completely removed, the clash-removing procedure had no side effect on the global topology of the initial structures. The RMSD change in most cases was <0.9 Å and the average RMSD to the native slightly improved.

Figure 11.4c shows the improvement in the hydrogen bonding (HB) score in REMO models over the I-TASSER models. Here, I-TASSER models refer to the models that had been generated by PULCHRA [71] for adding backbone atoms (N, C, O) and Scwrl3.0 [70] to build side chain atoms. HB-score is defined as the fraction of the common hydrogen bonds between model and the native structure. As shown in the figure, the HB-score of REMO models have dramatically improved in more than 80% (187/230) of the test proteins.

REMO was also used in blind CASP8 experiment for refining the reduced models generated by I-TASSER Monte-Carlo simulations. Based on the 172 released targets/domains, the average TM-score and GDT-score of the I-TASSER (as "Zhang Server") models are higher with a significant margin than that of other groups in the server section (see http://zhanglab.ccmb.med.umich.edu/casp8). In particular, the average HB-score of the I-TASSER, which partially reflects the quality of local structures, is also higher than all other groups, except SAM-08-server [6,72], while in CASP7 the HB-score of the I-TASSER models were much lower than most of other groups [41,42]. These data demonstrate a significant progress in reconstructing and refining atomic models using the REMO protocol. REMO simulations are now a part of I-TASSER methodology for generating atomic level model. The source code and online server of REMO is freely available at http://zhanglab.ccmb.med.umich.edu/REMO.

11.2.4. Function Prediction

One of the main impetuses for predicting the structure is to use it for structure-based functional annotation. To identify the functional homologs of a query protein, the generated models are structurally aligned by TM-align [50] with all known structures in the PDB library that have known functions. The resultant structural alignment is scored based on a Fh-score (Functional homology score), which is defined as [73]:

$$\text{Fh-score} = \text{nC-score} * \left(\text{TM-score} + \tfrac{1}{1+\text{RMSD}_{\text{ali}}} * \text{Cov}\right) + 3 * \text{ID}_{\text{ali}} * \text{Cov}, \qquad (4)$$

where nC-score is the normalized C-score and is defined as $\text{nC} - \text{score} = \dfrac{\text{C} - \text{score} + 5}{7}$, which stays in [0, 1] and estimates the quality of I-TASSER protein structure predictions; TM-score [51] measures the global structural similarity between the model and the template proteins; RMSD_{ali} is the RMSD of query model and template structure in the structurally aligned region; Cov represents the coverage of the structural alignment; and ID_{ali} is the sequence identity between query and template based on the alignment by TM-align. For every query protein, predicted functions include both the predicted enzyme commission (EC) number [74] and the Gene Ontology (GO) molecular function [75] terms. While EC number is a commonly used scheme for functional classification of enzymes, GO terms provide a consistent description of function for both enzymatic and nonenzymatic proteins. Accordingly, two independent protein libraries of about 5800 nonredundant enzymatic proteins (pairwise sequence identity <90%) and about 13,500 nonredundant proteins (pairwise sequence identity <90%) with known GO terms have been constructed and are biweekly updated.

Based on a large-scale benchmark test set of 317 non-homologous proteins, it was found that by using the predicted structures (modeled while excluding all the homologous proteins with sequence identity >30% to query protein), the first three digits of EC number and 50% of associated GO terms of query protein could be correctly identified in more than 55% of the test cases from the best identified functional homologs based on Fh-score. Moreover, the true and false positive predictions could be discriminated well and achieved an area of more than 0.80 under the receiver operating characteristic (ROC) curve for both the predictions. For the 196 enzymatic proteins that had another functional homolog (enzymes with same first three EC digits) in the library and having less than 30% sequence identity, Fh-score and PSI-BLAST were able to identify functional homologs with same first three digits of EC number for 107 and 77 proteins, respectively. These data show that the structure-based functional annotations using the I-TASSER models can be about 39% more accurate than the sequence-based approaches (such as PSI-BLAST) [73].

11.3. *AB INITIO* PREDICTION OF I-TASSER ON SMALL PROTEINS

To explore the ability of I-TASSER to fold proteins for which no homologous templates are detected in the PDB, I-TASSER was tested on three sets of non-

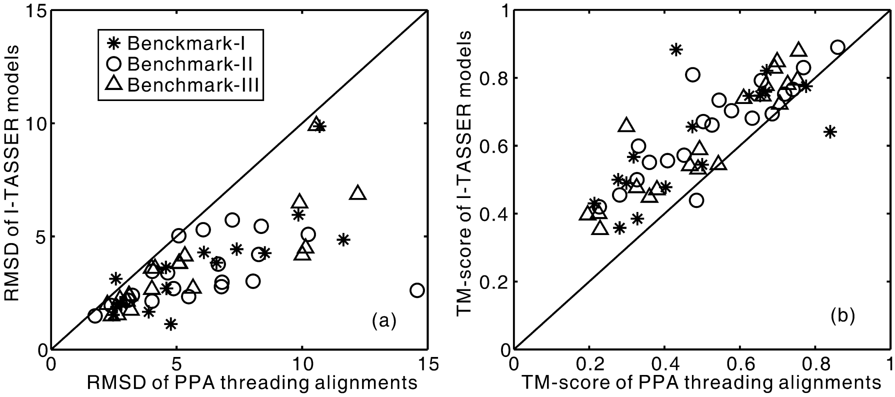

FIGURE 11.5 Comparison of I-TASSER models with the PPA threading alignment results. (a) C_α-RMSD to native of the I-TASSER models versus C_α-RMSD to native of the best threading alignment over the same aligned regions. (b) TM-score of the I-TASSER models versus TM-score of the best threading alignments. (From Wu, Skolnick, and Zhang [35]).

homologous proteins. The test proteins include: (i) Benchmark-I consisting of 16 proteins (<90 residues) that were used by Bradley et al. for testing ROSETTA [76]; (ii) Benchmark-II consisting of 20 proteins (<120 residues) that were used by Zhang et al for testing TOUCHSTONE II [64]; and (iii) Benchmark-III consisting of 20 proteins (<120 residues) selected from PDB [35].

The I-TASSER structure assembly started from PPA threading where all template proteins with a sequence identity >20% to the query or detectable by PSI-BLAST with an e-value <0.05 were excluded. Figure 11.5 shows the comparison of the best of the top five I-TASSER models with the initial PPA threading alignments in all three benchmark test sets. As seen in the figure, the global topology of the final models was significantly closer to the native structure than the threading alignments. In Benchmark-I, I-TASSER models have an average C_α-RMSD of 3.8 Å, with six of them having a high-resolution structure with the C_α-RMSD <2.5 Å. On the second set, (Benchmark-II), I-TASSER could fold four of them with a C_α-RMSD <2.5 Å. The average C_α-RMSD of the I-TASSER models in this set of test proteins was 3.9 Å. Average C_α-RMSD of 3.9 Å was obtained for the third benchmark set, with seven cases having a C_α-RMSD <2.5 Å. Overall, the first predicted models had an average C_α-RMSD ranging from 4.3 Å to 4.8 Å and the average TM-score ranged from 0.59 to 0.64 for the three benchmarks. For the best models in the top five predictions, the average C_α-RMSD ranged from 3.8 Å to 3.9 Å and the average TM-score ranged from 0.61 to 0.65.

The first set of proteins was also used for testing ROSETTA [76] and the best of the top five models by ROSETTA had an average RMSD of 3.8 Å; thus, the overall results between the two methods (ROSETTA and I-TASSER) are comparable, but the central processing unit (CPU) time required by

I-TASSER was much shorter (150 CPU days vs. 5 CPU hours). For the second test, the average RMSD by TOUCHSTONE-II [64] is 5.9 Å. These data, together with the significant performance of automated I-TASSER server (the Zhang Server) in the FM section of the CASP experiment [40], demonstrate a new progress in automated *ab initio* model generation.

11.4. BLIND TEST OF I-TASSER IN CASP EXPERIMENTS

CASP [46,77] is a biennial world-wide protein structure prediction experiment, where the organizers release a number of protein sequences for which structure is unknown. The participants are then asked by the organizers to predict the structures of these proteins and submit their predicted models before deadlines. Finally, the experts evaluate the predicted models by comparing them with the structures solved by the X-ray or nuclear magnetic resonance (NMR) experiments.

The seventh CASP experiment was held in 2006, where the performance of I-TASSER was tested in both the human (as "Zhang") and the server section (as "Zhang Server"). The procedure in the server and human predictions are essentially the same and follow the general I-TASSER methodology, except for that the human prediction involved domain border assignment based on visual inspection and used the server predictions from other groups for hard targets; meanwhile, the I-TASSER assembly simulations were done for a longer time in the human prediction.

Ninety-six proteins in CASP7 were split into 124 domains by the assessors. Depending on modeling difficulty (whether or not a good template is present in PDB), these domains can be categorized for simplicity into 105 template-based modeling (TBM) targets and 19 free modeling (FM) targets. Figure 11.6 shows a comparison of the first I-TASSER models and the best threading templates for all these targets in both server and human predictions. Here, the best template refers to the template of the highest TM-score to the native structure among all the templates exploited by I-TASSER. As shown in the figure, although there is a general tendency for better templates resulting in better models, in most cases I-TASSER was able to consistently improve the final model over the templates based on both RMSD and TM-score.

In the TBM category, I-TASSER reassembly resulted in a TM-score increase by ~14%, where about 10% is probably because of the topology reorientation of the secondary structure fragments and the rest may be due to the increase in the size of models when gaps are filled during the reassembly procedure. One of the major reasons for this improvement is because I-TASSER employs consensus spatial constraints from multiple templates that are usually of higher accuracy than that from individual templates. The second driving force for the structure improvement was the optimization of I-TASSER inherent potential from a collection of statistical terms from different resources [35,47,63,64].

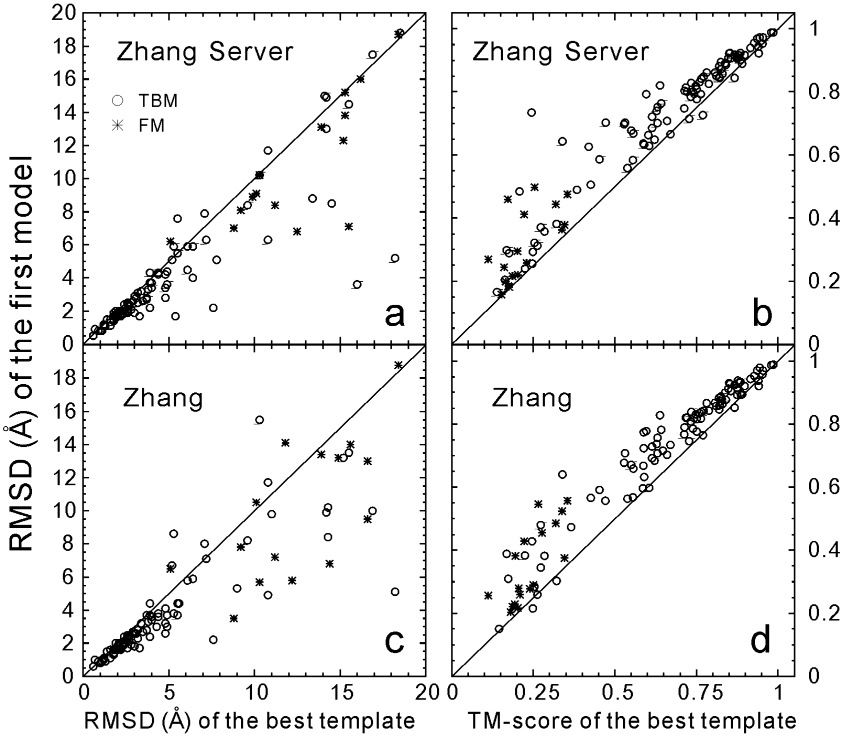

FIGURE 11.6 Comparison of the first predicted models by I-TASSER in human ("Zhang") and server ("Zhang-Server") sections of CASP7 with respect to the best exploited templates. The RMSD is calculated in the same set of aligned residues. The TM-score is calculated in the aligned regions for the templates and in full-length for the models (From Zhang, Y. Proteins 69 (2007): 112).

For the FM targets, I-TASSER was able to fold (RMSD < 6.5 Å or TM-score > 0.5) seven targets (about 1/3) that were up to 155 residues long. Figure 11.7 shows a more detailed analysis of a typical example (target T0382) of the I-TASSER predictions during CASP7. T0382 was a new fold protein (PDB ID: 2I9C) from *Rhodopseudomonas palustris* CGA009 crystallized by the structure genomics project. The topology of T0382 consists of five joggled α-helices. The left panel of Figure 11.7 shows the top five templates hit by the multiple threading programs used by I-TASSER, all having correct local second structure elements but incorrect global topologies with the best RMSD of 9.3 Å from 1xm9A1 (TM-score = 0.28). Contact prediction program generated 148 side chain contacts with 37 correct contacts (accuracy = 25%). The average error of the best predicted C_α distances is 2.2 Å. I-TASSER cuts the fragments from the template alignments and reassembles the topology under the guide of the predicted restraints and the inherent potential, which result in a model of full-length RMSD 3.6 Å and TM-score 0.66 (right panel of Figure

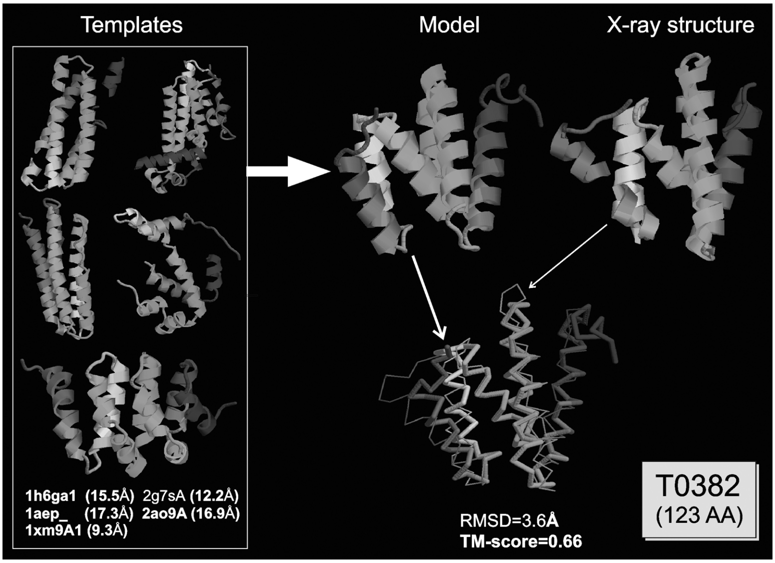

FIGURE 11.7 Structure comparisons of the threading templates, the final I-TASSER model, and the experimental structures for the CASP7 target T0382. Blue to red runs from N- to C-terminals (From Zhang [44]). (See color insert.)

11.7). The correlation of I-TASSER energy and the RMSD of the structure decoys is 0.72, which demonstrates the consistency of the external restraints and the inherent force field.

11.5. CONCLUDING REMARKS

The protein structure prediction problem can be solved in two ways. The first one is to fold all proteins by computationally recovering the nature's protein folding pathway. This task does not appear to be accomplished in foreseeable future, unless a detailed physicochemical description of the intra-protein and protein-solvent interactions are developed, not to mention the delicate interactions of proteins with the associated ligands and chaperones that will dramatically complicate the situation. The second solution is more of an engineering-oriented rather than scientific, in which a selected set of proteins are solved by experiments so that all proteins with unknown structure have at least one neighboring protein with known structure, which can be used as a template in CM; this has been the goal of the SG projects [20]. On the basis of about 40,000 structures in the PDB library (many are redundant) [18], it is estimated that 4million models/fold assignments can be obtained by a simple combination of the PSI-BLAST search and the CM technique [78]. Development of more sophisticated and automated computer modeling approaches will dramatically enlarge the scope of modelable proteins in the SG projects.

Despite intense efforts and considerable progress in the field [79], the accuracy of protein structure prediction is still largely dictated by the evolutionary distance between the target and the solved proteins in the PDB library. Robust methods that can model proteins that have no or weak structure homologous templates are lacking. Nevertheless, the most efficient approaches to model both homologous and non-homologous proteins are those that combine different algorithms of threading, fragment assembly, *ab initio* modeling, and structural refinements. I-TASSER is one of the successful examples of these composite approaches. The exploitation of multiple threading templates and the optimization of the composite knowledge-based energy terms constitute the two major factors contributing to the success of I-TASSER in refining individual template structures closer to the native.

However, since I-TASSER has a resolution limitation set by its inherent reduced potential, high-resolution models cannot be predicted for most of proteins when a good template is not present. One of the ongoing efforts in this regard is to extend the reduced I-TASSER modeling to the atomic representation with the goal of improving the modeling accuracy at the atomic-level [44]. REMO represents part of our recent efforts to refine atomic models by optimizing the hydrogen-bonding networks. The development of new physics-based force fields in combination with the current I-TASSER knowledge-based potentials as well as the development of the function prediction methodology will be of significant importance in increasing the accuracy and the applicability of these approaches to genome-wide structure and function predictions.

REFERENCES

1. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Research*, 37:D169–174, 2009.
2. S.R. Wiley. Genomics in the real world. *Current Pharmaceutical Design*, 4(5):417–422, 1998.
3. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
4. S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
5. J. Soding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–960, 2005.
6. K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.
7. C. Chothia and A.M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5(4):823–826, 1986.

8. T.C. Wood and W.R. Pearson. Evolution of protein sequences and structures. *Journal of Molecular Biology*, 291(4):977–995, 1999.
9. C.A. Wilson, J. Kreychman, and M. Gerstein. Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology*, 297(1):233–249, 2000.
10. D. Pal and D. Eisenberg. Inference of protein function from protein structure. *Structure*, 13(1):121–130, 2005.
11. W.R. Pearson. Effective protein sequence comparison. *Methods Enzymology*, 266:227–258, 1996.
12. W. Tian and J. Skolnick. How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology*, 333(4):863–882, 2003.
13. B. Rost. Enzyme function less conserved than anticipated. *Journal of Molecular Biology*, 318(2):595–608, 2002.
14. D. Eisenberg, E.M. Marcotte, I. Xenarios, and T.O. Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, 2000.
15. G. Lopez, A. Rojas, M. Tress, and A. Valencia. Assessment of predictions submitted for the CASP7 function prediction category. *Proteins*, 69(Suppl 8):165–174, 2007.
16. G.J. Kleywegt. Recognition of spatial motifs in protein structures. *Journal of Molecular Biology*, 285(4):1887–1897, 1999.
17. A.C. Wallace, R.A. Laskowski, and J.M. Thornton. Derivation of 3D coordinate templates for searching structural databases: Application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Science*, 5(6):1001–1013, 1996.
18. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.
19. M. Gerstein, A. Edwards, C.H. Arrowsmith, and G.T. Montelione. Structural genomics: Current progress. *Science*, 299(5613):1663, 2003.
20. J.M. Chandonia and S.E. Brenner. The impact of structural genomics: Expectations and outcomes. *Science*, 311(5759):347–351, 2006.
21. D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, 2001.
22. J. Skolnick, J.S. Fetrow, and A. Kolinski. Structural genomics and its importance for gene function analysis. *Nature Biotechnology*, 18(3):283–287, 2000.
23. P. Aloy, E. Querol, F.X. Aviles, and M.J. Sternberg. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *Journal of Molecular Biology*, 311(2):395–408, 2001.
24. D. Vitkup, E. Melamud, J. Moult, and C. Sander. Completeness in structural genomics. *Nature Structural & Molecular Biology*, 8(6):559–566, 2001.
25. Y. Zhang. Protein structure prediction: when is it useful? *Current Opinion in Structural Biology*, 19(2):145–155, 2009.
26. A. Sali and T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815. 1993.

27. A. Fiser, R.K.G. Do, and A. Sali. Modeling of loops in protein structures. *Protein Science*, 9(9):1753–1773, 2000.

28. J.U. Bowie, R. Luthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, 1991.

29. S. Wu and Y. Zhang. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, 72(2):547–556, 2008.

30. D.T. Jones, W.R. Taylor, and J.M. Thornton. A New Approach to Protein Fold Recognition. *Nature*, 358(6381):86–89, 1992.

31. Y. Xu and D. Xu, Protein threading using PROSPECT: Design and evaluation. *Proteins*, 40(3):343–354, 2000.

32. J. Skolnick, D. Kihara, and Y. Zhang. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins*, 56(3):502–518, 2004.

33. S. Wu and Y. Zhang. LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, 35(10):3375–3382. 2007.

34. P. Bradley, K.M.S. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005.

35. S. Wu, J. Skolnick, and Y. Zhang. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology*, 5:17, 2007.

36. A. Liwo, J. Lee, D.R. Ripoll, J. Pillardy, and H.A. Scheraga. Protein structure prediction by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences U S A*, 96(10):5482–5485, 1999.

37. K.T. Simons, C. Strauss, and D. Baker. Prospects for ab initio protein structural genomics. *Journal of Molecular Biology*, 306(5):1191–1199, 2001.

38. D. Kihara, H. Lu, A. Kolinski, and J. Skolnick. TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proceedings of the National Academy of Sciences U S A*, 98(18):10125–10130, 2001.

39. C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(96):223–230, 1973.

40. R. Jauch, H.C. Yeo, P.R. Kolatkar, and N.D. Clarke. Assessment of CASP7 structure predictions for template free targets. *Proteins*, 69(S8):57–67, 2007.

41. J. Kopp, L. Bordoli, J.N. Battey, F. Kiefer, and T. Schwede. Assessment of CASP7 predictions for template-based modeling targets. *Proteins*, 69(S8):38–56, 2007.

42. J.N. Battey, J. Kopp, L. Bordoli, R.J. Read, N.D. Clarke, and T. Schwede, Automated server predictions in CASP7. *Proteins*, 69(S8):68–82, 2007.

43. R. Das, B. Qian, S. Raman, R. Vernon, J. Thompson, P. Bradley, S. Khare, M.D. Tyka, D. Bhat, D. Chivian, D.E. Kim, W.H. Sheffler, L. Malmstrom, A.M. Wollacott, C. Wang, I. Andre, and D. Baker. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins*, 69(S8):118–128, 2007.

44. Y. Zhang. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins*, 69(8):108–117, 2007.

45. Y. Zhang. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9:40, 2008.

46. J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, and A. Tramontano. Critical assessment of methods of protein structure prediction-Round VII. *Proteins*, 69(8):3–9, 2007.

47. Y. Zhang and J. Skolnick. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences U S A*, 101(20):7594–7599, 2004.

48. Y. Zhang and J. Skolnick. SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*, 25(6):865–871, 2004.

49. Y. Li and Y. Zhang. REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins*, 76(3):665–676, 2009.

50. Y. Zhang and J. Skolnick. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005.

51. Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710, 2004.

52. D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–579, 1995.

53. S. Wu and Y. Zhang. ANGLOR: A composite machine-learning algorithm for protein backbone torsion angle prediction. *PLoS ONE*, 3(10):e3400, 2008.

54. P.J. Silva. Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis. *Proteins*, 70(4):1588–1594, 2008.

55. S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

56. D. Fischer, L. Rychlewski, R.L. Jr. Dunbrack, A.R. Ortiz, and A. Elofsson. CAFASP3: The third critical assessment of fully automated structure prediction methods. *Proteins*, 53 (6):503–516, 2003.

57. D. Fischer. 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor. *Proteins*, 51(3):434–441, 2003.

58. H. Zhou and Y. Zhou. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*, 58(2):321–328, 2005.

59. J. Shi, T.L. Blundell, and K. Mizuguchi. FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology*, 310(1):243–257, 2001.

60. K. Karplus, R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans, and R. Hughey. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, 53 (6):491–496, 2003.

61. Y. Zhang, D. Kihara, and J. Skolnick. Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins*, 48(2):192–201, 2002.

62. L.J. McGuffin, K. Bryson, and D.T. Jones. The PSIPRED protein structure prediction server. *Bioinformatics*, 16(4):404–405, 2000.

63. Y. Zhang and J. Skolnick. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophysical Journal*, 87:2647–2655, 2004.

64. Y. Zhang, A. Kolinski, and J. Skolnick. TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophysical Journal*, 85(2):1145–1164, 2003.

65. S. Wu and Y. Zhang. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, 24(7):924–931, 2008.

66. T.P. Hopp and K.R. Woods. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences U S A*, 78(6):3824–3828, 1981.

67. J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

68. J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, 8:113, 2007.

69. J. Cheng and P. Baldi. Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms. *Bioinformatics*, 21 (1):i75–i84, 2005.

70. A.A. Canutescu, A.A. Shelenkov, and R.L. Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 12(9):2001–2014, 2003.

71. M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, and C.L. Brooks. 3rd. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *Proteins*, 41(1):86–97, 2000.

72. K. Karplus, S. Katzman, G. Shackleford, M. Koeva, J. Draper, B. Barnes, M. Soriano, and R. Hughey. SAM-T04: What is new in protein-structure prediction for CASP6. *Proteins*, 61 (7):135–142, 2005.

73. A. Roy, A. Kucukural, S. Mukherjee, P.S. Hefty, and Y. Zhang. Large scale benchmarking of protein function prediction using modeled protein structures. *Journal of Molecular Biology*, 2010, submitted.

74. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme supplement 5 (1999). *European Journal of Biochemistry*, 264(2):610–650, 1999.

75. 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, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene ontology: Tool for the unification of biology. *The Gene Ontology Consortium. Nature Genetics*, 25(1):25–29, 2000.

76. P. Bradley, K.M. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005.

77. J.N. Battey, J. Kopp, L. Bordoli, R.J. Read, N.D. Clarke, and T. Schwede Automated server predictions in CASP7. *Proteins*, 69 (8):68–82, 2007.

78. U. Pieper, N. Eswar, F.P. Davis, H. Braberg, M.S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B.M. Webb, D. Eramian, M.Y. Shen, L. Kelly, F. Melo, and A. Sali. MODBASE: A database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, 34(Database issue):D291–D295, 2006.

79. Y. Zhang. Progress and challenges in protein structure prediction. *Current Opinion Structural Biology*, 18(3):342–348, 2008.