PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade

Jianyi Yang,[1,2] Wenxuan Zhang,[1,2] Baoji He,[1,2] Sara Elizabeth Walker,[1,2] Hongjiu Zhang,[1,2] Brandon Govindarajoo,[1,2] Jouko Virtanen,[1,2] Zhidong Xue,[1,2] Hong-Bin Shen,[1,2] and Yang Zhang[1,2]*

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109

[2] Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT

We report the structure prediction results of a new composite pipeline for template-based modeling (TBM) in the 11th CASP experiment. Starting from multiple structure templates identified by LOMETS based meta-threading programs, the QUARK *ab initio* folding program is extended to generate initial full-length models under strong constraints from template alignments. The final atomic models are then constructed by I-TASSER based fragment reassembly simulations, followed by the fragment-guided molecular dynamic simulation and the MQAP-based model selection. It was found that the inclusion of QUARK-TBM simulations as an intermediate modeling step could help improve the quality of the I-TASSER models for both Easy and Hard TBM targets. Overall, the average TM-score of the first I-TASSER model is 12% higher than that of the best LOMETS templates, with the RMSD in the same threading-aligned regions reduced from 5.8 to 4.7 Å. Nevertheless, there are nearly 18% of TBM domains with the templates deteriorated by the structure assembly pipeline, which may be attributed to the errors of secondary structure and domain orientation predictions that propagate through and degrade the procedures of template identification and final model selections. To examine the record of progress, we made a retrospective report of the I-TASSER pipeline in the last five CASP experiments (CASP7-11). The data show no clear progress of the LOMETS threading programs over PSI-BLAST; but obvious progress on structural improvement relative to threading templates was witnessed in recent CASP experiments, which is probably attributed to the integration of the extended *ab initio* folding simulation with the threading assembly pipeline and the introduction of atomic-level structure refinements following the reduced modeling simulations.

Proteins 2016; 84(Suppl 1):233–246.

Key words: protein structure prediction; CASP11; threading; I-TASSER; QUARK.

INTRODUCTION

The first template-based protein structure prediction can be traced back to 1969 when Browne and colleagues tried to build structural model of the bovine alpha-lactalbumin using the solved hen egg-white lysozyme structure as template.[1] The power of template-based modeling (TBM) have since then been significantly extended, which can be attributed to several factors. First, the invention of PSI-BLAST[2] and the consequent profile-to-profile alignment techniques[3–5] has significantly increased the accuracy of template identification and alignment, compared to the original single-sequence based or manual alignment approaches. Second, composite structure assembly simulations combine multiple templates identified by meta-server threading alignments,[6,7] which can drive individual templates considerably closer to the native structures.[8–12] Finally, the rapid accumulation of experimental sequence and structure databases

Grant sponsor: National Institute of General Medical Sciences; Grant numbers: R01GM083107 and R01GM084222.
*Correspondence to: Yang Zhang; Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109 or Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109. E-mail: zhng@umich.edu
Received 27 May 2015; Revised 13 August 2015; Accepted 31 August 2015
Published online 7 September 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24918

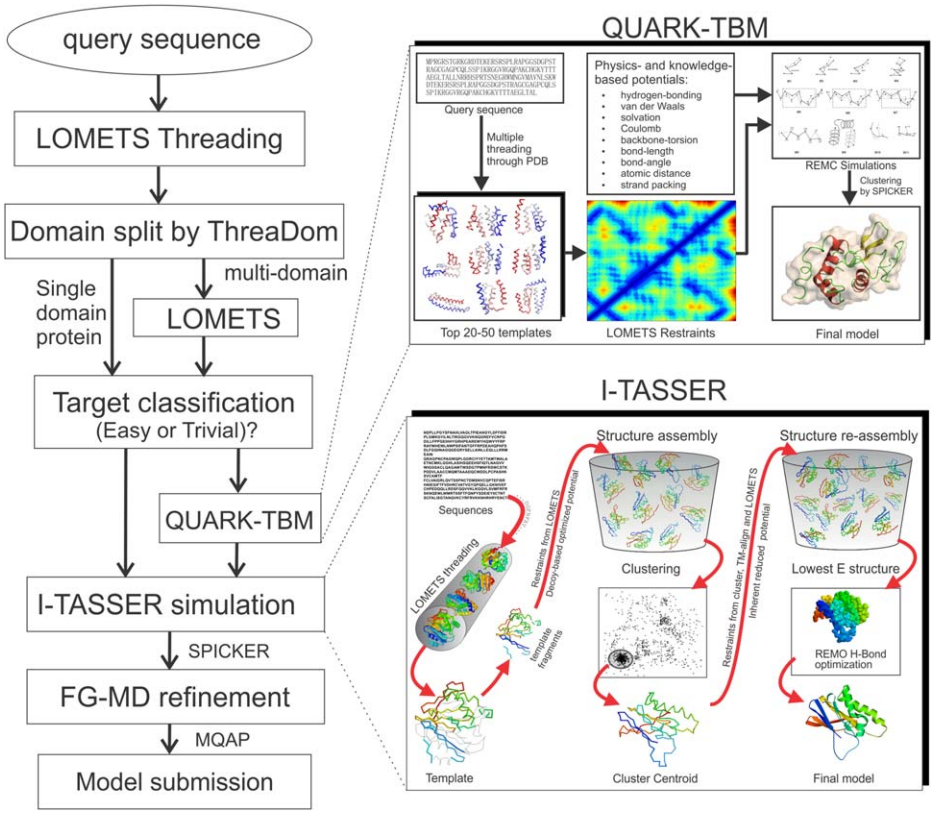

Figure 1

The flowchart of the template based modeling (TBM) by "Zhang-Server" in CASP11. Models by "Zhang" human group were generated similarly except that models from other groups in the Server Section were exploited in addition to the LOMETS templates.

converted many non- or distant-homology targets to homology ones by providing close homology templates.

Despite the progress, significant challenges still exist in distant-homology template detection and atomic-level structure refinement for TBM. To partially address these challenges, we developed a new pipeline specifically designed for composite template structure prediction that was tested in the TBM section of the 11th Critical Assessment of protein Structure Prediction (CASP) experiment. The major component of the pipeline is based on iterative threading reassemble refinement (I-TASSER),[9,13] where the *ab initio* folding method, QUARK,[14] was extended as an intermediate step for TBM structure refinement. The results showed promise for improving TBM accuracy by the integration of the extended *ab initio* folding process.

Three pipelines ("Zhang-Sever," "QUARK," and "Zhang") were tested in CASP11, our report will be mainly focused on the first model generated by "Zhang-Server," which implemented an automated pipeline of composite I-TASSER and QUARK-TBM as depicted in Figure 1. The "Zhang" group is a human group using exactly the same pipeline as "Zhang-Server" but with input including server predictions from other groups, where "QUARK" is an automated server predictor based on QUARK-TBM pipeline. Following the suggestion of the CASP organizers, at the end of the study we present a retrospective comparative study on the Zhang-Server models from the last five CASP experiments, which provides a unique opportunity to track possible progress (if any) of the same pipeline developed over the last decade.[15,16]

METHODS

The pipeline that we used for TBM in CASP11 is depicted in Figure 1, which can be generally divided into four steps of threading and domain parsing, template-based QUARK modeling, I-TASSER assembly simulation, and model selection. One of the major differences from the standard I-TASSER server is that QUARK-TBM is integrated into the current pipeline for multiple-step template structure reassembly.

Threading and threading-based domain structure determination

The query sequence of the target protein is first threaded through the PDB library by LOMETS,[7] a meta-threading approach containing multiple individual

threading programs to identify possible template structures as well as super-secondary structure segments. The domain structure of the query sequence is then determined by ThreaDom,[17] which was developed based on the distribution of a domain conservation score that counts for the gap/insertion distribution of LOMETS alignments and the domain boundaries of the PDB templates. The gap penalty score is measured by the number of gaps in multiple alignments of the template sequences and the domain boundaries of the template structures is defined based on the definition in CATH,[18] both of which are combined linearly with equal weight in the domain conservation score.[17] If ThreaDom deems the target to be a multidomain protein, LOMETS is used again to generate threading alignments for each domain. Based on the normalized Z-score and the degree of consensus of LOMETS alignments, the domains are categorized into four classes ["Trivial," "Easy," "Hard," and "Very-Hard," see Eq. (1) of Ref. 19]. A two-step simulation process, including QUARK-TBM and I-TASSER, is performed if the target is deemed as a "Trivial" or "Easy" target (Fig. 1).

QUARK-TBM

QUARK was originally developed for *ab initio* structure prediction by assembling the continuously distributed fragments excised from un-related PDB structures.[14] Here we extended it to template-based modeling, called QUARK-TBM, which is built on the same force field and Monte Carlo search engine. Rather than starting from random conformations, however, QUARK-TBM starts from the top threading templates identified by LOMETS. Meanwhile, spatial restraints collected from the template alignments, including Cα distance-map and side-chain contacts, are integrated with the generic QUARK potential (including hydrogen-bonding, van der Waals, solvation, Coulomb, backbone-torsion, bond-length and bond-angle, atomic distance, and strand pairing) to guide the Monte Carlo folding simulations.[14] Considering the extensive time request on large proteins, QUARK-TBM was used only on the domains with length below 300 residues. These models will be used as the input of the next step of I-TASSER simulations (Fig. 1).

I-TASSER

I-TASSER[9,20] was designed to construct protein structural models by reassembling continuous fragments excised from the top LOMETS threading alignments. In addition to spatial restraints from threading templates, here I-TASSER also has restraints taken from the full-length models generated from QUARK-TBM. Because the QUARK-TBM models are full-length, the residue in the middle of each loop is deleted so that the I-TASSER program can recognize and reassemble the secondary structure segments in the simulations. Since the QUARK-TBM simulations have been strongly constrained to the templates, the QUARK-TBM models are often closer to the templates than the I-TASSER simulations. We found in our benchmark tests that the inclusion of QUARK-TBM in the LOMETS templates as starting conformations can often improve the quality of local structural packing, which is particularly helpful when the quality of final models is measured by the GDT-HA score.

Here, although both simulations use homologous templates as restraints, the major difference between I-TASSER and QUARK-TBM lies at the structural representation and the force field employed. The I-TASSER simulations are built on a reduced C-alpha and side-chain of mass model, and the force field is a purely knowledge-based potential including multiple terms derived from the regularities of the PDB structures.[9,20,21] On the contrast, QUARK-TBM models contain atomic detail of backbone (Cα, C, O, N) plus Cβ and the side-chain center of mass.[14] The more detailed conformational representation in QUARK-TBM allows the consideration of more physics-based potentials, which includes van der Waals, Coulomb, backbone-oriented hydrogen bonding and solvation interactions as outlined in the last section, in addition to the knowledge-based components.[14,22] Meanwhile, QUARK-TMB has a stronger weight for the external spatial restraints than I-TASSER, which results in the final models with a closer similarity to the templates. These differences help generate models of complementary structural features when integrating the programs into a unified pipeline as depicted in Figure 1.

Decoys clustering, model selection and side-chain atom refinement

Following the QUARK-TBM and I-TASSER simulations, we cluster the structure decoys using the SPICKER program.[23] Fragment-guided molecular dynamics simulations (FG-MD)[24] is used to refine the SPICKER models at the atomic level. Finally, multiple Model Quality Assessment Programs (MQAPs) are used to select the best models for submission. The MQAP programs contain three classes of scores: (1) a structure consensus score that is defined as the average TM-score[25] of the target model to all other candidate models; (2) statistical potentials derived from the PDB structures (DOPE,[26] GOAP,[27] and RWplus[28]); (3) I-TASSER or QUARK-TBM confidence score (or C-score) that is calculated based on the product of the significance score of LOMETS alignments and the structure density of the SPICKER clusters.[29]

The decoy models are sorted by each of the MQAP scoring functions. The final rank score of each decoy model equals to the sum of the ranks from all the MQAP programs. The models with the lowest rank score are finally selected for submission to CASP. Apparently, models selected through this procedure should be reasonably favored by all the MQAPs since a low rank from one MQAP program can dramatically increase the overall

rank score and therefore remove the target model from selection.

If a target is determined by ThreaDom to consist of multiple domains, the full-length model is constructed by docking the models of individual domains using the full-length I-TASSER model as the template, where FG-MD simulation is used to remove possible clashes created during the domain docking simulation. The entire pipeline as shown in Figure 1 is fully automated without human intervention.

Here, we note that both I-TASSER and QUARK simulations are based on reduced models with the side-chain represented by a single point at the center of mass (but with backbone represented differently as outline above). Therefore, no atomic-level side-chain optimization is implemented in the structure assembly simulation process in the current pipeline. In I-TASSER simulations, the side-chain center position of the ith residue is computed on a local Cartesian system built on the three adjacent Cα atoms $(i-1,\ i,\ i+1)$ with the bond-length and angle parameters derived from high-resolution PDB structures that are specified by the secondary structure type (helix, coil and strand) and amino acid identity.[20] In QUARK, the side-chain position is calculated in a similar manner but with the parameters specified with 20 different amino acids and backbone torsion-angle pairs $(\varphi,\ \psi)$ that are divided into 72 bins from 6023 training PDB structures.[14]

Atomic-level side-chain refinements in our pipeline are performed after the I-TASSER and QUARK simulations, that is, the atomic details including backbone and side-chain atoms are first added by REMO[30] to the C-alpha traces which are then refined by the fragment-guided molecular dynamic simulations.[24] Since the molecular dynamic simulation in FG-MD is short (~30 CPU minutes), changes in backbone conformation at this step are modest. Therefore, one of the limitations of the current pipeline is that the side-chain conformations cannot be sufficiently optimized, because the full-atomic side-chain optimization often requires adjustments of backbone conformations that cannot be completed by the current short-term FG-MD simulations. One strategy that is on-going to address the issue is to incorporate the atomic-level side-chain rotamer optimization[31] into the I-TASSER and QUARK assembly simulations. Considering that the inclusion of full-atom details dramatically increases the simulation time, this can be implemented in the later stage of the Monte Carlo simulations, which should help optimize the side-chain rotamer conformations and their interactions with the backbone structures.

RESULTS AND DISCUSSION

Overall results

There are 82 domains from 68 protein entries, which were assigned by the assessors as TBM targets in the final assessment. These 82 domains contain 10 targets that have distant-homology templates in the PDB but are assumed to be difficult to detect by the assessors; these domains are named as "TBM-hard." The rest of 72 domains are referred as "TBM-easy" throughout this manuscript.

Improvement of final models over threading templates

One of the major goals of template-based protein structure prediction is to refine the initial templates and draw the structure closer to the native. In Figure 2, we present a head-to-head comparison between the first submitted models in Zhang-Server versus the best templates from LOMETS that were used by I-TASSER and QUARK-TBM. Figure 2(A) shows the RMSDs of templates and final models. Because models are full-length while template alignments usually contain gaps and insertions, we calculate the RMSD of models only on the regions that are aligned in the templates. For 67 of the 82 TBM domains (82%), the RMSD of the final models is lower than that of the best templates, indicating that the I-TASSER/QUARK-TBM simulations have drawn the templates closer to the native. Such improvement occurs on both TBM-easy and TBM-hard domains, showing that the ability to improve the protein structures does not depend on the type of protein target. The average RMSD reduction is 1.1 Å (5.8 vs. 4.7 Å for template and model, respectively). A summary of the numerical data of template vs. model comparison is also presented in Table I.

Since the RMSD values are often more sensitive to the local error than to the correctness of the global fold,[25] we present in Figure 2(B) the TM-score of final model versus the best templates. Again, the majority of the targets appear in the upper triangle of the plot, meaning that the final models have a higher TM-score than the best templates. For several targets, the final models have the TM-score increased significantly compared to the templates, including T0828-D1 (ΔTM-score=0.467), T0828-D2 (0.455), T0773-D1 (0.323), and T0827-D1 (0.318). As expected, most of the TBM-hard targets are distributed in the low TM-score range, where in 9 out of 10 cases the TM-score of initial templates was increased by the structural reassembly process. Overall, the average TM-score of the final models increases by 12% compared to the templates for all 82 TBM domains, or by 25% for the 10 TBM-hard domains.

For the targets that are categorized into TBM-hard targets, there may still be close templates existing in the PDB but not successfully detected by LOMETS. If we use TM-align[32] to match the target structure through the PDB library, we found that 8 out of the 10 cases have a template with a TM-score above 0.5, indicating correct fold[33]; the templates of the remaining two (T0814-D3 and T0848-D2) are approximately correct

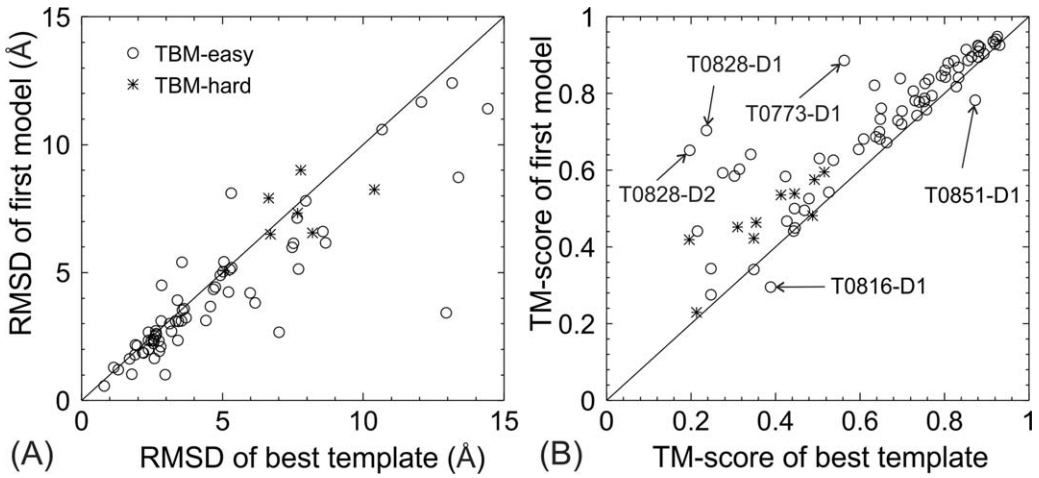

Figure 2

Comparison of the first Zhang-Server models and the best threading templates used. Stars and circles indicate TBM-easy and TBM-hard domains, respectively. (**A**) RMSD of the first model versus RMSD of the best templates in the threading aligned regions. (**B**) TM-score of the first model versus TM-score of the best templates.

with TM-score = 0.455 and 0.496, respectively. The average TM-score for the 10 TBM-hard targets is 0.585 that is 55% higher than the best LOMETS templates (0.378). Quite surprisingly, there are 4 targets (T0799-D3, T0814-D3, T0831-D1, and T0848-D2) whose best templates have a sequence identity >30% to the target sequences but none of the last three targets (T0814-D3, T0831-D1 and T0848-D2) have a LOMETS template with TM-score >0.4. A comparison between the sequence and structure alignments on the best templates showed that the target sequence was aligned to the completely different regions of the templates by the two alignments for these targets. This suggests that one reason for the failure might be that the current LOMETS programs have put a too strong weight on the evolutionary scores and appropriate structure-based scoring function are needed for improving the fold-recognition for these hard targets.

Atomic-level refinement by FG-MD

As outlined in Method, both I-TASSER and QUARK simulations are built on the reduced models; the atomic details are quickly added by REMO[30] after the simulations, which are then refined by FG-MD.[24] To have a quantitative assessment on the effect of FG-MD, we list in Table II a comparison of the models before and after running FG-MD on the 58 single-domain targets (– the multi-domain targets are not counted here because multiple REMO and FG-MD simulations were conducted on different stages of single-domain structure refinement and complex model assembly).

The upper rows of Table II show the parameters measuring the similarity of the models relative to the experimental structure. There is no significant difference between the REMO and FG-MD models in terms of RMSD, TM-score and GDT-HA, showing that the refinement on backbone structures is marginal. But the hydrogen-bonding networks (reflected by HB-score) and the side-chain orientation (by GDC-SC[34]) are obviously improved with a P values in student's t test below 10^{-12} and 10^{-2}, respectively.

The lower rows of Table II show the physical quality of the models assessed by MolProbity score.[35] Again, there is a significant improvement on the MolProbity score by FG-MD with a P values below 10^{-11}, where the major contributions of the improvements are from the reduction of atomic overlaps and the number of bond-length and bond-angle outliers.

Model selection by MQAP programs

The I-TASSER pipeline generally generated around 500 to 1500 structure decoys, depending on the difficulty of

Table I

Summary of the First Predicted Models Compared to Templates by LOMETS

Target type	*N* (Nm/Nt)	TMb_t/Rb_t	TM1_m/R1_m
TBM-easy	72 (60/12)	0.662/5.1 Å	0.735/4.0 Å
TBM-hard	10 (7/3)	0.378/10.4 Å	0.472/9.6 Å
TBM-all	82 (67/15)	0.627/5.8 Å	0.703/4.7 Å

N: Number of targets.
Nm: Number of targets for which the first model has a RMSD lower than the best starting template in the threading aligned region.
Nt: Number of targets for which the first model with a RMSD higher than the best starting templates in the threading aligned region.
TMb_t: Average TM-score of the best template.
Rb_t: Average RMSD of the best template.
TM1_m: Average TM-score of the first submitted model.
R1_m: Average RMSD of the first submitted model.

Table II
Comparison of Models Before and After Running FG-MD

	Parameters	Model before FG-MD	Model after FG-MD	*P* values
Comparison to the native structure	RMSD (Å)	7.132	7.103	0.10
	TM-score	0.664	0.665	0.40
	GDT-HA	0.430	0.431	0.08
	GDC-SC	0.243	0.249	0.02
	HB-score	0.443	0.519	1.01e-12
Physical quality assessment	MolProbity score	3.280	2.819	8.39e-11
	# Clashes	122.5	8.2	5.55e-17
	# Cβ_out	6.9	7.1	0.27
	# Rotamer_out	5.2	5.5	0.30
	# Ramachandran_out	16.3	14.6	0.12
	#Bond-length_out	10.7	0.2	3.50e-05
	#Bond-angle_out	25.6	2.0	7.01e-06

Data are from the first model submitted by Zhang-Server on the 58 single-domain proteins.

the targets, that is, more decoys were created for the hard than the easy targets (see http://zhanglab.ccmb.med.umich.edu/decoys/casp11/). In Table III, we listed the results of the first model selected by different MQAP programs. Again, for simplicity we only count the 58 single-domain protein targets, because the final model of the multi-domain targets involved two levels of single-domain and complex structure selections and the same model submitted for multi-domain targets contains selection results from different MQAP programs, inclusion of which will compromise the clearness of data comparison.

If we consider individual MQAP programs, the average performance of GOAP outperforms others in all scores of TM, GDT-HA and MolProbity. But a combination of the statistical potentials with the consensus and confidence scores makes the model selection more robust than the individual MQAP programs. Not surprisingly, however, the selected models are all far worse than the best decoys, highlighting significant rooms for further MQAP improvement.

The values in the parenthesis in Table III indicate the number of times when the MQAP rank-1 model is nearly the best decoy. Since the absolute best model has never been selected, a model is here defined as "nearly the best" if the model is within the top 10 best decoys and has the quality score (TM-score, GDT-HA and MolProbity) not worse than 95% of the score of the best decoy in each category. The number of cases with the near-best decoy selection is relatively low for all the MQAP selections. But this number by the combined MQAP scores is slightly higher than that by the individual MQAP programs, indicating again an enhanced robustness of the model selection through score combination.

Table III
Summary of MQAP Model Selections on the 58 Single-Domain Proteins

Mqap Programs	TM-score	GDT-HA	MolProbity
Best decoy	0.7183 (58)	0.4816 (58)	2.1643 (58)
MQAP_combined	0.6803 (6)	0.4439 (6)	2.8091 (12)
GOAP	0.6734 (4)	0.4389 (1)	2.8323 (4)
Consensus	0.6683 (3)	0.4356 (3)	3.0840 (1)
RWplus	0.6662 (3)	0.4293 (1)	2.8622 (7)
C-score	0.6629 (2)	0.4256 (1)	3.1011 (1)
DOPE	0.6544 (3)	0.4194 (2)	2.9386 (4)

Case studies reveal impact of secondary structure and domain orientation on final models

Despite the significant improvement of threading templates, there are 15 out of the 82 TBM domains (18%) where the I-TASSER/QUARK-TBM simulations deteriorated the templates by increasing the RMSD in the aligned regions. Even if we consider TM-score, which gives a slight favor to the final models, since they have a longer length, there are still 7 domains whose final models became worse, that is, having a lower TM-score than the templates. The most significant deterioration occurs for the targets of T0816-D1 (TM-score reduced by 0.093) and T0851-D1 (TM-score reduced by 0.09) [Fig. 2(B)].

T0816-D1

T0816-D1 is a small single-domain protein (68 residues) with a 4-helix bundle fold [Fig. 3(E)]. The X-ray structure shows four short helices from W4-I15, I18-L30, N39-M49, and L53-E67. However, the secondary structure prediction, which is from a combination of PSIPRED[36] and PSSpred,[9] resulted in only three helices, where there is only one residue break (instead of two in the X-ray structure) between the first and second helices [Fig. 3(A)]. As a result, the majority of the LOMETS templates have a two-helix bundle topology, with the best template from 1fewA that has a TM-score = 0.389 [Fig. 3(B)].

The I-TASSER simulations are dominated by the two-helix bundle topology, due to the population in the threading templates. The first model selected by SPICKER[23] thus has a two-helix bundle fold with a TM-score = 0.296 [Fig. 3(C)]. Nevertheless, there are 18% of the I-TASSER decoys that possess the correct fold of

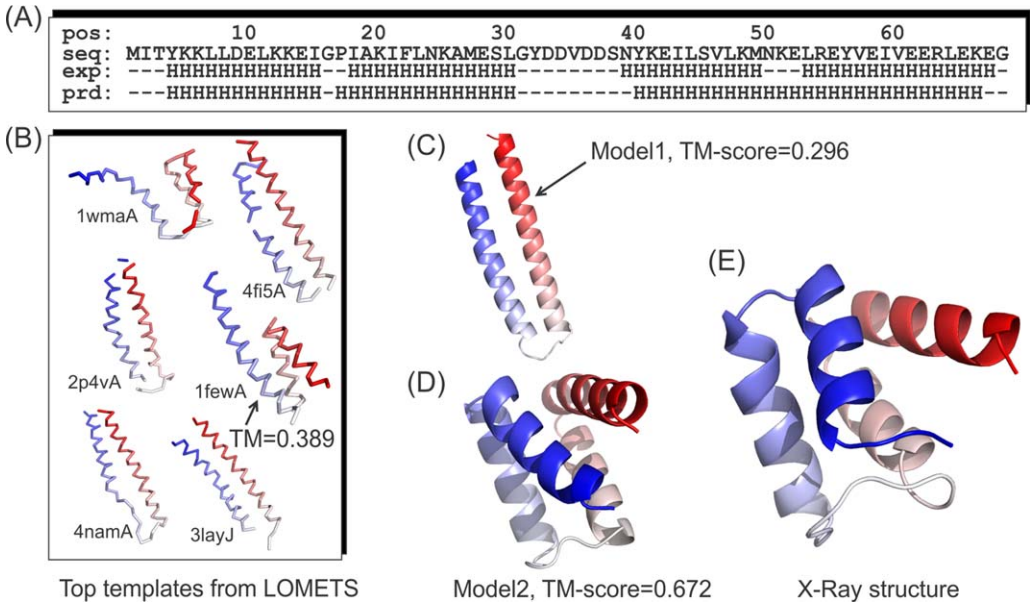

Figure 3

Case study on T0816-D1 where the first model has a TM-score lower than the fourth template. (**A**) Secondary structure prediction; (**B**) top six templates identified by LOMETS; (**C**) the first model submitted by Zhang-Server; (**D**) the second model submitted by Zhang-Server; (**E**) the X-ray structure of T0816-D1.

4-helix bundle; this resulted in the second submitted model that has a high TM-score = 0.672 [Fig. 3(D)]. This case represents a typical example where the error in secondary structure prediction propagates through threading and leads to incorrect model selection of the final structure prediction.

In Table IV, we showed the MQAP ranks on the two models for T0816-D1, which demonstrated an opposite tendency from consensus and statistical scores. Apparently, the first model of two-helix bundle is favored by the consensus score (rank = 2), where the second model is ranked as the 230th because there are much fewer modeling decoys with such topology. Interestingly, all the statistical potentials rank favorably on the second model, that is, with the model ranked as No. 1 by RWplus, No. 2 by GOAP and DOPE, where the first model has an unfavorable statistical rank (211,191, and 156). But the total MQAP rank score of the second model is still worse than the first model due to the unfavorable C-score and consensus scores. Here scores listed in Table IV have been normalized by the average score of all decoys.

Table IV
MQAP Model Selection on T0816-D1 and T0851-D1

Target	Scores	Model1	Model2
T0816-D1	TM-score	0.296	0.672
	Consensus (rank)	1.172 (2)	0.990 (230)
	RWplus (rank)	−0.985 (211)	−1.142 (1)
	GOAP (rank)	−1.064 (191)	−1.379 (2)
	DOPE (rank)	−1.017 (156)	−1.153 (2)
T0851-D1	TM-score	0.783	0.928
	Consensus (rank)	0.964 (246)	0.961 (258)
	RWplus (rank)	−1.025 (1)	−1.015 (28)
	GOAP (rank)	−1.088 (3)	−1.068 (19)
	DOPE (rank)	−1.032 (1)	−1.023 (19)

In both cases the second model has a higher TM-score than the first but failed to be selected by the total MQAP rank score. All MQAP scores have been normalized by the average of all decoys.

T0851-D1

T0851-D1 is a two-domain target of 456 residues, for which the first model, built on 3g79A, has a TM-score (0.783) that is considerably lower than the best template from 2y0cA (0.873). A closer look at the first model shows that the error was mainly due to the twist of domain orientation of the second domain, where the superposition of the second domain on the native results in a TM-score=0.893 [Fig. 4(A–C)].

The second submitted model was built based on the template that has a correct domain orientation; this results in a TM-score=0.928, slightly higher than that from 2y0cA [Fig. 4(D,E)]. Both 3g79A and 2y0cA consist of two dinucleotide binding Rossmann-like folds in the N- and C-terminal domains, where the subtle difference in the intermediate linker domain results in a small twist on the domain orientation.[37] However, the I-TASSER potential was unable to distinguish the domain orientations where nearly equal numbers of the structure decoys have been generated by the I-TASSER simulations, for the different domain orientations, as indicated by the

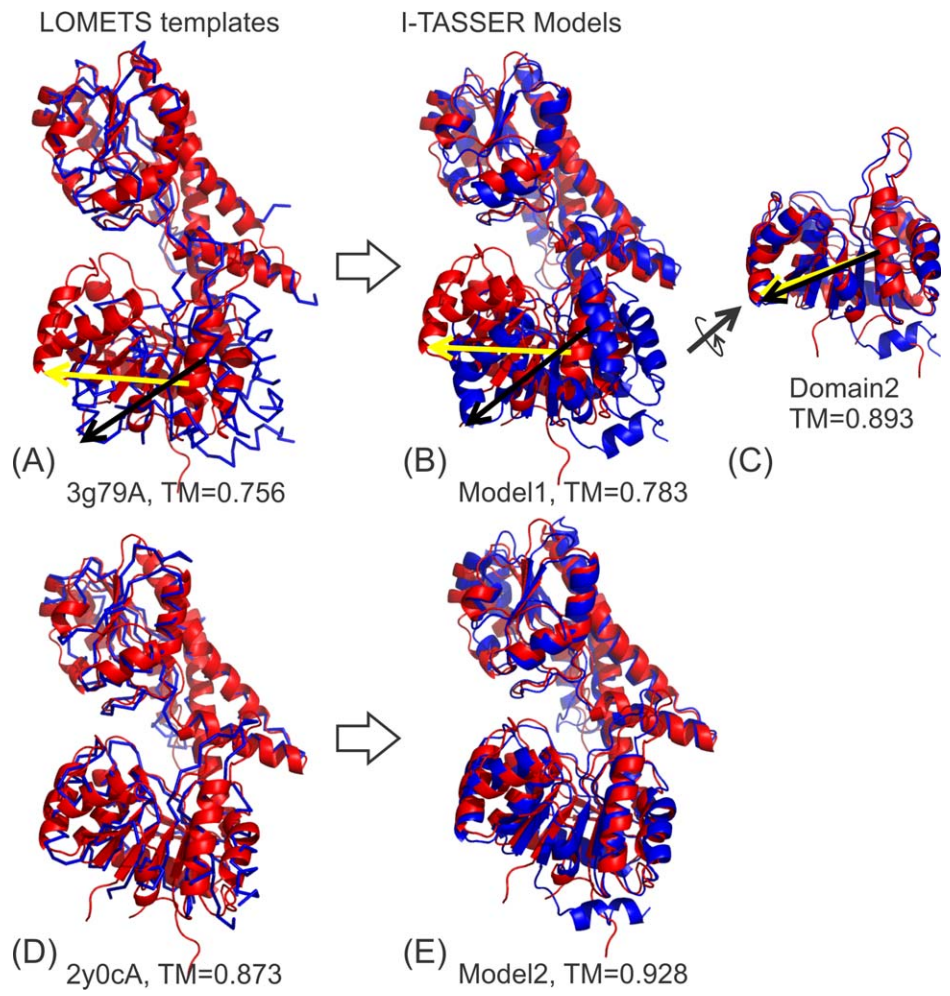

Figure 4

Structure prediction on T0851-D1. Red and blue represent X-ray structure and predicted models, respectively. Yellow and black arrows mark the domain orientations of X-ray and model structures, respectively. (**A**) Superposition of the X-ray structure and the LOMETS template (PDB ID: 3g39A) that is the closest to the first I-TASSER model; (**B**) Superposition of the first submitted model and the X-ray structure; (**C**) superposition of the second domain of the first submitted model and the X-ray structure; (**D**) superposition of the X-ray structure and the best LOMETS template (PDB ID: 2y0cA) that is the closest to the second submitted model; (**E**) superposition of X-ray structure and the second submitted model.

similar rank of consensus score shown in Table IV. The statistical potentials also failed to pick up the correct domains and the rank of the first model (with a wrong orientation) is better than the second model (with a correct orientation) in all the RWplus, GOAP and DOPE ranks (Table IV). This result indicates the modeling of domain orientation remains an open problem in the current structure prediction pipelines; this is particularly true when the domain orientation of the template structure is different from the target.

T0830-D1

In addition to the above cases where the TM-score of final models is lower than the template, there are also cases in which the RMSD of the models in the aligned region is much worse than the best template. T0830-D1 is such example that has the RMSD of the first model much higher than the template (8.2 vs. 5.3 Å) although the TM-scores of the model and template are comparable (0.482 vs. 0.485).

The T0830-D1 is the transmembrane domain of the UDP transferase protein, where PSSpred/PSIpred generated correct secondary structure predictions with the Q3 accuracy=86.3%. Most LOMETS programs detected the correct template (PDB ID: 3wajA) that has a topology similar to the target. However, the structure of 3wajA has two additional helices inserted in W276-V333 compared to the target structure (Fig. 5), where the majority of the threading programs mistakenly aligned these two helices on the target sequence (– only two threading programs generated correct alignments with the two helices

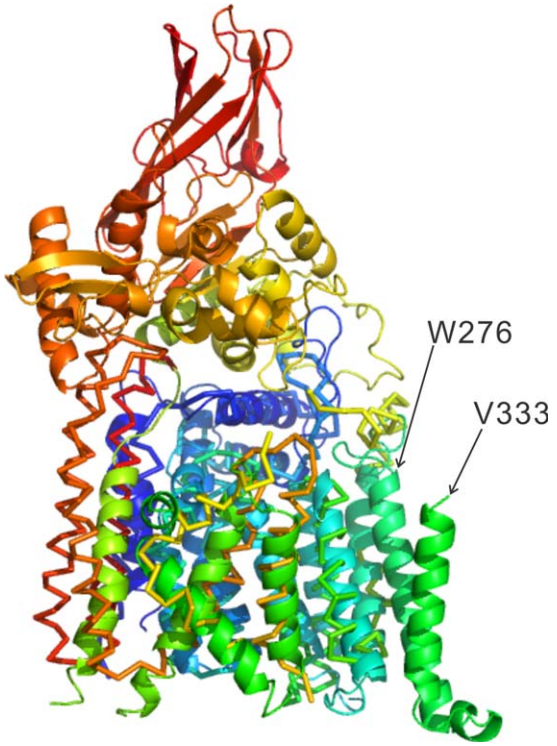

Figure 5

Structural superposition of the target structure of T0830-D1 (sticks) on the best template (PDB ID: 1wajA, cartoons) that is created by TM-align. Blue to red runs from N- to C-terminus of the structures. There is an insert of two helices (W276-V333) on the template structure that are missed on the target. But most threading programs failed to skip these two helices when aligning the target sequence onto the template structure.

skipped). Thus, the first I-TASSER model has the two helices incorrectly arranged due to the consensus but incorrect alignments. The second model by I-TASSER based on the correct alignments has a much better quality (TM-score = 0.683, RMSD = 4.2 Å in the aligned region). This example represents a typical case of failures in the I-TASSER model and template selections when the best template alignment is the minority; this case was witnessed and extensively discussed in the previous CASP reports.[19,38,39]

Comparison of QUARK-TBM and I-TASSER refinement

The QUARK-TBM simulations were used as an intermediate step of the Zhang-Server pipeline, which are conducted only on the proteins that have a size below 300 residues. To examine the effect of QUARK-TBM on the final models, we present in Table V the modeling results from three different pipelines (QUARK-TBM, I-TASSER, and I-TASSER+QUARK, the model finally submitted by the Zhang-Server group) on the 63 TBM domains that have fewer than 300 residues.

First, the data showed that all three pipelines have a significantly better modeling quality than the initial threading templates, in terms of TM-score, GDT-TS and GDT-HA scores. Second, I-TASSER+QUARK outperforms I-TASSER in TM-score and GDT-TS, which demonstrates a positive impact of QUARK-TBM on the final I-TASSER model predictions. Although the QUARK-TBM models on their own have a lower TM-score/GDT-TS than the I-TASSER models (P values = 0.011/0.016 in Student's t-test), such contribution to the improvement of the final models is statically significant as shown by the p-values between the I-TASSER+QUARK and I-TASSER models that is below 0.05 for both TM-score and GDT-TS.

A closer examination on the data shows that there are 43 cases whose TM-score of the I-TASSER+QUARK model is higher than that by I-TASSER, while the other 20 cases have the TM-score of the I-TASSER+QUARK models lower than the I-TASSER models. The TM-score improvement of I-TASSER+QUARK over I-TASSER models is smaller than 0.1 in almost all the cases, except for T0828-D1 (a 84-residue beta-barrel domain) and T0828-D2 (a 84-residue alpha-helix bundle domain) that have the TM-score increased by 0.14 and 0.22 respectively. In both of these two cases, the QUARK-TBM models alone have the TM-score 0.09 and 0.12 higher than the I-TASSER models, where a refinement further enhanced the model quality. These data indicates that the QUARK-TBM assists the I-TASSER pipeline by providing moderate but consistent improvement on multiple cases rather than in a few significant but anecdotal examples.

Interestingly, the QUARK-TBM models have a comparable GDT-HA score with the I-TASSER models, despite the fact that the I-TASSER models have a much better TM-score (or GDT-TS). For instance, the P-values of TM-score and GDT-TS differences between QUARK-TBM

Table V

Comparison of the First Model Generated by Different Pipelines

Pipeline	TM-score[a]	GDT-TS[a]	GDT-HA[a]	$TM_{template}$[b]
LOMETS	0.613	0.532	0.387	
QUARK-TBM	0.634	0.546	0.393	0.798 (0.644)
I-TASSER[c]	0.642 (0.011)	0.553 (0.016)	0.394 (0.396)	0.763 (0.637)
I-TASSER+QUARK[d]	0.650 (0.022)	0.561 (0.021)	0.401 (0.041)	0.773 (0.639)

[a]TM-, GDT-TS and GDT-HA scores of the first models compared to the native. The values in parenthesis are the P values compared to the models generated by the previous row (that is, P values of I-TASSER vs. QUARK-TBM, and I-TASSER-QUARK vs. I-TASSER).
[b]TM-score of the first model to the closest template. The value in parenthesis is the average TM-score of the first model to the top 30 closest templates.
[c]I-TASSER without using QUARK-TBM models.
[d]I-TASSER predictions using QUARK-TBM models as input. These models were finally submitted to CASP.

and I-TASSER models are both below 0.02; but the *P* values of the GDT-HA score difference is above 0.3. This is probably because of the fact that GDT-HA score is relatively more sensitive to the local structure accuracy due to its finer distance cutoffs, that is, GDT-HA counts only for residue pairs with a distance below 0.5, 1, 2, and 4 Å instead of 1, 2, 4, and 8 Å in GDT-TS (or all residues in TM-score). In QUARK-TBM, much stronger template-based restraints were used, which results in final models that are closer to the threading templates than those in the I-TASSER pipeline (see Column 5 of Table V). Since the threading templates in the PDB are obtained from experimental structures, the strong constraints restraining models to the templates can result in finer local structures that may favor a higher GDT-HA score, despite the less magnitude of refinement in global topology as measured by TM- and GDT-TS scores. To further examine the correlation, we made a post-CASP simulation on a set of 20 randomly selected proteins on QUARK-TBM while the weight of the threading restraints was reduced by two times. It was found that average TM-score between final model and template is reduced by 5.2% (meaning that models are less similar to the template), while the GDT-HA score decreases (by 2.6%) slightly faster than TM-score (by 1.7%). Such sensitivity of GDT-HA score to local structural quality probably explains part of the reason for the rank variations in CASP when the models are assessed by GDT-HA and GDT-TS (or TM-score), where the methods generating models based on single-template have often a finer local structure that favors GDT-HA while the methods based on multiple templates may benefit in GDT-TS or TM-score if the global topology is improved. In addition, compared to I-TASSER, the QUARK-TBM force field contains more detailed atomic-level energy terms. These physics-based energy terms help improve the physical realism of the local structures, which should also contribute to the improvement of the GDT-HA score of the final models.

Retrospect of Zhang-server in the last five CASP experiments

The community-wide blind CASP experiment provides a unique opportunity to assess the weaknesses and strengths of the current state-of-the-art techniques in protein structure prediction. However, it is non-trivial to quantitatively assess the progress of the community across the different CASP experiments.[15,16] One difficulty is on the definition of difficulty of modeling targets across different CASP experiments where template databases keep changing and template structures used by different predictors are also varying, which are not available to the assessors. A constructive approach was to define the target difficulty by the sequence and structural similarities of the target to the best template identified by structural alignment and then compare the quality of the model predictions in different CASPs for the targets of the same level of modeling difficulty.[15,16] However, the best templates, which are identified by structurally matching the target structure to the PDB library, are usually different from what the predictors used. With the increasing size of the structure databases, it becomes increasingly difficult to identify the absolutely best templates by using current threading approaches.[15]

Following the suggestion from the CASP organizer, here we try to assess the progress of the Zhang-Server group (essentially based on the I-TASSER pipeline) over the last five CASP experiments. One convenience over the community-wide progress assessment is that the templates used to construct the models are well documented, which allows a quantitative assessment of progress with regard to different steps of structure modeling, including template identification and template structure refinement.

In Figure 6, we present a summary of the threading templates identified by LOMETS and PSI-BLAST and the final models by I-TASSER across different CASP experiments. Here, we only consider single-domain proteins in order to isolate the data from errors in domain boundary prediction and domain splitting; this results in 368 single-domain targets with 69, 83, 89, 74, and 53 domains from CASP7, 8, 9, 10, and 11, respectively. The LOMETS and Zhang-Server models are collected from the original submissions. Starting from the target sequences and structures, PSI-BLAST[2] and TM-align[32] were used to thread through the PDB library to set up controls for LOMETS-based template identifications, where all templates solved after each CASP were excluded when running the PSI-BLAST and TM-align searches. The data in Figure 6 confirms the previously observed fact that the LOMETS threading programs significantly outperform PSI-BLAST in structure template identification [Fig. 6(A)], demonstrating the advantage of profile-profile alignments, which most threading programs are based on, over the sequence-profile alignment approach in PSI-BLAST.[40] It is also clear that the final models predicted by the I-TASSER pipeline, built on the assembly of multiple templates, are consistently closer to the native than the best individual templates [Fig. 6(B,C)].

To examine the progress of the modeling procedure, we present in Figure 7 the average quality of the models along with different CASP experiments. Here, targets are split into two categories based on the quality of the threading templates, that is, a target is defined as "Easy" if the TM-score of the third best LOMETS template is above 0.5 or as "Hard" otherwise. This resulted in 270 Easy and 98 Hard targets in total. For the Hard targets, PSI-BLAST templates are almost random with an average TM-score close to 0.17. However, TM-align can almost always identify correct fold with the average TM-score above 0.5 although the quality of the best templates for

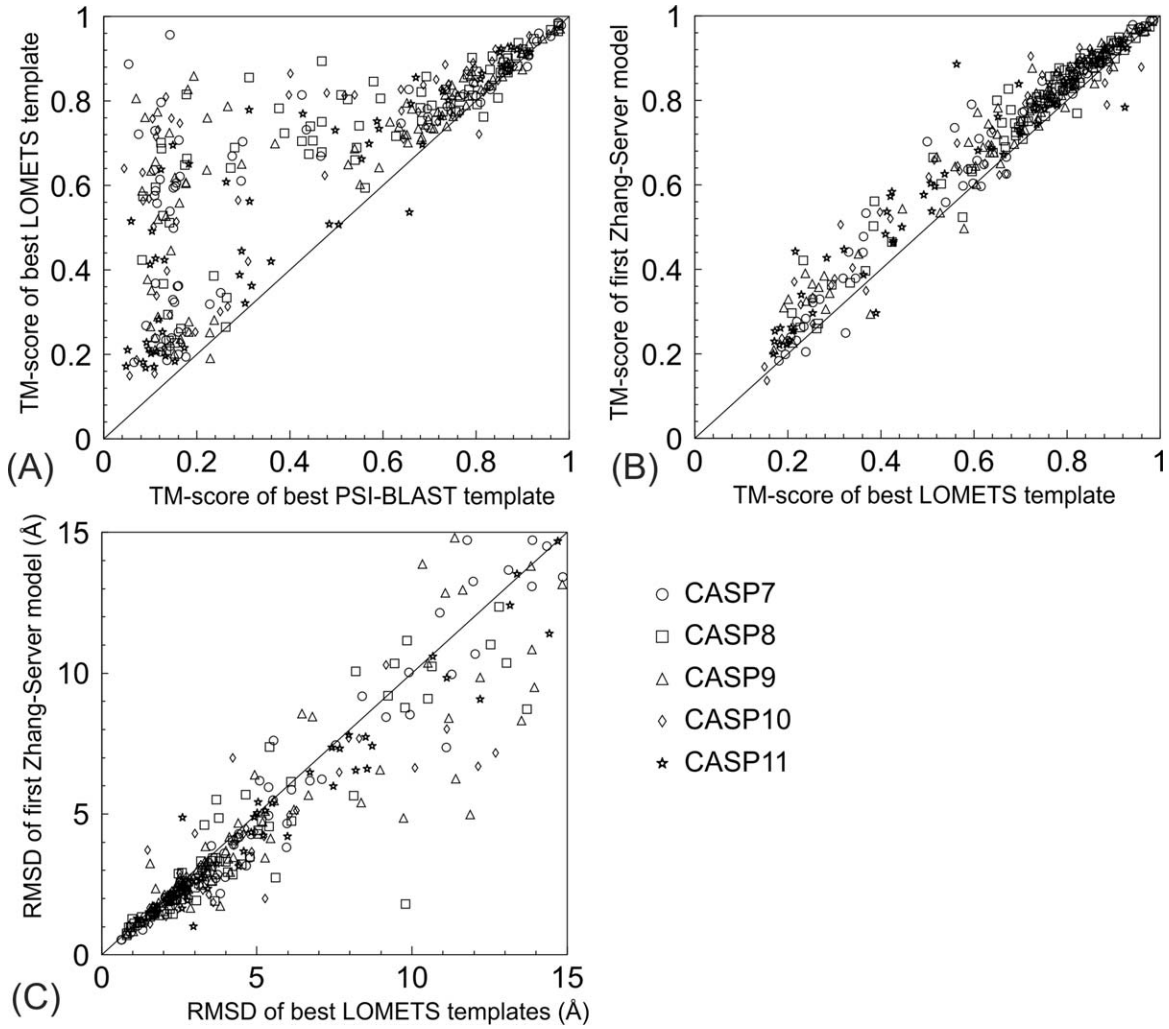

Figure 6

Summary of the structure prediction by Zhang-Server in the last five CASP experiments. (**A**) TM-score of the best templates identified by LOMETS versus that by PSI-BLAST; (**B**) TM-score of the first Zhang-Server models versus TM-score of the best LOMETS templates; (**C**) RMSD of the first Zhang-Server models versus RMSD of the best LOMETS templates, where RMSD was calculated in the same threading aligned regions.

the Hard targets is considerably lower than that for the Easy targets as expected. Overall, there is no obvious trend in the TM-score of the PSI-BLAST and TM-align templates from CASP7 to CASP11 [Fig. 7(A)], which suggests that the difficulty of targets is essentially unchanged through these experiments. The average TM-score of All targets has a noticeable reduction in CASP11 in both TM-align and PSI-BLAST alignments; this is probably due to the fact that the number of hard targets in CASP11 increases.[41] But the average TM-score of the Hard or Easy targets did not change significantly compared to former CASP experiments.

In Figure 7(B), we show the difference of TM-scores between LOMETS and PSI-BLAST. Since LOMETS collected most of the state-of-the-art threading programs developed by the community, this plot should roughly reflect the progress of threading results over PSI-BLAST alignments. Meanwhile, since LOMETS and PSI-BLAST searches are made through the same structure library, the calculation of the TM-score difference is not influenced by the effort of the database increase across different CASP experiments. From Figure 7(B), there seems to be no obvious difference between CASP7 and CASP11 in terms of the improvement of LOMETS over PSI-BLAST. There is a fair TM-score increase in CASP11 for the hard targets over PSI-BLAST; but the TM-score difference for the easy targets drops.

Finally, we present in Figure 7(C,D) the structural improvement of final models over the best LOMETS templates in terms of TM-score and RMSD, respectively. Here there seems to be a steady increase from CASP7 to CASP11 in both Easy and Hard targets. First, for the Hard targets a jump in model quality occurred in CASP8 while a small jump occurred in CASP10 for the Easy

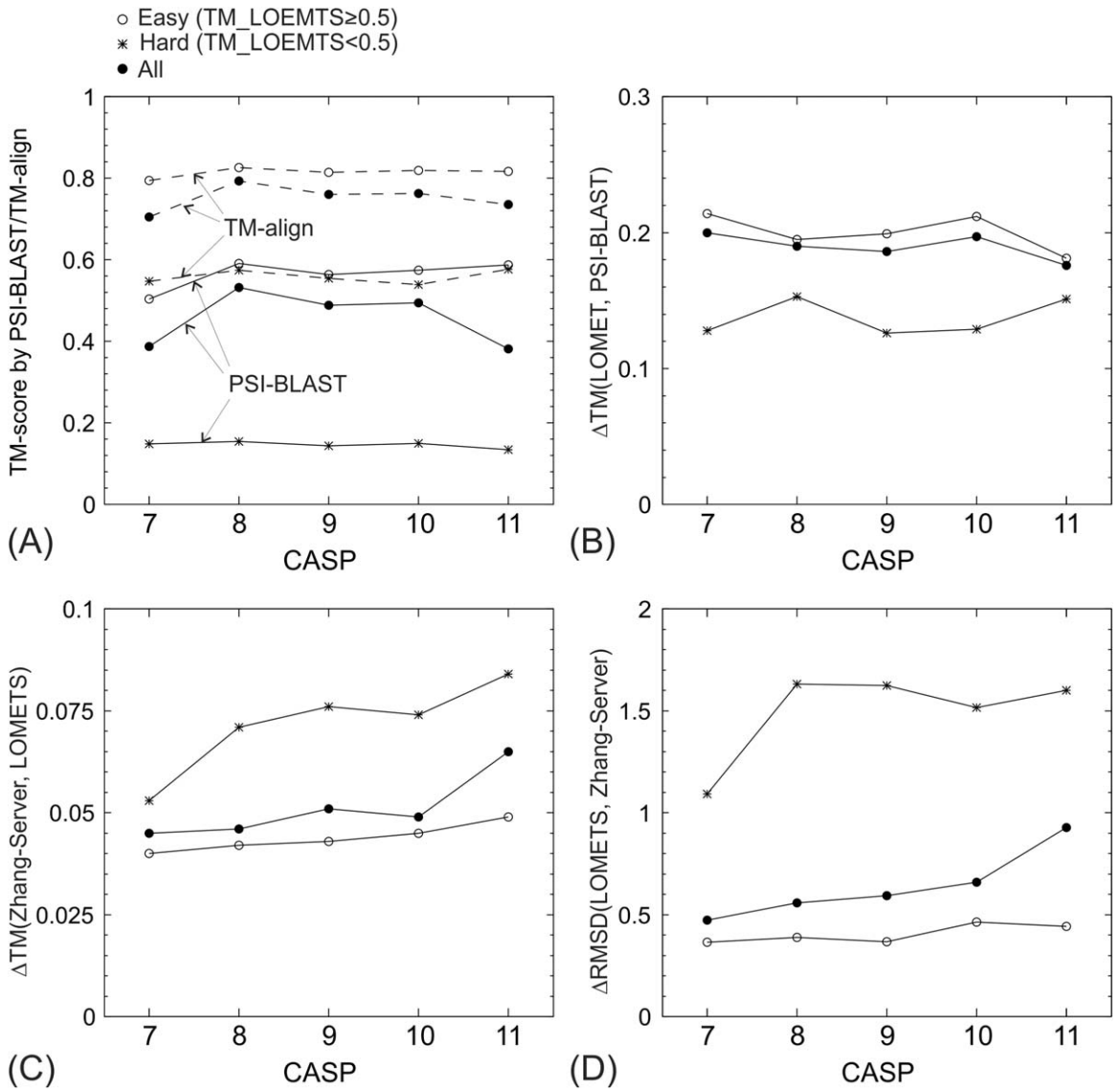

Figure 7

Zhang-Server modeling results in CASP7-11. Open circles, stars and solid circles indicate Easy, Hard, and All targets, respectively. $\Delta X(Y, Z) = X_Y - X_Z$. (**A**) TM-score of the best templates identified by PSI-BLAST and TM-align, respectively; (**B**) Improvement of LOMETS over PSI-BLAST in terms of TM-score; (**C**) Improvement of the first Zhang-Server model over LOMETS in terms of TM-score; (**D**) Improvement of the first Zhang-Server model over LOMETS in terms of RMSD in the threading aligned regions.

targets. A plausible explanation is probably that the improvement on the Hard targets was brought out by the integration of the *ab initio* (QUARK) and template-based (I-TASSER) modeling simulations introduced since CASP8,[38,39] while the improvement for the Easy targets was due to the recently introduced atomic-level structure refinement approaches (FG-MD) since CASP10.[19,24]

CONCLUDING REMARKS

We developed and tested a new template-based structure prediction pipeline in the TBM section of the 11th CASP experiment. In addition to traditional LOMETS threading and I-TASSER structure assembly simulation approaches, the QUARK-based *ab initio* folding simulation was extended to perform template-based simulations by integrating multiple threading alignments with the physics-based force field in QUARK. The results show that the inclusion of more physics-oriented fragment assembly modeling as an intermediate simulation step can improve the quality of the final models of the template-based prediction. Overall, considerable improvements were witnessed for the final models of the pipeline compared to the initial threading templates, where the TM-score of the first submitted model is 12% higher than the best

threading templates with the RMSD in the threading aligned regions reduced by 1.1 Å (that is, from 5.8 to 4.7 Å).

There are, however, 18% of the TBM cases where the final models are worse than the initial threading templates in terms of RMSD. Detailed analyses showed that errors in secondary structure prediction could propagate through and influence the template identification and final model selection processes. Second, modeling of domain orientations remains an open problem for multi-domain protein structure predictions, especially when the orientation of the templates is different from the targets. While the statistical potentials can help pick up correct folds for some targets, the large-scale benchmark and the CASP data showed that a combination of the statistical and consensus-based MQAP programs outperforms the statistical potential or consensus MQAP alone in final model selection. Among the 15 cases that have the RMSD of the final model higher than the RMSD of initial threading template, 6 cases (T0793-D3, T0781-D2, T0816-D1, T0830-D1, T0838-D1, and T0851-D1) have the RMSD difference above 1 Å. Out of the six cases, three cases (T0838-D1, T0793-D3, and T0781-D2) have the Q3 accuracy of secondary structure prediction below 80% (62.7, 79.3, and 77.1%, respectively) which have similar issue with T0816-D1 as shown in Figure 3. T0851-D1 has the domain orientation issue as highlighted in Figure 4. The last target (T0830-D1) represents the typical case of the I-TASSER failures in template selection when the best template alignment is minority (Fig. 5), which has been witnessed and discussed in previous CASP reports as well.[19,38,39]

To track the progress of the I-TASSER-based structure modeling pipelines, we presented a retrospective report of the Zhang-Server models in the last five CASP experiments. There is no clear improvement on the quality of the LOMETS threading templates over the PSI-BLAST templates from CASP7 to CASP11; but a clear trend in the ability of structure refinement was shown over the threading templates that the I-TASSER structure predictions are based on. This is probably due to the integration of the template-based modeling with the extended and more physics-oriented *ab initio* folding simulations and the introduction of the atomic-level structure refinement.

REFERENCES

1. Browne WJ, North AC, Phillips DC, Brew K, Vanaman TC, Hill RL. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol 1969;42:65–86.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
3. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
4. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960.
5. Wu S, Zhang Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 2008;72:547–556.
6. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
7. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucl Acids Res 2007;35:3375–3382.
8. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
9. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nature Methods 2015;12:7–8.
10. Joo K, Lee J, Sim S, Lee SY, Lee K, Heo S, Lee IH, Lee SJ, Lee J. Protein structure modeling for CASP10 by multiple layers of global optimization. Proteins 2014;82 Suppl 2:188–195.
11. Cheng J, Wang Z, Tegge AN, Eickholt J. Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins 2009;77 Suppl 9:181–184.
12. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 2006;103:5361–5366.
13. Roy A, Kucukural A, Zhang Y. ITASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5:725–738.
14. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80:1715–1735.
15. Kryshtafovych A, Fidelis K, Moult J. CASP10 results compared to those of previous CASP experiments. Proteins 2014;82 Suppl 2:164–174.
16. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins 2005;61 Suppl 7:225–236.
17. Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. Bioinformatics 2013;29:i247–i256.
18. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH–a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
19. Zhang Y. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins 2014;82 Suppl 2:175–187.
20. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007;5:17
21. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003;85:1145–1164.
22. Xu D, Zhang Y. Toward optimal fragment generations for ab initio protein structure assembly. Proteins 2013;81:229–239.
23. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.
24. Zhang J, Liang Y, Zhang Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 2011;19:1784–1795.
25. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
26. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
27. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 2011;101:2043–2052.
28. Zhang J, Zhang Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PloS One 2010;5:e15386

29. Zhang Y. ITASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9:40
30. Li Y, Zhang Y. REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins 2009;76:665–676.
31. Krivov GG, Shapovalov MV, Dunbrack RL. Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009; 77:778–795.
32. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
33. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 2010;26:889–895.
34. Keedy DA, Williams CJ, Headd JJ, Arendall WB, 3rd, Chen VB, Kapral GJ, Gillespie RA, Block JN, Zemla A, Richardson DC, Richardson JS. The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. Proteins 2009;77 Suppl 9:29–49.
35. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D 2010;66(Pt 1):12–21.
36. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
37. Rocha J, Popescu AO, Borges P, Mil-Homens D, Moreira LM, Sa-Correia I, Fialho AM, Frazao C. Structure of Burkholderia cepacia UDP-glucose dehydrogenase (UGD) BceC and role of Tyr10 in final hydrolysis of UGD thioester intermediate. J Bacteriol 2011;193: 3978–3987.
38. Zhang Y. ITASSER: Fully automated protein structure prediction in CASP8. Proteins 2009;77:100–113.
39. Xu D, Zhang J, Roy A, Zhang Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 2011;79 (Suppl 10):147–160.
40. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep 2013;3:2619
41. Grishin NV. CASP11 and CASP ROLL Domain Definition and Classification. Riviera Maya, Mexico; 2014.