

PREDICTION REPORT

I-TASSER: Fully automated protein structure prediction in CASP8

Yang Zhang*

Center for Bioinformatics, University of Kansas, Lawrence, Kansas 66047

Department of Molecular Bioscience, University of Kansas, Lawrence, Kansas 66047

ABSTRACT

The I-TASSER algorithm for 3D protein structure prediction was tested in CASP8, with the procedure fully automated in both the Server and Human sections. The quality of the server models is close to that of human ones but the human predictions incorporate more diverse templates from other servers which improve the human predictions in some of the distant homology targets. For the first time, the sequence-based contact predictions from machine learning techniques are found helpful for both template-based modeling (TBM) and template-free modeling (FM). In TBM, although the accuracy of the sequence based contact predictions is on average lower than that from template-based ones, the novel contacts in the sequence-based predictions, which are complementary to the threading templates in the weakly or unaligned regions, are important to improve the global and local packing in these regions. Moreover, the newly developed atomic structural refinement algorithm was tested in CASP8 and found to improve the hydrogen-bonding networks and the overall TM-score, which is mainly due to its ability of removing steric clashes so that the models can be generated from cluster centroids. Nevertheless, one of the major issues of the I-TASSER pipeline is the model selection where the best models could not be appropriately recognized when the correct templates are detected only by the minority of the threading algorithms. There are also problems related with domain-splitting and mirror image recognition which mainly influences the performance of I-TASSER modeling in the FM-based structure predictions.

Proteins 2009; 77(Suppl 9):100–113.
© 2009 Wiley-Liss, Inc.

Key words: protein structure prediction; threading; I-TASSER; CASP8; contact prediction; free modeling.

INTRODUCTION

When will computers beat humans in protein structure prediction? Or are there still any human insights that cannot be reproduced in automated approaches? During the CASP experiments, several groups^{1–3} demonstrated that intervention by human experts, who made use of biochemical information (function, family characteristics, mutagenesis, catalytic residues, etc.), can indeed help with template recognition, structural assembly, and final model selection. Nevertheless, fully automated algorithms have an advantage in genome-wide structure prediction^{4–6}; they also allow non-experts to generate structural models on their own or through internet services.^{7–9} Undoubtedly, with the rapid accumulation of genome-wide sequences, the development of fully automated computer-based structure prediction methods becomes unprecedentedly demanded.¹⁰

Recent years have witnessed significant progress in automated structure prediction.^{6,11} In CASP7, for example, it was stated in the assessors' reports^{12–14} that “the best prediction server (Zhang-Server) was ranked third overall, that is, it outperformed all but two of the human participating groups.” Actually, in the current framework of CASP, it is difficult to have an entirely fair assessment of the performance of automated versus human prediction because

Additional Supporting Information may be found in the online version of this article.

The author states no conflict of interest.

Grant sponsor: Alfred P. Sloan Foundation; Grant sponsor: NSF (Career Award); Grant number: DBI 0746198; Grant sponsor: National Institute of General Medical Sciences; Grant numbers: GM083107, GM084222.

*Correspondence to: Dr. Yang Zhang; Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, 2030 Becker Dr, Lawrence, KS 66047. E-mail: yzhang@ku.edu.

Received 31 March 2009; Revised 22 July 2009; Accepted 7 August 2009

Published online 21 August 2009 in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/prot.22588

human predictors can use all the models generated by servers and therefore have a better pool of initial templates to start with.

In CASP8, we participated in both human (as “Zhang”) and server (as “Zhang-Server”) predictions. For the purpose of the development and testing of automated structure prediction approaches, both Zhang and Zhang-Server used identical I-TASSER approaches.¹⁵ Compared with CASP7, new developments in I-TASSER include the employment of *de novo* sequence-based contact predictions¹⁶ and atomic-level hydrogen-bonding (H-bond) optimization.¹⁷ Because the only difference between Zhang and Zhang-Server is that the “human” prediction uses more templates (including those generated by other groups in the Server section), the difference between their performances may be viewed as a measure of the effect due to the different template pools used in human and server predictions.

RESULTS AND DISCUSSION

A total of 164 domains from 121 protein targets were eventually assessed in the Server Section, and 71 domains in the Human Section. Among the 164 domains, 50 are high-accuracy (HA), 102 are template-based modeling (TBM), and only 12 are free-modeling (FM, including TBM/FM) targets. Because more targets were tested in the server section and the methods used in our server and human predictions are essentially identical, our report will mainly focus on the server predictions. In particular, we summarize what went right and what were the major problems with our approach.

What went right?

I-TASSER pulls templates closer to the native conformation

As observed in both benchmark tests¹⁵ and previous CASP experiments,¹⁸ one of the most important advantages of I-TASSER is that the fragment assembly procedure can consistently drive the initial template structures closer to their native states. In Figure 1(a), we present the RMSD of the first I-TASSER server models versus the RMSD of the best threading templates used in I-TASSER for all 164 domains, with both the RMSDs calculated for the aligned regions of threading alignments. Although FM targets are supposed to have no appropriate templates, we show them in the plot because the I-TASSER procedure always starts from the top scoring templates obtained by threading no matter how weak the alignment scores are. In fact, even when the global topology of the templates is incorrect, the super-secondary structure segments are useful as structural building blocks. Apparently, I-TASSER simulations improve the template structure in the majority of test cases as measured by RMSD.

For 139 out of 164 domains, the RMSD of the final models is lower than that of the templates. In the remaining 22 (and 3) cases, the RMSD of the I-TASSER models is higher than (and equal to) that of the templates. Overall, the average RMSD of the best threading template is 5.54 Å for the aligned regions with an average alignment coverage of 91%; this RMSD is reduced to 4.24 Å by I-TASSER.

Because some threading alignments are very short and may consist of only a small piece of structure, a TM-score comparison should reflect more appropriately the improvement by I-TASSER in full-chain model construction from the templates. Figure 1(b) is a comparison of final models versus the best threading templates in terms of TM-score. Now, 150 targets have a final model with a higher TM-score than the templates, and 10 (4) have a final model with a lower (equal) TM-score than the templates. Noticeably, there are two domains, T0472_2 and T0474, where the first submitted models are significantly worse than the best templates. **T0472** has a duplicated $\beta_3\alpha$ two-domain structure with its closest structural template, 3bid, being a domain-swapped dimer. Because our threading library includes only single-chain proteins, most of the whole-chain threading templates have only the N-terminal domain aligned. The first submitted model by our I-TASSER server is based on the whole-chain modeling and has a reasonably good quality for the N-terminal domain (RMSD = 1.54 Å and TM-score = 0.731) but a low-quality C-terminal domain (TM-score = 0.605 for T0472_2). The second submitted model by the server for T0472 was built by modeling the domains separately, followed by domain docking as described in Methods; it has a TM-score of 0.767 for T0472_2, slightly higher than that of the template (TM-score = 0.755).

T0474 is a small protein of 80 residues solved by Structural Genomics Consortium and has a very extended structure (85.3Å from N to C terminus). All the three closest templates (2ay0, 2bj1, 2hza) are dimers, with the “necks” of the chains intertwined with each other. The individual chains are apparently unstable on their own, but our server attempted to fold the chain as an individual compact domain; this resulted in a much less extended structural model with a TM-score = 0.560. The second submitted model has a more extended structure with a TM-score = 0.683, which is still lower than the best template with TM-score = 0.726.

As threading algorithms usually generate substantial alignment errors, in Figure 1(c,d), we compare I-TASSER models with the best threading templates as used in Figure 1(a,b) but the alignments are regenerated by structurally aligning the templates to the native structures by TM-align.¹⁹ Because the native structure information is used, the structural alignment is more accurate than the threading, with the average RMSD reduced from 5.54 to 2.42 Å and the average TM-score increased from 0.633

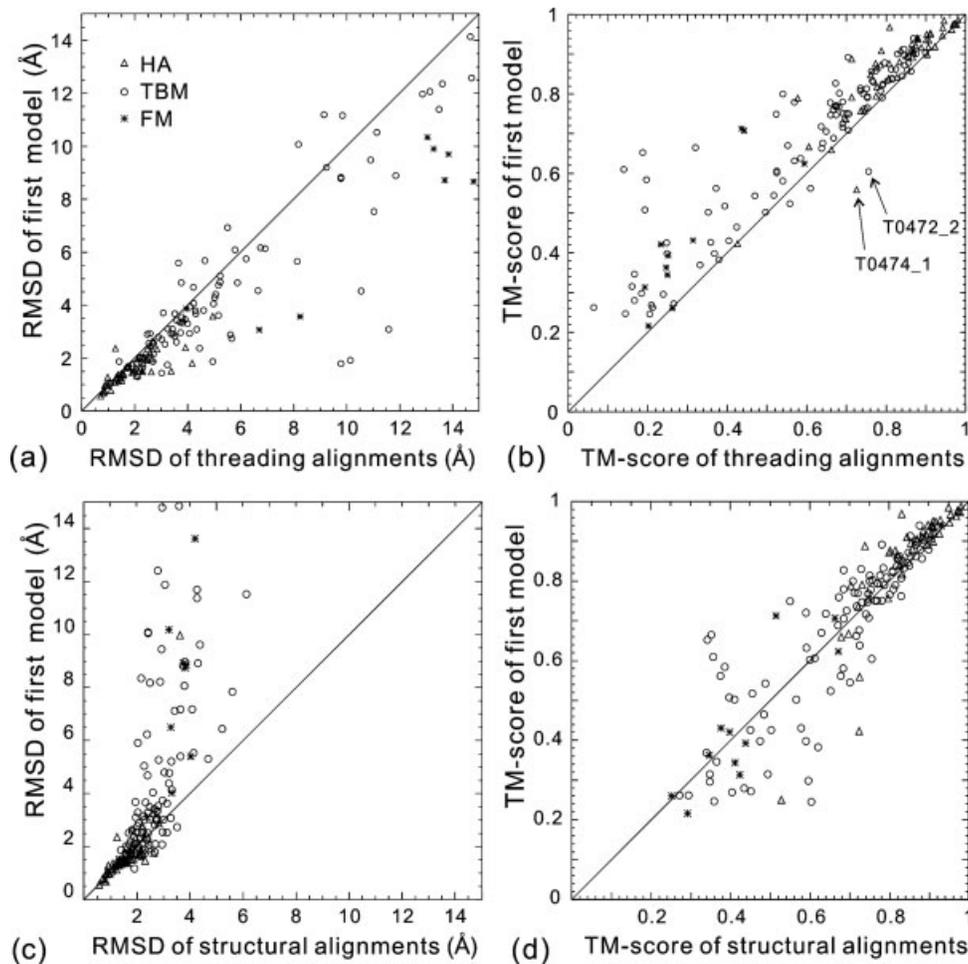


Figure 1

Comparison of the best templates with the first model predicted by the I-TASSER server. The alignments in (a) and (b) are from threading algorithms which have been used as input of I-TASSER simulations; the alignments in (c) and (d) are generated by structurally aligning the templates to the native by TM-align. RMSD for models is calculated in the same aligned region as the alignments in templates. The highlights in (b) are two domains where I-TASSER deteriorates the best templates.

to 0.709. In 47 (or 99) cases, the I-TASSER models have a lower RMSD (or a higher TM-score) than the TM-align alignments. Although the overall quality of the final I-TASSER models is still worse than the best structural alignments in terms of RMSD, the data shows that at least for part of the cases the model can be drawn by I-TASSER closer to the native than the best aligned template structures; these improvements come from the fragment rearrangement rather than from refining the threading alignments.

Restrictions from multiple templates cover a larger portion of the structure than those from the best single templates

One of the major driving forces of the structure refinement in I-TASSER is the high-quality consensus restraints

taken from multiple templates by MUSTER²⁰ or LOMETs.²¹ Five types of template-based restraints are used in I-TASSER: (1) side-chain contact restraints taken from the top N templates ($N = 20$ for easy targets, 30 for medium and 50 for hard targets); (2) $C\alpha$ contact restraints from the top N templates; (3) short-range $C\alpha$ distance-map for separation $|i - j| \leq 6$ with the average distance from the top N templates; (4) $C\alpha$ distance-map for separation >6 from the top four templates (i.e., each residue pair having up to four different distance restraints); and (5) pair-wise contact potential based on the frequency of the side-chain contacts appearing in the top N templates.²²

Although there has been a long-time belief that consensus restraints should have a better accuracy than those from single templates, there is no systematic comparison of the two based on the same set of templates in litera-

Table I

Comparison of Spatial Restraints Taken from Multiple Templates and from the Single Best Threading Template (the Latter Shown in Parentheses). A Detailed List is Shown in Table S1 in Supporting Information

	Side-chain contact restraints			C α contact restraints			Short distance ^d	Long distance ^e	RM ^f	TM ^g
	N ^a	Acc ^b	Cov ^c	N ^a	Acc ^b	Cov ^c				
HA	128.0	0.39 (0.45)	0.87 (0.79)	97.3	0.70 (0.79)	0.86 (0.77)	0.45 (0.41)	0.65 (0.92)	1.6	0.895
TBM	132.0	0.29 (0.31)	0.58 (0.48)	99.1	0.53 (0.55)	0.59 (0.50)	0.87 (0.84)	1.72 (2.66)	5.7	0.668
FM	60.0	0.17 (0.10)	0.11 (0.06)	40.2	0.26 (0.05)	0.13 (0.05)	1.32 (1.40)	3.14 (7.34)	9.0	0.380
All	125.5	0.31 (0.34)	0.64 (0.55)	94.2	0.56 (0.59)	0.64 (0.55)	0.77 (0.75)	1.50 (2.47)	4.7	0.712

^aNumber of contacts appearing in the native structure.

^bAccuracy of contact predictions: the number of correctly predicted contacts divided by the total number of contact predictions.

^cCoverage of contact predictions: the number of correctly predicted contacts divided by the number of contacts in the native structure.

^dError of short-range distance predictions ($|i-j| \leq 6$) relative to the native structure.

^eError of medium- and long-range distance predictions ($|i-j| > 6$) relative to the native structure.

^fRMSD (Å) of the first submitted model by Zhang-Server (best in top 5 shown for FM).

^gTM-score of the first submitted model by Zhang-Server (best in top 5 shown for FM).

ture. In Table S1 in Supporting Information, we present a detailed list of the accuracy and coverage of four restraint types taken either from multiple templates or from the best single threading template that has the highest TM-score to the native in the top N templates. Table I is a summary of Supporting Information Table S1 with an average accuracy of the restraints listed in each category of targets. In all categories of targets (i.e., HA, TBM, and FM), the consensus contact predictions have a higher coverage, that is, more correct contacts are predicted. However, somewhat contrary to expectation, the accuracy of the contacts based on single templates is slightly higher than that of the consensus ones, which is probably due to the fact that we are using the best individual template from threading. In fact, if we use the first template (as ranked by threading rather than TM-score), the accuracy of the contact prediction is similar to that of consensus contacts, but the coverage is lower than when the best threading template (i.e., with the highest TM-score) is used. Here, we compare consensus restraints to the best templates because we try to highlight the possible reason that I-TASSER improves the quality of the best templates as shown in Figure 1. Overall, the average accuracy/coverage for side-chain and C α contact predictions are 0.34/0.55 and 0.59/0.55 from the best single template, compared to 0.31/0.64 and 0.56/0.64 from multiple templates. One reason for the apparently higher accuracy of C α contacts in comparison with side-chain contacts is that side-chain contacts are more variable due to rotamer conformations, and are therefore more difficult to predict.

The eighth and ninth columns of Table I and Supporting Information Table S1 show the errors of short- and long-range C α distance predictions, respectively. For short-range distance prediction, single-template-based prediction has a slightly smaller average error than the multiple-template-based one. But for the long-range distance prediction, the distance error from multiple templates is much smaller than that from the best single template. Here, each residue pair has four distance pre-

dictions collected from the first four MUSTER/LOMETS templates and we report the best of the four predictions in the tables. Moreover, as the major advantage of using multiple templates, multiple-template-based predictions cover again a larger portion of the structure. Overall, the multiple-template based prediction produces on average 1302/2563 short/long-range distance predictions while single-template prediction produces only 1099/2243 short/long-range predictions.

Interestingly, there are some targets for which the accuracy and coverage of contact predictions is apparently high, but the quality of the final models is still poor. For example, two FM targets (T0476_1 and T0482_1) have C α contact predictions with both accuracy and coverage >0.5 (see Table S1 in Supporting Information). However, all 11 correctly predicted contacts in T0476_1 are concentrated in two β -hairpins (one at the tail and another in the middle, both being short-range), and are actually not helpful for assembling the global topology. On the contrary, the side-chain contact predictions have a lower accuracy but cover a larger portion of the structure. A similar situation is seen with T0482_1 as well. In fact, the correlation coefficient (calculated for all 164 domains) between the TM-score of the final models and the product of accuracy and coverage of side-chain contacts is 0.87, while the same quantity for C α contacts is 0.79, which indicates that side-chain contact predictions are more important for the structure assembly.

Sequence-based contact predictions help both FM and TBM modeling

In addition to the consensus restraints from multiple templates, the second important contribution to the I-TASSER template structural refinement is the sequence-based contact prediction from SVMSEQ.¹⁶ Our original purpose when developing SVMSEQ was to improve the I-TASSER structure assembly only for FM targets, because for TBM/HA targets, the overall accuracy of

Table II

Summary of Sequence-Based Contact Predictions (by SVMSEQ) Compared with the Template-Based Contact Predictions (by LOMETS)

		Side-chain contacts				C α contacts				C β contacts			
		Tem ^a	S6 ^b	S7 ^c	S8 ^d	Tem ^a	S6 ^b	S7 ^c	S8 ^d	S6 ^a	S7 ^b	S8 ^c	Com ^e
HA	NP ^f	114.9	15.2	21.9	28.3	84.7	17.6	28.6	35.5	11.2	22.4	31.9	10.8
	ACC ^g	0.39	0.219	0.32	0.403	0.7	0.26	0.397	0.475	0.162	0.328	0.442	0.228
	COV ^h	0.87	0.136	0.199	0.25	0.86	0.162	0.265	0.333	0.099	0.202	0.292	0.381
	NN ⁱ		0.7	1.8	5.4		1.1	9	15.6	0.7	7.2	14.4	4.7
	CON ^j		0.007	0.017	0.049		0.013	0.082	0.144	0.007	0.062	0.129	0.153
TBM	NP ^f	78.5	15.5	22.7	28.6	61.8	20.4	33.3	42.1	14.5	26.5	36.9	20.3
	ACC ^g	0.29	0.216	0.322	0.389	0.53	0.274	0.419	0.494	0.209	0.366	0.472	0.325
	COV ^h	0.58	0.135	0.203	0.247	0.59	0.167	0.276	0.352	0.128	0.225	0.307	0.253
	NN ⁱ		2.7	4.8	8.5		3.6	11.9	19.1	2.5	9.5	16.6	8.4
	CON ^j		0.024	0.045	0.072		0.033	0.102	0.163	0.024	0.082	0.143	0.105
FM	NP ^f	6.2	9	13	15.9	5.8	9.9	15.1	18.8	7.8	13.3	17.8	14.5
	ACC ^g	0.17	0.211	0.312	0.381	0.26	0.238	0.363	0.448	0.199	0.32	0.422	0.448
	COV ^h	0.11	0.164	0.245	0.299	0.13	0.179	0.277	0.353	0.151	0.24	0.332	0.267
	NN ⁱ		6.8	9.5	11.5		7.6	11.8	15.5	5.3	10.1	14.5	10.2
	CON ^j		0.124	0.176	0.218		0.133	0.215	0.293	0.101	0.178	0.271	0.187
AAll	NP ^f	84.3	15	21.8	27.6	64.7	18.8	30.5	38.4	13	24.3	34	17
	ACC ^g	0.31	0.216	0.321	0.392	0.56	0.267	0.408	0.485	0.194	0.351	0.459	0.305
	COV ^h	0.64	0.138	0.205	0.252	0.64	0.167	0.273	0.347	0.121	0.219	0.304	0.311
	NN ⁱ		2.4	4.2	7.8		3.1	11	17.8	2.2	8.9	15.7	7.4
	CON ^j		0.026	0.046	0.075		0.034	0.104	0.167	0.024	0.083	0.148	0.132

^aContact predictions from multiple threading templates by LOMETS²⁰.^bContact prediction from SVMSEQ¹⁶ with a cutoff of 6 Å.^cContact prediction from SVMSEQ with a cutoff of 7 Å.^dContact prediction from SVMSEQ with a cutoff of 8 Å.^eContact prediction by taking consensus of predictions from CASP8 servers.^fTotal number of predictions.^gAccuracy of contact predictions: the number of correctly predicted contacts divided by the total number of contact predictions.^hCoverage of contact predictions: the number of correctly predicted contacts divided by the number of contacts in the native structure.ⁱNumber of true-positive predictions which are not generated by the template-based predictions.^jCoverage of novel predictions: NN divided by the number of contacts in the native structure.

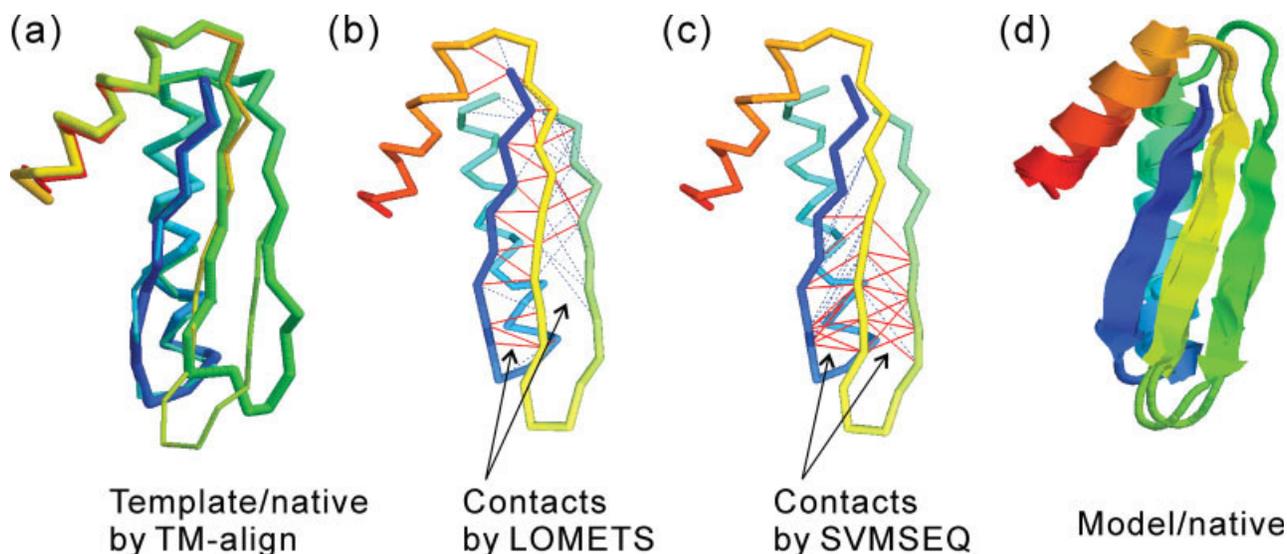
SVMSEQ is lower than that of the template-based contact prediction.¹⁶ However, we found that the SVMSEQ prediction also improves the quality of models for the TBM targets.

In Table II, we present a summary of the SVMSEQ contact prediction for both side-chain and C α contacts. As expected, the sequence-based contact predictions have the highest impact on FM targets. For these targets, the average accuracy of the side-chain contacts by LOMETS is only 17%, covering 11% of all native contacts. But the SVMSEQ prediction on side-chain contacts (with a 8 Å cutoff distance) has an accuracy of 38.1%, with a coverage of 29.9% of all contacts in the native structure; out of this coverage, 21.8% are newly predicted contacts that are not generated by LOMETS. If we look at C α contacts, the average accuracy of SVMSEQ predictions is 44.8%, compared with 26% by LOMETS. This covers 35.3% of all native contacts, with 29.3% being new. The C β predictions have similar results to C α . These sequence-based “*de novo*” predictions are of great value for I-TASSER in the case of FM target predictions.

In Figure 2, we show one example of successful modeling by the I-TASSER server on an FM target, **T0416_2**. I-TASSER first runs LOMETS on the whole chain (332 residues), which yields alignments dominated by 3crmA and

2qgnA. However, there is a middle region spanning 87 residues (L112-T198) that has no alignment with any of the top 20 templates. The server then automatically defines this region as a new domain and runs LOMETS again on the domain, which results in a number of weakly scoring hits. Although none of these templates for the small domain has a correct fold, some have close fragments, which provides building blocks for I-TASSER assembly (Row 3 of Fig. 2). Out of the top 29 side-chain contact predictions by SVMSEQ, 13 (45%) are correct, covering 46% of all native contacts (Row 4 of Fig. 2). Under the guidance of these restraints, I-TASSER finally assembles a model for T0416_2 (S124-K180, as defined by the assessors) with a RMSD = 3.4 Å and a TM-score = 0.53.

The accuracy of SVMSEQ predictions for HA/TBM targets is similar to that for FM targets. However, the coverage and accuracy of the contacts by LOMETS are much higher than SVMSEQ predictions for these targets. Nevertheless, SVMSEQ still generates a considerable number of correct contacts which cannot be generated by template-based predictions. The SVMSEQ-based C α contact predictions with a 8 Å cutoff, for example, provide 14.4 and 16.3% of new true-positive contact predictions for HA and TBM targets, respectively. These restraints are useful in modeling the regions lacking threading

**Figure 3**

SVMSEQ contact predictions improve the modeling of T0437_1. (a) Structural superposition of the target (thin backbone) on the best template 2jz5A (thick backbone) with structural alignment generated by TM-align¹⁹ (RMSD = 1.34Å, TM-score = 0.838). (b) Backbone structure of the native with lines between residues indicating C α contact prediction from LOMETS.²¹ Red solid lines are true-positive and green dashed ones are false-positive. There is no true-positive contact in the lower part of the second β -hairpin. (c) Same as (b) but contacts are from SVMSEQ¹⁶ with 10 true-positive predictions in the lower part of the second β -hairpin. (d) Superposition of the I-TASSER server model on the native with a RMSD = 1.13 Å and a TM-score = 0.885. The image is generated by MVP.²³

alignments as well as improving the global topology. It is worth mentioning that when we use the SVMSEQ-predicted contacts in the I-TASSER assembly, a large percentage of them are false positive. However, these false positive predictions do not necessarily affect the modeling of the regions with good templates because the consensus restraints from LOMETS are strong and dominating in those regions compared with the weak noise from SVMSEQ predictions. For the weakly aligned regions, however, the false-positive rate of SVMSEQ is lower than that of LOMETS, and therefore becomes helpful.

Figure 3 is one such example of a TBM-HA target, T0437_1, demonstrating the positive contribution of SVMSEQ to homology-based modeling. The LOMETS threading alignments are dominated by the template 2jz5A, which has a sequence identity of 32% to the target. The best threading alignment generated by HHsearch²⁴ has an RMSD = 2.30 Å and TM-score = 0.778. If we structurally align 2jz5A to the experimental structure by TM-align,¹⁹ the RMSD is 1.34 Å with TM-score = 0.838 [Fig. 3(a)]. Although the global topology of 2jz5A matches the target well, there is a major mismatch in the region V49-T60 [the lower part of the second β -sheet, Fig. 3(a)]. Correspondingly, there is no correct contact prediction from LOMETS in this region [Fig. 3(b)]. The sequence-based SVMSEQ contact prediction, however, generates 10 correct C α contact predictions in this region [two others are false positive, Fig.

3(c)]. These restraints help I-TASSER generate models with a correct β -sheet structure in this region. The RMSD of the overall model is 1.13 Å, which is even closer than the best structural alignment [Fig. 3(d)]. In this example, although the overall accuracy of the SVMSEQ prediction is still lower than LOMETS, the novel contacts from the sequence-based prediction improve the quality of local structures. In other regions (e.g., the N-terminal β -sheet), SVMSEQ generates a number of false positive contact predictions. As the LOMETS predictions provide strong consensus restraints, these weak false-positive predictions did not reduce the modeling accuracy in those regions.

In the last column of Table II, we also list a consensus prediction taken from 6 CASP8 servers including LEE-SERVER, MULTICON-CMFR, MUProt, SAM-T08-2stage, RR_FANG_1, and Parings. A consensus contact is collected if it is predicted by more than half of the servers. These contacts were used in our human predictions. Somewhat unexpectedly, the consensus prediction from multiple servers does not outperform the prediction from the single program SVMSEQ. For FM targets, the consensus prediction has a slightly higher accuracy than SVMSEQ but a lower coverage. The overall accuracy of consensus contact prediction for all targets is lower than SVMSEQ but the coverage is similar. The SVMSEQ server also participated in CASP8 contact prediction,²⁵ but it submitted predictions obtained by combining

Table III
Comparison of REMO¹⁷ and Pulchra²⁷ on 149 Domains

	RMSD (Å)	TM-score	HBscore (all-atom)	HBscore (backbone)	N_{clash}
REMO + combo	4.50	0.725	0.496	0.643	1.6
Pulchra + cloc	4.75	0.708	0.380	0.520	3.5
Pulchra + combo	4.51	0.716	0.390	0.531	34.3

results from SVMSEQ and LOMETS. Although this combination helps increase the accuracy for TBM/HA targets, it substantially decreases the accuracy of the original SVMSEQ predictions for FM targets; the FM targets was eventually assessed in the contact prediction section of CASP8.

Atomic-level structure refinement improves hydrogen-bonding networks

The SPICKER program²⁶ clusters the structure decoys from I-TASSER and generates two types of reduced models: the cluster centroid (as “combo”) obtained by averaging the coordinates of all clustered decoys and the decoy closest to the centroid (as “cloc”). Combo structures are usually closer to the native but have more structural clashes than the cloc models. When constructing the full-atomic models, REMO¹⁷ has the advantage to eliminate clashes from combo and optimize the hydrogen-bonding network, over a number of other similar algorithms.^{27–29}

In Table III, we compare the REMO models of 149 domains (corresponding to 117 targets) with the full-atom models regenerated by Pulchra²⁷ based on the same set of cloc and combo models. The models of these 149 domains have been generated by the I-TASSER server without domain splitting, and we selected them for these comparisons so that we can eliminate the possible influence of the domain docking procedure. Clearly, the models by Pulchra based on combo have a better TM-score and HBscore compared with that on cloc. However, Pulchra could not remove the steric clashes in the combo models. Here, HBscore is defined as the number of H-bonds appearing in both model and native divided by that in the native structure, with H-bonds defined by HBPLUS3.0.³⁰ The final models generated by REMO have on average a better TM-score and HBscore than both the Pulchra models. The average number of steric clashes of the REMO models is 1.6, which is close to the average in the experimental structures in the PDB.¹⁷

Human and automated server predictions are consistent

Figure 4 is a head-to-head comparison of Zhang-Server and Zhang in terms of TM-score and RMSD for the first models of 71 domains that have been tested in both the Server and the Human sections. There are slightly more targets with the human model having a higher TM-score than the server prediction, which results in a 1.8% overall increase in TM-score. Because the strategies of human and server predictions are identical, this difference reflects the gain from using multiple threading

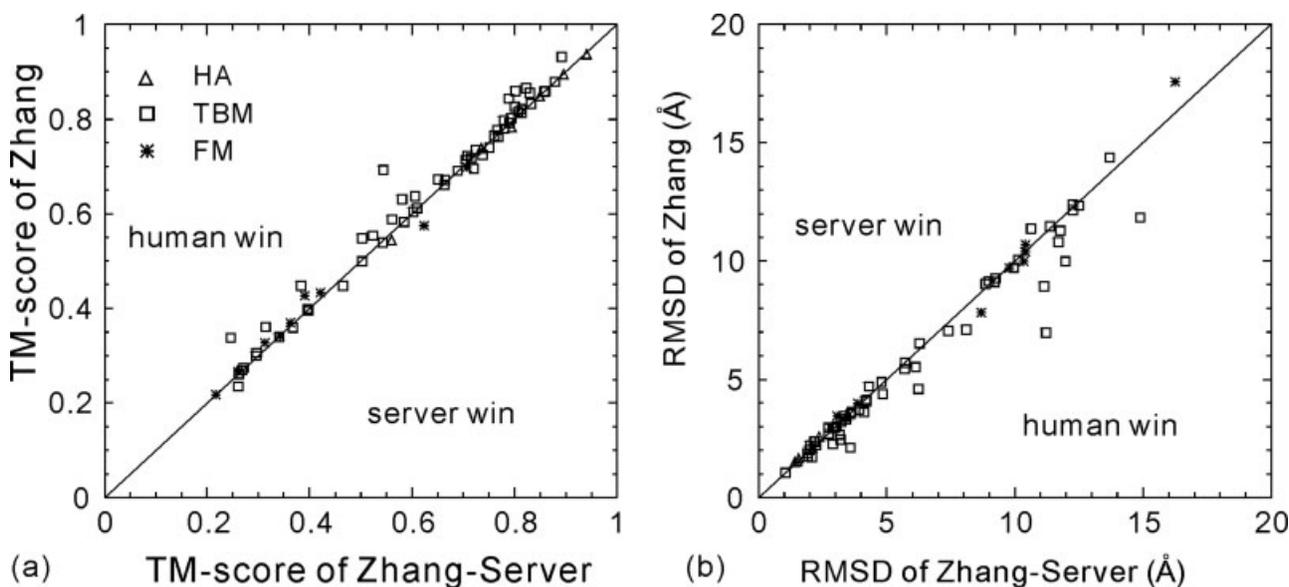


Figure 4
Comparison of the first models predicted by human (as “Zhang”) and server (as “Zhang-Server”) for all 164 domains.

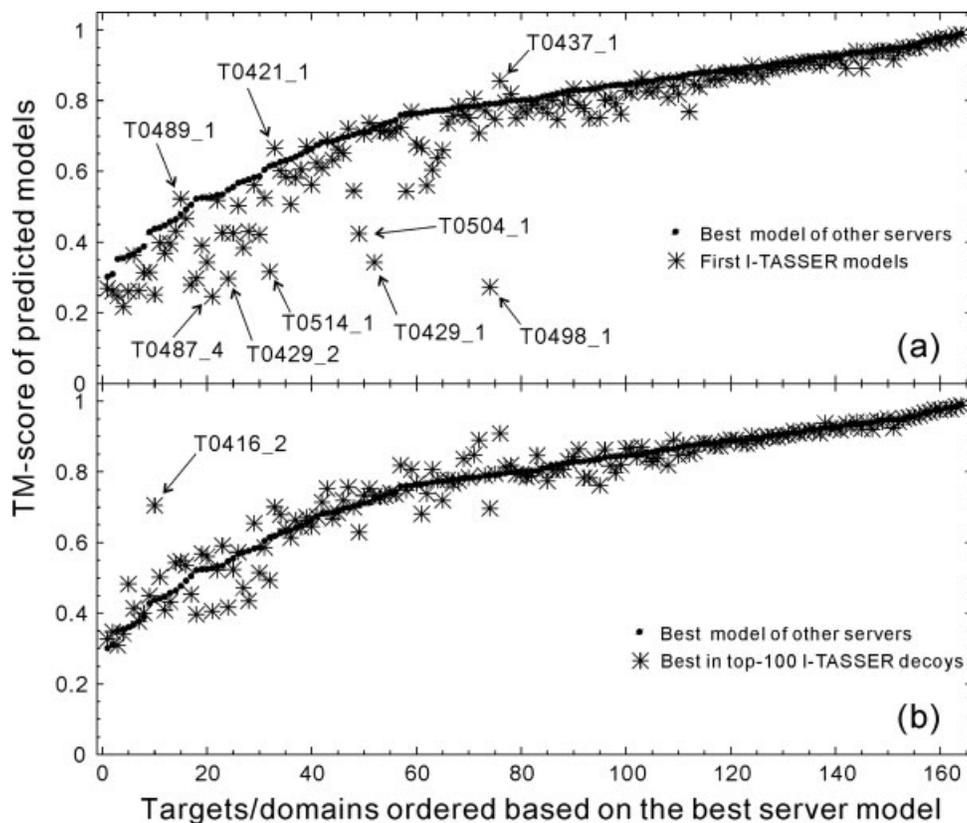


Figure 5

TM-score of the I-TASSER server prediction (stars) in control with the best model (solid spheres) predicted by other servers in CASP8. (a) The first model by I-TASSER. (b) The best in top 100 models in I-TASSER simulation.

programs from other servers in addition to LOMETS. However, the “human-won” targets are mainly in the TBM and FM categories. For HA targets, the average TM-score of the server models is actually 0.6% higher than that of human-predicted models. This shows that at least for the easy targets, human interventions are not necessary.

What went wrong?

I-TASSER fails to select non-consensus correct folds

To help highlight the problems of the I-TASSER structure modeling and especially to identify the targets which I-TASSER failed to generate good models for, we use the best model generated by the servers in CASP8 other than Zhang-Server as the reference. All models were downloaded from http://predictioncenter.gc.ucdavis.edu/download_area/CASP8/server_predictions. In Figure 5(a), we compare, for each target, the TM-score of the first model predicted by the I-TASSER server with that of the best model generated by other servers. Although there are

several targets where I-TASSER generates better models than all others, the I-TASSER models are worse than the best models from other servers for most targets in the TBM/FM categories. The average TM-score of the I-TASSER models, calculated for all 164 domains, is 0.712 versus 0.765 for the best of other servers.

In Figure 5(b), we list the best (by TM-score) of the top 100 (as ranked by SPICKER) models generated by the I-TASSER simulations with reference to the best models from other servers. These models were generated by I-TASSER but many of them were ranked low by SPICKER and not selected for submission. The average TM-score of these models is 0.765, equal to that of the best models by other servers. This data on one hand demonstrates that most of the good quality structures have been already generated in the I-TASSER simulations; on the other hand, the difference highlights a major problem of the I-TASSER pipeline: the model selection. The top 100 I-TASSER models for each target are available at <http://zhang.bioinformatics.ku.edu/casp8/decoys>; these will serve as a benchmark set for the next stage of model selection development.

I-TASSER builds models as guided by the consensus restraints from multiple threading templates. The consensus information is reinforced in the final step when the structures are clustered by SPICKER. These procedures are based on the assumption that a consensus template structure, ranked high by different scores of multiple threading programs, should be of better quality than those hit only by individual threading algorithms because there are much more ways for a threading program to pick up a wrong alignment than a right one.⁶ For some targets, this assumption does not hold, and the selection based on consensus fails to select the correct fold. This turns out to be the major reason for the failure of I-TASSER model selection, especially for most of the cases highlighted in Figure 5(a).

For example, **T0498_1** is a designed protein which was designed to have a high sequence similarity (95%) with **T0499_1**, but to have a different fold, that is, T0498_1 has a 3 α -fold while T0499_1 has an $\alpha\beta$ -fold.³¹ Among all LOMETS programs, only MUSTER²⁰ has a correct but weakly scoring hit on the template 2fs1A with a 3 α conformation and a TM-score = 0.67. However, because of the high sequence and profile similarity, the majority of the high-scoring alignments are with the $\alpha\beta$ -fold templates from 2igd, 1zxhA, 1mhxA, and 2i2yA. Thus, although I-TASSER did generate models with TM-score >0.70 in this case, the correct 3 α -fold was ranked low, and the selection preferred the incorrect $\alpha\beta$ -fold.

While T0498_1 is a special challenge for modeling and ranking which probably occurs very rarely in nature, **T0504_1** is another example of a similar ranking problem. T0504 is a three-domain protein but I-TASSER modeled T0504_1 and T0504_2 together because these regions were aligned simultaneously. T0504_3 was successfully modeled, with the first model having an RMSD = 1.77Å. The best template for T0504_1 and T0504_2 is 2g3r which is hit only by HHsearch,²⁴ with a low rank. The majority of LOMETS programs detect 2gf7A as a template, which has a similar architecture of two domains, both having a two- β -hairpin wound structure [Fig. 6(b)]. Interestingly, domains in 2gf7A swap one β -hairpin with each other, which results in a different topology from T0504 [Fig. 6(a)]. This situation is similar to oligomer domain swapping³² but the swap here occurs within a single protein chain. This may reflect a new evolutionary mechanism where oligomer domain swapping is followed by gene fusion. Correspondingly, the first I-TASSER model has a similar architecture to the target [Fig. 6(c)] but the TM-scores of both T0504_1 and T0504_2 are low because of the different orientation of the β -hairpins.

T0514_1 is another type of inaccurate I-TASSER ranking. The difference from T0499_1 and T0504_1 is that LOMETS has no strong hit on any of the templates. I-TASSER is usually good at assembling fragments from multiple weakly hit templates.¹⁵ However, in this example,

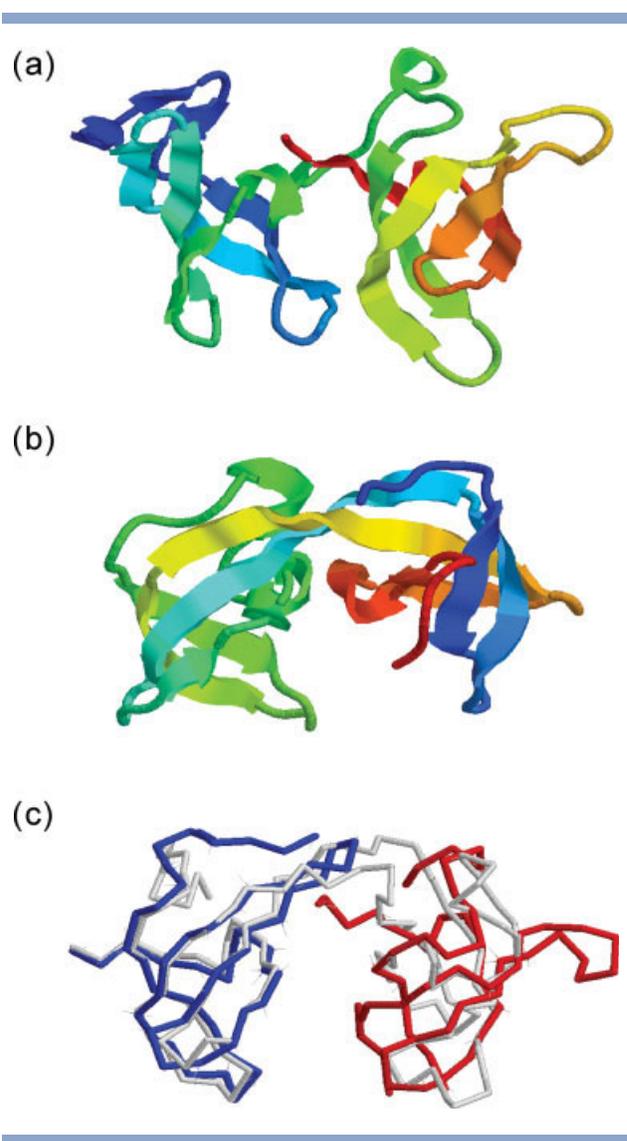


Figure 6

Structural modeling for T0504. (a) The experimental structure of the first two domains of T0504. (b) The template structure of 2gf7A detected by LOMETS which has the β -hairpin swapped and may reflect a new evolutionary mechanism from the target. (c) Superposition of the native on the I-TASSER model (white backbone). The native structures of T0504_1 and T0504_2 are in blue and red. The architecture of the model and the native is similar but with different orientation of β -hairpins.

the I-TASSER server failed to rank the best model as the first. The third submitted model has a TM-score = 0.490 while the first model is a mirror image of the third model and has a TM-score = 0.316 (see discussion below).

Problem in domain splitting

Inappropriate domain assignment is the second major reason for the failure of I-TASSER modeling. This can happen in two scenarios. The first is when each individual domain has good templates from different proteins

but the threading programs fail to detect them when whole-chain sequences are used. The difficulty in this scenario is that we do not have an efficient algorithm for domain prediction. One such case is **T0429**, which is a two-domain protein. The first domain T0429_1 has an alignment with template 2f5kA hit by HHsearch with a TM-score = 0.85, and the second domain T0429_2 has a hit from 1oi1A by MUSTER with a TM-score = 0.47. However, because of the failure of domain splitting, I-TASSER attempted to fold the target based on *ab initio* modeling, which resulted in models significantly worse than the best model by other servers which was based on the correct templates (Figure 5a).

The second scenario occurs when one of multiple domains has no strong alignment while other domains have strong templates. If we model the target as a whole chain, the final clustering will be dominated by the well-aligned regions, which will result in the weakly-aligned domains having insufficient sampling because the structures of those domains are more diverse. One such example is **T0487** which is a 685-residue target consisting of five domains. The sequences of all five domains are strongly aligned with the template 1yvuA, except for **T0487_4** which is a 87-residue domain (S178-V264) with no correct alignment with 1yvuA. Because the target is big, I-TASSER does not have sufficient sampling in this region, and the SPICKER clustering is dominated by the other well-aligned regions. As a result, the model of T0487_4 has a much worse quality than the best of other servers which obviously split the target into domains and hit the correct templates (1r4kA and 1si2A) for this domain (information obtained from the head of the models). This problem was noticed in the CASP7 experiment¹⁸ and we have attempted to split the sequence into domains and model the domains separately. However, this does not always work better than folding the whole-chain sequence because the corresponding chain connectivity restraints and interactions with partner domains are lost in the individual domain modeling. One solution to the problem may be to fold the easy domains first and then fold the remaining domains while keeping the structures of the other domains frozen.

Potential function fails to recognize mirror image fold for FM targets

The predicted distance map and contact restraints have no ability to distinguish mirror image structures because both the right model and the mirror can satisfy the restraints equally well. This is one of the problems of I-TASSER in free modeling when the models are generated from scratch and no template can be used to guide the model selection. **T0405_1** is one such example, which is the first domain (N2-E73) of a two-domain target T0405 (see Figure 7). The I-TASSER server correctly recognized the target as having two domains but incorrectly split the

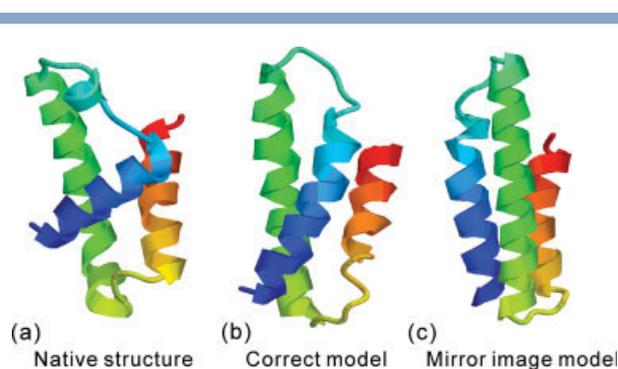


Figure 7

The I-TASSER modeling for T0405_1 (a), where the mirror image structure (c) is ranked higher than the correct model (b).

first domain as M1-L101. As expected, the accuracy of the contact predictions from LOMETS is low (11% for side-chain and 0% for C α contacts, see Table S1); but SVMSEQ predictions have an accuracy of 25% for side-chain contacts and 20% for C α contacts. The I-TASSER server generated two types of models for T0405_1 which are mirror images of each other with a distance-RMSD = 2.1 Å [Fig. 7(b,c)]. However, the incorrect mirror image was finally picked up by SPICKER [Fig. 7(c)]. There are several other big, hard targets where the mirror image structure was also ranked higher than the correct one. For example, in the above-mentioned target T0514, which is a 154-residue protein with a β -sandwich topology, I-TASSER ranks the mirror image structure as the first model and the one with the correct image as the third.

MATERIALS AND METHODS

The I-TASSER prediction pipeline includes four general steps: template identification, structure reassembly, atomic model construction, and final model selection.

Template identification

Target sequences are threaded through a non-redundant PDB structure library for identifying appropriate global-structure templates (for TBM targets) or local fragments (for FM targets). Threading is done by MUSTER,²⁰ which uses an extended sequence profile-profile alignment algorithm with the alignment score enhanced by secondary structure match, fragment structure profile, solvent accessibility, backbone torsion angle, and hydrophobic scoring matrix. The fragment structure profile refers to a frequency matrix of the template proteins which are calculated from a set of nine-residue fragments that have a similar local structure and depth to the templates.^{20,33} For hard targets, additional templates are

used that are identified by LOMETS,²¹ a local meta-threading server including FUGUE,³⁴ HHSEARCH,²⁴ PROSPECT,³⁵ PPA,¹⁵ and SP3.³³ In human prediction, we include additionally the models generated by other groups in the Server Section in the template pool. Having more threading templates is the only source of differences between Zhang and Zhang-Server predictions.

Structure assembly

Continuous fragments excised from the threading templates are used to assemble full-length models^{15,36} with unaligned loop regions built by *ab initio* modeling in a lattice system.³⁷ The structure assembly process consists of two sets of simulations.¹⁵ The first set uses the threading templates as initial structures. In the second set, the simulations start from the cluster centroids generated by SPICKER²⁶ which clusters all the trajectories from the first set of simulations. Spatial restraints, which are collected from the PDB structures hit by TM-align¹⁹ using the cluster centroids as query structures, are also incorporated in the I-TASSER simulations. The purpose of the second stage is to refine the local geometry as well as the global topology of the SPICKER centroids.

Energy force field

The structure assembly simulations (for both the threading-aligned and the *ab initio* modeled regions) are guided by a unified knowledge-based force field, which includes three components: (1) general knowledge-based statistics terms from the PDB ($C\alpha$ /side-chain correlations,³⁷ H-bonds³⁸ and hydrophobicity³⁹) (2) spatial restraints from threading templates,²¹ and (3) sequence-based contact predictions from SVMSEQ.¹⁶

The last energy term is relatively new in comparison with the force field used in the previous CASP experiment.¹⁸ SVMSEQ is a support-vector-machine (SVM)-based residue-residue contact predictor that only uses sequence information.¹⁶ It was trained using local window features (position-specific scoring matrices, secondary structure, and solvent accessibility predictions) and in-between segment features (residue separations, secondary structure of the contacting residues, and state distributions of the contacting residues). Nine sets of contact predictions are generated, which are based on three atom types ($C\alpha$, $C\beta$, and side-chain center); each atom type has three types of contact cutoffs (6, 7, and 8 Å). All nine predictions are used in I-TASSER simulation as restraints with weights proportional to their confidence.

Atomic model construction

The SPICKER cluster centroids from I-TASSER are reduced models with each residue represented by its $C\alpha$ and side-chain center. The full-atomic models are built by REMO,¹⁷ a new protocol we developed for construct-

ing full-atomic models from $C\alpha$ traces by optimizing the H-bond networks. The basic backbone fragments ($C\alpha$, C, N, O) are matched from a secondary structure specific backbone isomer library which consist of a total of 68,206 non-redundant isomers from high-resolution PDB structures. The driving force in the REMO refinement protocol includes H-bonding, clash/break-amendment, I-TASSER restraints, and the CHARMM22 potential. On the basis of a test set of 230 nonhomologous proteins, REMO has the ability of removing steric clashes while retaining a topology score (e.g., TM-score) similar to that of cluster centroids. Moreover, the H-bond network was improved in more than 80% (187/230) of test proteins by REMO.¹⁷

Model selection

The reduced models from I-TASSER are ranked based on the structure density in SPICKER clusters.²⁶ For each reduced model, atomic models from REMO are selected based on an empirical scoring function, which is equal to the sum of the number of H-bonds divided by the target length, the TM-score⁴⁰ of the model with the SPICKER cluster centroid, and the average TM-score of the model with the initial templates (used for easy target only). The weights of the empirical score have been trained in benchmark tests. The highest scoring models are finally submitted.

Multiple-domain proteins

The procedure to deal with multiple-domain proteins is similar to what we used in CASP7.¹⁸ If a segment of the target sequence with >80 residues has no aligned residues in the top two threading templates, the target is treated as a multiple domain protein, and domain boundaries are automatically assigned based on the boundaries of the large gaps. The I-TASSER simulations are run for the full chain as well as the separate domains. The final full-length models are generated by docking the models of all domains together through a quick Metropolis Monte Carlo simulation, where the simulation energy is defined as the RMSD of the domain models to the full-chain models plus the reciprocal of the number of inter-domain steric clashes. This procedure is only applied to proteins that have some domains not aligned in the top-scoring templates. If multiple-domain templates are available with all domains aligned, the whole-chain will be modeled in I-TASSER simultaneously.

CONCLUSIONS

The I-TASSER pipeline was tested in the CASP8 experiment. The success mainly comes from the fact that the algorithm manages to make use of information from multiple templates to assemble models with an optimized

knowledge-based potential³⁷ to accommodate the global and local structural packing. The multiple template information is represented in I-TASSER as consensus spatial restraints and rigid structural fragments. The consensus restraints have a similar accuracy to those from the top individual templates but cover a larger portion of the structure and a larger fraction of native contacts. The rigid structure fragments excised from the PDB template structures help reduce the entropy of the conformational search and increase the fidelity of local structures. Encouragingly, the procedure has been made fully automated and generates models with a quality close to the human predictions for at least the close homology modeling.

For the first time, the sequence-based contact predictions from machine-learning techniques¹⁶ are found helpful in both TBM and FM 3D structure assembly. In TBM, although the overall accuracy is most desirable, the key factor that determines the usefulness of the *de novo* contact predictions is the complementarity to the template-based predictions, that is, only those contacts that are novel relative to the templates are essential. The false-positive predictions in the well-aligned regions are mostly neutralized by the strong template-based restraints. However, special treatment of the false-positive predictions, for example, removing the sequence-based contacts involving the well-aligned regions while keeping those in weakly aligned or unaligned regions, may further eliminate possible side effects of the *de novo* contact predictions in TBM. Progress has also been made in atomic-level structural refinement which optimizes the hydrogen-bonding network and improves local structural packing.¹⁷

Nevertheless, one of the major issues of the current I-TASSER approach lies in the selection of correct models. This is especially the case when the best templates are hit only by a minority of threading algorithms and ranked low in the scoring function. External statistical and physics-based atomic potentials may be borrowed to deal with this issue in combination with the I-TASSER potentials and SPICKER clustering. Another related issue is the mirror image recognition for free modeling, for which chirality-dependent energy terms need to be introduced in I-TASSER. Finally, incorrect domain splitting turns out to be the major issue influencing the quality of the I-TASSER models for multiple-domain targets. As both separate domain modeling and simultaneous modeling of multiple domains have defects, that is, individual domain modeling misses the restraint information from partners while simultaneous modeling suffers from insufficient sampling for small and weakly aligned domains, one solution may be to model the domain structures in a sequential order while keeping the other domains frozen. All these issues highlighted in the CASP8 experiment will be of high priority in the development of the next generation of I-TASSER.

ACKNOWLEDGMENTS

The author thanks Drs. S. Wu, Y. Li, and A. Roy for assistance in CASP8, Dr. A. Szilagyí for reading the manuscript.

REFERENCES

1. Murzin AG, Bateman A. CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. *Proteins* 2001;Suppl 5:76–85.
2. Ginalski K, Rychlewski L. Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins* 2003;53 Suppl 6:410–417.
3. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmstrom L, Wollacott AM, Wang C, Andre I, Baker D. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* 2007;69 Suppl 8:118–128.
4. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;294:93–96.
5. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. *Nat Biotechnol* 2000;18:283–287.
6. Zhang Y. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* 2008;18:342–348.
7. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 2008;9:40.
8. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 2005;33:W244–W248.
9. Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* 2009;4:363–371.
10. Zhang Y. Protein structure prediction: when is it useful? *Curr Opin Struct Biol* 2009;19:145–155.
11. Kryshchukovych A, Fidelis K, Moult J. Progress from CASP6 to CASP7. *Proteins* 2007;69 Suppl 8:194–207.
12. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. Automated server predictions in CASP7. *Proteins* 2007;69 Suppl 8:68–82.
13. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins* 2007;69 Suppl 8:38–56.
14. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. *Proteins* 2007;69 Suppl 8:57–67.
15. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* 2007;5:17.
16. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics (Oxford England)* 2008;24:924–931.
17. 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.
18. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* 2007;69 Suppl 8:108–117.
19. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 2005;33:2302–2309.
20. Wu ST, Zhang Y. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* 2008;72:547–556.
21. Wu ST, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucl Acids Res* 2007;35:3375–3382.
22. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. *Proc Natl Acad Sci USA* 2005;102:1029–1034.

23. Xu D, Zhang Y. MVP: macromolecular visualization and processing. <http://zhang.bioinformatics.ku.edu/MVP>. 2009.
24. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 2005;21:951–960.
25. Wu ST, Zhang Y. Protein residue contact prediction by SVMSEQ and LOMETS servers. CASP8 Abstract, 2008;114.
26. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* 2004;25:865–871.
27. Rotkiewicz P, Skolnick J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J Comput Chem* 2008;29:1460–1465.
28. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, Soto CS, Goldsmith-Fischman S, Kernytzky A, Schlessinger A, Koh IY, Alexov E, Honig B. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins* 2003;53 Suppl 6:430–435.
29. Holm L, Sander C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J Mol Biol* 1991;218:183–194.
30. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. *J Mol Biol* 1994;238:777–793.
31. He Y, Chen Y, Alexander P, Bryan PN, Orban J. NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc Natl Acad Sci USA* 2008;105:14412–14417.
32. Bennett MJ, Schlunegger MP, Eisenberg D. 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci* 1995;4:2455–2468.
33. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* 2005;58:321–328.
34. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 2001;310:243–257.
35. Xu Y, Xu D. Protein threading using PROSPECT: design and evaluation. *Proteins* 2000;40:343–354.
36. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci USA* 2004;101:7594–7599.
37. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J* 2003;85:1145–1164.
38. Zhang Y, Hubner I, Arakaki A, Shakhnovich E, Skolnick J. On the origin and completeness of highly likely single domain protein structures. *Proc Natl Acad Sci USA* 2006;103:2605–2610.
39. Chen H, Zhou HX. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res* 2005;33:3193–3199.
40. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;57:702–710.