PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Prediction Report

Template-based modeling and free modeling by I-TASSER in CASP7

Yang Zhang*

Center for Bioinformatics, Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas 66047

ABSTRACT

We developed and tested the I-TASSER protein structure prediction algorithm in the CASP7 experiment, where targets are first threaded through the PDB library and continuous fragments in the threading alignments are exploited to assemble the global structure. The final models are obtained from the progressive refinements started from the last round structure clusters. A majority of the targets in the template-based modeling (TBM) category have the templates drawn closer to the native structure by more than 1 Å within the aligned regions. For the free-modeling (FM) targets, I-TASSER builds correct topology for 7/19 cases with sequence up to 155 residues long. For the first time, the automated server prediction generates models as good as the human-expert does in all the categories, which shows the robustness of the method and the potential of the application to genome-wide structure prediction. Despite the success, the accuracy of I-TASSER modeling is still dominated by the similarity of the template and target structures with a strong correlation coefficient (~0.9) between the root-mean-squared deviation (RMSD) to native of the templates and the final models. Especially, there is no high-resolution model below 2 Å for the FM targets. These problems highlight the issues that need to be addressed in the next generation of atomic-level I-TASSER development especially for the FM target modeling.

Proteins 2007; 69(Suppl 8):108–117.

Key words: CASP; I-TASSER; threading; template refinement; free modeling.

INTRODUCTION

Probably the most noteworthy effort in recent years' protein structure determination is the structure genomics that aims to obtain 3D models of all proteins by an optimized combination of experimental structure solution and computer-based structure prediction.[1–5] Two factors will dictate the success of structure genomics: Experimental structure determination of optimally selected proteins and efficient computer modeling algorithms. On the basis of 37,000 structures in the PDB library (many are redundant),[6] four million models/fold-assignments can be obtained by a simple combination of the PSI-Blast search and the comparative modeling technique.[7] Development of more sophisticated and automated computer modeling approach will dramatically enlarge the scope of modelable proteins in the structure-genomics project.[8] The critical problems/efforts in the field include the following: (1) for the sequences of strong homologies in PDB, how to build up high-accuracy structures at a resolution level useful for virtual ligand screening[9,10] and biological function inference[5,11]; (2) for the sequences with weakly/distant homologous templates, how to identify the correct templates[12,13] and how to refine the templates closer to native by computational simulations.[14] Typical to what is often found is that, the final models are closer to the templates rather than to the native structures[15,16]; (3) for the sequences without appropriate solved template structures, how to build models of correct topology/fold from scratch. Current successes of the *ab initio* modeling are limited to small proteins.[17–21] Progress along all these directions is assessed in the CASP7 experiment under the categories of high accuracy (HA), template-based modeling (TBM), and free modeling (FM), respectively.

The authors state no conflict of interest.
Grant sponsor: KU Start-Up Fund; Grant number: 06194; Grant sponsor: NFGRF; Grant number: 2302003.
*Correspondence to: Yang Zhang, Center for Bioinformatics, University of Kansas, 2030 Becker Dr, Lawrence, KS 66047. E-mail: yzhang@ku.edu
Received 15 February 2007; Revised 16 June 2007; Accepted 22 June 2007
Published online 25 September 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21702

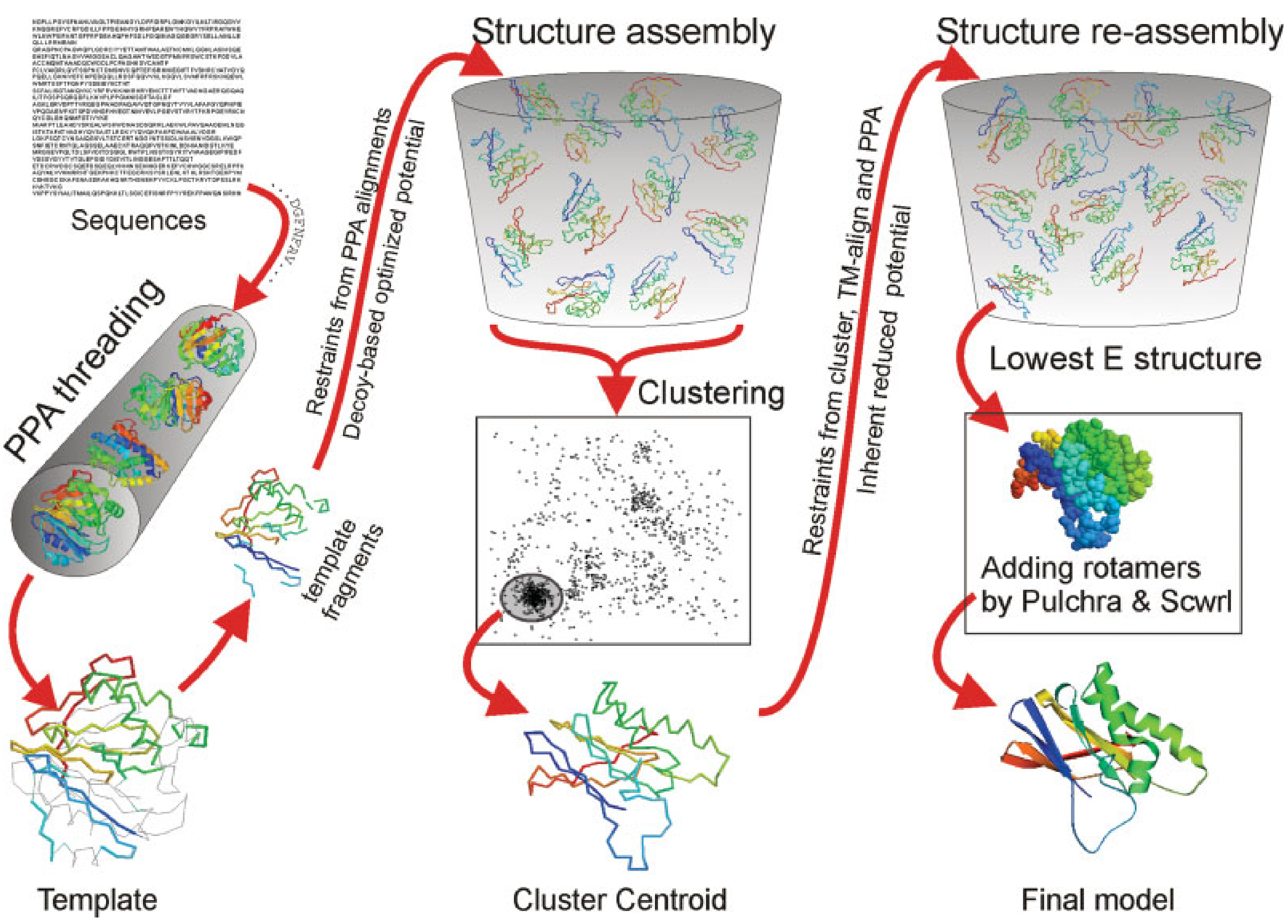

Figure 1

Flowchart of the I-TASSER protocol. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

We have developed a hierarchical approach, Threading/ASSEmbly/Refinement (TASSER), to the protein tertiary structure prediction problem.[14,22] TASSER has been tested in CASP6[23] with the threading templates generated from PROSPECTOR_3.[24] Recently, we developed a new version of structure modeler, called I-TASSER,[21] by progressively implementing the TASSER simulations, where template alignments are generated by four simple variants of the profile–profile alignment (PPA) method with different combinations of the hidden Markov model (HMM) and PSI-Blast profiles with the Needleman-Wunsch (NW) and Smith-Waterman (SW) alignment algorithms. In CASP7, we tested the I-TASSER method in both the human expert (as "Zhang") and automated server (as "Zhang-Server") sections. In this article, we will summarize the result of I-TASSER modeling of all CASP7 targets. Emphasis will be made on the template refinement for the TMB targets and the *ab initio* modeling for the small FM targets. Progress of I-TASSER compared with TASSER since CASP6 and the advantage/disadvantage of human expert over automated server prediction will be discussed.

MATERIALS AND METHODS

The I-TASSER algorithm consists of three consecutive steps of threading, fragment assembly, and iteration. A flowchart is presented in Figure 1.

Threading

PPA is a simple sequence Profile-Profile Alignment approach confined with the secondary structure matches. The alignment score between ith residue of the query sequence and jth residue of the template structure is defined as

$$\mathrm{Score}(i,j) = \sum_{k=1}^{20} F_{\mathrm{query}}(i,k) P_{\mathrm{template}}(j,k) + c_1 \delta\big(s_{\mathrm{query}}(i), s_{\mathrm{template}}(j)\big) + c_2, \quad (1)$$

where $F_{\mathrm{query}}(i,k)$ is the frequency of kth amino acid at ith position of the multiple sequence alignment searched by

PSI-Blast[25] or HMM[26] for the query sequence against a nonredundant sequence database (ftp://ftp.ncbi.nih.gov/blast/db/nr.00.tar.gz and ftp://ftp.ncbi.nih.gov/blast/db/nr.01.tar.gz); $P_{\text{template}}(j,k)$ is the summed log-odds to *k*th amino acid from the multiple sequence alignment by the PSI-Blast or HMM at *j*th position of the template sequence; $s_{\text{query}}(i)$ is the secondary structure prediction combined from PSIPRED[27] and SAM[26] for *i*th residue of the query sequence; and $s_{\text{template}}(j)$ is the secondary structure assignment by DSSP[28] for *j*th residue of the template. The combination of PSIPRED and SAM is done by summing up the raw probabilities predicted by these two programs on the helix/strand/coil states and then selecting the state of the highest probability which is followed by the smoothing of the singular secondary structure states along the sequence. The NW[29] or SW[30] dynamic programming algorithm is used to identify the best match between query and template sequences. The four parameters, c_1, c_2 in Eq. (1), the gap opening penalty (c_3), and the gap extension penalty (c_4) are decided by trial and error on the ProSup benchmark.[31] Depending on the profiles generated from PSI-Blast or HMM search and the alignment search by the NW global or SW local dynamic programming algorithms, four complementary PPA threading alignments are used in the consequent I-TASSER assembly. The target sequences will be automatically categorized by the significance of the PPA alignments: An Easy target is defined when at least two PPA alignments have a *Z*-score higher than 8; if there is no alignment with a *Z*-score > 7, the target will be defined as a Hard target; others will be Medium targets.

Structure assembly simulation

On the basis of PPA threading alignments, target sequences are divided into aligned and unaligned regions. The fragments in the aligned regions are directly excised from the template structures and allowed to rotate and translate in an off-lattice system.[14] The unaligned regions are modeled by *ab initio* simulations in a cubic lattice system of grid size 0.87 Å.[20] The global topology is decided by the relative reorientation of the continuous fragments while the on-lattice modeling serves as the linkage of the rigid-body fragment movements. Protein conformations are represented by a trace of C_α atoms and side-chain centers of mass (SC). The force field consists of a variety of knowledge-based energy terms describing SC pair-wise interactions and short-range C_α correlations,[20,32] propensity to the consensus secondary structures predictions from PSIPRED[27] and SAM,[26] residue-based solvent accessibility by neural network training,[21,33] secondary structure specific backbone hydrogen-bonding,[34] and the consensus SC contact and C_α distance constraints extracted from the multiple threading alignments. Weighting balances between the energy terms are trained in the Easy/Medium/Hard categories separately by the maximization of the total energy-TM-score correlation based on an ensemble of continuously distributed structure decoys.[20] The structure assembly procedure is driven by a modified replica-exchanged Monte Carlo simulation[35,36] and the trajectories in low temperature replicas are clustered by SPICKER.[37] The cluster centroids are obtained by averaging the coordinates of all clustered decoys and are ranked based on the structure density.

Iteration

Starting from the selected SPICKER cluster centroids, we implement the TASSER assembly refinement simulation again. While the inherent I-TASSER potential keeps unchanged in the second run, the external constraints are pooled from the initial high-confident restraints from PPAs, the restraints taken from the cluster centroid structures, and the restraints from the PDB structures searched by the structural alignment program TM-align.[38] The purpose of the iteration is to remove the steric clashes of cluster centroids and to refine the topology as well.[21] The conformations of the lowest energy in the second round are selected. Finally, Pulchra[39] is used to add backbone atoms (N, C, O) and Scwrl_3.0[40] to build side-chain rotamers.

Multiple domain proteins

If any region with >80 residues has no aligned residues in at least two strong PPA alignments of *Z*-score $>$ 8, the target will be judged as a multiple domain protein and domain boundaries are automatically assigned based on the borders of the large gaps. As a defect, this multiple-domain assignment does not include the cases which have all domains simultaneously aligned. I-TASSER simulations will be run for the full chain as well as the separate domains. The final full-length models are generated by docking the model of domains together. The domain docking is performed by a quick Metropolis Monte Carlo simulation where the energy is defined as the RMSD of domain models to the full-chain model plus the reciprocal of the number of steric clashes between domains. The goal of the docking is to find the domain orientation that is closest to the I-TASSER full-chain model and has the minimum steric clashes. The final models docked from I-TASSER domains are submitted to CASP7.

Predictions in human section

The above I-TASSER modeling procedure is fully automated and used for the predictions in the server sections. The human section prediction uses essentially the same procedure, except for the following differences: (1) the domain border assignment has been made based on visual view of the 1D threading sequence alignments and 3D template structures, which are further adjusted by the CASP7 domain server predictions from Robetta-Ginzu[41] and MaOPUS-DOM; (2) for the hard targets that have no strong PPA hit with a *Z*-score > 7, additional alignments from the

 DOI 10.1002/prot

Table I
Average Results of I-TASSER Predictions in Both Human and Server Sections

				Best template		First model			Best model			Constraints	
	Type	N_{target}	Size	$R_{ali}{}^{a}$/Fra^{b} (%)	TM	$R_{ali}{}^{a}$	$R_{all}{}^{c}$	TM	$R_{ali}{}^{a}$	$R_{all}{}^{c}$	TM	Ac^{d}/Cov^{e} (%)	Er^{f}
Zhang-Server	TBM	105	161	5.0/90	0.659	4.0	4.7	0.729	3.4	3.9	0.750	0.40/156	1.5
	FM	19	131	13.5/81	0.211	12.3	13.2	0.302	10.2	10.9	0.364	0.20/92	3.2
	ALL	124	157	6.3/89	0.591	5.3	6.0	0.664	4.5	5.0	0.691	0.38/146	1.8
Zhang	TBM	105	161	4.9/91	0.670	3.8	4.4	0.740	3.3	3.8	0.761	0.42/159	1.4
	FM	19	131	13.2/82	0.239	11.8	12.7	0.341	9.5	0.3	0.378	0.18/130	3.0
	ALL	124	157	6.2/90	0.603	5.0	5.7	0.679	4.3	4.8	0.702	0.39/151	1.7

[a] R_{ali}, RMSD (in Å) to native in the threading aligned regions.
[b] Fra, Fraction of the aligned residues relative to the query sequence.
[c] R_{all}, RMSD (in Å) to native in full-length.
[d] Ac, Accuracy of the predicted contacts.
[e] Cov, Number of predicted contacts divided by the number of native contacts.
[f] Er, Error (in Å) of the best in up to four predicted long-range distance restraints for each C_α pair.

CASP7 servers, including FUGUE,[42] HHpred,[43] mGenThreader,[44] and SP3,[45] are exploited as I-TASSER starting structures; (3) I-TASSER simulations run within a relatively longer CPU time in the human section.

RESULTS

Summary

Ninety-six effective targets in CASP7 have been split into 124 domains by the assessors which include 28 HA-TBM, 77 TBM, 4 TBM/FM, 15 FM, and 1 decoration targets. For conciseness, we will divide our analysis in two big categories of TBM (including HA-TBM and TBM) and FM (including TBM/FM and FM).

In Table I, we present a summary of the average performance of Zhang-Server and Zhang compared with the best threading templates used by I-TASSER. Column 5 is the average RMSD and the alignment coverage of the best threading template in different categories. Here, the best template refers to the template of the highest TM-score to the native structure among all the templates exploited by I-TASSER. It is usually worse than the real best template by the structural alignment in the PDB library, identification of which needs the native structure information.[38] Obviously, the PPA threading identified much better alignments for the TBM targets than that for the FM targets. On average, the PPA alignments have a RMSD 5.0 Å over 90% aligned regions for the TBM targets and a TM-score 0.66. For the FM targets, the templates have an average RMSD 13.5 Å in 81% aligned regions. The average TM-score (0.21) is close to that expected for the random structure matches (0.17),[46] understandable because by definition there is no appropriate templates in PDB for the FM targets. Overall, the incorporation of the CASP7 servers as taken in the human prediction results in a slightly better set of threading alignments in both TBM and FM categories. Here, many TBM targets are also categorized as Medium/Hard targets by the PPA system and the threading alignments from the CASP7 servers are therefore exploited. The average TM-score of all templates increases from 0.591 to 0.603 (by 2%).

Column 7 is the average RMSD to native of the first I-TASSER models calculated in the same aligned region as in templates. The RMSD decreases (~1 Å) compared with the templates in these regions therefore reflects the improvement purely by the I-TASSER reorientation of the secondary structure fragments. It should be mentioned that I-TASSER does not attempt to "re-tune" the alignments because the local fragments are kept rigid during the simulations. The fragment repacking is driven by the inherent I-TASSER force field and the external consensus restraints. The columns 9 and 12 show the TM-score of the first and the best I-TASSER models. On average, I-TASSER reassembly results in a TM-score increase by ~14% in the TBM category. On the basis of the previous statistics,[23] a simple loop connection can lead to a TM-score increase of 3.5% because of the length elongation. Therefore, about 10% of the TM-score increase may be due to the topology improvements. For the FM targets, the TM-score increase is about 70%, more significant than that for the TBM targets, since the low TM-score templates have much more space for improvement. In contrast, the RMSD improvement from 13 Å to 10–12 Å for the FM targets sounds marginal, partially because RMSD is not an appropriate quality for distinguishing the topology in this range of accuracy.[46]

Column 13 is the consensus contact constraints collected from the PPA threading alignments (or PPA plus CASP7 threading servers for the Medium/Hard targets in the human predictions). For the Easy/Medium/Hard targets, top 20/30/50 templates are employed with a contact occurring frequency cutoff of 0.2/0.1/0.1. Because of the differences in the alignment quality, the average accuracy and coverage of contact restraints in TBM is much higher than that in the FM category. Even for the FM targets, the contact is still much better than the random prediction. (Wu ST, Zhang Y. Could the sequence-based

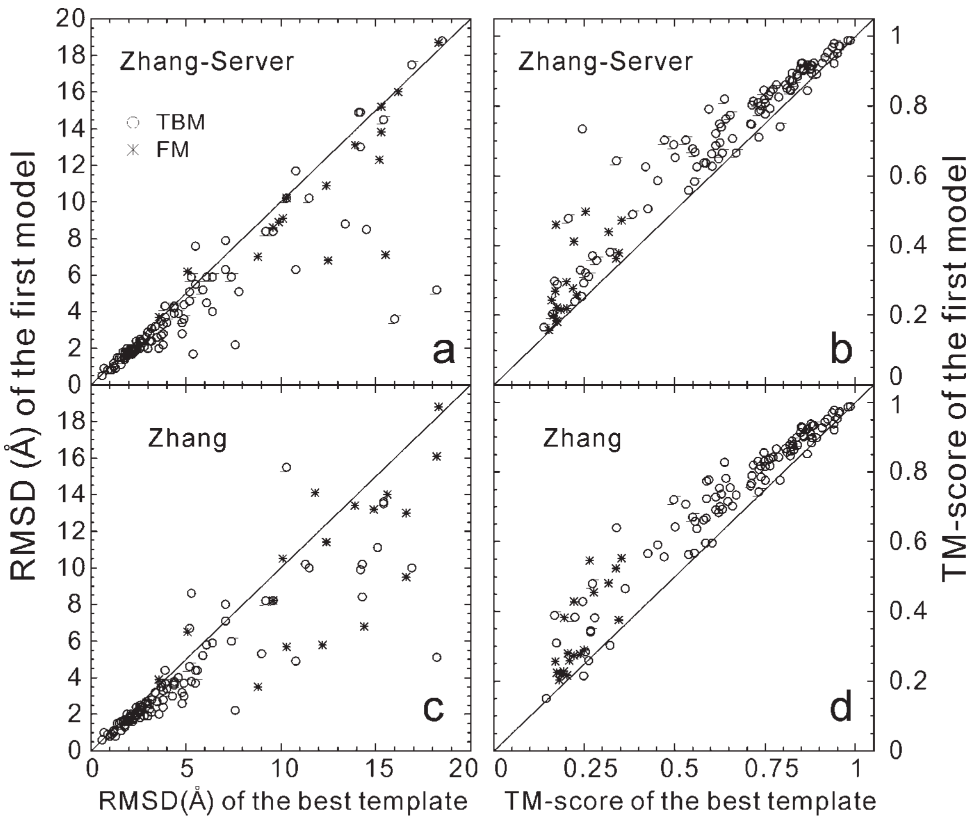

Figure 2

Comparison of the first predicted models by human ("Zhang") and server ("Zhang-Server") with respect to the best exploited templates. The RMSD is calculated in the same set of aligned residues. The TM-score is calculated in the aligned regions for the templates and in full-length for the models.

contact prediction be useful for protein tertiary structure modeling? Submitted for publication 2007.) Having in mind that a set of contact prediction with an average accuracy higher than 0.22 will be helpful for *ab initio* MC simulations to drive the topology at the correct direction,[20] it is estimated that in about half of the FM cases the employment of the restraint prediction should be better than not using them in the *ab initio* modeling. It should be mentioned that the purpose of the contact collections is to provide helpful constraints for the I-TASSER simulation rather than to generate the most accurate contact prediction. Certainly, if we collect the contacts only from the most confident templates and based on a higher frequency cutoff, the accuracy of the contacts may be higher and the coverage will be lower. But we found the current setting of the template number and cutoff parameters work the best for I-TASSER in our benchmark test. For each of the 10 residues, we generate up to four distance predictions for the long-range C_α pairs (with $|i-j| > 6$). Column 14 shows the average difference between the native C_α distances and the best predicted distances. Obviously, the distance map prediction of TBM is again more accurate than that of the FM targets.

In Figure 2, we present the comparison of the first I-TASSER models and the best threading templates for both server and human predictions. There is a consistent improvement of final models over templates based on RMSD and TM-score. There is no systematic difference in template refinements with regard to the targets from TMB or FM targets and to the models by Zhang or Zhang-Server. One notable exception is T0258 at Figure 2(c), where the RMSD of the final model by human is 3.3 Å worse than the best template (from 5.3 Å to 8.6 Å). The main reason is that our human prediction combines the threading alignments from the CASP7 servers with wrong templates for the target although our in-house PPA threading programs hit the best template of 2a2pA. The mixture of bad server templates results in the biggest cluster having a wrong orientation at the C-terminal, although the third human model has a correct topology of full-length RMSD 5.3 Å. There are also some cases where the big differences of RMSD in templates and models may not be entirely due to the structural topology improvement. In T0347_1, for example, RMSD of the template is reduced from 18.2 Å to 5.2 Å mainly because the misorientated tails in the template has been corrected by the I-

DOI 10.1002/prot

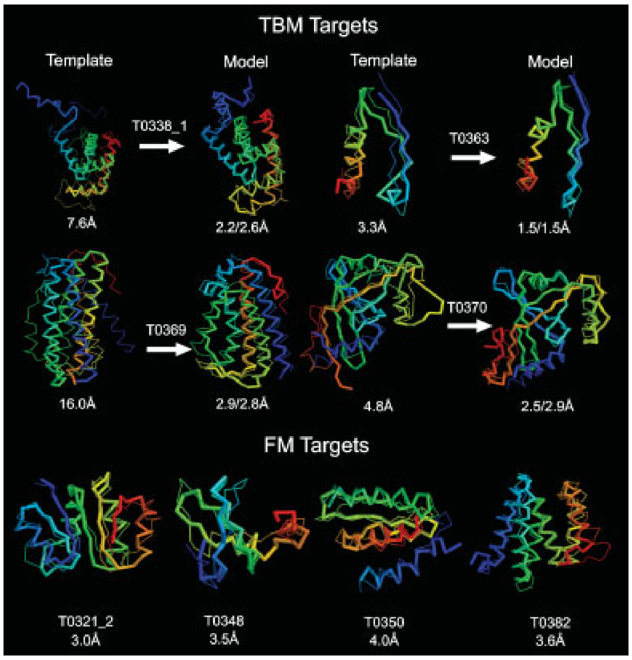

Figure 3

Representative examples for the TBM (upper panel) and FM (lower panel) targets. The thin lines represent the backbone of the experimental structures and the thick lines are the threading templates or the final models. The two number under the TBM models are the RMSD to native in the threading aligned regions and the RMSD of the full-length. Blue to red runs from N- to C-terminals. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

TASSER reassembly. But the core region does not change much in the I-TASSER modeling and the overall TM-score increases only by 0.12 in this case. Because the templates from the CASP7 servers are sometime better than our in-house PPA templates, the RMSD improvement over the templates in the human prediction appears less dramatic in some of those cases [see Fig. 2(c)].

Examples

In the upper panel of Figure 3, we show four representative examples where I-TASSER successfully refines the templates from high RMSD (3.3–16 Å) to low RMSD (1.5–3 Å). In all these cases, the consensus contact predictions have a high accuracy and coverage, that is T0338_1 with 0.5/142%, T0363 with 0.41/162%, T0369 with 0.36/198%, T0370 with 0.43/173%. The consensus restraints combined with the optimized I-TASSER inherent potential serve as the major driven force for the refinement of the templates. In all of the four cases, the accuracy and coverage of the constraints are higher than that extracted from the best individual template (data not shown), which helps to refine the loops and tails and sometime the global topology such as T0369.

In the lower panel of the Figure 3, we also show four FM examples where I-TASSER builds models of correct topology with a RMSD of 3–4 Å. Figure 4 shows a more detailed analysis of the typical example of T0382. It is a new fold protein from *Rhodopseudomonas palustris* CGA009 crystallized by the structure genomics project.[47] The topology of T0382 consists of six joggled α-helixes. The left panel of the Figure 4 shows the top five templates hit by the multiple threading programs used by I-TASSER, all having correct local second structure elements but incorrect global topologies with the best RMSD of 9.3 Å from 1xm9A1 (TM-score = 0.28). Our contact prediction program generates 148 side-chain contacts with 37 contacts correct (accuracy 25%). The average error of the best predicted C_α distances is 2.2 Å. I-TASSER cuts the fragments from the template alignments and reassembles the topology under the guidance of the predicted restraints and the inherent potential, which results in a model of full-length RMSD 3.6 Å and TM-score 0.66 (right panel of Fig. 4). The correlation of I-TASSER energy and the RMSD of the structure decoys is 0.72 which demonstrates the consistency of the external restraints and the inherent force field.

Human versus server predictions

The data in Table I have shown that the overall performance of our human predictions is slightly better than the automated server prediction. The improvement mainly occurs in the FM category where the average TM-score of the first model of the human prediction increases from 0.302 to 0.341 by 13% compared with the server. The increase of TBM targets is modest from 0.729 to 0.740 by 1.5%, which lead to an overall TM-score increase by 2.3% (0.664 to 0.679) for the first model. The overall increase of the best in top-five models for all targets (1.6%) is lower than that of the first model (2.3%), which indicates that the employment of multiple CASP7 servers tends to improve the ranking of the model rather than the best topology.

In Figure 5, we present a detailed comparison of the human versus server predictions for the first model. Similar to the tendency of Table I, for the high quality models (mainly TBM targets), for example, the models of RMSD < 5 Å or TM-score < 0.75, there is no notable difference between human and server. But for the hard targets, there is a tendency that the human prediction generates more models with better scores than the server.

There are several reasons for the human prediction outperforming the server prediction. (1) For hard targets when PPA programs have no confident hit, we exploited multiple threading templates from the CASP7 servers. Figure 6(a) represents one example where a better template 1kk1A hit by HHsearch[43] has been exploited by the human prediction, which results in a TM-score increase from 0.36 to 0.45. (2) Human visual view of the multiple threading alignments and the template tertiary structures usually leads to a better domain parser than that by the

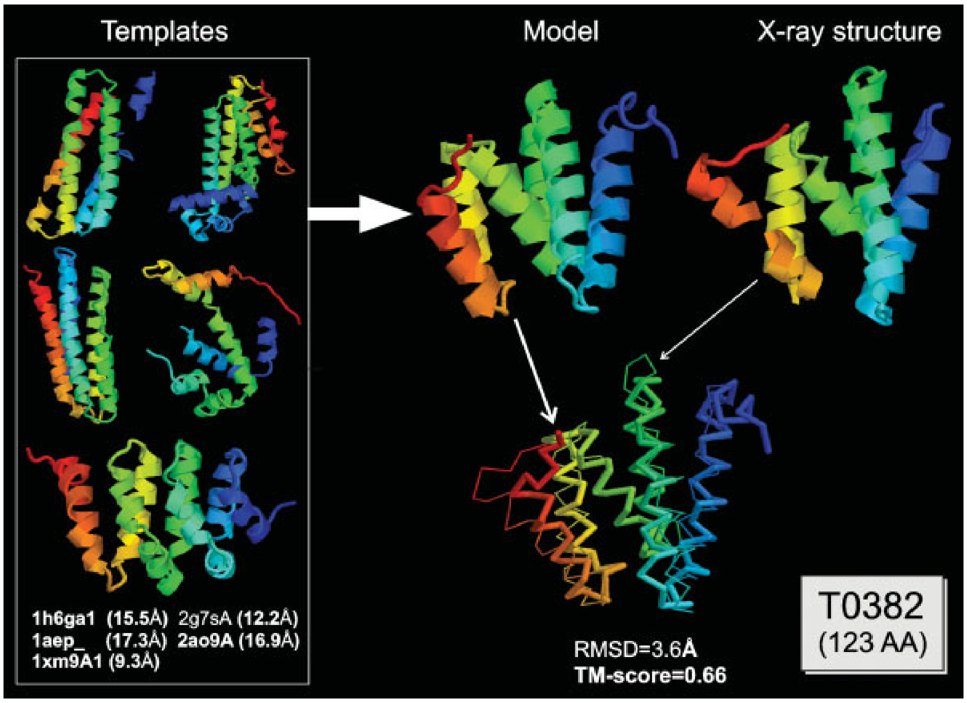

Figure 4

Structure comparison of the threading templates, the final model, and the experimental structures for the target T0382. Blue to red runs from N- to C-terminals.

simple domain assignment procedure used by server (see section "Materials and Methods"). For example, T0289 is a two-domain protein and PPA threading that hit both domains of 2bconA with high *Z*-scores. The server prediction fails to split the domains based on the sequence alignments and therefore folds the entire chain together. In the human prediction, by viewing the template structures, we correctly split the target into two domains at Residue ILE224 and fold the domains separately. As a result, the quality of both domains has been improved since I-TASSER tends to handle better the simulation of small single proteins partially due to the conformational entropy reduction.[21] The TM-score of the final human prediction increases from 0.68 to 0.7 for T0289_1 and from 0.37 to 0.51 for T0289_2. Figure 6(b) is the structure superposition for T0289_2. (3) The human prediction can benefit from the longer CPU running time. Because of the current limited computing power at our lab, most of the hard targets in the server prediction did not run sufficient trajectories as in benchmark. Figure 6(c) shows an example of T0382 where I-TASSER needs to reconstruct the models from wrong templates. In the server prediction, the average energy of the largest cluster is −3024 kT. But by a longer run, the human simulation reached a cluster of average energy of −3264 kT, which results in a TM-score improvement from 0.54 to 0.66.

There are also some cases where the human prediction can be worse than the server prediction. Figure 6(d) is a typical example from the T0358 where the server prediction generate better models because our in-house PPA threading programs consistently hit the correct template from 2a2pA but with weak *Z*-scores. In the human prediction, we exploit the multiple templates from the CASP7 servers that actually have hit worse templates. The incorporation of incorrect templates can result in

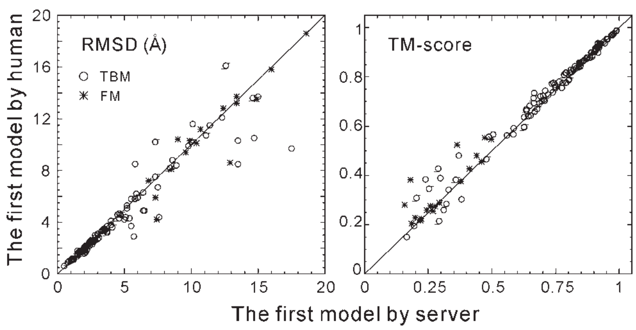

Figure 5

Comparison of the human and the server predictions for all 124 domains/ targets.

 DOI 10.1002/prot

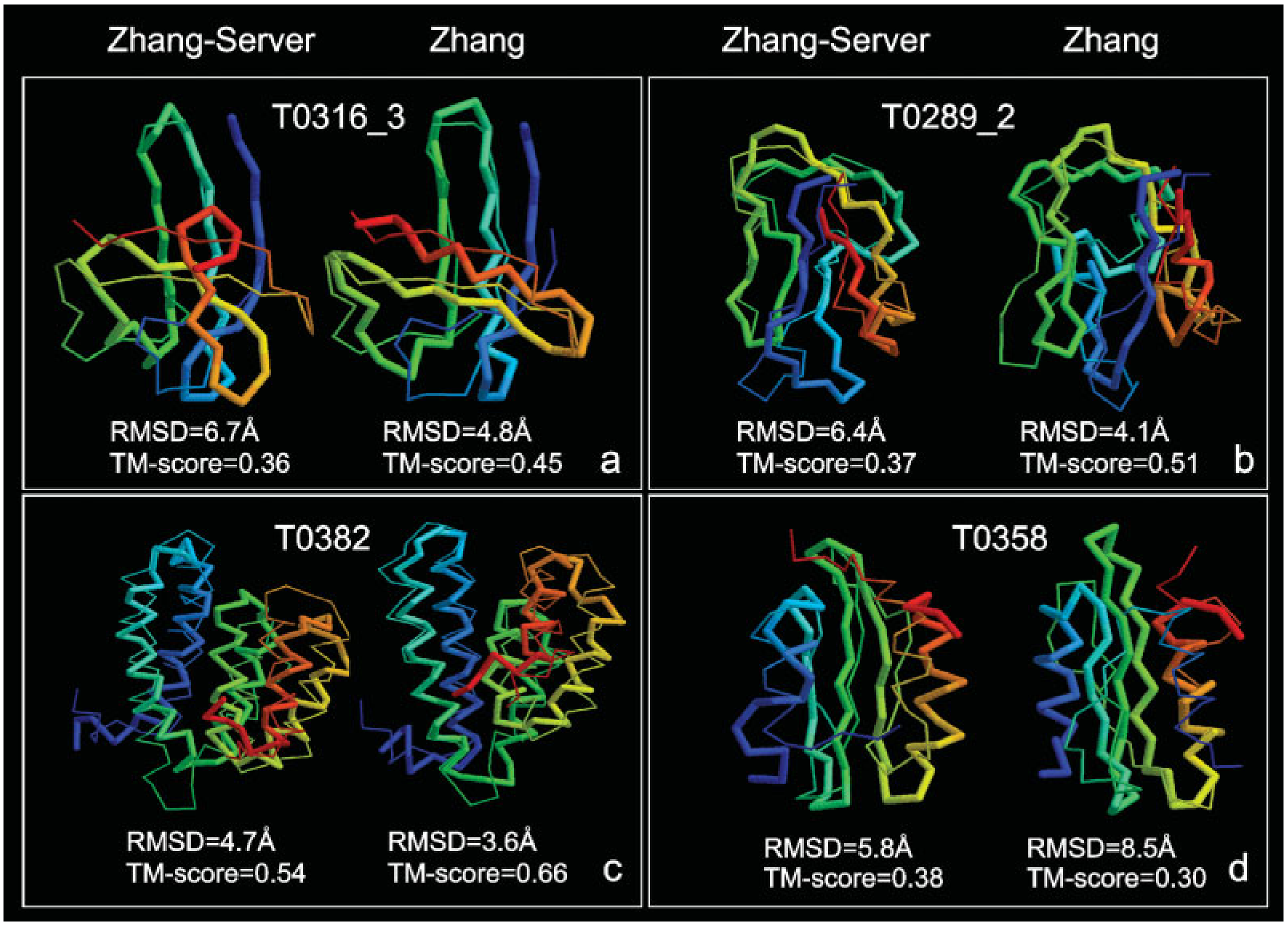

Figure 6

*The examples where the human predictions generate better (**a–c**) and worse (**d**) models than that by the server predictions. The thin lines represent the backbone of the experimental structures and the thick lines are the final models. Blue to red runs from N- to C-terminals. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

worse constraint data and therefore reduce the performance of I-TASSER modeling. For the first model, the server prediction has a RMSD of 5.8 Å but the human has a RMSD 8.5 Å with a disoriented C-terminal.

DISCUSSION

We have developed and tested a new version of I-TASSER algorithm at the CASP7 experiment. Compared with the original version of TASSER,[14,22,23] new components include: (1) a new set of secondary structure confined PPA threading programs are developed; (2) new energy terms including neural network solvent accessibility predictions are incorporated and reparameterized on the basis of structure decoys in different categories; (3) a two-round progressive assembly simulation is developed for removing structure clashes and refining models.

What went right?

One of the most important highlights of the I-TASSER simulations is the ability of template refinement. Overall, about 1 Å RMSD reduction can be obtained within the aligned residues compared with the best threading alignment. The full-length TM-score increases by about 14% where about 10% is probably because of the topology reorientation of the secondary structure fragments and the rest may be due to the size elongation by filling the unaligned gaps. One of the major contributions to the structure improvement is the employment of consensus spatial constraints of multiple templates which is usually of higher accuracy than that from individual templates. Our benchmark test[21] shows that the combination of four PPA threading alignments performs better than that with the best PPA-I program from PSI-Blast profile and NW global alignment. The CASP7 results demonstrate that including other threading resources can result in further improvements, especially for the hard targets. The second driven force for the structure improvement is the optimized I-TASSER inherent potential. The off-line analysis shows that in almost all the successful cases there is a strong correlation between the inherent potential and the external restraints. For the less successful targets, this correlation is weak. Finally, the requirement for the chain connectivity also helps to improve the reassembly of the fragments from some structurally unphysical threading alignments. Overall, in comparison with the physics based structural modeling approaches, the success of the I-TASSER method is largely due to the successful utilization of the evolutionary relations of the target and the solved proteins where both the spatial constraints and the knowledge-based reduced potential of I-TASSER come from the target–template alignments and the statistics of the PDB structures.

The procedures of our human and the server predictions are essentially the same. If ignoring the minor effects from taking the multiple CASP7 servers for the hard targets, the overall performance of Zhang and Zhang-Server is almost indistinguishable as shown in Table I and Figure 5. One goal of the I-TASSER development is to release the heavy human intervention from the structure modeling procedure. The automatization and robustness of the algorithms are particularly important for the application to the large-scale automated structure predictions. The I-TASSER server is freely available at our website: http://zhang.bioinformatics.ku.edu/I-TASSER.

What went wrong?

Among the 19 free-modeling targets, I-TASSER generates correct topology for seven of them (about 1/3) up to 155 residues long with RMSD < 6.5 Å or TM-score $>$ 0.5. Despite the success on some of the FM targets, the overall quality of I-TASSER modeling is strongly correlated by the quality of the templates with a Pearson correlation coefficient of 0.89 for RMSD and 0.95 for TM-score in the server section (from [Fig. 2(a,b)]). For several small FM proteins below 120 residues, I-TASSER still failed to generate the correct topology. The failure often occurs when we tried to fold the small hard domains together with another big strong-hit domain. The structure phase space of small domains has not been sufficiently explored because most of the Monte Carlo movements are devoted to the bigger domain regions in these cases. Moreover, when the multiple domain structure decoys are clustered, the structures of the big domain part will dominate the RMSD matrix and therefore the lowest free-energy state of the small domains can not be identified by normal SPICKER program.[37] So it will be helpful to develop robust domain parser programs to split the domains correctly and fold the individual domain separately. Second, the more essential reason for the failure is that the I-TASSER potential and the external restraints cannot provide appropriate long-range interaction information for the FM targets. We are in the process of examining the possibility of exploiting long-range contact predictions from other resources (Wu ST, Zhang Y. Could the sequence-based contact prediction be useful for protein tertiary structure modeling? Submitted for publication 2007.).

Another issue of our modeling is the suboptimal secondary structures for several small hard proteins, for example T0304. The main reason is that current I-TASSER modeling is based on a reduced C_α and side-chain center of mass model where the hydrogen-binding is only considered approximately based on the backbone C_α atoms. The other atoms are added by external programs of Pulchra[39] and Scwrl[40] after the simulation and clustering. While for the hard targets, the goal of I-TASSER is to generate correct topology, no effort has been made for the optimization of the hydrogen-bonding network except for that of backbone C_α atoms.[34] The development of an atomic I-TASSER, which embodies all heavy atoms in the modeling and aims to optimize the hydrogen-bonding of both backbone and side-chain atoms, is under progress.

ACKNOWLEDGMENTS

The CASP7 calculations of our lab have been performed on the KUCB and KU-ITTC clusters where help from Drs. V. Frost, A. Hock, A. Tovchigrechko, and I. Vakser are gratefully acknowledged. We also thank Dr. J. Skolnick for general supports, Drs. S. Lorenzen and S. Wu for help.

REFERENCES

1. Terwilliger TC, Waldo G, Peat TS, Newman JM, Chu K, Berendzen J. Class-directed structure determination: foundation for a protein structure initiative. Protein Sci 1998;7:1851–1856.
2. Brenner SE, Levitt M. Expectations from structural genomics. Protein Sci 2000;9:197–200.
3. Stevens RC, Yokoyama S, Wilson IA. Global efforts in structural genomics. Science 2001;294:89–92.
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. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
7. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2006;34:D291–D295.
8. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–566.
9. Klebe G. Virtual ligand screening: strategies, perspectives and limitations. Drug Discov Today 2006;11:580–594.

 DOI 10.1002/prot

10. Hubbard RE, editor. Structure-based drug discovery, 1st ed. Royal Society of Chemistry; Cambridge, UK, 2006.
11. Pal D, Eisenberg D. Inference of protein function from protein structure. Structure 2005;13:121–130.
12. 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.
13. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
14. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
15. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53(Suppl 6):352–368.
16. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005;61(Suppl 7):27–45.
17. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
18. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
19. Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci USA 1999;96:5482–5485.
20. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003;85:1145–1164.
21. Wu ST, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007;5:17.
22. Zhang Y, Skolnick J. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophys J 2004; 87:2647–2655.
23. Zhang Y, Arakaki A, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 2005;61(Suppl 7):91–98.
24. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Protein 2004;56:502–518.
25. 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.
26. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
27. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
28. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
29. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
30. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
31. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. J Mol Biol 2000;297:1003–1013.
32. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci 1997;6:676–688.
33. Chen H, Zhou HX. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 2005;33:3193–3199.
34. 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.
35. Swendsen RH, Wang JS. Replica Monte Carlo simulation of spin glasses. Phys Rev Lett 1986;57:2607–2609.
36. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002;48:192–201.
37. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.
38. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
39. Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL, III. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 2000;41:86–97.
40. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001–2014.
41. Kim DE, Chivian D, Malmstrom L, Baker D. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 2005;61(Suppl 7):193–200.
42. 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.
43. Soding J. Protein homology detection by HMM–HMM comparison. Bioinformatics 2005;21:951–960.
44. McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 2003;19:874–881.
45. 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.
46. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
47. Cymborowski MT, Evdokimova E, Kagan O, Chruszcz M, Savchenko A, Edwards A, Minor W, Joachimiak A. Crystal structure of the protein RPA1889 from *Rhodopseudomonas palustris* CGA009, doi 10.2210/pdb2i9c/pdb.