TASSER: An Automated Method for the Prediction of Protein Tertiary Structures in CASP6

Yang Zhang, Adrian K. Arakaki, and Jeffrey Skolnick*
Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York

ABSTRACT **The recently developed TASSER (Threading/ASSembly/Refinement) method is applied to predict the tertiary structures of all CASP6 targets. TASSER is a hierarchical approach that consists of template identification by the threading program PROSPECTOR_3, followed by tertiary structure assembly via rearranging continuous template fragments. Assembly occurs using parallel hyperbolic Monte Carlo sampling under the guide of an optimized, reduced force field that includes knowledge-based statistical potentials and spatial restraints extracted from threading alignments. Models are automatically selected from the Monte Carlo trajectories in the low-temperature replicas using the clustering program SPICKER. For all 90 CASP targets/domains, PROSPECTOR_3 generates initial alignments with an average root-mean-square deviation (RMSD) to native of 8.4 Å with 79% coverage. After TASSER reassembly, the average RMSD decreases to 5.4 Å over the same aligned residues; the overall cumulative TM-score increases from 39.44 to 52.53. Despite significant improvements over the PROSPECTOR_3 template alignment observed in all target categories, the overall quality of the final models is essentially dictated by the quality of threading templates: The average TM-scores of TASSER models in the three categories are, respectively, 0.79 [comparative modeling (CM), 43 targets/domains], 0.47 [fold recognition (FR), 37 targets/domains], and 0.30 [new fold (NF), 10 targets/domains]. This highlights the need to develop novel (or improved) approaches to identify very distant targets as well as better NF algorithms. Proteins 2005;Suppl 7:91–98. © 2005 Wiley-Liss, Inc.**

Key words: comparative modeling; threading; ab initio prediction; TASSER; PROSPECTOR_3

INTRODUCTION

Since their start in 1994, the biennial CASP experiments have stimulated progress in protein structure prediction,[1,2] with the following general trends apparent in the recent CASP experiments. First, despite considerable progress on the ab initio/new fold (NF) predictions,[3–5] comparative modeling (CM),[6] and threading/fold-recognition (FR)[7,8] remain the only methods that provide reliable and accurate models. Second, generating models that are closer to native than the template alignments remains a challenging problem for most structural refinement approaches.[9] Third, for the template-based modeling (including CM and FR), metapredictions,[10,11] which combine consensus information from different algorithms/servers, consistently outperform the predictions from individual algorithms/servers. Finally, human-expert knowledge combined with biochemical information (function, mutations, catalytic residues, etc.) could be helpful in both structural assembly and template/model selection.[12,13]

On the other hand, the rapidly increasing gap between the number of known sequences and known structures creates the crucial need to develop robust, automated computational methods for proteome-scale structure predictions.[14,15] Despite their advantages, current metaserver approaches need to coordinate and exploit the computational resources of different laboratories. This makes the automation of large-scale protein structure prediction difficult to achieve because of differences in the available computational resources among different laboratories and the difficulty in collecting large-scale predictions from disparate groups. Similarly, it is not feasible to apply human-expert based approaches on a proteome scale.

Recently, we developed a new methodology, Threading/ASSembly/Refinement (TASSER),[16] for automated tertiary structure prediction, that generates full-length models by rearranging continuous fragments identified by our threading algorithm PROSPECTOR_3.[17] The method was tested on large-scale benchmarks, with templates generated by both threading[16,18] and structural alignments.[19] For weakly/nonhomologous proteins, about two thirds of single-domain proteins could be folded. Often, the final models are considerably closer to native than the templates. Since the methodology employs templates selected from our in-house PROSPECTOR_3 program, TASSER has been used for the genome-scale protein structure modeling on *Escherichia coli* and other genomes.[16,20] In this round of CASP, we implemented TASSER for all targets in a similar way as we did in the benchmarks.[16,18]

Grant sponsor: National Institutes of Health, Division of General Medical Sciences; Grant numbers: GM-37408 and GM-48835.

*Correspondence to: Jeffrey Skolnick, Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington Street, Buffalo, NY 14203. E-mail: skolnick@buffalo.edu

Received 29 March 2005; Accepted 22 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20724

This article was originally published online as an accepted preprint. The “Published Online” date corresponds to te preprint version.

Special emphasis is put on the comparison of final models and threading templates. The results obtained by the automated implementation of our method in CASP6 therefore allow an independent assessment of the quality of the models generated by TASSER.

METHODS AND MATERIALS

Since TASSER methodology has been described previously,[16–18,21] here we just outline the essentials.

For a given target, we first thread the sequence through a representative template library (at a 35% pairwise sequence identity cutoff) of the Protein Data Bank (PDB).[22] Threading is done by PROSPECTOR_3,[17] an iterative sequence–structure alignment approach whose scoring function consists of sequence profiles, secondary structure propensities from PSIPRED,[23] and consensus contact predictions generated from the alignments in the prior threading iterations. Targets are categorized as Easy/Medium/Hard on the basis of the score significance and alignment consistency.[17]

Based on the threading alignments, target sequences are split into threading template aligned regions and unaligned regions. parallel hyperbolic Monte Carlo sampling[24] (PHS) is exploited to assemble full-length protein models by rearranging the continuous aligned fragments (building blocks) excised from the threading templates. PHS logarithmically flattens local high-energy barriers, allowing the simulation to tunnel more efficiently through energetically inaccessible regions to low-energy valleys. During assembly, the building blocks are kept rigid and off-lattice to retain their geometric accuracy; unaligned regions are modeled on a cubic lattice by an ab initio procedure[5] and serve as the linkage points of the rigid-body rotations. Movements are guided by an optimized force field,[5] which includes knowledge-based statistical potentials describing short-range backbone correlations, pairwise interactions, hydrogen bonding, secondary structure propensities from PSIPRED,[23] and consensus contact restraints extracted from PROSPECTOR_3 identified templates/alignments. For a given template, an initial full-length model is built up by connecting the continuous template fragments (greater than or equal to five residues) by a random walk of C_α–C_α bond vectors of variable lengths from 3.26 Å to 4.35 Å. If a template gap is too big to be spanned by a specified number of unaligned residues, a big C_α–C_α bond will remain at the end of the random walk, and a springlike force that acts to draw sequential fragments close will be applied in subsequent Monte Carlo simulations, until a physically reasonable bond length is achieved.

Up to the 10 top-scoring templates in Easy (high confidence) targets and up to the 20 top-scoring templates for the Medium/Hard (low confidence) targets are used in TASSER. Depending on size, 40–80 replicas are exploited in PHS (larger proteins need more replicas). The 14 low-temperature replica trajectories are clustered by SPICKER,[21] and the five highest structural density clusters are selected. Since TASSER models include only C_α and side-chain centers of mass, the remaining backbone and side-chain atoms are added by PULCHRA.[25]

RESULTS

Sixty-four targets were assessed in CASP6; these were split into 90 targets/domains by the assessors.

Overall Results

Table I summarizes TASSER predictions for all 90 targets/domains, together with the threading alignments from PROSPECTOR_3. Columns 5–7 show the root-mean-square deviation (RMSD) to native of PROSPECTOR_3 templates and final models in the same aligned regions. In the majority of cases, the final models have lower RMSD than the threading templates. On average, the threading templates have 79% of the residues aligned, with an average RMSD to native of 8.4 Å. For the same aligned residues, the average RMSD of the best models (rank 1 model) is 5.4 (6.4) Å, which demonstrates that TASSER brings the threading templates closer to native by ~2–3 Å. If we only consider the 38 targets whose threading template RMSD is below 6 Å, the average RMSDs of threading templates and final models are 3.2 Å and 2.3 Å respectively, in the aligned regions. Columns 8 and 9 are the RMSD to native of full-length TASSER models. On average, the difference between the rank 1 model and the best of five models is about 1 Å, which shows that there is still room for improvement in the model selection strategy.

RMSD is usually not sensitive to the global topology, because some local errors (e.g., tail misorientation) can give a high RMSD.[26] In columns 10–12 of Table I, we list the Template Modeling Score (TM-score) of the PROSPECTOR_3 templates and the final models.[27] The value of TM-score is between [0, 1], with a TM-score = 1 indicating an identical structure pair and a TM-score < 0.17 indicating random structure pairs.[27] Since smaller distances are weighted more strongly than larger distances, the TM-score is more sensitive to global topology than to local modeling errors.[27]

The cumulative TM-score of threading templates, the rank 1 models, and the best submitted models are 39.44, 49.32, and 52.53, respectively, which indicates an overall TM-score improvement of 24–33% in the TASSER refinement procedures. A trivial part of the TM-score improvement is because of the length extension of the final models created by filling in the gapped regions. In a recent unpublished test of 1489 benchmark proteins, as used in Zhang and Skolnick,[16] we used MODELLER[6] to fill the gaps of 1489 threading templates from PROSPECTOR_3. The cumulative TM-scores of the 1489 PROSPECTOR_3 templates and MODELLER models are 706.8 and 731.7, respectively. Since MODELLER builds the full-length models by optimally satisfying the spatial restraints from templates and the topology of threading aligned regions are essentially unchanged, an increase of 3.5% (~731.7/706.8–1) represents the portion of TM-score improvement due to the length extension. Therefore, TASSER modeling results in more than a 20% TM-score increase relative to the threading templates, due to the improvement in model accuracy by fragment reassembly.

In Figure 1(a and b), we compare the threading templates and final models using both the RMSD and TM-score. Again, we see the improvements of final models over threading templates in the majority of the cases, with no obvious dependence on target difficulty. But the overall quality of Easy targets [e.g., CM targets are mainly located at the right-top corner of Fig. 1(b)] is still clearly better than that of hard targets [e.g., NF targets mainly at the left-below corner of Fig. 1(b)]. In Figure 1(c and d), we also compare the RMSD and TM-score of the best structure alignment of the threading template to the best model for the set of residues identified by the structural alignment algorithm SAL.[28] In around half of the cases (57/90), the RMSD of the models is lower than that of the best structural alignments in the aligned regions. This indicates that the improvements in those targets come from the fragment rearrangement rather than from refining the threading alignments. This represents significant progress relative to the previous CASPs, where improvement over the best template alignment was not observed.

Loop Modeling

There are 348 unaligned loops (defined as an unaligned region in the PROSPECTOR_3 alignment) in the 64 target chains. We only count those loops where the native coordinates of both the loop and stem regions (five neighboring residues on both sides) are available. In general, loop modeling results consistent with the benchmark[18] are found, with accuracy decreasing with increasing length. An RMSD below 6 Å for all loops defined both with respect to the stems and considering the internal loop geometry was found.

Examples

Although all TASSER models are automatically generated in CASP6, as indicated in Table I, TASSER obviously performed better for some targets than others. In Figure 2, we select five successful examples from the NF, FR, and CM categories where TASSER has significantly improved the aligned regions of threading templates and the overall quality of TASSER models are the best, or among the best, of all CASP6 groups (see http://bioinformatics.buffalo.edu/new_buffalo/people/zhang6/casp6/ or http://predictioncenter.genomecenter.ucdavis.edu/casp6/).

T0201 is an NF target with 94 residues. PROSPECTOR_3 has seven inconsistent template hits. As shown in Figure 2(a), the best TM-score alignment is from 1irsA, which is 13.3 Å from native for the 63 matched residues. Although the template topologies are quite different from native, the local secondary structures in all templates are correct. TASSER takes the continuously aligned fragments from all the templates and rearranges them. Fifty-four percent of the trajectories belong to the first cluster, with an average RMSD of 5.1 Å to the cluster centroid, which gives a C-score = 0.21. The C-score is an indicator of likelihood of success for TASSER models; higher C-score values indicate higher confidence in the quality of the model (the definition of C-score is given in Eq. (1) in Zhang and Skolnick[16]). Based on the PDB benchmark results,[18] 74% of cases with this C-score have an RMSD to native below 6.5 Å. As shown in Figure 2(a), the RMSD of the first submitted model for T0201 is 4.9 Å (4.8 Å over the threading aligned regions) and a TM-score = 0.51.

T0212 is a 216-residue FR/A (analogy) target. PROSPECTOR_3 does not have significant hits, with four weakly scoring templates. The highest TM-score alignment is from 1rouA, which has an RMSD to native of 13 Å [Fig. 2(b)]. Using TASSER, 32% of the structures are in the first cluster, with an average RMSD of 5.9 Å to the centroid. The divergence of the trajectories gives a C-score = −0.5; 52% of targets with this C-score have an RMSD to native below 6.5 Å. Indeed, the best submitted model is 6.1 Å from native (4.7 Å in the threading aligned regions). While the topology of the final model is drawn significantly closer to native than that in the threading templates, the loop and tail regions still need improvement.

T0251 is a 102-residue FR/H (homology) target. PROSPECTOR_3 finds seven templates: 1h75A, 1b4qA, 1ego_, 1eejA, 1j0fA, 1h75A, and 1ego_, with the best alignment from 1b4qA having a TM-score = 0.44 (RMSD = 5.8 Å over 83 matched residues). PROSPECTOR_3 categorizes T0251 as an Easy target, and the global topology of the alignments is correct except for some errors around the loops [Fig. 2(c)]. TASSER takes the consensus contact restraints from all the templates and rearranges the fragments. The Monte Carlo trajectories highly converge; 78% of the structures belong to the first cluster, with an average RMSD of 2.4 Å to the centroid. This results in a C-score = 1.9: 98% of targets with this C-score in the PDB benchmark have an RMSD < 6.5 Å.[18] The actual RMSD of the first model is 3.1 Å in both full-length and aligned regions. The TM-score = 0.67, which is more than 10% higher than the best prediction of other groups. This is a typical example of a successful prediction where the threading alignments are in consensus, and TASSER achieves significant further refinement.

T0267, a 175-residue CM/Hard target, is a good case to examine TASSER's ability to refine loops. PROSPECTOR_3 provides 10 different alignments from five templates, the best of which is from 1vhsA, with a 4.4 Å RMSD over 160 residues. The main alignment errors are in four loop regions: Loop I (T22–L39, 8.9 Å to native), Loop II (P55–P58, 7.0 Å), Loop III (A82–Y88, 10.9 Å), and Loop IV (E154–K160, 5.2 Å). Here, the loop RMSD is calculated based on the global TM-score superposition. After TASSER reassembly, all the loops improve with the RMSD of the loops in the final model: 2.3 Å (Loop I), 3.8 Å (Loop II), 4.1 Å (Loop III), and 3.2 Å (Loop IV) [Fig. 2(d)]. The average RMSD of these loops is reduced from 8.0 Å (in the template) to 3.3 Å (in the model). The overall RMSD to native of the submitted TASSER model is 2.5 Å, and the TM-score is 0.87.

T0231 is a CM/Easy target (with 142 residues). PROSPECTOR_3 has strong consensus alignments from 1ahq_, 1m4jA, and 1hqz1, with Z scores > 30, the best of which is from 1ahq_, with an RMSD of 2.8 Å over 128 aligned residues. After TASSER refinement, the best final model has a RMSD to native of 1.25 Å (1.1 Å in the aligned

TABLE I. Summary of TASSER Models of 90 CASP6 Targets/Domains

				RMSD to native					TM-score		
ID	Type	Lch/Ln	Cov	R_Ta	R_M1a	R_MBa	R_M1	R_MB	TM_T	TM_MI	TM_MB
T0204	CM/easy	351/297	0.98	4.8	3.7	3.4 (3)	3.7	3.4 (3)	0.825	0.866	0.866 (3)
T0229_1	CM/easy	138/24	1.00	0.8	0.7	0.7 (1)	0.9	0.9 (1)	0.524	0.530	0.530 (1)
T0229_2	CM/easy	138/102	0.93	2.1	2.1	2.0 (2)	2.2	2.2 (1)	0.776	0.798	0.798 (1)
T0231	CM/easy	142/137	0.93	2.8	1.3	1.1 (3)	1.4	1.3 (3)	0.770	0.929	0.943 (3)
T0233_1	CM/easy	362/66	1.00	1.4	1.3	1.2 (4)	1.3	1.2 (4)	0.840	0.864	0.877 (4)
T0233_2	CM/easy	362/265	0.96	2.1	1.6	1.6 (4)	1.7	1.7 (4)	0.893	0.951	0.952 (4)
T0240	CM/easy	90/90	0.81	8.3	5.3	5.3 (1)	6.6	6.6 (1)	0.378	0.492	0.492 (1)
T0244	CM/easy	301/296	0.86	8.8	8.1	7.0 (5)	8.4	7.3 (5)	0.642	0.767	0.778 (5)
T0246	CM/easy	354/354	0.99	2.3	1.9	1.5 (2)	2.0	1.5 (2)	0.905	0.931	0.955 (2)
T0247_1	CM/easy	364/150	0.99	4.0	4.2	3.9 (2)	4.3	4.3 (2)	0.785	0.785	0.789 (2)
T0247_2	CM/easy	364/135	0.95	2.0	1.9	1.9 (1)	2.0	2.0 (1)	0.845	0.881	0.881 (1)
T0247_3	CM/easy	364/76	0.95	2.2	2.0	2.0 (5)	2.5	2.4 (5)	0.768	0.772	0.776 (5)
T0264_1	CM/easy	294/116	0.94	2.4	1.9	1.9 (5)	2.0	2.0 (5)	0.810	0.877	0.877 (1)
T0266	CM/easy	152/150	0.96	2.5	1.6	1.6 (1)	1.6	1.6 (1)	0.836	0.904	0.904 (1)
T0268_1	CM/easy	285/172	0.98	1.6	1.1	1.1 (1)	1.2	1.2 (5)	0.906	0.952	0.952 (5)
T0268_2	CM/easy	285/109	0.96	1.5	1.5	1.4 (3)	1.6	1.6 (1)	0.869	0.896	0.896 (1)
T0269_1	CM/easy	250/158	0.96	2.4	2.0	1.8 (2)	2.1	1.9 (2)	0.866	0.904	0.916 (2)
T0271	CM/easy	161/161	0.88	2.2	2.2	2.1 (4)	4.9	4.9 (3)	0.766	0.798	0.802 (3)
T0274	CM/easy	159/156	0.96	3.1	3.2	3.0 (3)	3.3	3.2 (3)	0.834	0.868	0.874 (2)
T0275	CM/easy	137/135	0.99	3.5	2.6	2.6 (1)	2.7	2.7 (1)	0.733	0.829	0.829 (1)
T0276	CM/easy	184/168	1.00	2.7	2.3	2.3 (3)	2.3	2.3 (3)	0.816	0.854	0.855 (3)
T0277	CM/easy	119/117	0.94	3.1	1.6	1.6 (2)	1.6	1.6 (2)	0.795	0.899	0.900 (2)
T0280_1	CM/easy	208/113	0.96	2.8	2.8	2.7 (3)	2.9	2.7 (3)	0.757	0.753	0.772 (5)
T0282	CM/easy	332/323	0.82	4.3	2.3	2.3 (5)	4.6	3.9 (4)	0.707	0.844	0.848 (4)
T0235_1	CM/easy	499/309	0.68	11.4	4.3	3.3 (4)	4.7	3.7 (4)	0.462	0.747	0.851 (3)
Ave/Cum	CM/easy	256/167	0.94	3.4	2.5	2.4 (3)	2.9	2.7 (3)	19.108	20.691	20.913 (3)
T0196	CM/hard	116/89	0.96	9.5	4.6	4.4 (3)	4.7	4.5 (3)	0.715	0.780	0.781 (2)
T0199_1	CM/hard	338/74	0.85	3.5	3.2	1.6 (4)	3.3	1.8 (4)	0.573	0.707	0.797 (4)
T0200	CM/hard	255/255	0.80	7.0	4.9	4.8 (5)	8.1	8.1 (1)	0.547	0.680	0.680 (5)
T0205	CM/hard	130/103	0.80	2.6	1.8	1.8 (1)	7.5	3.3 (4)	0.641	0.726	0.726 (1)
T0208	CM/hard	357/344	0.76	12.5	9.5	7.5 (4)	10.4	10.1 (5)	0.411	0.619	0.689 (4)
T0211	CM/hard	144/136	0.91	4.5	3.7	3.6 (4)	4.0	3.9 (4)	0.629	0.725	0.725 (1)
T0222_1	CM/hard	373/264	0.87	5.0	3.6	3.6 (3)	4.2	4.2 (1)	0.686	0.822	0.822 (1)
T0223_1	CM/hard	206/114	0.77	16.8	3.0	3.0 (1)	3.3	3.3 (1)	0.209	0.735	0.735 (2)
T0226_1	CM/hard	290/182	0.50	27.7	3.6	3.3 (5)	12.6	11.4 (2)	0.130	0.606	0.642 (5)
T0232_1	CM/hard	236/81	0.94	3.0	2.4	2.3 (2)	2.6	2.4 (2)	0.716	0.773	0.793 (2)
T0232_2	CM/hard	236/146	0.86	3.4	6.7	4.1 (5)	8.2	4.8 (5)	0.631	0.496	0.648 (5)
T0234	CM/hard	165/135	0.91	6.9	4.1	3.9 (3)	4.3	4.1 (3)	0.602	0.680	0.691 (3)
T0264_2	CM/hard	294/173	0.77	5.0	3.9	3.4 (3)	5.1	4.9 (2)	0.606	0.694	0.701 (3)
T0265	CM/hard	109/102	0.93	7.5	6.7	6.0 (4)	8.2	7.4 (4)	0.599	0.627	0.650 (5)
T0267	CM/hard	175/174	0.91	4.3	2.5	2.0 (5)	2.6	2.5 (5)	0.672	0.847	0.870 (5)
T0269_2	CM/hard	250/61	0.90	7.1	7.4	6.2 (2)	8.7	8.0 (2)	0.415	0.398	0.420 (4)
T0279_1	CM/hard	261/127	0.92	3.4	2.7	2.6 (3)	3.0	2.9 (3)	0.720	0.771	0.776 (3)
T0279_2	CM/hard	261/121	0.96	3.2	2.5	2.5 (2)	2.5	2.5 (2)	0.688	0.774	0.774 (5)
Ave/Cum	CM/hard	233/149	0.85	7.4	4.3	3.7 (3)	5.7	5.0 (3)	10.190	12.460	12.920 (3)
T0197	FR/H	179/166	0.20	6.8	6.9	6.1 (3)	17.5	15.0 (2)	0.114	0.234	0.364 (2)
T0199_2	FR/H	338/134	0.78	9.0	9.3	8.9 (4)	11.0	9.7 (4)	0.416	0.486	0.549 (4)
T0202_1	FR/H	249/123	0.85	15.2	10.0	10.0 (1)	9.8	9.8 (1)	0.224	0.546	0.546 (1)
T0203	FR/H	382/365	0.64	7.7	13.7	7.5 (2)	14.5	10.6 (2)	0.424	0.531	0.599 (2)
T0206	FR/H	220/138	0.62	15.0	10.1	10.1 (1)	14.1	11.5 (4)	0.149	0.224	0.249 (4)
T0213	FR/H	103/103	0.97	14.0	5.3	5.3 (1)	5.3	5.3 (1)	0.200	0.544	0.544 (1)
T0214	FR/H	110/110	0.91	13.3	12.5	9.2 (2)	13.6	9.9 (2)	0.224	0.214	0.279 (2)
T0222_2	FR/H	373/64	0.00	0.0	0.0	0.0 (4)	13.8	2.1 (4)	0.000	0.160	0.775 (4)
T0223_2	FR/H	206/92	0.00	0.0	0.0	0.0 (4)	10.9	10.9 (1)	0.000	0.217	0.282 (4)
T0224	FR/H	87/87	0.97	5.5	4.2	4.2 (1)	4.3	4.3 (1)	0.513	0.627	0.627 (1)
T0227	FR/H	121/84	0.73	11.3	9.7	9.4 (4)	12.3	12.0 (4)	0.208	0.233	0.249 (4)
T0228_1	FR/H	429/157	0.71	20.2	9.3	8.6 (2)	8.7	8.4 (2)	0.135	0.484	0.484 (1)
T0228_2	FR/H	429/235	0.72	14.9	5.2	5.2 (1)	17.1	12.9 (5)	0.218	0.547	0.547 (1)
T0237_1	FR/H	445/149	0.74	19.6	20.4	16.2 (5)	20.1	17.5 (5)	0.120	0.170	0.197 (5)
T0237_2	FR/H	445/101	0.35	10.8	10.5	9.2 (4)	14.0	14.0 (3)	0.137	0.190	0.192 (3)
T0237_3	FR/H	445/55	0.00	0.0	0.0	0.0 (4)	7.0	7.0 (1)	0.000	0.352	0.352 (1)

TABLE I. Continued

				RMSD to native					TM-score		
ID	Type	Lch/Ln	Cov	R_Ta	R_M1a	R_MBa	R_M1	R_MB	TM_T	TM_MI	TM_MB
T0243	FR/H	93/88	0.27	5.6	1.4	1.4 (1)	3.7	3.7 (1)	0.172	0.617	0.617 (1)
T0249_1	FR/H	209/73	0.89	3.5	2.4	2.4 (1)	2.4	2.4 (1)	0.603	0.702	0.702 (1)
T0249_2	FR/H	209/77	0.86	13.2	5.2	4.1 (3)	4.9	4.1 (3)	0.180	0.533	0.591 (3)
T0251	FR/H	102/99	0.82	5.8	3.1	3.1 (1)	3.1	3.1 (1)	0.444	0.672	0.672 (1)
T0262_2	FR/H	256/97	0.89	17.6	13.6	5.8 (4)	14.9	7.2 (4)	0.164	0.206	0.468 (4)
T0263	FR/H	101/97	0.93	5.0	3.7	3.7 (1)	3.8	3.8 (1)	0.482	0.665	0.665 (1)
Ave/Cum	FR/H	251/122	0.63	9.7	7.1	5.9 (2)	10.3	8.4 (2)	5.127	9.154	10.550 (2)
T0198	FR/A	235/225	0.48	14.5	9.2	9.2 (1)	25.6	17.4 (2)	0.187	0.347	0.347 (1)
T0199_3	FR/A	338/82	0.59	8.7	9.1	9.1 (3)	12.3	11.2 (4)	0.173	0.178	0.236 (4)
T0209_1	FR/A	239/108	0.94	15.8	16.6	13.7 (5)	16.5	14.2 (5)	0.166	0.212	0.212 (1)
T0212	FR/A	126/124	0.77	13.0	6.5	4.7 (2)	7.8	6.1 (2)	0.232	0.342	0.482 (2)
T0215	FR/A	76/53	0.98	8.1	5.1	3.6 (2)	5.2	3.9 (2)	0.259	0.423	0.500 (2)
T0230	FR/A	104/102	0.85	6.8	9.2	5.5 (4)	10.0	5.8 (4)	0.423	0.564	0.564 (1)
T0235_2	FR/A	499/43	0.91	14.9	13.4	13.4 (1)	13.2	13.2 (1)	0.119	0.191	0.286 (5)
T0239	FR/A	98/98	0.95	14.1	9.6	9.1 (5)	9.8	9.3 (5)	0.183	0.326	0.350 (2)
T0248_1	FR/A	294/79	1.00	8.8	11.8	7.7 (2)	11.8	7.7 (2)	0.279	0.361	0.414 (3)
T0248_3	FR/A	294/87	0.97	9.0	14.9	8.0 (3)	14.9	7.9 (3)	0.257	0.270	0.317 (3)
T0262_1	FR/A	256/72	0.85	10.3	10.1	9.3 (2)	10.6	9.7 (2)	0.261	0.264	0.264 (1)
T0272_1	FR/A	211/85	0.00	0.0	0.0	0.0 (4)	10.1	8.6 (2)	0.000	0.236	0.314 (2)
T0272_2	FR/A	211/99	0.29	11.3	5.1	4.4 (2)	13.5	13.5 (1)	0.147	0.229	0.229 (1)
T0280_2	FR/A	208/51	0.08	0.6	0.9	0.5 (4)	8.9	8.9 (1)	0.073	0.202	0.233 (4)
T0281	FR/A	70/70	0.99	8.1	10.4	7.8 (5)	10.5	8.2 (5)	0.324	0.335	0.381 (5)
Ave/Cum	FR/A	217/92	0.71	9.6	8.8	7.1 (3)	12.0	9.7 (3)	3.083	4.480	5.129 (2)
T0201	NF	94/94	0.67	13.3	4.8	4.8 (1)	4.9	4.9 (1)	0.262	0.508	0.508 (1)
T0209_2	NF	239/57	1.00	13.2	10.8	10.3 (2)	10.8	10.3 (2)	0.249	0.239	0.305 (3)
T0216_1	NF	435/209	0.64	28.1	21.2	19.1 (5)	22.9	19.6 (5)	0.104	0.153	0.213 (5)
T0216_2	NF	435/213	0.79	23.5	20.0	17.9 (5)	20.1	18.2 (3)	0.177	0.171	0.228 (4)
T0238	NF	251/181	0.83	19.6	17.0	17.0 (1)	20.7	19.8 (2)	0.189	0.238	0.292 (5)
T0241_1	NF	237/117	0.76	17.5	12.5	12.5 (1)	13.0	13.0 (1)	0.179	0.216	0.217 (3)
T0241_2	NF	237/119	0.94	20.0	17.3	15.2 (3)	17.4	15.3 (3)	0.149	0.229	0.252 (4)
T0242	NF	116/115	0.98	20.3	13.1	12.1 (5)	13.2	12.1 (5)	0.177	0.239	0.289 (5)
T0248_2	NF	294/87	0.87	8.6	13.5	7.8 (5)	14.1	8.2 (5)	0.270	0.265	0.350 (3)
T0273	NF	187/186	0.84	16.1	15.4	11.3 (2)	15.4	11.7 (2)	0.176	0.285	0.361 (2)
Ave/Cum	NF	253/138	0.83	18.0	14.6	12.8 (3)	15.3	13.3 (3)	1.932	2.543	3.015 (4)
Ave/Cum	All	243/137	0.79	8.4	6.4	5.4 (3)	8.2	6.9 (3)	39.440	49.328	52.527 (3)

ID, target or domain identification; Type, categories of each target/domain: new fold (NF), fold recognition/analogy (FR/A), fold recognition/homology (FR/H), comparative modeling/hard (CM/Hard), and comparative modeling/easy (CM/Easy); Lch, number of residues in the released sequences modeled by TASSER; Ln, number of residues in the solved structures for the targets/domains; Cov, fraction of the aligned residues defined with respect to Ln in PROSPECTOR_3 alignments; R_Ta, RMSD of the best initial template; R_M1a, RMSD of the first submitted model (calculated in the aligned regions); R_MBa, RMSD of the best submitted model (calculated in the aligned regions), with rank of the best model in parentheses; R_M1, RMSD of the first submitted model (for the entire chain); R_MB, RMSD of the best submitted model (for the entire chain), with the rank of the best model in parentheses; TM_T, TM-score of the best initial template; TM_M1, TM-score of the first submitted model; TM_MB, TM-score of the best submitted model; Ave/Cum, average or cumulative score of all targets/domains.

regions) and a TM-score = 0.943. The major modeling error is in the loop (Y75–S83), with an RMSD of 3.4 Å to native (4.1 Å in the template). In Figure 2(e), we also present the best structural alignment between native and the PROSPECTOR_3 template by TM-align,[26] which has an RMSD = 1.8 Å. Therefore, the best final model is even closer to native than the structural alignment.

DISCUSSION

TASSER has been exploited to automatically generate models for all categories of CASP6 targets. Consistent with the large-scale PDB benchmark tests,[16,18] over all target categories, the final models are often closer to native than the best of the threading templates (sometimes even better than the best structural alignment between the target and template). One of the reasons contributing to the improvement over the template alignment is the long-range tertiary restraints taken from the consensus of multiple threading templates, which are generally of higher accuracy than the individual template alignments. Second, the knowledge-based energy terms (which include hydrogen bonding, secondary structure predictions, short-range correlations, and pairwise side-chain interactions) were optimized based on a large number of difficult decoys,

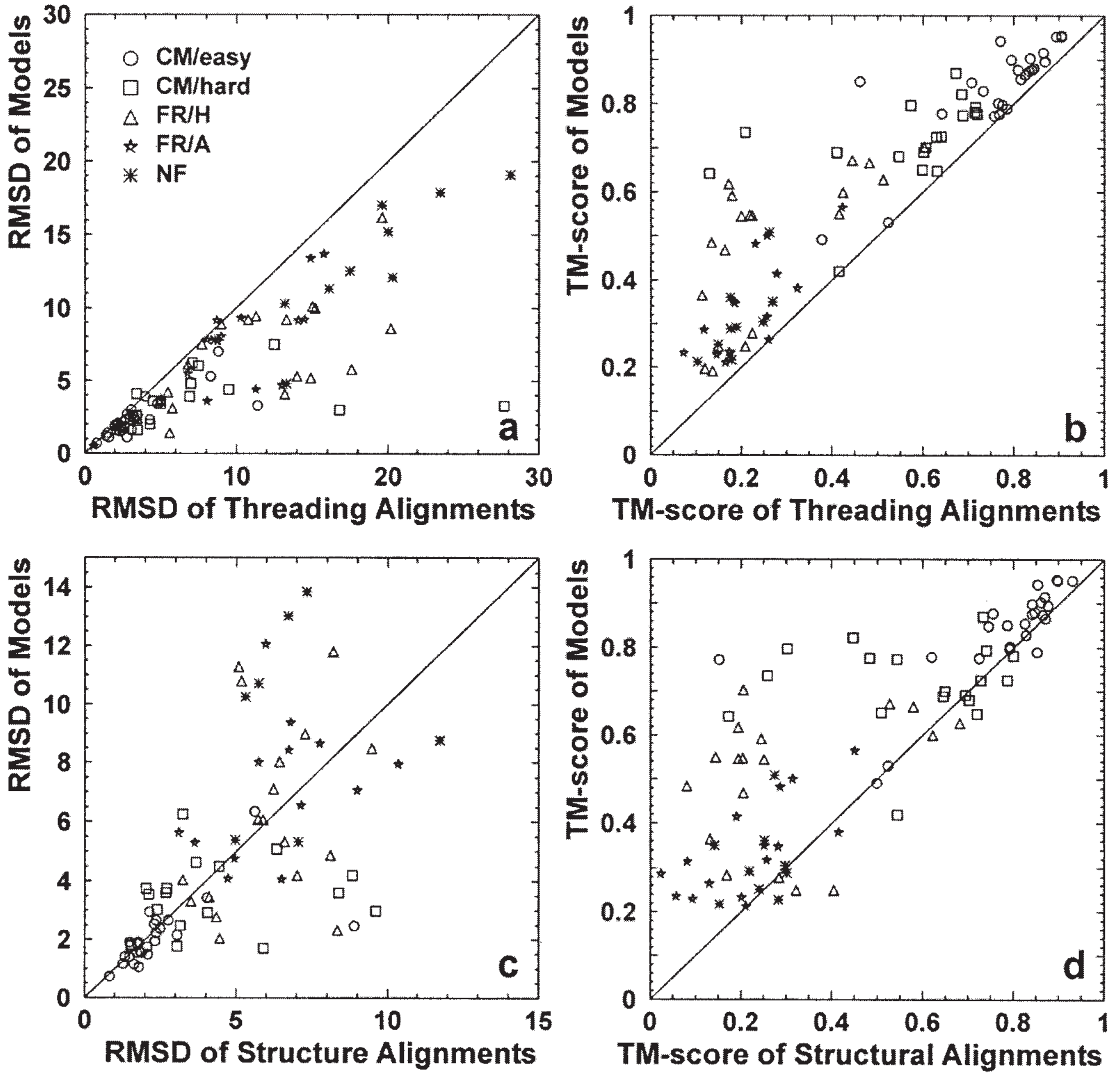

Fig. 1. Comparison between the final TASSER models and the threading templates with alignments from both threading (by PROSPECTOR_3) and structural alignments (by SAL). (**a**) RMSD to native of the best submitted models versus RMSD to native of the best threading alignments, calculated over the same threading aligned regions. (**b**) TM-score of the best submitted models versus TM-score of the best threading alignments. (**c**) RMSD to native of the best models versus RMSD to native of the best structural alignments, calculated in the same structural aligned regions. (**d**) TM-score of the best models versus TM-score of the best structural alignments.

and substantial correlation (an average correlation coefficient of ~.7) of total energy and RMSD was achieved.[5] The interaction between the predicted side-chain contact restraints and the inherent potentials is the major driving force for fragment rearrangement and the improvement in the overall quality of the structural prediction.

One of the problems of TASSER, although not assessed in CASP6, is that it fails to correctly predict the relative orientation of protein domains when the domain orientation in the threading template is incorrect. In a recent benchmark test on 258 multidomain PDB proteins,[18] 172 (67%) of them have the structure of individual domains correctly predicted (< 6.5 Å); however, only 112 (43%) have the correct relative domain orientation. A similar example in CASP6 is T0198 [see Fig. 3(a)], where the local conformations of the individual domains in the first TASSER model are well predicted (both below 5 Å) but the global RMSD is 26 Å from native (TM-score = 0.35) because of a mistake in the domain orientation. TASSER does generate a model in the top 10 clusters that has the correct orientation [Fig. 3(b)], but its free energy is too high to be selected by SPICKER. In this context, including pairwise statistical potentials specific for the domain interface might help.

Another problem is that we often failed to split multidomain targets into individual domains. The danger of threading multidomain sequences is that one domain may dominate the alignment scoring function; therefore, the algorithm will fail for other domains if these domains belong to different template structures. For example, T0223 has two structurally similar domains: the N-terminal (T0223_1) is a CM target, and the C-terminal (T0223_2) is a FR target. Based on TM-align,[26] the best structural analogs of both domains should be from 1nox_. However, because the N-terminal domain dominates the

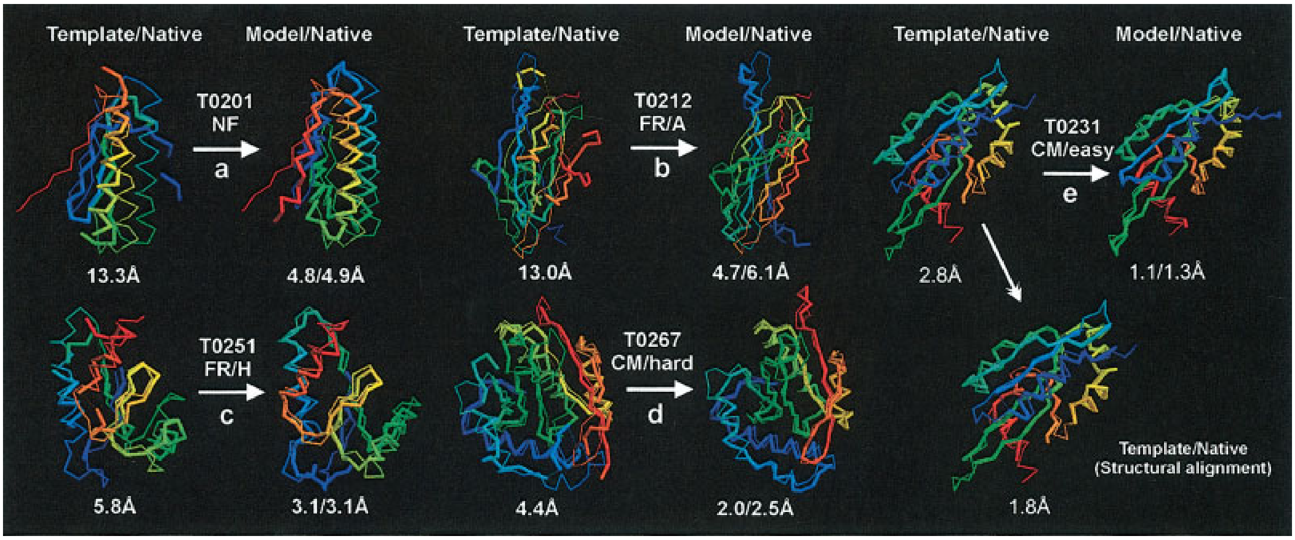

Fig. 2. Successful examples of TASSER modeling in different categories. For each target, on the left is the superposition of the threading template (thick backbone) and native (thin backbone); on the right is the final model (thick backbone) and native (thin backbone). Blue to red goes from the N- to the C-terminus. The numbers below the superposition are the RMSD over the aligned regions and RMSD over the full-length molecule, respectively.

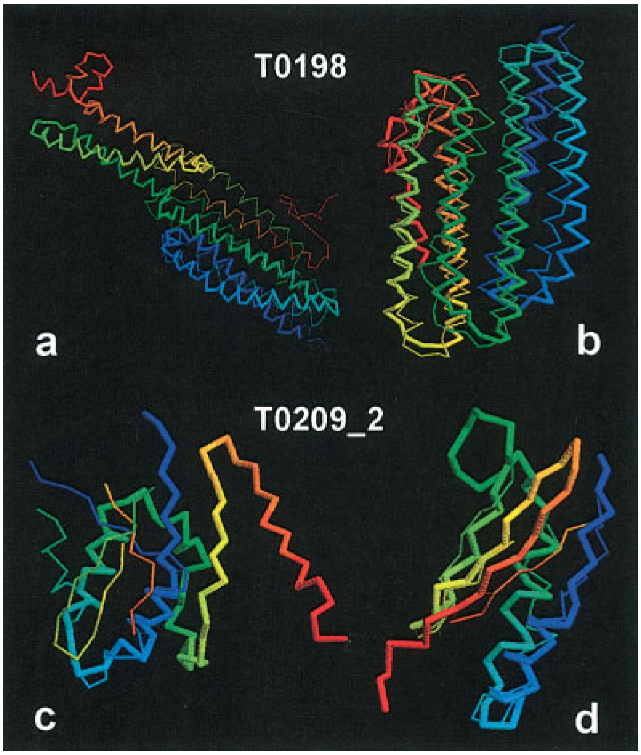

Fig. 3. Two failed examples of TASSER modeling. Superposition is shown for model (thick backbone) and native (thin backbone), with blue to red running from N- to C- terminus. (**a**) The first submitted model of T0198 (TM-score = 0.35), where the domain orientation is mispredicted. (**b**) Prediction for T0198 with correct domain orientation (TM-score = 0.53) but not selected by SPICKER because of its high free-energy. (**c**) The first submitted model for T0209_2 (C-terminal domain of T0209), where TASSER models the two domains together. (**d**) The predicted model for T0209_2 after CASP6 when TASSER models the C-terminal domain separately, under the assumption that we had known the domain border of the target.

sequence profile, PROSPECTOR_3 failed to provide alignments for the C-terminal. Domain parsing problems also influence the performance of TASSER in ab initio modeling. Figure 3(c) shows an example for T0209_2 (N-terminal domain of T0209, an NF target) where in CASP6 we folded it together with T0209_1 (C-terminal domain), giving a TM-score = 0.24. When we reran T0209_2 separately under the assumption that we had correctly parsed the domains, the model was much better, with a TM-score = 0.53 [Fig. 3(d)]. To partly address this issue, one ongoing approach is to iterate PROSPECTOR_3 by redoing the alignments for the missed domains in the subsequent steps; the hope is to provide TASSER with alignments and restraints of all domains, as well as reliable domain parsing information.

In addition to the domain parsing/orientation problem, TASSER also has problems with generating high-resolution models for large single-domain proteins (e.g., > 130 residues) when threading fails to provide reasonable alignments. This partially highlights the inability of the inherent TASSER force field to handle larger proteins. For example, for T0197 (179 residues, single domain), PROSPECTOR_3 provides an alignment for only 36 residues, and TASSER needs to generate the conformation of the remaining 141 residues from scratch. This results in a TASSER model 15 Å away from native (TM-score = 0.36).

We also noticed that in around 10% of the targets (e.g., T0226, T0249, T0272, and T0282), there are local distortions of secondary structure in the structurally diverse regions because we submitted the models from the cluster centroid. A simple solution is to use the individual decoy closest to the cluster centroid instead of the cluster centroid itself, since all the individual decoys are proteinlike, without local distortions. In practice, there are no observable differences between the TM-scores of these choices. We are also developing methods to refine the TASSER models using atomic potentials that could help to remove the problem of local distortion of cluster centroids. Algorithms for the structure-based detection of biologically relevant sites (e.g., active sites[29]) could in principle benefit from CM models with less geometrical errors.

Overall, there is considerable improvement in the performance of TASSER compared with our TOUCHSTONE predictions in CASP5.[30] The advantage of the current approach is that TASSER directly exploits and manipulates the continuous fragments from templates in an off-lattice system; this reduces the conformational entropy and yet helps retain the geometric accuracy of the well-aligned fragments. Moreover, the relative orientation of the fragments is flexible, which allows for global improvement over the template when a reasonable force field is used. This strategy gave obviously improved performance in both the CM and FR categories. Improvements also arise from use of PROSPECTOR_3 compared with the previous generation of the algorithm,[31] because of the introduction of a variety of more specific pair potentials and rigorous scoring cutoffs. Furthermore, TASSER has a better hydrogen bond scheme and implementation of tertiary restraints and pair potentials that provides much higher specificity and accuracy. The clustering approach is also better and is now designed to identify the lowest free energy state. Nevertheless, as indicated in both PDB benchmark tests and in CASP6, the success rate for weakly/nonhomologous single-domain proteins is only around two thirds. To successfully predict the structure of the remaining one third of single-domain proteins (essentially those lacking reasonable alignments to solved structures), as well as to deal with multidomain proteins, are the major issues that must be addressed in future TASSER development.

REFERENCES

1. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins 2001;Suppl 5:2–7.
2. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 2003;53(Suppl 6):334–339.
3. Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kazmierkiewicz R, Oldziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye YJ, Scheraga HA. Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proc Natl Acad Sci USA 2001;98:2329–2333.
4. 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.
5. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003;85:1145–1164.
6. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
7. 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.
8. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
9. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53(Suppl 6):352–368.
10. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
11. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441.
12. Murzin AG, Bateman A. CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. Proteins 2001;Suppl 5:76–85.
13. 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.
14. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotechnol 2000;18:283–287.
15. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
16. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
17. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins 2004;56:502–518.
18. Zhang Y, Skolnick J. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophys J 2004;87:2647–2655.
19. 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.
20. Zhang Y, Skolnick J. Structure modeling of 907 G protein-coupled receptors in the human genome. 2004. Submitted for publication.
21. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.
22. 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.
23. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
24. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002;48:192–201.
25. 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.
26. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on TM-score. Nucl Acid Res 2005;77(7):2302–2309.
27. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
28. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol 2003;334:793–802.
29. Arakaki AK, Zhang Y, Skolnick J. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. Bioinformatics 2004;20:1087–1096.
30. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D. TOUCHSTONE: a unified approach to protein structure prediction. Proteins 2003;53(Suppl 6):469–479.
31. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 2001;42:319–331.