WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

PREDICTION REPORT

Analysis of TASSER-based CASP7 protein structure prediction results

Hongyi Zhou, Shashi B. Pandit, Seung Yup Lee, Jose Borreguero, Huiling Chen, Liliana Wroblewska, and Jeffrey Skolnick*

Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30318

ABSTRACT

An improved TASSER (Threading/ASSEmbly/Refinement) methodology is applied to predict the tertiary structure for all CASP7 targets. TASSER employs template identification by threading, followed by tertiary structure assembly by rearranging continuous template fragments, where conformational space is searched via Parallel Hyperbolic Monte Carlo sampling with an optimized force-field that includes knowledge-based statistical potentials and restraints derived from threading templates. The final models are selected by clustering structures from the low temperature replicas. Improvements in TASSER over CASP6 involve use of better templates from 3D-jury applied to three threading programs, PROSPECTOR_3, SP^3, and SPARKS, and a fragment comparison method for better model ranking. For targets with no reliable templates, a variant of TASSER (chunk-TASSER) is also applied with potentials and restraints extracted from ab initio folded supersecondary chunks of the target to build full-length models. For all 124 CASP targets/domains, the average root-mean-square-deviation (RMSD) from native and alignment coverage of the best initial threading models from 3D-jury are 6.2 Å and 93%, respectively. Following TASSER reassembly, the average RMSD of the best model in the template aligned region decreases to 4.9 Å and the average TM-score increases from 0.617 for the template to 0.678 for the best full-length model. Based on target difficulty, the average TM-scores of the final model to native are 0.904, 0.671, and 0.307 for high-accuracy template-based modeling, template-based modeling, and free modeling targets/domains, respectively. For the more difficult targets, TASSER with modest human intervention performed better in comparison to its server counterpart, MetaTASSER, which used a limited time simulation.

Proteins 2007; 69(Suppl 8):90–97.
© 2007 Wiley-Liss, Inc.

Key words: template-based modeling; TASSER; MetaTASSER; PROSPECTOR_3; SPARKS; SP^3; 3D-jury; fold recognition; ab initio structure prediction.

INTRODUCTION

With the rapid growth in both genome sequencing and the experimental determination of tertiary structures, template-based protein structure prediction approaches are becoming increasingly useful as they are the only methods that provide reasonably reliable protein structure predictions.[1] In practice, about two-thirds of genome sequences <300 residues can be modeled on the basis of templates whose structures are found in the PDB.[1–4] For the remaining one-third of sequences, most, if not all, have structurally related templates in the PDB; however, current fold recognition methods can hardly recognize them.[5] Various approaches have been developed to detect such distantly related proteins including pairwise sequence[6–8] to sequence–profile[9] and

The authors state no conflict of interest.
Grant sponsor: Division of General Medical Sciences of the National Institutes of Health; Grant numbers: GM-37408, GM-48835.
*Correspondence to: Jeffrey Skolnick, Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, GA 30318.
E-mail: skolnick@gatech.edu
Received 21 February 2007; Revised 25 April 2007; Accepted 21 May 2007
Published online 17 August 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21649

profile–profile comparisons,[10–12] sequence to structure threading,[2,13] machine learning approaches,[14,15] and the use of Metaservers.[16–18] Another issue in template-based modeling (TBM) had been the inability to refine the template-based structures such that the resulting models are closer to the native structure than the starting template alignment.[19–23] Furthermore, although the PDB is growing faster than ever, the gap between the number of sequences and solved structures remains large.[24] This necessitates the need for a robust automated prediction method for proteome-scale structure prediction. Advances in TBM can be achieved through the development of more sensitive fold recognition approaches, the generation of more accurate sequence–structure alignments, and methods for model refinement along with a better way of selecting the best models.

Over the past several years, we have developed the TASSER[19] methodology for automated tertiary structure prediction that generates full-length models by rearranging the continuous fragments identified by threading. Based on TASSER's performance in CASP6, the lessons learned were[25]: (1) TASSER's performance depends considerably on the quality of the initial threading models; (2) we failed to correctly predict the relative orientation of multiple domain proteins; (3) the submitted models which are derived from clustering often have unrealistic bond lengths and bond angles with many distance clashes; (4) TASSER's performance was unsatisfactory on free modeling (FM) targets.

For CASP7, we attempted to improve the methodology by incorporating better initial threading template structures and initial alignments, better selection of near native structures, and for the targets with no reliable templates, we implemented a new method referred to as chunk-TASSER. For selection of better initial templates, we used our in-house 3D-jury Metaserver from three state-of-the-art threading procedures: PROSPECTOR_3,[26] SPARKS,[27] and SP^3.[28,29] We participated in CASP7 in the server category as MetaTASSER, which employed selection of templates by 3D-jury, a limited time TASSER simulation followed by clustering of structures and submission of the top five models. For the human prediction category, as the TASSER group we used 3D-jury templates followed by multiple instances of TASSER simulations for template-based targets and chunk-TASSER for targets with no reliable templates. The resulting structures are clustered and the top five models are submitted.

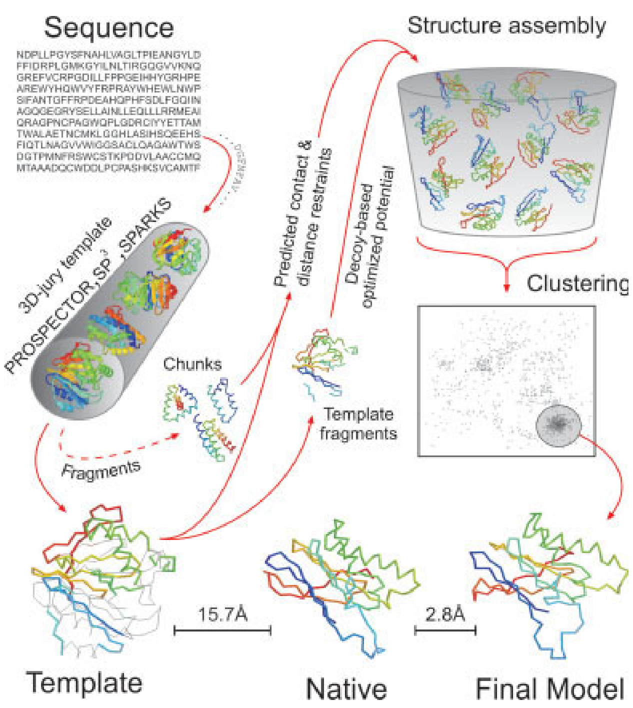
Sequence
NDPLLPGYSFNAHLVAGLTPIEANGYLD
FFIDRPLGMKGYILNLTIRGQGVVKNQ
GREFVCRPGDILLFPPGEIHHYGRHPE
AREWYHQWVYFRPRAYWHEWLNWP
SIFANTGFFRPDEAHQPHFSDLFGQIIN
AGQGEGRYSELLAINLLEQLLLRRMEAI
QRAGPNCPAGWQPLGDRCIYYETTAM
TWALAETNCMKLGGHLASIHSQEEHS
FIQTLNAGVVWIGGSACLQAGAWTWS
DGTPMNFRSWCSTKPDDVLAACCMQ
MTAAADQCWDDLPCPASHKSVCAMTF
3D-jury template
PROSPECTOR,SP3,SPARKS
Chunks
Fragments
Predicted contact & distance restraints
Decoy-based optimized potential
Template fragments
Structure assembly
Clustering
15.7Å
2.8Å
Template
Native
Final Model

Figure 1

Flowchart of the TASSER methodology.

METHODS

Figure 1 shows the flow chart of the TASSER methodology. We employed the 3D-jury approach[18] to select templates from the PROSPECTOR_3,[26] SPARKS,[27] and SP^3[28,29] threading algorithms for subsequent use by TASSER. In practice, for the 3D-jury approach, the 10 top-scoring templates from each threading method are compared using the structural alignment algorithm, TM-align,[30] with the TM-score[31] used as the similarity measure. The 3D-jury score is the sum of pairwise TM-scores for each template and is used to rank the templates. The obtained consensus templates provide the aligned fragments, tertiary distance restraints and contact restraints used in TASSER full-length structure assembly. Targets are classified as Easy, Medium, and Hard according to the template similarities of the top models from each threading method. When the top models have a TM-score > 0.5 with each other, the target is defined as Easy and is likely to have a structurally related template with a good alignment; when they have a TM-score < 0.4 with each other, the target is defined as Hard and the template is likely to be incorrect; all other cases are defined as Medium. The classification procedure is empirical and may be optimized in the future but provides a basis of the expected accuracy of the prediction as well as dictates the particular TASSER variant that is chosen.

TASSER[25] represents the protein by a C_α and side chain center-of-mass representation in both off- and on-lattice space. The initial full-length model is built by connecting the continuous template-provided fragments (off-lattice building blocks) by a random walk confined to lattice bond vectors. If the specified number of unaligned residues cannot span the gap, a long C_α—C_α bond remains, and a spring-like force draws sequential fragments together until a physically reasonable bond length is achieved. Parallel Hyperbolic Monte Carlo sampling[32] with replica exchange explores conformational space by rearranging the continuous fragments excised from the template. During assembly, the template fragments are kept rigid and off-lattice to retain their geometric accuracy; unaligned regions are modeled on a cubic lattice by an ab initio procedure and serve as linkage points for rigid body fragment rotations. Conformations are selected using an optimized force field, which includes knowledge-based statistical potentials describing short-range backbone correlations, pairwise interactions, hydrogen bonding, secondary structure propensities, consensus C_α and side-chain center of mass contacts, and short and long distance restraints for C_α atoms.

Chunk-TASSER (Biophy J, in press) is a variant of TASSER that utilizes ab initio folded sets of three consecutive secondary structural elements that are used to extract consensus sequence specific contact potentials and distance restraints for use by TASSER to build full length models. It is designed to address the situation when appropriate templates cannot be identified from threading, that is, for most Medium and all Hard targets. In this case, final model quality is mostly dictated by the quality of the selected chunk models rather than by the threading template quality as in TASSER.

For each target, MetaTASSER uses 3D-jury template selection followed by TASSER simulation and structure clustering using SPICKER.[33] The side chains for the top five cluster centroids are rebuilt using PULCHRA (manuscript in preparation), and these models are submitted. In TASSER human predictions, for template-based targets, the human interventions were: parsing the sequence into possible domains; for high sequence identity targets, running TASSER with manually selected templates in addition to 3D-jury automated selection; manual deletion of possible bad templates selected by our in-house 3D-jury procedure based on the fact that they are quite dissimilar to the high confidence templates. In addition, we used SPICKER[33] to rank models from multiple TASSER simulations with different inputs.

RESULTS

Ninety-five targets were assessed in CASP7 that were split into 124 targets/domains by the assessors. The detailed results of TASSER and MetaTASSER for all targets are available at http://cssb.biology.gatech.edu/skolnick/files/casp7/.

Overall results

Table I summarizes TASSER predictions for the 79 TBM and 16 FM targets/domains together with the accuracy of the threading alignments. The average fractions of aligned residues for the best threading template (3D-jury) with respect to the native structures are 93% (92%) for all (TBM) targets, respectively (data not shown). The template aligned residues have higher coverage in comparison to the CASP6 targets by PROSPECTOR_3.[25,26] On average, the root-mean-square-deviation (RMSD) from native of the best threading template (Column 2) are 6.2 and 6.1 Å for all and TBM targets, respectively. After TASSER refinement, the average RMSDs of the best model, over the same aligned residues, for all targets and TBM targets are 4.9 and 4.7 Å, respectively. Thus, TASSER can refine the models over the initial template alignment by ~1 Å for TBM targets. Column 4 shows the RMSD to native of the full-length TASSER models for the best-ranked full-length models.

We next analyzed the quality of the models using the TM-score.[31] The cumulative TM-scores[31] for the best threading model (Column 5), first and best submitted models (Column 6) are 76.47, 82.32, and 84.11, respectively. The relative TM-score improvement of the TASSER best model over the threading aligned regions relative to the threading template is 5.3%. Comparison of the best TASSER models (Column 6) with those by TM-align on the best threading templates (Column 7) shows that for 52 of the 79 TBM targets/domains, the TASSER model is better than the model provided by TM-align on the best threading template. We have also analyzed the performance of threading methods for the alignment accuracy and identification of the best possible template. The average TM-score of the structural alignment by TM-align for the best threading template is 0.668 (Column 7) in comparison to the average TM-score of 0.617 (Column 5) for the threading alignment, which suggests that alignment accuracy could be improved in threading. Columns 8 and 9 show the TM-score for the best template that could be identified using LGA and TM-align, respectively. The average TM-score of the best template identified using TM-align (LGA) is 0.739 (0.727) in comparison to the average TM-score of 0.668 for the best threading template alignment, which indicates that the better templates could be identified with improved threading methods.

In Figure 2(A,B), using both RMSD and TM-score, we compare the best threading templates and final best models. For most targets, TASSER improves over the threading models. However, for a few targets, the final model is slightly worse. We can see from the TM-score plot [Fig. 2(B)] that the absolute improvement does not depend on target difficulty. Figure 3 shows a comparison result of the best model from TASSER and MetaTASSER and the inset of the figure shows a similar comparison for models from the corrected MetaTASSER protocol. There was a technical error in the MetaTASSER protocol

Table I
Summary of TASSER Models for 79 TBM and 16 FM CASP Targets/Domains

ID	R_Ta	R_Mba	R_MB	TM_Ta	TM_MB	TM_Ta_TMaln	TM_LGA	TM_TMaln
TBM targets								
T0283	17.2	4.8 (1)	4.8 (1)	0.338	0.593 (2)	0.370 (1nfn_)	0.441 (2b2j_A)	0.596 (1vdza)
T0284	2.4	2.0 (1)	2.2 (1)	0.89	0.911 (1)	0.906 (1muma)	0.908 (1oqf_A)	0.906 (1muma)
T0285	11.6	10.2 (1)	10.3 (1)	0.244	0.316 (4)	0.367 (1e7ka)	0.637 (1p0z_G)	0.648 (1p0za)
T0286	4.4	3.7 (4)	6.3 (2)	0.682	0.762 (4)	0.755 (1yzfa)	0.791 (1esd_#)	0.792 (1esc_)
T0289_D1	7.2	7.1 (3)	7.3 (1)	0.647	0.683 (3)	0.704 (1yw4a)	0.704 (1yw4_A)	0.704 (1yw4a)
T0289_D2	6.4	7.1 (1)	7.1 (1)	0.344	0.377 (1)	0.531 (2bcoa)	0.559 (1vdz_A)	0.587 (1ghk_)
T0293_D1	3.8	4.0 (1)	4.0 (1)	0.701	0.731 (5)	0.768 (1t43a)	0.754 (1nv9_A)	0.768 (1t43a)
T0297	4.2	4.0 (1)	4.5 (1)	0.806	0.839 (1)	0.808 (1es9a)	0.806 (1bwp_#)	0.808 (1es9a)
T0298_D1	3.8	2.5 (5)	2.5 (5)	0.776	0.851 (4)	0.824 (2g17a)	0.824 (2g17_A)	0.831 (1ys4a)
T0298_D2	3	1.8 (5)	1.8 (5)	0.709	0.903 (5)	0.731 (2cvoa)	0.897 (1pqu_A)	0.895 (1gl3a2)
T0299_D1	10.9	9.4 (5)	10.2 (5)	0.3	0.336 (2)	0.377 (1shtx)	0.688 (2cg8_C)	0.675 (2bmba1)
T0299_D2	13.6	14.3 (4)	14.1 (4)	0.212	0.255 (5)	0.282 (1ido_)	0.599 (1rjj_A)	0.610 (1diqa)
T0301_D1	7	6.8 (4)	8.1 (4)	0.547	0.639 (2)	0.540 (1w61a)	0.540 (1w61_A)	0.610 (1tm0a)
T0301_D2	5.8	5.1 (2)	6.2 (2)	0.587	0.661 (2)	0.650 (1w61a)	0.652 (1w62_A)	0.650 (1w61a)
T0306	9.7	7.8 (2)	8.2 (2)	0.252	0.352 (4)	0.259 (1wf8a)	0.536 (1d7q_A)	0.550 (1s1hl)
T0312	13.6	14.5 (5)	14.9 (5)	0.204	0.273 (1)	0.233 (1ozga2)	0.655 (1xv2_B)	0.656 (1xv2a)
T0316_D1	4	3.2 (3)	5.6 (3)	0.624	0.725 (1)	0.660 (1kora)	0.665 (1kh3_C)	0.715 (2c5sa)
T0316_D3	19	16.0 (2)	15.7 (4)	0.19	0.210 (2)	0.338 (1kora)	0.686 (1wb3_B)	0.714 (1d2ea)
T0318_D1	9.3	9.2 (5)	9.9 (4)	0.386	0.458 (3)	0.575 (1lam_)	0.594 (1vhu_A)	0.595 (1hjza)
T0318_D2	1.8	1.8 (1)	1.9 (1)	0.914	0.935 (3)	0.922 (1gyta)	0.924 (1gyt_L)	0.922 (1gyta)
T0320_D1	11.6	5.7 (5)	7.7 (5)	0.612	0.706 (5)	0.643 (1sur_)	0.643 (1sur_#)	0.653 (1zuna)
T0320_D2	16.9	13.6 (1)	13.6 (1)	0.146	0.160 (3)	0.288 (1ni5a)	0.278 (1dik_#)	0.458 (1coja)
T0321_D1	13.4	11.9 (2)	11.9 (2)	0.218	0.299 (3)	0.283 (1ps0a)	0.633 (1f9c_A)	0.625 (2mnr_)
T0322	4.5	3.2 (4)	3.2 (4)	0.774	0.823 (3)	0.795 (1q4sa)	0.832 (2h4u_D)	0.795 (1q4sa)
T0323_D1	4.6	3.5 (3)	3.7 (3)	0.559	0.665 (3)	0.334 (1ko9a)	0.602 (1yqm_A)	0.526 (1rrqa)
T0323_D2	1.9	1.5 (4)	1.5 (4)	0.821	0.884 (4)	0.823 (1mpga)	0.835 (1diz_A)	0.829 (1ko9a)
T0325	5.3	4.4 (4)	6.3 (4)	0.612	0.686 (3)	0.695 (1v6ta)	0.695 (1v6t_A)	0.699 (1uuqa)
T0327	4.2	4.1 (1)	4.4 (1)	0.638	0.689 (4)	0.677 (1jgsa)	0.682 (1lnw_F)	0.686 (1tqia)
T0329_D1	2.1	1.5 (3)	1.5 (3)	0.865	0.913 (3)	0.883 (1feza)	0.885 (1rdf_B)	0.883 (1feza)
T0329_D2	6.2	4.3 (5)	5.1 (5)	0.563	0.634 (5)	0.495 (1feza)	0.511 (1rql_A)	0.601 (1qyia)
T0330_D1	2.5	2.2 (1)	2.4 (1)	0.754	0.838 (1)	0.813 (2ah5a)	0.813 (2ah5_A)	0.813 (2ah5a)
T0330_D2	3.3	2.5 (1)	2.5 (1)	0.64	0.713 (1)	0.679 (1lvha)	0.686 (1lvh_B)	0.679 (1lvha)
T0331	5.1	4.5 (1)	4.7 (1)	0.702	0.732 (1)	0.731 (1t9ma)	0.727 (1ty9_A)	0.734 (2a2ja)
T0333_D1	5	3.9 (2)	5.4 (2)	0.656	0.735 (2)	0.685 (1iira)	0.696 (1rrv_B)	0.685 (1iira)
T0333_D2	3.7	3.4 (4)	3.3 (4)	0.713	0.760 (4)	0.803 (2c1xa)	0.794 (1rrv_B)	0.803 (2c1xa)
T0335	3.9	2.8 (3)	2.8 (3)	0.391	0.505 (2)	0.444 (1ywma2)	0.644 (1y1u_A)	0.775 (1lvfa)
T0338_D1	5.4	2.9 (4)	4.6 (4)	0.661	0.747 (4)	0.706 (1vin_)	0.754 (1jkw_#)	0.754 (1jkw_1)
T0338_D2	3.4	3.0 (1)	3.7 (1)	0.663	0.732 (1)	0.666 (1vin_)	0.673 (1n4m_A)	0.685 (1zp2a)
T0339_D1	2.6	2.5 (1)	3.1 (1)	0.741	0.800 (1)	0.765 (1p3wb)	0.739 (1eg5_B)	0.765 (1p3wb)
T0341_D1	1.4	1.2 (2)	1.3 (2)	0.888	0.932 (2)	0.888 (2c4na)	0.887 (1zjj_B)	0.888 (2c4na)
T0341_D2	2.6	2.3 (4)	2.3 (4)	0.81	0.822 (2)	0.889 (1wvia)	0.889 (1wvi_B)	0.889 (1wvia)
T0342	2.5	2.6 (2)	2.9 (2)	0.76	0.796 (2)	0.774 (2g0qa)	0.774 (2g0q_A)	0.774 (2g0qa)
T0347_D1	11.1	3.9 (3)	4.0 (3)	0.656	0.627 (3)	0.692 (1vk1a)	0.692 (1vk1_A)	0.692 (1vk1a)
T0349_D1	1.7	1.7 (4)	2.0 (5)	0.656	0.746 (4)	0.756 (1yj7a1)	0.784 (1yj7_D)	0.756 (1yj7a1)
T0351	8.4	5.9 (3)	5.9 (3)	0.288	0.396 (3)	0.418 (1jfba)	0.430 (1cs1_C)	0.553 (1k2yx1)
T0354	10.3	8.0 (1)	7.8 (1)	0.311	0.524 (2)	0.405 (1wp9a3)	0.623 (2be3_A)	0.623 (2be3a)
T0356_D2	21.4	19.2 (2)	20.0 (2)	0.152	0.181 (4)	0.301 (2fn0a)	0.562 (1eje_A)	0.562 (1ejea)
T0357	4.4	3.1 (5)	3.7 (5)	0.536	0.693 (1)	0.630 (1zyma2)	0.709 (1aco_#)	0.709 (1aco_)
T0358	12.2	6.6 (3)	6.7 (3)	0.264	0.331 (5)	0.395 (1nj1a2)	0.566 (1dgd_#)	0.592 (1jqga1)
T0360	4.9	5.0 (1)	5.0 (1)	0.679	0.587 (1)	0.723 (1dvoa)	0.723 (1dvo_A)	0.723 (1dvoa)
T0362	2.4	2.1 (3)	3.3 (1)	0.782	0.842 (1)	0.791 (1z54a)	0.809 (2gf6_A)	0.809 (2gf6a)
T0363	3.6	2.5 (5)	2.6 (5)	0.498	0.612 (1)	0.566 (1vjka)	0.678 (2bb6_A)	0.692 (2bb5a)
T0364	2.6	2.5 (2)	3.9 (1)	0.748	0.810 (1)	0.750 (1s5ua)	0.777 (2av9_B)	0.775 (2av9a)
T0365	3.2	2.7 (1)	2.9 (1)	0.729	0.815 (1)	0.742 (1xwma)	0.742 (1xwm_A)	0.742 (1xwma)
T0368	3.5	2.7 (1)	2.8 (1)	0.683	0.801 (2)	0.619 (2fbna)	0.694 (2c2l_C)	0.712 (1hz4a)
T0369	4.5	3.3 (2)	3.7 (2)	0.635	0.748 (5)	0.701 (2f22a)	0.730 (1rxq_A)	0.730 (1rxqa)
T0370	7.7	2.6 (5)	2.6 (5)	0.649	0.819 (5)	0.746 (1nrga)	0.752 (1vl7_B)	0.818 (2arza)
T0371_D1	2.1	2.0 (1)	2.7 (1)	0.785	0.859 (1)	0.792 (1wvia)	0.801 (1vjr_A)	0.793 (1zjja)
T0371_D2	2.5	2.6 (4)	2.6 (4)	0.755	0.768 (1)	0.760 (1zjja)	0.760 (1zjj_A)	0.760 (1zjja)
T0372_D1	4.7	3.9 (3)	4.2 (3)	0.521	0.602 (3)	0.670 (1ozpa)	0.706 (1ro5_A)	0.706 (1ro5a)
T0372_D2	4.4	4.3 (1)	4.4 (1)	0.658	0.676 (1)	0.762 (1ne9a)	0.764 (1xf8_A)	0.762 (1ne9a)
T0373	4.3	3.1 (5)	3.2 (5)	0.589	0.722 (5)	0.667 (2fbia)	0.679 (1s3j_B)	0.727 (2a61a)
T0374	3.6	3.1 (4)	3.1 (4)	0.747	0.821 (3)	0.783 (1tiqa)	0.783 (1tiq_A)	0.802 (1s7fa)

(*Continued*)

Table I
(Continued)

ID	R_Ta	R_Mba	R_MB	TM_Ta	TM_MB	TM_Ta_TMaln	TM_LGA	TM_TMaln
T0375	3.2	3.4 (3)	3.5 (3)	0.812	0.829 (2)	0.846 (1rkd_)	0.847 (2fv7_B)	0.846 (1rkd_)
T0376	2.5	2.5 (3)	2.5 (3)	0.839	0.893 (4)	0.864 (1xkya)	0.863 (1xxx_B)	0.864 (1xkya)
T0378_D1	3.7	3.5 (1)	3.5 (1)	0.685	0.671 (5)	0.766 (1ipaa)	0.766 (1ipa_A)	0.766 (1ipaa2)
T0378_D2	2.5	2.2 (4)	2.3 (4)	0.808	0.865 (4)	0.840 (1ipaa)	0.873 (1gz0_C)	0.880 (1x7oa)
T0379_D1	3	2.1 (5)	2.2 (2)	0.787	0.852 (2)	0.802 (2b0ca)	0.827 (1zd5_A)	0.810 (2gfha)
T0379_D2	3.9	3.2 (4)	3.2 (4)	0.552	0.597 (4)	0.623 (2b0ca)	0.623 (2b0c_A)	0.623 (2b0ca)
T0380	5.6	2.3 (1)	2.2 (1)	0.672	0.845 (1)	0.763 (2a2ja)	0.789 (2fhq_A)	0.794 (2arza)
T0381_D1	0.7	0.7 (1)	0.7 (1)	0.941	0.941 (3)	0.941 (2g7ua)	0.941 (2g7u_A)	0.941 (2g7ua)
T0381_D2	1.8	1.6 (4)	1.7 (5)	0.894	0.909 (4)	0.894 (1mkma)	0.901 (2g7u_D)	0.903 (1ysqa)
T0383	6.8	5.4 (1)	6.2 (1)	0.377	0.565 (1)	0.667 (1fx3a)	0.671 (1qyn_B)	0.667 (1fx3a)
T0384_D1	3.9	3.2 (1)	3.2 (1)	0.823	0.862 (5)	0.879 (1h6da)	0.873 (1ydw_A)	0.879 (1h6da)
T0385	2.6	2.0 (3)	2.0 (3)	0.835	0.874 (2)	0.846 (1moja)	0.853 (1jgc_B)	0.853 (1jgca)
T0386_D1	5.5	5.1 (1)	7.7 (5)	0.569	0.649 (5)	0.642 (2f6sa)	0.651 (2g03_A)	0.642 (2f6sa)
T0304*	14.3	11.4 (5)	11.3 (5)	0.159	0.304 (5)	0.371 (1ppn_)	0.583 (2gnx_A)	0.606 (1ztua)
T0348*	12.1	5.9 (2)	7.6 (2)	0.501	0.422 (4)	0.506 (1p91a)	0.528 (1rfs_#)	0.550 (1g8kb)
T0382*	14.9	4.2 (2)	5.6 (2)	0.29	0.559 (2)	0.424 (1orja)	0.660 (1kps_B)	0.660 (1kpsb)
Average (TBM)	*6.1*	*4.7*	*5.1*	*0.600*	*0.671*	*0.654*	*0.716*	*0.729*
FM targets								
T0321_D2*	11.5	8.7 (1)	8.9 (1)	0.48	0.532 (1)	0.577 (1llua)	0.601 (1kxz_E)	0.647 (1ibja)
T0287	12.6	14.3 (2)	14.5 (2)	0.277	0.309 (2)	0.386 (1sumb)	0.419 (1v55_N)	0.495 (1ti2a)
T0296	23.7	18.1 (3)	18.4 (3)	0.191	0.353 (3)	0.255 (1xwl_)	0.721 (1r9e_B)	0.720 (1cm5a)
T0300	5.2	9.0 (3)	12.1 (1)	0.346	0.354 (1)	0.476 (1vp7a)	0.469 (16vp_A)	0.594 (1h2rl)
T0307	14.5	14.1 (3)	14.1 (3)	0.239	0.256 (1)	0.377 (1s1ea)	0.529 (1g3n_G)	0.529 (1g3nc1)
T0309	12.3	12.1 (3)	12.1 (3)	0.269	0.240 (4)	0.449 (1wytb2)	0.352 (1gqf_A)	0.513 (2fp3a)
T0314	16.3	12.7 (2)	12.7 (2)	0.178	0.263 (5)	0.297 (1f08a)	0.331 (1vdv_B)	0.460 (1fo4a)
T0316_D2	14.3	13.6 (5)	13.6 (5)	0.191	0.188 (5)	0.256 (1k92a)	0.510 (1aqf_H)	0.545 (1zpsa)
T0319	15.3	13.9 (2)	14.1 (2)	0.187	0.258 (5)	0.242 (1r62a)	0.320 (1j78_B)	0.458 (1uwka)
T0347_D2	14.9	7.8 (2)	7.9 (2)	0.231	0.291 (1)	0.292 (1vk1a)	0.595 (1h8h_A)	0.586 (1e79a1)
T0350_D1	14.1	9.5 (2)	9.8 (2)	0.284	0.321 (3)	0.339 (1orja)	0.483 (1tdh_A)	0.509 (1cfr_)
T0353	10.2	8.2 (2)	8.2 (2)	0.314	0.438 (2)	0.426 (1nbua)	0.535 (2bab_A)	0.583 (2ffla)
T0356_D1	14	19.1 (3)	21.1 (3)	0.194	0.188 (3)	0.266 (1qdla)	0.397 (1w66_A)	0.443 (1pc3a)
T0356_D3	14.4	13.1 (4)	12.8 (4)	0.182	0.363 (3)	0.303 (1gk8a)	0.323 (1yk3_F)	0.473 (1bu6o)
T0361_D1	17.7	10.0 (3)	10.5 (3)	0.159	0.330 (3)	0.252 (1o6sa2)	0.519 (1ufb_C)	0.517 (1ufba)
T0386_D2	13.2	10.9 (5)	12.2 (5)	0.192	0.222 (5)	0.393 (1xi8a)	0.496 (1y14_D)	0.527 (1xdwa2)
Average (FM)	*14.0*	*12.2*	*12.7*	*0.245*	*0.307*	*0.349*	*0.475*	*0.537*
Average (ALL)	*6.2*	*4.9*	*5.3*	*0.617*	*0.678*	*0.668*	*0.727*	*0.739*

ID, target or domain identifier; R_Ta, RMSD of the best initial template; R_Mba, RMSD of the best submitted model in the aligned region with the ranks in the parentheses; R_MB, RMSD of the best submitted model for the entire chain with the ranks in the parentheses; TM_Ta, TM-score of the best template to the native; TM_Ta_TMaln, TM_LGA, and TM_TMaln are the TM-scores for the structural alignment of the best threading template, the best LGA identified template, and the best TM-Align identified template, respectively; TM_MB, TM-score of the best submitted model. Targets marked with (*) are classified in TBM/FM or FM/TBM category.

that was fixed after target T0328 was submitted. In the inset, we report results from the fixed protocol. This resulted in better models in comparison to the submitted models, consistent with the performance of MetaTASSER on other targets. The average TM-score of the best fixed MetaTASSER models is 0.634 in comparison to the average TM-score of 0.601 for the best submitted models for targets T0283–T0328. Based on Figure 3, the overall improvement of TASSER models over MetaTASSER is for Medium/Hard targets with a global RMSD > 6 Å, a regime where the chunk-TASSER method was mostly used.

Representative examples

Figure 4 shows some successful examples of TASSER for the different target categories. We discuss below some examples in detail.

T0326 is a HA modeling target with a target–template sequence identity of 48% and alignment coverage of 89%. The first 16 residues are unaligned in all 3D-jury selected threading models. Moreover, residues (17–26) of the target are misaligned, mostly due to their low sequence identity (4%) in this region that is also responsible for the high RMSD (7.2 Å) of the initial threading template to native. Although we mainly used chunk-TASSER for modeling Medium/Hard targets, we also used it for this target by including the ab initio folded first chunk for the missing residues and used only the highest sequence identity template in the TASSER refinement procedure. Using the combined approach, the overall RMSD improves from 7.2 to 1.8 Å (3.1 Å) over the aligned region (full-length model). In this case, final model quality depends on the success of modeling the unaligned N-terminus and the misaligned residues 17–26.

DOI 10.1002/prot

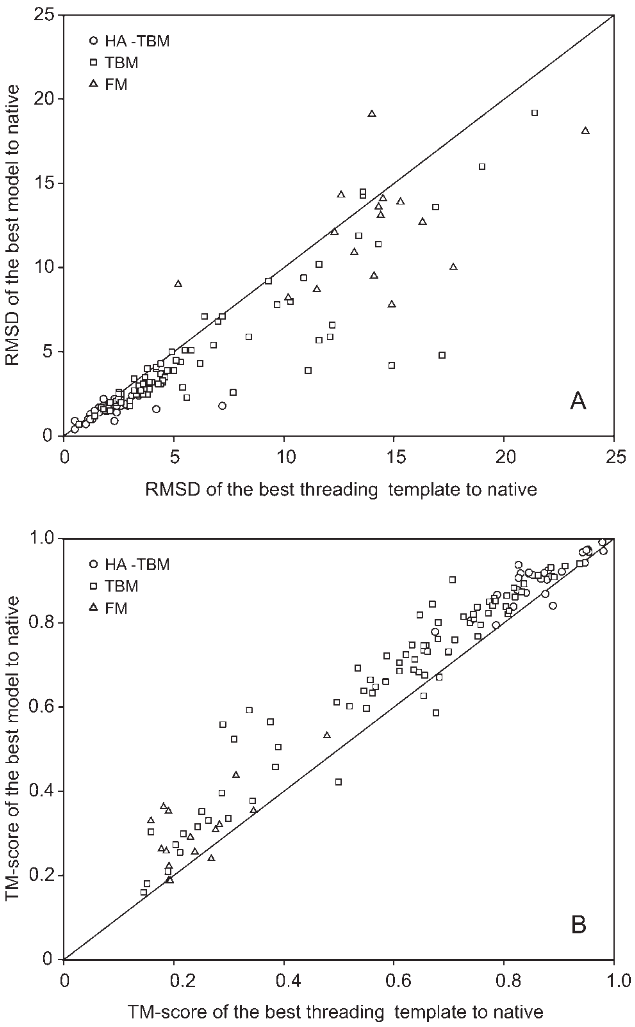

Figure 2

*Comparison between the best TASSER model and the best threading template over the threading aligned region. (**A**) Scatter plot of the RMSD of the best model (aligned region) to the native structure versus the RMSD of the best threading template to the native structure. (**B**) Scatter plot of the TM-score of the best model to the native structure versus the TM-score of the best threading template to the native structure.*

T0298_D2, a TBM target, has an alignment coverage of 84% in the best threading template (2cvoA). Based on their threading *Z*-scores, there are many high confidence templates such as 1gl3A and 1ys4A. TASSER refines these initial threading models. In the aligned region, the best final model is closer to native with a RMSD of 1.8 Å when compared with the initial model's 3.0 Å. The best full-length model has an RMSD from native of 1.8 Å as well.

T0283 and **T0382** are TBM targets. However, our 3D-jury procedure did not find any confidently identified, good templates and classified them as Medium/Hard targets. Both are lower contact order, all α proteins. For both, we employed the chunk-TASSER ab initio method. For T0283, over the threading aligned residues, the RMSD improves from 17.2 to 4.8 Å, mostly by fixing the orientation of the N and C termini, with a final full-length model RMSD to native of 4.8 Å. For T0382, the initial RMSD of the threading model is 14.9 Å that improves to 4.2 Å, with a full-length RMSD of 5.6 Å. In this case, the improvement in model quality is over the entire molecule. These results are consistent with our benchmark test results that show chunk-TASSER has a higher success rate for α proteins than for proteins belonging to other secondary structural classes.

T0370 is a TBM target and is one of the few targets with a 100% covered best threading template alignment to native. The problem with the threading template (1t9mA) is that the helix spanning residues 98–106 are positioned far away in the template structure and there is a huge gap between residues 97 and 98 that TASSER corrected. This orients the helix to the correct position and decreases the global RMSD from 7.7 to 2.6 Å over the threading aligned residues.

T0347 is a two-domain protein with T0347_D1 as a TBM target and T0347_D2 as a FM target. However, it is classified as a Hard target by our 3D-jury procedure. The 3D-jury ranked 1vk1A and 1vz0A as the best threading templates. Closer examination of the threading models suggested that the first domain could be modeled from the templates; however, the second domain needs to be modeled by ab initio procedure. Regular TASSER was used for the prediction of the first domain and chunk-

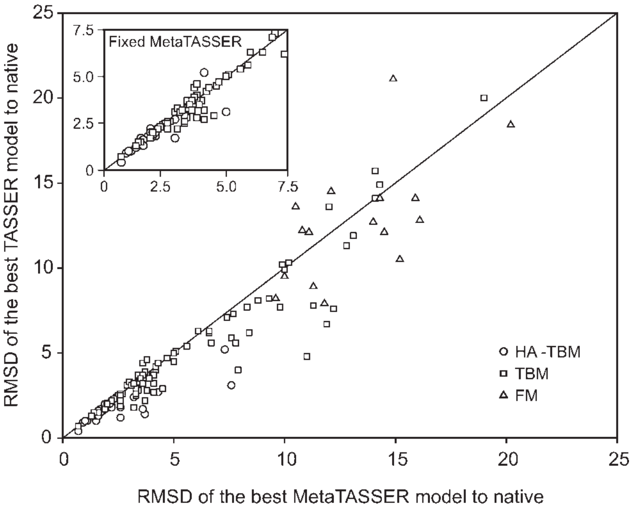

Figure 3

Comparison between TASSER and MetaTASSER of the RMSD of the best model to the native structure. The figure inset shows a similar comparison of the RMSD to native for models from the fixed MetaTASSER protocol (only models with a RMSD < 7.5 Å are shown).

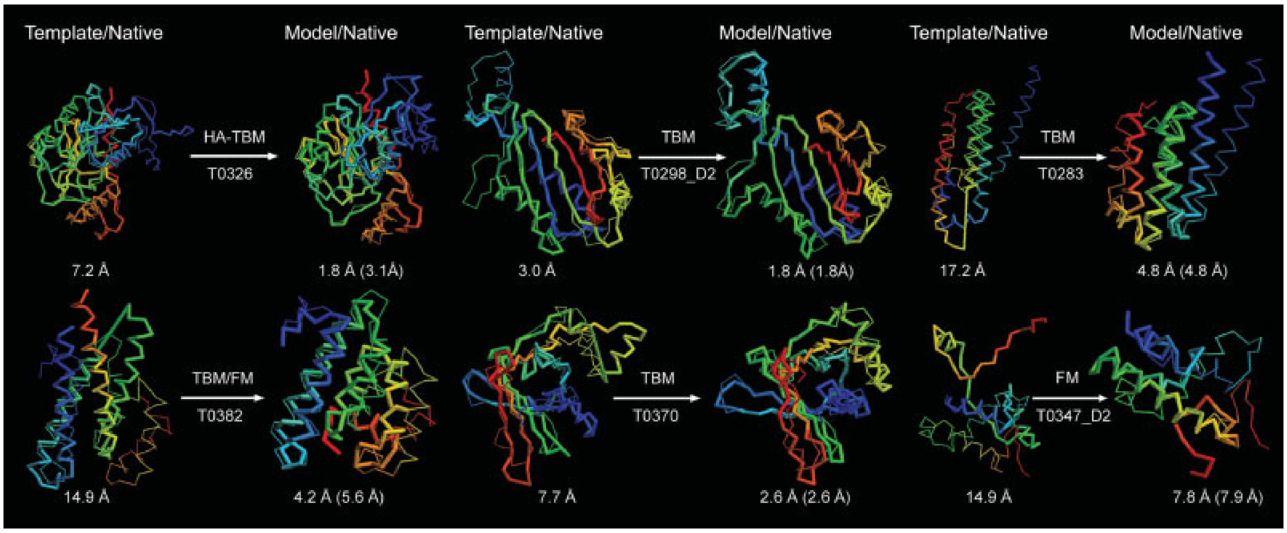

Figure 4

Successful examples of TASSER modeling for the different target categories. For each target, on the left is the superposition of the best threading template (thick backbone) onto the native structure (thin backbone); on the right is the superposition of the final model (thick backbone) onto the native structure (thin backbone). Blue to red goes from the N- to the C-terminus. The numbers below the structural superposition are the RMSD over the aligned regions (entire chain), respectively.

TASSER was applied to second domain. The global RMSD of the N-terminal domain (not shown) is 4.0 Å; while for the C-terminal domain, it is 7.9 Å for the best TASSER model. Also, for the second domain, the RMSD over the aligned region improves from 14.9 to 7.8 Å.

We also analyzed TASSER models for the targets **T0299_D1**, **T0299_D2**, **T0338**, **T0312**, and **T0316_D3** (not shown) that have a significantly poorer TM-score/RMSD with respect to the best-submitted models from other groups. Often this is because we failed to split multidomain proteins into their individual domains. The issue in threading of multidomain targets is that one domain may dominate the alignment scoring function; therefore, the algorithm will fail for the others. Hence, parsing domains would improve both template identification and refinement. For example, for **T0316_D3**, if we use the domain boundary information and thread it separately, we could find the correct template 1exmA. The same holds true for **T0299_D1**. For **T0312**, 3D-jury failed to appropriately rank the best threading templates as first. While on average 3D-jury is a major factor responsible for the improvement in the performance of TASSER in CASP7 relative to CASP6, it failed to rank the best threading template as first in some trivial targets with good initial templates.

DISCUSSION

TASSER and MetaTASSER have been used to generate final models for all CASP7 targets. Consistent with previous CASP6[25] and large scale benchmark[19] results, TASSER shows an improvement over the initial template alignments and generates final models closer to native. In CASP7, we have improved the performance of TASSER by providing a better initial template using three threading methods with a 3D-jury selection protocol. Over the template aligned regions, on average, the final models show an improvement of ~1 Å over the initial templates. The success of TASSER and MetaTASSER can be attributed to the long-range tertiary restraints taken from the consensus of multiple threading models and a reasonably satisfactory nonrestraint knowledge based potential. A somewhat improved method of ranking the final models was also implemented; the average rank of TASSER's best models based on their TM-score to native for the 124 targets/domains is 2.63 compared with 2.75 of 90 targets/domains in CASP6.[25] Another important improvement has been the development of chunk-TASSER for predicting tertiary structures of those targets for which no confident template could be identified by threading.

Comparison of the fully automated procedure MetaTASSER with TASSER showed that modest human intervention in the case of TASSER resulted in better models for FM targets and some the TBM targets, which are classified as Medium/Hard by our 3D-jury method. Thus, for most targets, the limited time simulation of MetaTASSER is adequate.

As noted by the assessors, among the problems of the models generated by TASSER are incorrect side-chain conformations and poor hydrogen bonding patterns in the final models. This is partly because of the on-lattice modeling of the unaligned regions by TASSER and the unphysical geometry of the SPICKER[33] cluster centroid structure when it was used as final model. We are currently developing methods to refine TASSER models using atomic potentials that could provide a high-resolution model that better reproduces the finer structural details.

As anticipated,[5] all single domain TBM and FM targets actually have structurally related folds in the PDB with an average TM-score of 0.54 and a minimum of 0.44 for FM targets (see Table I last column and http://cssb.biology.gatech.edu/skolnick/files/casp7/structaln.html); that is, at least for the CASP7 targets, there were no new folds. For some weakly homologous targets, we could not recognize these related template structures. Thus, we need to improve upon existing methods to detect such remote homologues/structural analogues as well as to improve the accuracy of TASSER in the FM limit. We also need to improve our 3D-jury procedure to select the best threading template as the first rank among the top 20 templates. The modeling of multidomain proteins needs to be addressed by parsing the sequence into domains and then assembling the domains. Finally, we note that the academic version of the TASSER executable is available for download from http://cssb.biology.gatech.edu/skolnick/files/tasser/. The web server is publicly available at http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER.

ACKNOWLEDGMENT

We thank Dr. Adrian Arakaki for help in preparation of the figures.

REFERENCES

1. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
2. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
3. Skolnick J, Fetrow J, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotechnol 2000;18:283–287.
4. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
5. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci USA 2006;103:2605–2610.
6. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
7. Altschul SF, Gish W, Miller W, Myers E, Lipman D. Basic local alignment tool. J Mol Biol 1990;215:403–410.
8. Vingron M, Waterman MS. Sequence alignment and penalty choice. Review of concepts, case studies and implications. J Mol Biol 1994;235:1–12.
9. Altschul SF, Madden TL, Schffer AA, Zhang J, Zhang Z, Miller W, Lipman D-J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
10. Godzik A. Fold recognition methods. Methods Biochem Anal 2003;44:525–546.
11. Jaroszewski L, Rychlewski L, Godzik A. Improving the quality of twilight-zone alignment. Protein Sci 2000;9:1487–1496.
12. Jaroszewski L, Rychlewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
13. David R, Korenberg MJ, Hunter IW. 3D-1D threading methods for protein fold recognition. Pharmacogenomics 2000;1:445–455.
14. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
15. Lundsröm J, Rychlewski L, Bunnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2362.
16. Wallner B, Fang H, Elosson A. Automatic consensus-based fold recognition using Pcons, Pro Q, and Pmodeller. Proteins: Struct Funct Genet Suppl 2003;6:534–541.
17. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441.
18. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
19. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
20. Qian B, Ortiz AR, Baker D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proc Natl Acad Sci USA 2004;101:15346–15351.
21. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 2004;103:5361–6366.
22. Offman MN, Fitzjohn PW, Bates PA. Developing a move-set for protein model refinement. Bioinformatics 2006;22:1838–1845.
23. Valencia A. Protein refinement: a new challenge for CASP in its 10th anniversary. Bioinformatics 2005;21:277.
24. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, Webb B, Greenblatt D, Huang CC, Ferrin TE, Sali A. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 2004;32:D217–D222.
25. 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.
26. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins 2004;56:502–518.
27. Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2004;55:1005–1013.
28. 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.
29. Zhou H, Zhou Y. SPARKS^2 and SP^3 servers in CASP 6. Proteins (Suppl CASP Issue) 2005; (Suppl 7):152–156.
30. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
31. Zhang Y, Skolnick J. A scoring function for the automated assessment of protein structure template quality. Proteins 2004;57:702–710.
32. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002;48:192–201.
33. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein fold. J Comput Chem 2004;25:865–871.