PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Improving threading algorithms for remote homology modeling by combining fragment and template comparisons

Hongyi Zhou and Jeffrey Skolnick*

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

ABSTRACT

In this work, we develop a method called fragment comparison and the template comparison (FTCOM) for assessing the global quality of protein structural models for targets of medium and hard difficulty (remote homology) produced by structure prediction approaches such as threading or ab initio structure prediction. FTCOM requires the C_α coordinates of full length models and assesses model quality based on fragment comparison and a score derived from comparison of the model to top threading templates. On a set of 361 medium/hard targets, FTCOM was applied to and assessed for its ability to improve on the results from the SP^3, SPARKS, PROSPECTOR_3, and PRO-SP^3-TASSER threading algorithms. The average TM-score improves by 5–10% for the first selected model by the new method over models obtained by the original selection procedure in the respective threading methods. Moreover, the number of foldable targets (TM-score $\geq$ 0.4) increases from least 7.6% for SP^3 to 54% for SPARKS. Thus, FTCOM is a promising approach to template selection.

Proteins 2010; 78:2041–2048.

Key words: protein structure prediction; threading; quality assessment prediction; SP^3; PROSPECTOR; SPARKS; PRO-SP^3-TASSER.

INTRODUCTION

Despite numerous determined efforts, threading or fold recognition algorithms are still the only reliable approaches for protein structure prediction.[1–4] One of the key tasks in threading is assessing the quality of a predicted protein structural model and ranking it accordingly. Ranking models or templates for medium/hard targets (remote homology modeling) is particularly difficult because of the weak evolutionary signal of the target to its relevant templates. In this regime, existing threading methods have much room for improvement.[3] Traditionally, the *Z*-score or *E*-value, a statistical measure based on threading scores, is used by most threading algorithms for model selection.[5–9] There are also threading algorithms that use machine-learning approaches such as a neural network[10,11] or a support vector machine (SVM)[12–14] to rank models/templates. Some meta-threading methods use consensus information from models/templates provided by individual threading components.[15–17] In principle, any good quality assessment prediction method[14,18–26] can be used for ranking/selecting models generated by threading methods. However, in practice, it is by no means guaranteed that a particular quality assessment prediction method can improve on existing threading algorithms, especially for methods that depend on consensus among the assessed models. Furthermore, as threading algorithms and their associated model building approaches often produce models with reduced representations such as C_α or main chain atoms, methods that require a detailed atomic model, (e.g., physics-based or knowledge-based all–atom energy function methods[22]), will not be useful. Methods depending on a single target–template alignment[12,13] will not work for models generated from multiple target–template alignments, for example, models built by (chunk-)TASSER simulations[27,28] or models obtained from ab initio folding.[29]

We have previously developed a method, TASSER-QA,[23] which combines fragment comparison and a consensus C_α contact potential derived from the models to be assessed for global model quality assessment prediction. TASSER-QA was used in our recently developed PRO-SP^3-TASSER[30,31] (with the fragment comparison only) and METATASSER[32] servers and for TASSER human prediction[32] in past Critical Assessment of Techniques for Protein Structure Prediction (CASPs) for initial and final model selection. TASSER-QA only performs well when most of the assessed models are among the good ones (this gives rise to a good consensus C_α contact potential). This is always true for "easy" targets[7,8,23,33] where threading provides good templates/alignments. However, for

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

*Correspondence to: Jeffrey Skolnick, Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, 250 14th Street, N.W., Atlanta, GA 30318. E-mail: skolnick@gatech.edu.

Received 3 December 2009; Revised 11 February 2010; Accepted 3 March 2010
Published online 18 March 2010 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22717

medium/hard targets (where, at best, the structural alignments of the template to the native structure are good, whereas the threading alignments themselves are generally of poor quality), good models are not always among the majority. Then, the consensus C_α contact potential is not useful. In this study, we show that for medium/hard targets, a score derived from comparison of the model to the top threading templates can replace the consensus C_α contact potential. These templates can be provided by the same threading methods that provide the input structures. Alternatively, given a set of arbitrarily generated models (e.g., from ab initio folding), then templates from SP^3 will be used. By combining fragment comparison and the template comparison (FTCOM), the new method can assess a single structure because it does not depend on the consensus of the model to other models. Furthermore, the new method, similar to TASSER-QA, depends only on the C_α coordinates. Thus, it can be readily used in threading/fold recognition methods as well as for ranking models generated by refinement protocols.[34,35] Here, we applied and tested FTCOM on models for a test set of 361 medium/hard targets provided by the SP^3,[7,33] PROSPECTOR_3,[8,27] and SPARKS[6,33] threading methods as well as those initial models generated by the latest version of PRO-SP^3-TASSER.[30,31] SP^3,[7,33] PROSPECTOR_3,[8,27] and SPARKS[6,33] all use *Z*-scores to rank models/templates. PRO-SP^3-TASSER uses short chunk-TASSER[28] simulations to build full length initial models, followed by TASSER-QA[23] (the fragment comparison score only) to select initial models as input for further TASSER[27] refinement. By using the new method FTCOM to select models, for medium/hard targets, we show that it consistently improves the TM-score[36] by 5–10% over models selected by the original threading methods.

METHOD

Fragment library generation and fragment comparison

This part of FTCOM is similar to that used in the TASSER-QA[23] approach. SP^{3}[7] threading generates the library for fragment comparison. We then calculate the local sequence similarity between the target and template sequences by computing and recording the alignment score at each target sequence position aligned to each template by threading. The position-specific score is smoothed by averaging over a sliding nine residue window. For each target position, these nine residue long fragments of the top 25 scoring templates form the fragment library for subsequent fragment comparison. This fragment library is similar to that used by ROSETTA,[29] but it is generated and used differently. Fragment comparison is done as follows: for the target sequence, at each residue position in the model, a nine residue fragment with the given residue in the middle (with appropriate adjustments for end effects) is compared with the 25 corresponding fragments in the fragment library according to their pairwise root-mean-square deviation (RMSD). The fragment comparison score E_{frg} is the average RMSD over these 25 fragments and over all model residue positions.

$$E_{frg} = \left(\sum_{j=1}^{N_r}\sum_{i=1}^{25} RMSD[\text{fragment}(i,j)\text{v.s. model(j)}]\right)/(25N_r), \quad (1)$$

where N_{r} is the number of residue of the target, *i* is index of the 25 fragments for each target position, and *j* is target position of the fragment and model. E_{frg} is positive semidefinite, with a smaller value indicative of a better model.

Comparison of the model to the templates

The model to be assessed is compared to the top templates (see below) using the TMalign[37] structural alignment program with the TM-score[36] normalized by model length. Models are assumed to be full length (no missing C_α coordinates). Two scenarios are investigated: (a) templates and assessed models come from the same threading algorithm; (b) if an algorithm provides only models without templates, for example, ab initio generated models, then the top templates from SP^3 (or its enhanced version SP^3-SVM) threading will be used. The template comparison score is obtained by:

$$E_{\mathrm{temp}} = \sum_{n=1}^{N_{\mathrm{t}}} \text{TM-score}(\text{template}_n \text{vs model})/N_{\mathrm{t}}, \quad (2)$$

where N_{t} is number of templates used in the comparison and E_{temp} is their average TM-score. A larger number means a better model. In this work, we used $N_{\mathrm{t}} = 5$ for single method threading. For meta-threading methods, $N_{\mathrm{t}} = 5 \times$ number of individual threading methods; that is, the top five templates from each individual threading are used. The score used for assessing/ranking a model is a simple linear combination of E_{frg} and E_{temp}:

$$E = E_{\mathrm{temp}} - E_{\mathrm{frg}}. \quad (3)$$

Testing dataset and threading algorithms

We used the same benchmark set as in PRO-SP^3-TASSER.[30] The original 723 protein set has 375 medium/hard targets as classified by the *Z*-score of the top template identified in SP^3 threading:[7] Targets whose top template has a *Z*-score ≥ 6.0 are classified as easy, those with a *Z*-score ≤ 4.5 as hard, and those having a $4.5 <$ *Z*-score < 6.0 as medium targets, respectively. All targets have a length <250 residues. After visually checking the target structures, we removed a few single/double helix structures; this gave a subset of 361 medium/hard targets that were used for testing. The threading library structures were released before the target structures. The target and threading template libraries may be found at http://cssb.biology.gatech.edu/skolnick/files/ftcom/.

All tested threading algorithms, SP^3,[7] SPARKS,[6] PROSPECTOR_3,[8] and PRO-SP^3-TASSER,[30] were previously developed by the authors. Full length models for SP^3, SPARKS, and PROSPECTOR_3 are built by MODELLER[38] for the top 100 templates of each target in each method. FTCOM selects models from these 100 models. Limited time chunk-TASSER[28] simulations were used to construct up to 155 full length models per target in PRO-SP^3-TASSER initial stages before full TASSER refinement.

SP^3,[7] SPARKS,[6] PROSPECTOR_3,[8] and PRO-SP^3-TASSER[30] are single method threading algorithms, whereas the threading in PRO-SP^3-TASSER[30] is a meta-threading algorithm with five embedded individual threading approaches. The resulting models are not tied to a single target–template alignment. We chose these threading algorithms because they performed well in past CASP experiments.[27,31–33] In practice, it is not easy to test algorithms developed by other people because of issues associated with template library and sequence profile generation control. Nevertheless, the tested threading algorithms should be representative, and FTCOM should also be applicable to other threading algorithms as well. The FTCOM executable and associated documentation may be downloaded from http://cssb.biology.gatech.edu/skolnick/files/ftcom/.

RESULTS

We first tested FTCOM using the following scenario: (a) the structures are ranked from the same threading method that provides the templates. Table I shows the results for the average TM-score[36] per target and the number of targets having selected models with TM-score > 0.4 to native (foldability) for SP^3, SPARKS, PROSPECTOR_3, and PRO-SP^3-TASSER on the 361 target dataset. For the selected first models, the least improved method, PRO-SP^3-TASSER, has a 5.7% TM-score improvement over the original ranking method (E_{frg} in TASSER-QA[23]). In terms of foldability, SP^3 shows little improvement for the first model but the improvement (8.3%) is obvious for the best of five models. As for the other methods, improvements in foldability are significant except for the best model of PRO-SP^3-TASSER. Notice that the original results of PRO-SP^3-TASSER are much better than those of the other three single method threading methods. This is partly due to its meta-threading feature and partly due to its use of limited time chunk-TASSER[28] simulations for building full length models of medium/hard targets. We note that chunk-TASSER has significantly better performance than the original TASSER[34] approach, whereas TASSER performs much better than MODELLER[38] for medium/hard targets.

Table I

Average TM-Score of Selected Models and Number of Targets Having Models With TM-Score > 0.4 to Native on the 361 Protein Dataset[a]

	First model				Best of top five models			
Method	Original		+FTCOM		Original		+FTCOM	
SP^3	0.290	66	0.309	67	0.346	96	0.353	104
SPARKS	0.268	50	0.295	63	0.314	75	0.332	85
PROSPECTOR_3	0.283	61	0.309	78	0.316	79	0.342	106
PRO-SP^3-TASSER initial model	0.348	98	0.368	120	0.397	148	0.404	150

[a] Compared templates are from the same threading method that provides the models used for ranking. In each cell, the first number is the average TM-score; the second number is the number of targets having a model with TM-score > 0.4 to native.

The quality of FTCOM selected models in PRO-SP^3-TASSER (the average TM-score of the first/best models are: 0.368/0.404) from models built by short chunk-TASSER runs already matches that of the much longer, original PRO-SP^3-TASSER approach[30] (the average TM-scores of first/best models are: 0.364/0.402). Figure 1 shows the scatter plot of TM-scores of the first models with and without FTCOM. While most targets are improved by FTCOM, some become worse. The TM-score of a few targets improves from 0.2–0.3 to 0.5–0.7. Figure 2 shows the cumulative histograms of the first model TM-scores, viz., the number of targets having a TM-score of the first model to native greater than the given cutoffs. At all cutoff levels, results with FTCOM show a clear improvement over the original threading algorithms.

We next test FTCOM using scenario (b): the top five templates from our SP^3 method are used in ranking models for the other threading algorithms. This is particularly useful when ranking ab initio generated models or when template information is inaccurate or missing. The results are compiled in Table II. Notice that the results for PRO-SP^3-TASSER are slightly worse than those in scenario (a) (see Table I) due to the poor quality templates of SP^3 as compared with PRO-SP^3-TASSER. The reverse is true for SPARKS and PROSPECTOR_3 because SP^3 provides better templates than SPARKS and PROSPECTOR_3. These results indicate that if we can use improved templates, we can further improve the selection.

Toward this goal, we have implemented a SVM approach for SP^3 (SP^3-SVM) similar to the one developed by Xu.[13] The SVM itself can improve template quality by on average around 6% in TM-score for SP^3; this is about the same improvement as that of FTCOM for SP^3 (see Table I). However, as the SVM features for a given model are calculated based on known target–tem-

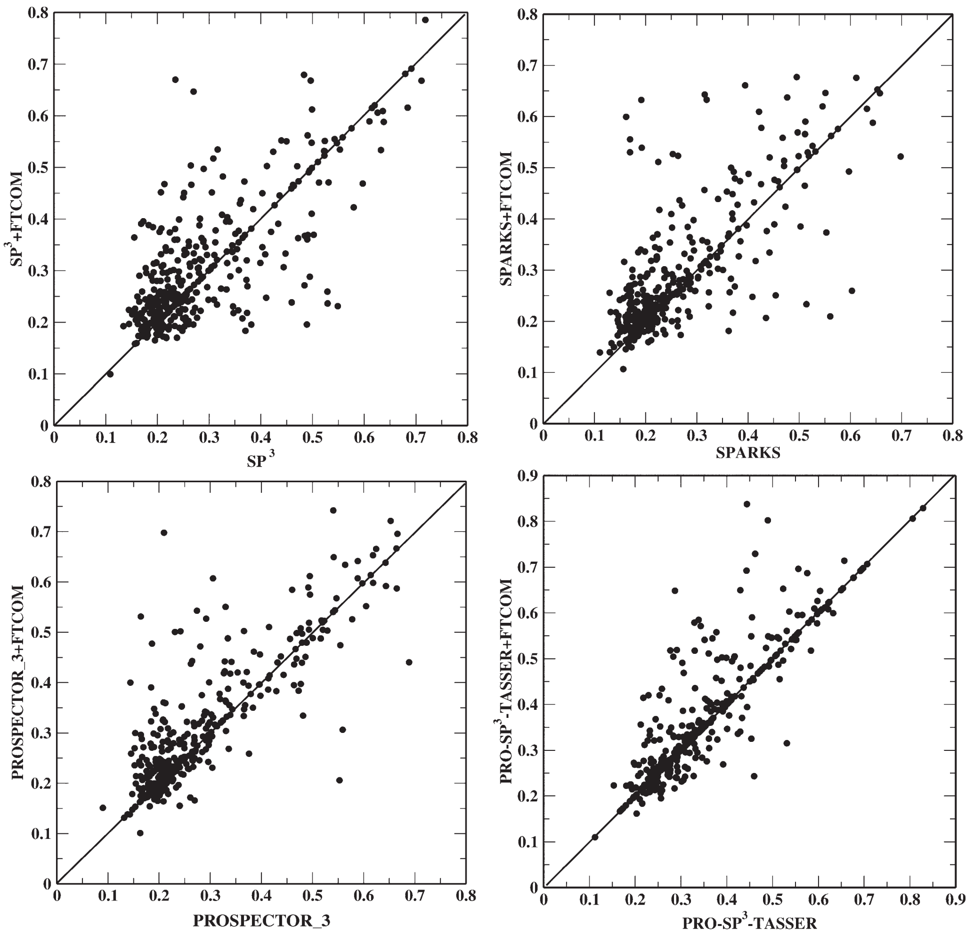

Figure 1

Scatter plots of first model TM-scores with versus those without FTCOM. Compared templates are from the original threading approach.

plate alignments provided by threading, it is not feasible when the target–template alignment is not available, for example, when we rank models generated from ab initio folding or built by the chunk-TASSER component of PRO-SP^3-TASSER. Results using templates provided by SP^3-SVM for the template comparison are shown in Table III. There are clear improvements over results using SP^3 templates. The TM-score of SP^3 now improves by 8.4% for the first models, whereas the improvements for SPARKS, PROSPECTOR_3, and PRO-SP^3-TASSER are 13.5, 11.7, and 5.2%, respectively, for the first model. Notable is the foldability improvements of PROSPECTOR_3 in Tables I–III. Even though its average TM-scores after FTCOM are slightly worse than or the same as those for SP^3, the number of foldable targets is consistently better than SP^3. The reason for the difference is that within the ~ 100 models to be ranked, PROSPECTOR_3 has 9.7% of its models with a TM-score > 0.4 to native, whereas SP^3 has only 5.0%. (SPARKS has 2.5% and PRO-SP^3-TASSER has 22.5%).

Using the compared templates from each threading algorithm itself [scenario (a)], we now examine the separate contributions of the two terms E_{frg} and E_{temp} by looking at the average TM-scores of the selected first models and the average Pearson correlation coefficient (per target) between the ranking scores and the real TM-scores of the ranked models. Table IV compiles the average TM-scores of the first models selected by E_{frg}, E_{temp}, and $E_{\text{temp}} - E_{\text{frg}}$ and their correlation with TM-scores. Each component does not always improve the TM-score. However, the combined score $E_{\text{temp}} - E_{\text{frg}}$ has a consistently better TM-score and Pearson correlation coefficient. For SP^3, E_{temp} contributes more than E_{frg}, whereas in other methods, the reverse is true. Although the correlation coefficients are small, they are all better than the average Z-score versus TM-score correlation coefficient of 0.17 in SP^3.

The first models selected by FTCOM have on an average a ~6% better TM-score than those models selected by either E_{temp} or E_{frg} (E_{frg} is the original method in

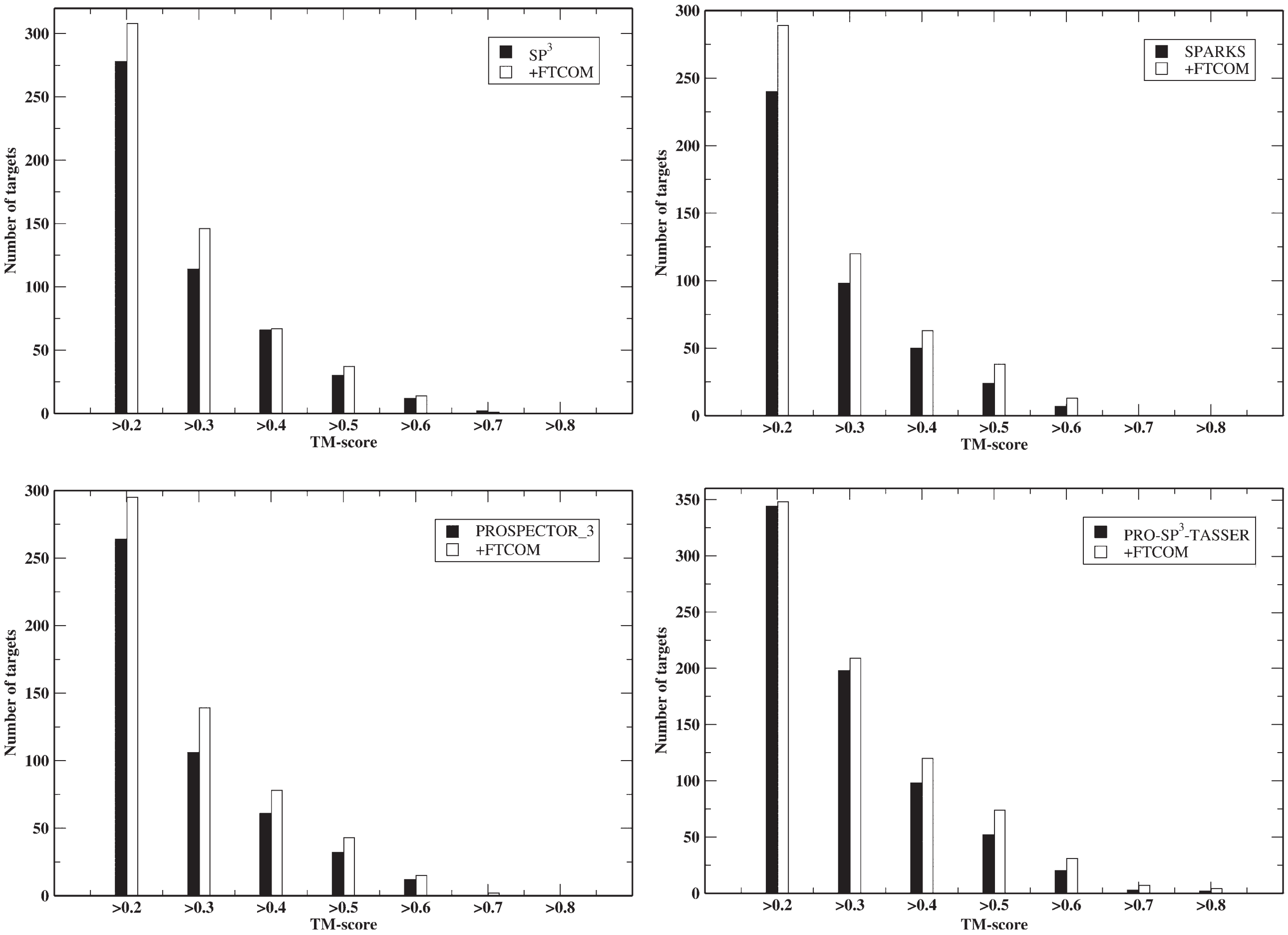

Figure 2

Cumulative histograms of first model TM-scores at various TM-score cutoffs. Compared templates are those given by the original threading algorithm.

PRO-SP^3-TASSER for initial model selection). In all the tested threading algorithms, the combination of E_{frg} and E_{temp} in FTCOM effectively lowers the false positive rate of the individual score for ranking. For example, for target 2qzg_A in PRO-SP^3-TASSER ranking, E_{frg} picks a model whose TM-score = 0.46, whereas E_{temp} picks a model whose TM-score = 0.47, with relatively large (poor) E_{frg} value. The combined score picks a model whose TM-score = 0.74. There are also cases that either E_{frg} or E_{temp} picks the same template as the combined score. For example, for 2qho_B in PROSPECTOR_3 ranking, FTCOM selects the same model as E_{frg} does, whereas for 2jxp_A, FTCOM selects the same model using E_{temp}.

In SP^3, SPARKS, and PROSPECTOR_3, each model is built from a single template alignment (i.e., each model has a corresponding template). For a given template, the quality of model built from it depends on threading alignment accuracy. FTCOM improves the selected model

Table II

Same as Table I But Compared With Templates From SP^3

	First model				Best of top five models			
Method	Original		+FTCOM		Original		+FTCOM	
SPARKS	0.268	50	0.300	67	0.314	75	0.341	98
PROSPECTOR_3	0.283	61	0.314	84	0.316	79	0.356	117
PRO-SP^3-TASSER initial model	0.348	98	0.362	115	0.397	148	0.400	147

Table III

Same as Table I But With Compared Templates From SP^3 -SVM

	First model				Best of top five models			
Method	Original		+FTCOM		Original		+FTCOM	
SP^3	0.290	66	0.315	71	0.346	96	0.360	110
SPARKS	0.268	50	0.304	77	0.314	75	0.343	99
PROSPECTOR_3	0.283	61	0.316	84	0.316	79	0.359	123
PRO-SP^3-TASSER initial model	0.348	98	0.367	122	0.397	148	0.405	151

Table IV

Average TM-Score of First Model Selected by E_{frg}, E_{temp}, *and* $E_{temp} - E_{frg}$ and Average (Per Target) Correlations of Them to TM-Scores on the 361 Dataset[a]

Method	Original	E_{frg}		E_{temp}		$E_{temp} - E_{frg}$	
SP^3	0.290	0.285	−0.29	0.293	0.29	0.309	0.38
SPARKS	0.268	0.278	−0.34	0.266	0.25	0.295	0.38
PROSPECTOR_3	0.283	0.301	−0.38	0.284	0.25	0.309	0.41
PRO-SP^3-TASSER initial model	0.348	0.348	−0.39	0.342	0.22	0.368	0.44

[a]Compared templates are from the same threading method that provide models for ranking. In each cell, the first number is the TM-score, whereas the second number is Pearson's linear correlation coefficient.

Table V

Average TM-Score of Templates and Average Threading Alignment Accuracy of Model Selected By Original Threading and By FTCOM on the 361 Dataset for the Three Single Threading Methods

	First model				Best of top five models			
Method	Original		+FTCOM[a]		Original		+FTCOM[a]	
SP^3	0.400	0.179 0.264 0.310	0.409	0.186 0.273 0.326	0.462	0.252 0.359 0.416	0.468	0.245 0.348 0.395
SPARKS	0.378	0.140 0.220 0.265	0.394	0.163 0.253 0.298	0.435	0.215 0.305 0.357	0.443	0.216 0.317 0.371
PROSPECTOR_3	0.375	0.179 0.258 0.304	0.400	0.197 0.276 0.323	0.428	0.221 0.300 0.348	0.451	0.236 0.330 0.385

[a]Scenario (a) is used: that is, templates in FTCOM template comparison are from the same threading method that provides models for ranking. The first column in each cell is the average TM-score and the numbers in second column are the average threading alignment accuracies defined by the ratio of (the number of threading aligned residues exactly, within ±1 and ±2 residues match to those of the structural alignment by TM-align)/(number of total structurally aligned residues).

quality partly by selecting better templates (those with a better TM-score to native as assessed by their structural alignment) and partly by selecting better alignments (same template quality but better threading alignment). Table V shows the average TM-score calculated by the structural alignment algorithm, TM-align[37] of templates corresponding to the selected model and the average threading alignment accuracy defined by the ratio of the number of threading aligned residues that exactly, and within ±1 and ±2 residues, match the structural alignment to the total number of structurally aligned residues for a given template. A ratio closes to 1 means that the threading alignment is close to structural alignment. In Table V, the improvements of the average template TM-scores (by structural alignment) of the first model are 2.2, 4.5, and 6.7% for SP^3, SPARKS, and PROSPECTOR_3, respectively. Thus, FTCOM selects better templates than the original threading methods. Table V also shows that except for the case of best of top five models in SP^3, FTCOM improves alignment accuracy.

By comparing the average TM-scores of templates in Table V to those of the selected models in Tables I–IV, we notice that there are still big gaps between threading generated alignments and the best possible alignments identified by TM-align. This suggests the further application of FTCOM to select models built from alternative/better alignments of existing threading templates. We are currently investigating this possibility.

It is also informative to compare TASSER-QA with FTCOM. As TASSER-QA depends on consensus information of models to be assessed, threading methods that provide enriched good models will work well with TASSER-QA. Table VI shows the average TM-scores and foldability of models identified by threading methods using FTCOM and TASSER-QA for model selection. In all threading methods, FTCOM is better than TASSER-QA, especially for foldability. As shown earlier, PROSPECTOR_3 and PRO-SP^3-TASSER have more good models (larger percentage of models having a TM-score $>$ 0.4 to native) within the models to be ranked; thus, the differences between TASSER-QA and FTCOM are relatively smaller compared with those in SP^3 and SPARKS. Therefore, FTCOM demonstrates an improvement over TASSER-QA, an approach that performed reasonably well in the past CASPs.[31,32] Table VI also confirms our earlier statement that a good quality assessment prediction method like TASSER-QA that depends on consensus model information is not necessarily able to improve individual threading methods like SP^3 and SPARKS.

So far, our tests of FTCOM are on models generated by MODELLER (using SP^3, SPARKS, and PROSPECTOR_3) or chunk-TASSER (in PRO-SP^3-TASSER) that are mostly based on templates. It is not clear if FTCOM

Table VI

Average TM-Score and Foldability of Models by Threading Methods Using FTCOM and Threading Using TASSER-QA for Model Selection on the 361 Dataset

	First model				Best of top five models			
Method	+FTCOM[a]		+TASSER-QA		+FTCOM[a]		+TASSER-QA	
SP^3	0.309	67	0.289	48	0.353	104	0.339	83
SPARKS	0.295	63	0.268	38	0.332	85	0.313	65
PROSPECTOR_3	0.309	78	0.302	70	0.342	106	0.339	100
PRO-SP^3-TASSER initial model	0.368	120	0.364	114	0.404	150	0.396	139

[a]Scenario (a) is used.

Table VII

Average TM-Score of Models Selected by FTCOM With SP^3 Templates, SPICKER, and TASSER-QA From the 5000 Models Generated by ROSETTA on 351 Targets[a]

	SP^{3b}		SPICKER		TASSER-QA		FTCOM	
Medium (146)	0.319	0.366	0.277	0.315	0.269	0.308	0.298	0.337
Hard (205)	0.267	0.330	0.299	0.334	0.301	0.342	0.315	0.355
No top five SP^3 template model with TM-score > 0.4 (258)	0.234	0.277	0.275	0.312	0.276	0.312	0.283	0.318

[a] Number in brackets are number of targets. The first number in each cell is for the first model; the second number is for best of top five models.
[b] SP^3 threading template model for comparison of ab initio and threading template models.

performs well on template-free models such as those generated with ROSETTA.[39] Furthermore, since FTCOM has some dependence on template quality, does FTCOM perform well when threading templates are poor, as is the case for hard targets? For hard targets, ab initio methods such as ROSETTA[39] on average do better than threading methods. To answer the above questions, we used ROSETTA to generate 5000 models for each target (successfully generated for 351 targets) and compared FTCOM with the conventional clustering method SPICKER[40] and the TASSER-QA[23] ranking method. In Table VII, we show the performance of FTCOM using SP^3 templates along with SPICKER clustering and TASSER-QA ranking on the ab initio models generated by ROSETTA.[39] For comparison, we also show the SP^3 threading predictions based on templates. For medium targets, FTCOM is around 8% (7%) better than both SPICKER and TASSER-QA in terms of TM-scores for the first (best of top five) models. However, predictions with FTCOM, SPICKER, and TASSER-QA are all worse than the template models by SP^3 for these medium targets. For hard targets where SP^3 template models have a >10% worse TM-score than the ab initio predictions, FTCOM with the SP^3 templates is still ~ 5% (4%) better than when SPICKER or TASSER-QA is used to select the first (best) models. In the extreme case when all top five SP^3 threading models have TM-score < 0.4 to native, FTCOM is slightly better than SPICKER/TASSER-QA (>2%). These results indicate that FTCOM performs well on ab initio models even when template models are much worse than the ab initio predictions.

The above results also show that FTCOM is better than SPICKER for selecting model from ab initio generated models. In Ref. [23], we have shown that TASSER-QA performs similar to SPICKER for selecting models from TASSER trajectories. In this study, we also find that FTCOM performs the same as SPICKER when selecting models from chunk-TASSER[28] trajectories (data not shown). A simple explanation is that the chunk-TASSER generated trajectories are populated near the consensus structure of the input threading templates and fragments that makes it easier for clustering methods like SPICKER to find the consensus structure. FTCOM also selects models that are mostly compatible with the templates and fragments as well as the consensus structures.

CONCLUSIONS AND DISCUSSIONS

We have shown that threading algorithms that use FTCOM to rerank template models perform consistently better than the corresponding original methods for medium/hard targets. FTCOM uses fragment and template comparisons. The former depends on the local quality of the evaluated structure models, whereas the latter depends more on the global structure as it compares the whole structure to the whole template. FTCOM offers the advantage that templates can be from the same threading method that provides models to be ranked/assessed or from default SP^3 provided templates when the methods that provide models have no template information. To further improve model ranking/assessment, templates identified by SP^3-SVM can also be applied. As FTCOM does not use consensus information from models to be ranked/assessed, it can assess the global quality of a single structural model for medium/hard targets. To achieve this goal, an accurate mapping between the score $E = E_{\text{temp}} - E_{\text{frg}}$ and the actual TM-score is needed. Work in this direction is underway.

REFERENCES

1. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
2. Skolnick J, Fetrow J, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotechnol 2000;18:283–287.
3. Wang G, Jin Y, Dunbrack RJ. Assessment of fold recognition predictions in CASP6. Proteins 2005;61(Suppl. 7):46–66.
4. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69:38–56.
5. Godzik A. Fold recognition methods. Methods Biochem Anal 2003;44:525–546.
6. 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.
7. 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.
8. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins 2004;56:502–518.
9. Karplus K, Barrett C, Hughey R. Hidden markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.

10. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
11. Xu Y, Xu D, Olman V. A practical method for interpretation of threading scores: an application of neural networks. Stat Sin Spec Issue Bioinformatics 2002;12:159–177.
12. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. J Bioinformatics Comput Biol 2003;1:95–117.
13. Xu J. Fold recognition by predicted alignment accuracy. IEEE/ACM Trans comput Biol Bioinformatics 2005;2:157–165.
14. Cheng J, Baldi P. A machining learning information retrieval approach to protein fold recognition. Bioinformatics 2006;22:1456–1463.
15. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
16. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441.
17. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 2005;15:900–913.
18. 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.
19. 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.
20. Wang Z, Tegge A, Cheng J. Evaluating the absolute quality of a single protein model using structural features and support vector machines. Proteins 2009;75:638–647.
21. McGuffin L. ModFOLD server for the quality assessment of protein structural models. Bioinformatics 2008;24:586–587.
22. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Sci 2006;15:1653–1666.
23. Zhou H, Skolnick J. Protein model quality assessment prediction by combining fragment comparisons and a consensus C_α contact potential. Proteins 2007;71:1211–1218.
24. Paluszewski M, Karplus K. Model quality assessment using distance constraints from alignments. Proteins 2009;75:540–549.
25. Benkert P, Tosatto S, Schomburg D. QMEAN: a comprehensive scoring function for model quality assessment. Proteins 2008;71:261–277.
26. Cheng J, Wang Z, Tegge A, Eickholt J. Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins 2009;77(Suppl. 9):181–184.
27. 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.
28. Zhou H, Skolnick J. Ab initio protein structure prediction using chunk-TASSER. Biophys J 2007;93:1510–1518.
29. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306:1191–1199.
30. Zhou H, Skolnick J. Protein structure prediction by pro-sp3-TASSER. Biophys J 2009;96:2119–2127.
31. Zhou H, Pandit S, Skolnick J. Performance of the Pro-sp3-TASSER server in CASP8. Proteins 2009;77:123–127.
32. Zhou H, Pandit SB, Lee S, Borreguerro J, Chen H, Wroblewska L, Skolnick J. Analysis of TASSER based CASP7 protein structure prediction results. Proteins 2007;69(S8):90–97.
33. Zhou H, Zhou Y. SPARKS 2 and SP3 servers in CASP6. Proteins 2005;61(Suppl. 7):152–156.
34. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
35. 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.
36. Zhang Y, Skolnick J. A scoring function for the automated assessment of protein structure template quality. Proteins 2004;57:702–710.
37. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
38. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
39. 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.
40. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein fold. J Comput Chem 2004;25:865–871.