PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Protein model quality assessment prediction by combining fragment comparisons and a consensus C_α contact potential

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 fully automated method for the quality assessment prediction of protein structural models generated by structure prediction approaches such as fold recognition servers, or ab initio methods. The approach is based on fragment comparisons and a consensus C_α contact potential derived from the set of models to be assessed and was tested on CASP7 server models. The average Pearson linear correlation coefficient between predicted quality and model GDT-score per target is 0.83 for the 98 targets, which is better than those of other quality assessment methods that participated in CASP7. Our method also outperforms the other methods by about 3% as assessed by the total GDT-score of the selected top models.

Proteins 2008; 71:1211–1218.

Key words: model quality assessment prediction; TASSER; SP³.

INTRODUCTION

Selecting the best quality models from a set of predicted structures is an essential part of protein structure prediction.[1,2] In the latest Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP7)[3] (http://predictioncenter.org/casp7/), a new prediction category that judges the quality of models and the reliability of predicting certain residues in the structure was implemented. There are a number of methods that address this issue that can be conceptually divided into the following three categories: (1) statistical methods, (2) machine learning methods, and (3) energy function-based methods. Examples of statistical methods include those based on clustering in ab initio structure prediction[4,5] and the 3D-jury[6,7] approach used for meta-servers. Their outcome depends on the population statistics of the models. Statistical methods do not depend on the details of the models, whereas energy function-based methods require a detailed molecular description. Such energy function-based methods include physics-based and knowledge-based energies for discriminating the native structure from decoys, near native structure selection and for the assessment of protein models.[8–21] Some meta-servers use machine learning approaches to select models from individual servers.[22–24] Eramian *et al.* studied 24 assessment scores in the literature and used support vector machine (SVM) regression to combine some of these machine learning approaches and energy function-based approaches for predicting errors in protein structure models.[25] In addition to methods that predict the overall quality of protein structure models, there are also alternative methods that assign a local quality score to each residue. This could be useful for constructing hybrid protein structure models.[7,26,27]

In this work, we have developed a knowledge-based energy function method, which employs a score function based on fragment comparisons in combination with a statistical potential to predict the quality of protein models. The approach only uses the C_α coordinates of the models. We tested our method in CASP7 (submitted prediction under name TASSER-QA), where it was shown to have the best average Pearson linear correlation coefficient and was the top ranked among the participating methods in its ability to select the best structures from the CASP7 server models. The method was also used by the TASSER group for selecting models for submission.[28]

METHOD

Fragment library generation and fragment comparison

The SP³ threading method[29,30] was used to generate the fragment library for fragment comparison. The details of SP³ were published elsewhere.[29] Here, we reoptimized

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, 250 14th Street, N.W., Atlanta, GA 30318. E-mail: skolnick@gatech.edu
Received 31 May 2007; Revised 9 August 2007; Accepted 5 September 2007
Published online 14 November 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21813

the parameters with a full grid search on the five dimensional parameter space. The new optimal solution (w_0, w_1, $w_{secondary}$, w_{struc}, w_{shift}) is (3.5, 0.1, −1.50, 0.5, 0.7). This resulted in the one-to-one match alignment accuracy of 66.1% against the ProSup structure alignment benchmark[31] compared to the original accuracy of 65.3%. Another change made to SP^3 that increases its sensitivity is the inclusion of profiles generated by PSI-BLAST[32] with a looser *e*-value cutoff of 1.0. To the target sequence, the sequence profile is replaced by the average of two profiles with *e*-value cutoffs 0.001 and 1.0, and to the templates, the structurally derived profile is replaced by the average of original and the PSIBLAST profile with an *e*-value cutoff of 1.0.

We extend the SP^3 threading method[29] to compute local sequence similarity between query and template sequences by computing and recording the alignment score at each query sequence position aligned to each template during threading. The position-dependent score is then smoothed by averaging over a nine-residue-wide sliding window. For each position, nine-residue-long fragments of the top 25 scoring templates are selected to form the fragment library used for subsequent fragment comparison. Fragment comparison is done in the following way: for each residue position in the model for the query sequence, a nine-residue fragment with the given residue in the middle (less in the N- or C-terminus, e.g., the fragment for the first residue will be residues 1–5) 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 is the average RMSD over the 25 fragments and over all model residue positions.

Consensus C_α contact potential

The consensus C_α contact potential is constructed from the set of models to be assessed using a similar procedure as was applied to TASSER.[33,34] For the set of models to be assessed (in practice the top scoring models of the CASP7 servers), a protein-specific consensus C_α contact potential between C_αs is calculated as:

$$E_{\text{contact}} = w_{r3}\Theta_5(p_{ij} - p^0)\sum_{j>i}\Theta_5(r_{ij} - 6\ \text{Å}) + w_{r4}\Theta_6\left[\Theta_5(p_{ij} - p^0)\sum_{j>i}\Theta_6(r_{ij} - 6\ \text{Å}) - N_{\text{cp}}\right] \quad (1)$$

where $\Theta_5(x)$ and $\Theta_6(x)$ are the step functions defined as

$$\Theta_5(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0; \end{cases} \qquad \Theta_6(x) = \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{if } x < 0; \end{cases} \quad (2)$$

p_{ij} is the fraction of models that the ith residue C_α is in contact with the jth residue C_α in the models, and r_{ij} is the C_α distance between residues i and j. p^0 defines the minimal fraction threshold that the two residues are in contact, and 6 Å is the distance cutoff that defines whether a given pair of residues are in contact. In this work, p_{ij} is predicted from all the models to be assessed. For example, if the ith residue C_α is in contact with the jth residue C_α (distance $<$ 6 Å) in n_1 models of total n assessed models, then $p_{ij} = n_1/n$. When $p_{ij} > p^0 = 0.3$, we consider residues i and j to be involved in a real contact in the native structure and the Θ_5 terms are effective in Eq. (1). The first term in Eq. (1) favors pairs predicted as being in contact that are within 6 Å, whereas the secondary term penalizes predicted contact pairs that are farther apart than 6 Å when the total violation exceeds a threshold value of N_{cp}. The weights w_{r3} and w_{r4}, and N_{cp} are taken from TASSER.[33,34]

The score used for predicting model quality is a simple combination of E_{frg} and E_{contact}:

$$E_{\text{q}} = E_{\text{frg}} + [w_{\text{c}} \times (E_{\text{contact}}/N_{\text{r}})], \quad (3)$$

where w_{c} is a relative weight of the two terms that will be determined from optimization on a training set (see below). N_{r} is the number of residues in the model. Because the value of E_{q} is not between 0 and 1, we transform it by the following logistic function so that the ranking score is in the range 0–1.

$$E_{\text{q}}^{\text{t}} = \frac{1}{1 + \exp(E_{\text{q}})} \quad (4)$$

Training and testing datasets

The only free parameter in the current approach is w_{c}. We optimized it on all the server models for the 40 easy targets (classified as those having a SP^3 Z-score $\geq$5.6) in CASP6.[30] The object function is the total TM-score[35] of the selected top models with respect to their native structures. To mimic the real prediction situation, the template library used for generating the fragment library in the optimization was built from structures released before May 28, 2004, which was before the CASP6 prediction season. The optimized result for w_{c} is 0.2.

We tested and compared our method with other methods on the server models of the 98 targets in CASP7.[3] The fragment libraries for all the targets were generated during the CASP blind test. We only evaluated the first models from each server and only those models with full length structures (no missing residues). Predictions by other quality assessment methods were downloaded from the CASP7 website. We only examined predictions for tertiary structures (not for alignment models) so that we could also calculate the model GDT-score[36] and TM-score.[35]

Table I
Average Correlation Coefficients and Total GDT-Scores of the Top Quality Assessment Predictors in CASP7

Method	Number of targets	Pearson		Spearman		GDT-score	
		Average[a]	*P*-value[b]	Average[c]	*P*-value	Total (average)[d]	*P*-value
125 (TASSER-QA)[e]	98	**0.834**		0.734		**58.06 (0.592)**	
TASSER-QA-all[f]	98	0.828	0.51	**0.791**	8.6×10^{-8}	57.22 (0.584)	0.08
634 (Pcons)	98	0.811	0.16	0.757	0.18	55.05 (0.562)	3.7×10^{-7}
556 (LEE)	96	0.797	0.049	0.747	0.67	53.86 (0.561)	2.4×10^{-6}
713 (Circle-QA)	98	0.730	2.7×10^{-14}	0.662	1.7×10^{-5}	56.06 (0.572)	1.9×10^{-4}
633 (ProQ)	98	0.719	1.1×10^{-12}	0.597	6.4×10^{-12}	54.15 (0.553)	4.5×10^{-6}
038 (GeneSilico)	88	0.712	1.4×10^{-11}	0.621	1.4×10^{-8}	49.09 (0.558)	2.1×10^{-5}
692 (ProQlocal)	98	0.711	2.5×10^{-11}	0.591	8.1×10^{-12}	54.03 (0.551)	3.1×10^{-6}
178 (Bilab)	98	0.699	1.7×10^{-12}	0.585	2.9×10^{-10}	54.53 (0.556)	1.2×10^{-4}
704 (QA-ModFOLD)	98	0.675	5.9×10^{-15}	0.600	3.9×10^{-10}	53.92 (0.550)	2.0×10^{-6}
699 (ABIpro-h)	98	0.674	2.5×10^{-12}	0.629	6.7×10^{-7}	56.49 (0.576)	7.8×10^{-3}
091 (Ma-OPUS)	97	0.662	1.6×10^{-17}	0.608	5.5×10^{-9}	53.25 (0.549)	5.1×10^{-5}
013 (Jones-UCL)	81	0.647	2.3×10^{-16}	0.544	7.7×10^{-16}	45.28 (0.559)	3.5×10^{-7}
703 (QA-ModCHECK)	71	0.624	1.9×10^{-13}	0.575	2.7×10^{-8}	30.61 (0.431)	1.4×10^{-8}
717 (CaspIta-FRST)	90	0.586	1.4×10^{-25}	0.518	3.9×10^{-21}	48.49 (0.539)	6.6×10^{-10}
025 (Zhang-Server)[g]	*98*		—		—	*57.35 (0.585)*	*0.07*
Best[h]	*98*		—		—	*62.00 (0.633)*	

The structural similarity of model to native is measured by GDT-score.[36]
Bold values are the best values among QA methods.
[a]Pearson linear correlation coefficient. The results are the average per target. They may be slightly different from those in CASP7 website; we did not include models where the alignments alone are given. We also calculate the GDT-scores based on target chains rather than domains.
[b]*P*-values are given for the differences of results between TASSER-QA and other methods. Difference with *P*-value of <0.05 is considered significant at 95% confidence level.
[c]Spearman rank correlation coefficient. The results are the average per target.
[d]Total GDT-score of the server models that are ranked the first by quality assessment prediction methods. Numbers in parenthesis are averages per target.
[e]This work using only first models from servers. Descriptions of all other methods can be found at the CASP7[3] website http://predictioncenter.org/casp7/.
[f]This work using all models from servers.
[g]Zhang-Server is not a quality-assessment (QA) prediction method. It is the best server whose models were used by QA methods and used here as a baseline for comparison.
[h]Ranked by GDT-score as another baseline.

RESULTS

We examined the performance of our approach (we participated in CASP7 as the TASSER-QA, group ID 125) as well as several other methods that participated in CASP7 by using as structural similarity measure GDT-score[36] on the 98 targets. Because we submitted predictions in CASP7 with only ranking and no predicted quality scores, the official assessors excluded our method as well as several other methods from their assessment, which mostly focused on correlation analysis rather than the quality of the selected top models. Here, we are able to use the predicted quality scores given by Eq. (4) for correlation analysis and compare TASSER-QA with other methods.

In Table I, we compile the results of the first prediction (named with "Target ID" QA "group ID"_**1**) using the GDT-score[36] as the structural similarity measure. The methods compared in Table I (also the following tables) include only those that have predicted quality scores and have average correlation coefficients of both Pearson and Spearman rank greater than 0.5 in our analysis. In Table I, TASSER-QA has the highest average (per target) Pearson linear correlation coefficients of 0.834; nevertheless, it is insignificantly different from Pcons' results. TASSER-QA also has the highest total GDT-score (58.06) of the first ranked models, which are about 3% better than the next best method ABIpro-h, which has total GDT-score of 56.49. The Pcons and LEE methods have the best average Spearman rank correlation coefficients[37] among all methods, with TASSER-QA ranking third. However, their differences are not statistically significant. TASSER-QA is only marginally (with an insignificant *P*-value of 0.07) better in total GDT-score than the best Zhang-Server whose models are used by TASSER-QA and other methods, and it is about 6% less than the best possible models as ranked by GDT-score (see Table I).

We further examine the performance of the compared methods on easy, medium, and hard targets as classified by our in-house 3D-jury approach[28] (see http://cssb.biology.gatech.edu/skolnick/files/tasser-qa/ for classification list). Table II shows the results for the 67 easy targets, and Table III shows the results for the 31 medium/hard targets. In both Tables II and III, the GDT-score is used as the structural similarity measure. For the easy targets, the LEE method has the best Pearson correlation coefficient and the best Spearman correlation coefficient, whereas TASSER-QA has the second and third best results, respectively, on the basis of the Pearson and Spearman correlation coefficients. TASSER-QA has a significantly higher total GDT-score than those of all other

Table II
Average Correlation Coefficients and Total GDT-Scores of the Top Quality Assessment Predictors in CASP7 for the 67 Easy Targets

Method	Number of targets	Pearson		Spearman		GDT-score	
		Average	*P*-value	Average	*P*-value	Total (average)	*P*-value
125 (TASSER-QA)	67	0.926		0.772		**47.32 (0.706)**	
634 (Pcons)	67	0.895	0.04	0.796	0.18	45.12 (0.673)	9.5×10^{-6}
556 (LEE)	65	**0.954**	0.01	**0.842**	2.9×10^{-4}	44.50 (0.685)	4.0×10^{-4}
713 (Circle-QA)	67	0.817	9.0×10^{-24}	0.701	5.6×10^{-4}	46.17 (0.689)	1.2×10^{-3}
633 (ProQ)	67	0.802	1.2×10^{-13}	0.618	2.2×10^{-9}	44.33 (0.662)	1.7×10^{-5}
038 (GeneSilico)	60	0.799	3.4×10^{-13}	0.651	8.2×10^{-7}	39.94 (0.666)	1.4×10^{-4}
692 (ProQlocal)	67	0.801	9.0×10^{-14}	0.618	2.1×10^{-9}	44.35 (0.662)	2.0×10^{-5}
178 (Bilab)	67	0.784	8.5×10^{-11}	0.614	1.4×10^{-7}	44.43 (0.663)	3.3×10^{-4}
704 (QA-ModFOLD)	67	0.773	5.4×10^{-14}	0.659	6.3×10^{-7}	44.20 (0.660)	3.3×10^{-5}
699 (ABIpro-h)	67	0.771	1.4×10^{-9}	0.685	4.3×10^{-4}	46.02 (0.687)	2.0×10^{-4}
091 (Ma-OPUS)	66	0.747	3.2×10^{-17}	0.654	4.7×10^{-6}	44.13 (0.669)	1.7×10^{-3}
013 (Jones-UCL)	56	0.734	1.9×10^{-16}	0.586	7.3×10^{-12}	37.72 (0.674)	1.8×10^{-6}
703 (QA-ModCHECK)	47	0.673	3.8×10^{-15}	0.587	5.4×10^{-7}	24.09 (0.512)	3.1×10^{-7}
717 (CaspIta-FRST)	61	0.662	1.8×10^{-22}	0.550	2.2×10^{-15}	39.55 (0.648)	2.7×10^{-8}
025 (Zhang-Server)	*67*		—		—	*47.14 (0.704)*	*0.34*

quality assessment methods. For the medium/hard targets, TASSER-QA surpasses the LEE method and has the highest Pearson correlation coefficient and the second highest Spearman correlation coefficient next only to that of Pcons (see Table III). The total GDT-score (10.74) of TASSER-QA is highest, but is almost the same as that of method ABIpro-h (10.47). The difference in total GDT-score between TASSER-QA and many other methods is statistically indistinguishable probably due to small data size of 31 targets.

In Table IV, we list the numbers of targets for which server models are ranked the first by TASSER-QA and by actual quality GDT-score. The top three servers by TASSER-QA are the same as those by GDT-score. However, Zhang-Server and MetaTasser are over-selected by TASSER-QA: 33 by TASSER-QA versus 24 by GDT-score for the Zhang-Server and 17 by TASSER-QA versus 12 by the GDT-score for the MetaTasser. Although Zhang-Server contributes one-third of the top ranked models by TASSER-QA and it is significantly better than other servers in CASP7, the total GDT-score (57.64) by TASSER-QA changes only very little when models from Zhang-Server are eliminated from the selection process.

Another analysis we carried out is the effects of the two terms in Eq. (3). When the contact potential term E_{contact} is set to zero, we get the result (Pearson, Spearman, total GDT-score) = (0.763, 0.648, 56.80). They are

Table III
Average Correlation Coefficients and Total GDT-Scores of the Top Quality Assessment Predictors in CASP7 for the 31 Medium/Hard Targets

Method	Number of targets	Pearson		Spearman		GDT-score	
		Average	*P*-value	Average	*P*-value	Total (average)	*P*-value
125 (TASSER-QA)	31	**0.636**		0.651		**10.74 (0.346)**	
634 (Pcons)	31	0.628	0.86	**0.674**	0.58	9.93 (0.320)	0.01
556 (LEE)	31	0.470	1.2×10^{-3}	0.546	5.6×10^{-3}	9.36 (0.302)	1.7×10^{-3}
713 (Circle-QA)	31	0.543	9.8×10^{-3}	0.579	0.012	9.89 (0.319)	0.04
633 (ProQ)	31	0.541	8.8×10^{-3}	0.551	7.5×10^{-4}	9.82 (0.317)	0.07
038 (GeneSilico)	28	0.527	0.01	0.558	6.1×10^{-3}	9.15 (0.327)	0.06
692 (ProQlocal)	31	0.514	9.1×10^{-3}	0.533	1.1×10^{-3}	9.68 (0.312)	0.04
178 (Bilab)	31	0.515	1.7×10^{-3}	0.522	6.6×10^{-4}	10.09 (0.326)	0.15
704 (QA-ModFOLD)	31	0.464	3.2×10^{-4}	0.473	1.4×10^{-4}	9.72 (0.314)	0.02
699 (ABIpro-h)	31	0.463	3.0×10^{-4}	0.508	3.7×10^{-4}	10.47 (0.338)	0.58
091 (Ma-OPUS)	31	0.482	4.5×10^{-4}	0.508	3.9×10^{-4}	9.11 (0.294)	0.01
013 (Jones-UCL)	25	0.453	1.1×10^{-3}	0.451	3.3×10^{-5}	7.55 (0.302)	0.02
703 (QA-ModCHECK)	24	0.527	0.03	0.552	0.01	6.52 (0.272)	3.8×10^{-3}
717 (CaspIta-FRST)	29	0.426	4.3×10^{-6}	0.451	5.7×10^{-7}	8.94 (0.308)	5.5×10^{-3}
025 (Zhang-Server)	*31*		—		—	*10.21 (0.329)*	*0.13*

Table IV
Numbers of Targets for Which the Server Models Are Ranked the First by TASSER-QA and by GDT-Score

Server[a]	TASSER-QA	GDT-score
Zhang-Server	33	24
MetaTasser	17	12
Pmodeller6	6	6
ROBETTA	5	3
HHpred3	5	5
Bilab-ENABLE	4	1
SAM_T06_server	3	3
RAPTOR	3	2
BayesHH	3	4
SPARKS2	2	0

[a]We refer to the CASP7 website http://predictioncenter.org/casp7/ for abstracts of individual server methods. Only top ten servers by TASSER-QA are presented.

all significantly worse than those given by the full terms in Eq. (3) that are (0.834, 0.734, 58.06). If the fragment comparison term E_{frg} is set to zero, we get (0.831, 0.698, 56.60). The correlation coefficients change slightly, but the total GDT-score changes significantly. These results show that both E_{frg} and $E_{contact}$ are important to the better performance of TASSER-QA method.

In CASP7, TASSER-QA only considered the first model given by each individual server (i.e., model names ending with "**_TS1**"). This is similar to the best-model-mode of the 3D-jury method.[6] It is of interest to know how much worse TASSER-QA will be if this strategy is not implemented or cannot be used because of the absence of ranking information from the individual servers. By including all models (model names ending with "**_TS1**" to "**_TS5**") in the assessment prediction, TASSER-QA will have (Pearson, Spearman, total GDT-score) = (0.828, 0.791, 57.22) compared to the original (0.834, 0.734, 58.06) (see also Table I). The Pearson correlation coefficient and total GDT-score are still the best of the compared methods. The Spearman rank correlation coefficient of TASSER-QA moves from the third to the top position. However, the strategy of prefiltering possibly worse models other than the first models (ending with "**_TS1**") by individual servers helps in the selection of good models by TASSER-QA as assessed by the total GDT-score. A similar result was also observed in the 3D-jury method.[6] That is because both TASSER-QA and 3D-jury use the similarity information between models, which depends on the ensemble of models. This may also be true for other methods that depend only on the properties of individual models, for example, Pcons will have (Pearson, Spearman, total GDT-score) = (0.820, 0.696, 55.62) if only first models by individual servers are used compared to (0.810, 0.757, 55.05) when all top five server models are assessed. The fact that TASSER-QA using all models is indistinguishable from using first models in Pearson's correlation coefficient and total GDT-score (see Table I) indicates that it can be reliably applied to models with no ranking information like those from CASP set-up.

From these results, a general trend for the correlation coefficients is also observed: the more models that are assessed, the better is the Spearman rank correlation coefficient and the worse is the Pearson correlation coefficient. This may be attributable to the intrinsic mathematical properties of the two kinds of coefficients and may have nothing to do with the properties of assessment prediction scores.

In Figure 1, we show some examples of the linear correlation between the TASSER-QA predicted scores and the actual GDT-scores of the models. The correlation coefficient can be as high as 0.996 (0.961 after excluding the three apparently outliers of very low GDT-scores) for target T0334. As shown in Table II, the average linear correlation coefficient for easy targets is 0.926. It can also be very poor for some hard targets. For example, the correlation coefficient for target T0356 is −0.164 and for target T0361, it is 0.083 (not shown in Fig. 1), even though TASSER-QA has the best average linear correlation coefficient of 0.636 for the 31 medium/hard targets.

We next consider the application of TASSER-QA to situations not like a CASP set-up in which there are many servers, and almost all of them provide internally ranked models for the same protein target. The first such application is one when there are only a few servers, with no internal ranking information. In Table V, we show the results for models from only two moderately performing servers ROBETTA and SP[3] in CASP7, and we assume no rank information among the ten (five from each server) models for each target. We have also assumed that the predictions of all other compared methods do not depend on the ensemble of models so that we can directly use their CASP7 predictions on the subset of models without having to re-do the predictions (although TASSER-QA's dependence on the ensemble has been taken into account through recalculating the consensus potential). TASSER-QA is applied using all ten models from the two servers. We see in Table V that TASSER-QA is still the best for both correlation methods and total GDT-score. The selected top model quality by many methods is better than that of ROBETTA or SP[3], but is more than 5% worse than that of the best models.

The second application is on models generated by ab initio or refinement modeling methods such as ROSETTA and TASSER. In this situation, there are a large number models with no a priori ranking information and the usual way of selecting models is to use clustering methods such as SPICKER.[4] To show TASSER-QA also works well in this situation, we apply TASSER-QA on TASSER-generated models for CASP7 targets with 16,000 models for each target and compare it with SPICKER on the same set of models. The results are compiled in Table VI. The *P*-values that characterize the statistical significances of the difference between the total

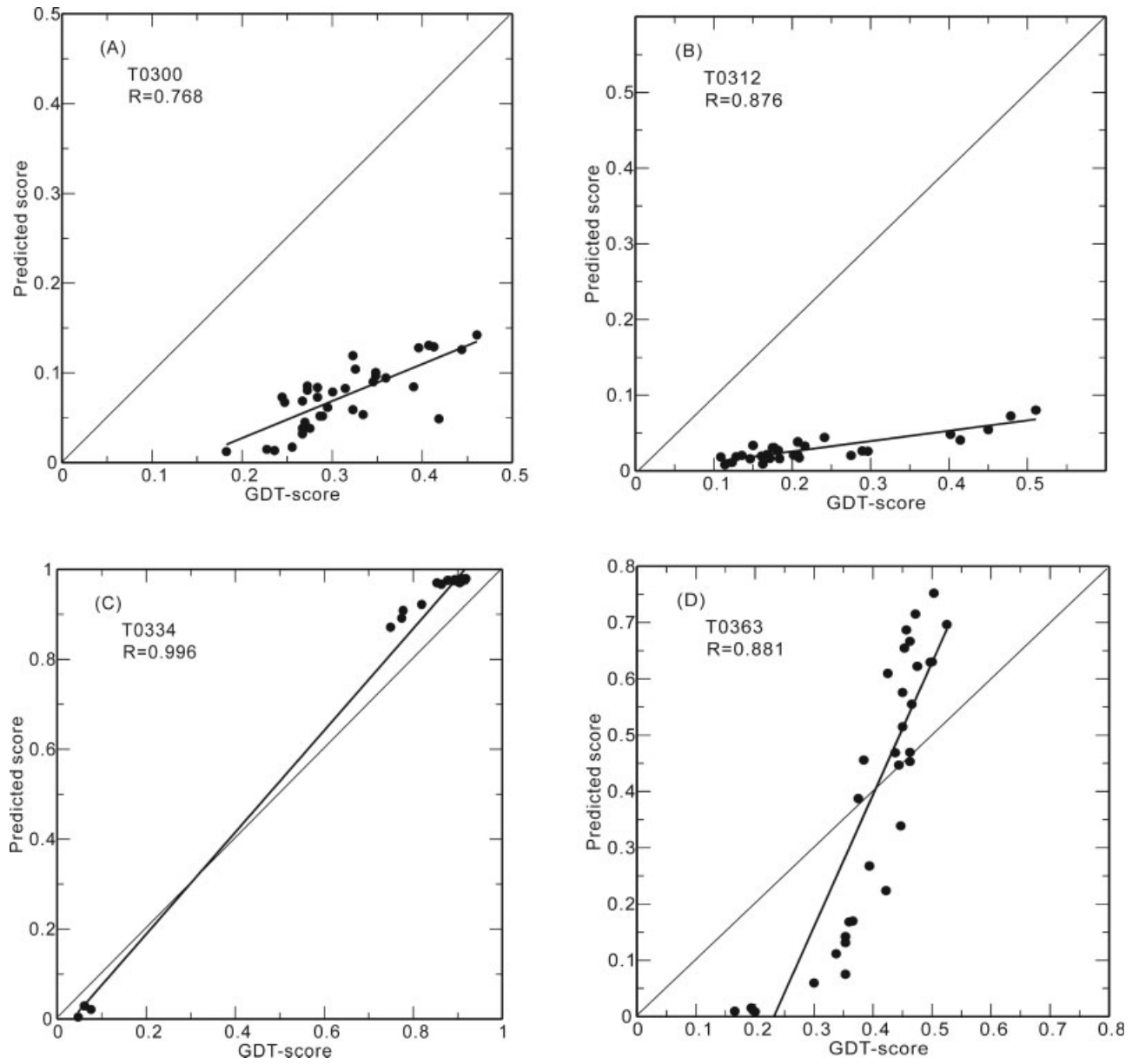

Figure 1

*Examples of correlation between the predicted model quality scores and model GDT-scores. (**A**) T0300 and (**B**) T0312 are hard targets. (**C**) T0334 is an easy target with a correlation coefficient of 0.961 after excluding the three low GDT-score outliers; (**D**) T0363 is medium target.*

GDT-scores of TASSER-QA and SPICKER selected models are all >0.05 except for one case when the fragment comparison term is set to zero and best of top five selected models are used for GDT-score computation. That means TASSER-QA has comparable performance with SPICKER on TASSER-generated models. Table VI also shows that the fragment comparison term is slightly more transferable to this kind of models, and the combination of fragment comparison and consensus potential does not give better results.

CONCLUSIONS

In this work, we presented a simple but accurate model quality assessment prediction method that is comparable to or even better than the state-of-the-art methods available to date. The method combines fragment comparisons and a consensus C_α contact potential. The fragment library is obtained by an extension of the SP^3 threading method. The consensus C_α contact potential is derived from the models to be assessed using the same

Table V
Average Correlation Coefficients and Total GDT-Scores of the Top Quality Assessment Predictors in CASP7 for Models from ROBETTA and SP³ Servers Only

Method	Number of targets	Pearson		Spearman		GDT-score	
		Average	*P*-value	Average	*P*-value	Total (average)	*P*-value
125 (TASSER-QA-all)	98	**0.677**		**0.609**		**55.11 (0.562)**	
634 (Pcons)	98	0.577	0.03	0.577	0.40	54.04 (0.551)	0.13
556 (LEE)	96	0.526	1.3×10^{-3}	0.476	4.2×10^{-3}	52.71 (0.549)	0.09
713 (Circle-QA)	98	0.569	4.4×10^{-3}	0.527	0.02	54.42 (0.555)	0.22
633 (ProQ)	98	0.430	5.4×10^{-7}	0.391	5.4×10^{-6}	53.98 (0.551)	0.10
038 (GeneSilico)	80	0.491	1.4×10^{-3}	0.454	3.3×10^{-3}	44.94 (0.562)	0.16
692 (ProQlocal)	97	0.418	3.6×10^{-7}	0.383	5.1×10^{-6}	53.88 (0.555)	0.15
178 (Bilab)	98	0.537	5.7×10^{-7}	0.489	9.2×10^{-4}	54.38 (0.555)	0.15
704 (QA-ModFOLD)	98	0.555	1.4×10^{-3}	0.513	5.4×10^{-3}	54.20 (0.553)	0.19
699 (ABIpro-h)	98	0.568	1.9×10^{-3}	0.526	6.8×10^{-3}	54.14 (0.552)	0.07
091 (Ma-OPUS)	97	0.549	7.2×10^{-4}	0.505	5.7×10^{-3}	52.18 (0.538)	6.8×10^{-3}
013 (Jones-UCL)	81	0.439	9.6×10^{-7}	0.408	3.5×10^{-5}	44.76 (0.553)	0.06
703 (QA-ModCHECK)	71	0.546	0.11	0.502	0.16	38.37 (0.540)	0.48
717 (CaspIta-FRST)	90	0.473	4.2×10^{-5}	0.485	4.9×10^{-3}	49.20 (0.547)	0.02
ROBETTA	*98*		—		—	*52.87 (0.539)*	
SP³	*98*		—		—	*51.58 (0.526)*	
Best	*98*		—		—	*58.01 (0.592)*	

approach as in TASSER.[33,34] This consensus C_α contact potential takes into account the effect of similarities among the models, which behaves in some sense like 3D-jury.[6] Both terms in Eq. (3) are important for the current method to be successful. This approach is fully automated and is useful for selecting the best possible models from a set of structures provided by other methods. The resulting selected models can also be used as a starting point for further refinement. The current methodology was used by TASSER group in CASP7[28] in selecting the final models for submission. In practice, it can also be used by TASSER[34] or other refinement methods in selecting initial models from other servers (e.g., all server models in CASP7) for refinement. The fact that this method has a very good Pearson correlation coefficient for easy targets makes it a suitable approach for near native structure selection. For medium/hard targets, although the Pearson correlation coefficient is worse, it is still better than other existing approaches.

Table VI
Comparison of TASSER-QA with SPICKER on TASSER-Generated Models for the 98 CASP7 Targets

	GDT-score of first model[e]	GDT-score of best of top five
	98 Targets	
SPICKER[a]	53.93	55.31
TASSER-QA-all[b]	54.29 (0.17)	55.23 (0.75)
TASSER-QA-all-frg[c]	54.35 (0.18)	55.74 (0.16)
TASSER-QA-all-ca[d]	53.81 (0.64)	54.79 (0.04)
	67 Easy targets	
SPICKER	45.31	45.96
TASSER-QA-all	45.31 (1.0)	45.90 (0.73)
TASSER-QA-all-frg	45.57 (0.21)	46.31 (0.09)
TASSER-QA-all-ca	45.19 (0.42)	45.66 (0.08)
	31 Medium/hard targets	
SPICKER	8.62	9.35
TASSER-QA-all	8.98 (0.09)	9.32 (0.45)
TASSER-QA-all-frg	8.78 (0.49)	9.43 (0.74)
TASSER-QA-all-ca	8.63 (0.97)	9.13 (0.26)

[a]Clustering method in Ref. 4.
[b]This work using all models.
[c]This work using all models with the consensus term in Eq. (3) set to zero.
[d]This work using all models with the fragment comparison term in Eq. (3) set to zero.
[e]Numbers in parentheses are *P*-values for the differences between TASSER-QA and SPICKER.

We note that for most of the easy targets, the correlations between predicted quality score and actual GDT-score are very good, but the slopes of the linear correlations are not close to one. That means the prediction is good for the relative quality of the models within a given target, but not good enough for absolute quality of the models that can be compared between different targets. This is true for TASSER-QA as well as for other top performing methods.

The current method can also be extended to assign a local quality measure of individual residues by simply considering the contribution of individual residues to Eq. (4). How well the extended prediction works on identifying high quality regions of the predicted structure needs further investigation.

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. Moult J, Fidelis K, Hubbard T, Kryshtafovych A, Rost B, Tramontano A. 7th Critical Assessment of Techniques for Protein Structure Prediction; http://predictioncenter.org/casp7/.

4. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein fold. J Comput Chem 2004;25:865–871.
5. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306:1191–1199.
6. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
7. Sasson I, Fischer D. Modeling three-dimensional protein structures for CASP5 using 3D-SHOTGUN meta-predictors. Proteins Struct Funct Genet 2003;53:389–394.
8. Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci 2002;11:430–448.
9. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
10. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1998;288:477–487.
11. Melo F, Feytmans E. Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 1998;277:1141–1152.
12. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001; 44:223–232.
13. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.
14. Petrey D, Honig B. Free energy determinants of tertiary structure and the evaluation of protein models. Protein Sci 2000;9:2181–2191.
15. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300:171–185.
16. Dominy BN, Books CL. Identifying native-like protein structures using physics-based potentials. J Comput Chem 2002;23:147–160.
17. Tobi D, Elber R. Distance-dependent, pair potential for protein folding: results from linear optimization. Proteins 2000;41:40–46.
18. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:83–85.
19. Eisenberg D, Luthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzym 1997;277: 396–404.
20. Zhang C, Liu S, Zhou H, Zhou Y. An accurate residue-level pair potential of mean force for folding and binding based on the distance-scaled ideal-gas reference state. Protein Sci 2004;13:400–411.
21. Zhang C, Liu S, Zhou Y. Accurate and efficient loop selections using DFIRE-based all-atom statistical potential. Protein Sci 2004;13:391–399.
22. 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.
23. Wallner B, Fang H, Elosson A. Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. Proteins Struct Funct Genet Suppl 2003;6:534–541.
24. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003;12:1073–1086.
25. 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.
26. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 2005;15:900–913.
27. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441.
28. 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(Suppl 8):90–97.
29. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005;58321–328.
30. Zhou H, Zhou Y. SPARKS 2 and SP3 servers in CASP 6. Proteins (Suppl CASP issue) 2005;7 (Suppl):152–156.
31. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. J Mol Biol 2000;297:1003–1013.
32. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman D-J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res 1997;25:3389–3402.
33. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003; 85:1145–1164.
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. Zhang Y, Skolnick J. A scoring function for the automated assessment of protein structure template quality. Proteins 2004;57:702–710.
36. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;3:22–29.
37. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes: the art of scientific computing. Cambridge: Cambridge University Press; 1989.