Scoring Function for Automated Assessment of Protein Structure Template Quality

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

ABSTRACT **We have developed a new scoring function, the template modeling score (TM-score), to assess the quality of protein structure templates and predicted full-length models by extending the approaches used in Global Distance Test (GDT)[1] and MaxSub.[2] First, a protein size-dependent scale is exploited to eliminate the inherent protein size dependence of the previous scores and appropriately account for random protein structure pairs. Second, rather than setting specific distance cutoffs and calculating only the fractions with errors below the cutoff, all residue pairs in alignment/modeling are evaluated in the proposed score. For comparison of various scoring functions, we have constructed a large-scale benchmark set of structure templates for 1489 small to medium size proteins using the threading program PROSPECTOR_3 and built the full-length models using MODELLER and TASSER. The TM-score of the initial threading alignments, compared to the GDT and MaxSub scoring functions, shows a much stronger correlation to the quality of the final full-length models. The TM-score is further exploited as an assessment of all 'new fold' targets in the recent CASP5 experiment and shows a close coincidence with the results of human-expert visual assessment. These data suggest that the TM-score is a useful complement to the fully automated assessment of protein structure predictions. The executable program of TM-score is freely downloadable at http://bioinformatics.buffalo.edu/TM-score. Proteins 2004;57:702–710.** © 2004 Wiley-Liss, Inc.

INTRODUCTION

The canonical comparative modeling/threading-based protein structure prediction procedure consists of two steps: (i) finding a solved structure related to the target sequence (i.e. template)[3–6] and (ii) building a full-length model based on the template.[7–10] The quality of the resulting full-length model is usually assessed by the root mean square deviation (RMSD)[11,12] between equivalent atoms in the model and native structures. However, RMSD alone is not sufficient to estimate the quality of the initial templates because the alignment coverage can be very different in different approaches.[3–6] Obviously, a template with a 2 Å RMSD to native having 50% alignment coverage is not necessarily better for structure modeling than one with a RMSD of 3 Å but having 80% alignment coverage. While the template aligned regions are better in the former because fewer residues are aligned, the resulting full-length model might be of poorer quality. The template assessment problem becomes particularly relevant during the development of efficient fold recognition algorithms, since different sequence–structure alignment schemes or parameters can result in various levels of alignment confidence with an associated loss or gain of alignment coverage.[3–6] Therefore, a single assessment score that has an appropriate balance of alignment accuracy and coverage and that is strongly related to the quality of the final full-length model is essential. Equally important, it must differentiate between a random and a statistically significant prediction.

Highly related to the above problems, several interesting scoring functions have been developed for the purpose of sequence-dependent comparison of two structures of different lengths (in contrast to sequence-independent structure alignment algorithms).[1,2,13–15] For example, with MaxSub,[2] Siew and coworkers tried to identify the maximum substructure in which the distances between equivalent residues of two structures after superposition are below some threshold value, such as 3.5 Å. Since the MaxSub scoring function only counts those residues included in the substructure, the spatial information of the templates outside the substructure is omitted. For example, Figure 1(a) shows the MaxSub superposition of the native structure of 2sas_ and the template alignment (94% coverage, which is the ratio of the number of aligned residues to the number of target residues) obtained from the threading program PROSPECTOR_3,[6] where residue pairs of distance <3.5 Å are highlighted in red (50% coverage), with the remainder of the aligned residues in yellow. The templates in Figure 1(b) (original alignment having 94% coverage) and Figure 1(c) (the 'well-aligned' part with 50% coverage) therefore have the same MaxSub-score, which is only associated with the set of red residues. However, the facility of the templates for the final full-length structure modeling can be significantly different. Using the structure building program MODELLER,[7,10] for example, the template in Figure 1(b) results in a full-length model with a RMSD from native of 4.4 Å, while the template in Figure 1(c) results in a full-length model

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

Received 30 December 2003; Revised 16 March 2004, 25 May 2004; Accepted 9 June 2004

Published online 8 October 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20264

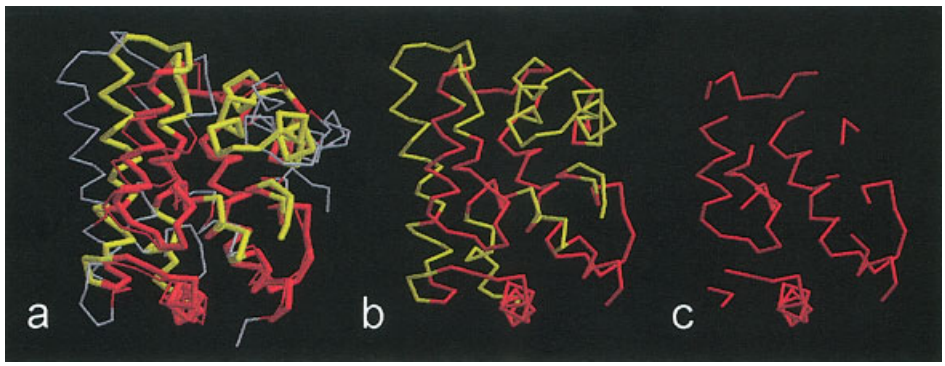

Fig. 1. (a) Superposition between the template and the native structure of 2sas from MaxSub. The native structure is shown as thin backbone in white; the template is shown as thick backbone in yellow, which is from Chain-A of 2scp in the PDB library[17] hit by PROSPECTOR_3.[6] The residues with pairwise distances lower than 3.5 Å are highlighted in red for both template and native. (b) The whole template alignment has 94% coverage. (c) The substructure of the template alignment with a distance to native of less than 3.5 Å ('well-aligned' part) has 50% coverage. The templates in (b) and (c) have the same MaxSub score of 0.434. However, using MODELLER,[7,10] the template in (b) results in a full-length model of RMSD 4.4 Å, while the template in (c) results in a full-length model of RMSD 12.5 Å.

having a RMSD from native of 12.5 Å. By way of further illustration, in a large-scale benchmark set of PROSPECTOR_3 alignments (see below), there are 81 cases having MaxSub-scores between 0.4 and 0.45. The RMSD values of the final full-length models built from MODELLER vary from 3.5 to 35.7 Å with a standard deviation of 4.8 Å. Thus, there is no apparent correlation between the Maxsub-score and the quality of the resulting full-length model.

In their GDT_TS scoring function, Zemla and coworkers[1,13] further identify multiple maximum substructures associated with several different threshold cutoffs (e.g. 1, 2, 4, and 8 Å as used in the recent CASP5 experiment[16]). The GDT_TS-score is defined as the average coverage of the target sequence of the substructures with the four different distance thresholds. Since the GDT_TS-score focuses only on the size of the substructures, the detailed match information of templates/models and native structures is partially missed (e.g. residues with deviations ranging from 4.1–8 Å from native have identical contributions to the scoring function). Zemla[13] further addressed this problem by introducing more distance thresholds.

Another problem associated with these score functions is the dependence of the score magnitudes on the evaluated proteins' size. In other words, one must address the issue of what the corresponding score value of a pair of randomly related structures would be. In Figure 2, we plot the average MaxSub- and GDT-scores as a function of the length of protein for random structure pairs in the Protein Data Bank (PDB)[17] that have pairwise sequence identity of less than 30%. These scores show a power-law dependence on the protein size. Obviously, a given absolute score, such as GDT = 0.4 or MaxSub = 0.3, can reflect a significant alignment for a target of 400 residues, but it is close to a random selection in the PDB for a target of 40 residues. This significant size dependence renders the absolute magnitude of these scoring functions meaningless.

The significant protein size dependence of structure similarity for randomly related structure pairs has also been observed by many authors when structure similarity is measured by structure alignment[18,19] or RMSD calculation.[20,21] To eliminate the dependence on protein size, Levitt and Gerstein[18] and Ortiz and coworkers[19] converted their structure alignment score into a statistical significance score, called the *P*-value, on the basis of the statistics of their random structure database. For their relative RMSD, Betancourt and Skolnick[20] normalized the RMSD by the average RMSD from random structure pairs with similar size and radii of gyration. In the RMSD-100 score,[21] Carugo and Pongor divided the RMSD by a factor of $1 + \sqrt{N}/100$, with N representing the protein length.

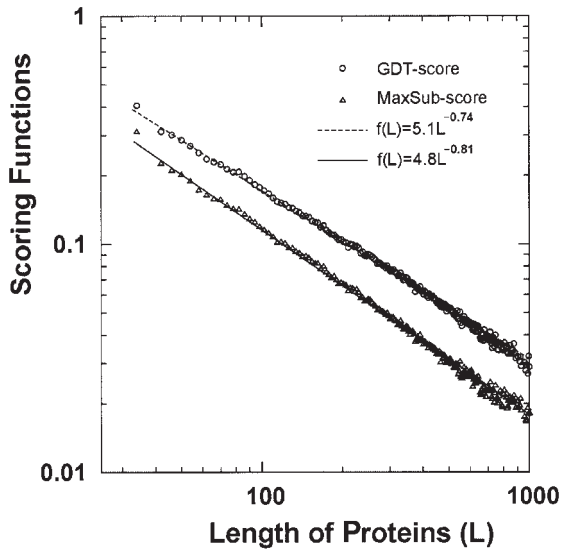

Fig. 2. The average MaxSub-scores (triangles) and GDT-scores (circles) as a function of protein size. The data are calculated from all pairs of 3656 proteins taken from PDB library[17] that have <30% pairwise sequence identity. The statistical error bars are smaller than the size of the points. The solid and dashed lines are the nonlinear least square Marquardt–Levenberg fit of the MaxSub and GDT data to a power-law equation of $f(L) = aL^b$, where L is the length of the smaller protein of the corresponding structure pair. The fit parameters a and b are as indicated.

In this article, we extend the above approaches and develop a new scoring function for the assessment of threading templates, which we call the template modeling (TM) score. One of our purposes is to re-scale the structure modeling errors so that the score value is independent of protein size for randomly related structure pairs. Since one of the most important uses of threading templates is to facilitate the final structure modeling, our second goal here is to have the scores of initial templates strongly correlated with the quality of final full-length models. Certainly, as has been noted by many authors,[2,19,20] RMSD is not a perfect indicator of full-length model quality. Besides the significant size dependence of random structure pairs, when others parts of a model have large prediction errors, the RMSD cannot identify well-predicted substructures. There are numerous other measurements for protein modeling quality in the literature.[1,2,15,19,20,22,23] With MAMMOTH[19] for example, Ortiz and coworkers assess structures by comparing both local and global similarities. Here, as one of many possible choices, we use a Z-score like expression of the relative RMSD (rRMSD) to score the quality of the final full-length models. We consider a large-score benchmark protein set that covers the current PDB[17] at 35% sequence identity for all proteins of less than 200 residues. The TM-score, as well as the Maxsub and GDT_TS scores, of the initial templates are evaluated on the basis of their correlations to the quality of the final full-length models in the benchmark targets built by the widely-used protein modeling software MODELLER.[7,10]

MATERIALS AND METHODS

Scoring Function

Our scoring function is a variation of the Levitt–Gerstein (LG) score,[18] which was first used for sequence-independent structure alignments:[24]

$$\text{TM-score} = \text{Max}\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \qquad (1)$$

where L_N is the length of the native structure, L_T is the length of the aligned residues to the template structure, d_i is the distance between the ith pair of aligned residues and d_0 is a scale to normalize the match difference. 'Max' denotes the maximum value after optimal spatial superposition. The value of the TM-score always lies between (0, 1], with better templates having higher TM-scores. A similar formula is also used in MaxSub,[2] but the summation is limited only to those residues with $d_i < d_0$. Here, the summation is over all of the template-aligned residues. In LiveBench,[15] Rychlewski and coworkers define a three-dimensional-score that has a similar function but with a different format from the LG-score. In their S-score,[14] Cristobal and coworkers use the non-normalized LG-score, including the gap penalty. As shown below, the number of gaps in the template alignments has no correlation with the quality of the final models.

The value of d_0 has been taken to be constant in all of the above approaches. For example, d_0 = 3.5 Å in MaxSub,[2] d_0 = 5 Å in the S-score[14] and the original LG-score.[18,24] As shown in Figure 2, these treatments result in a power-law dependence of the score on the size of the proteins in random protein pairs. In Figure 3, we calculate the average TM-score for 3656 protein structures from the PDB that have pairwise sequence identity <30%. If we take d_0 = 5 Å, which we call the 'raw TM-score' (rTM-score), we find that it has a similar power-law dependence on protein size as the MaxSub- and GDT-scores.

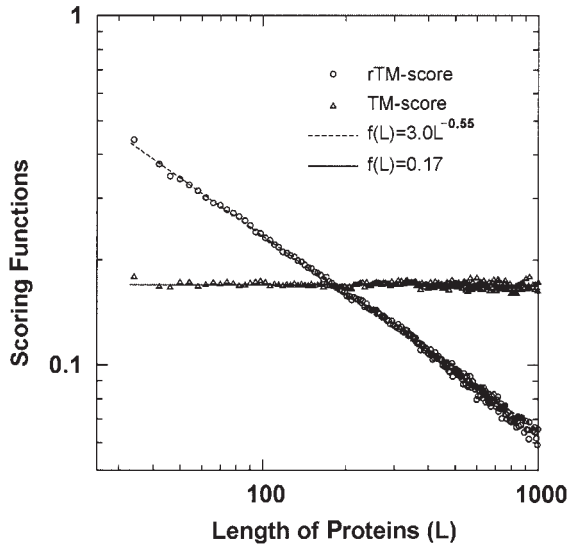

Fig. 3. The average 'raw TM-score' (rTM-score) and TM-score of random protein pairs as a function of protein size. For the rTM-score, d_0 = 5 Å; for the TM-score, d_0 is defined as in eq. (5). The data are calculated from all pairs of 3656 PDB structures of <30% sequence identity. The statistical error bars are smaller than the size of the points. The dashed line is a nonlinear least square Marquardt–Levenberg fit of the rTM-score data to a power-law equation $f(L)$, where L is the length of the smaller protein of the corresponding structure pairs. The solid line denotes the horizontal line of TM-score = 0.17.

To rule out protein size dependence in the rTM-score, let's first make an approximate estimation of the average structure match difference of random related structures. In general, the RMSD of two structures (A and B) of identical length of L_N, can be written as

$$\text{RMSD} = \sqrt{R_A^2 + R_B^2 - 2\frac{\sum_i \mathbf{r}_{Ai}\cdot\mathbf{r}_{Bi}}{\sqrt{\sum_i r_{Ai}^2 \sum_i r_{Bi}^2}} R_A R_B} \qquad (2)$$

where R_A (R_B) is the radius of gyration for structure A (B), $\mathbf{r}_{Ai}$ ($\mathbf{r}_{Bi}$) is the coordinate vector after global superposition. From the calculation of around 1300 non-homologous PDB structures, Betancourt and Skolnick[20] observed that the average correlation coefficient for randomly related structure pairs follows as

$$c = \left\langle \frac{\sum_i \mathbf{r}_{Ai}\cdot\mathbf{r}_{Bi}}{\sqrt{\sum_i r_{Ai}^2 \sum_i r_{Bi}^2}} \right\rangle = 0.42 - 0.05 L_N e^{-L_N/4.7} + 0.63 e^{-L_N/37}. \qquad (3)$$

Keeping in mind that the average radius of gyration of globular protein structures has a power-law dependence

on the length[8] (i.e. $\langle R \rangle \approx {L_N}^{0.39}$), we can estimate the average distance of corresponding residue pairs of random related proteins in the TM-score superposition:

$$\langle d_i \rangle \sim L_N^{0.39} \sqrt{1 - 0.42 + 0.05 L_N e^{-L_N/4.7} - 0.63 e^{-L_N/37}} - h \,. \quad (4)$$

Here the constant h is introduced because the average distance of the optimal local structure matches in TM-score calculations is always smaller than that of the global matches in RMSD calculations (see below). When $h = 0.75$, eq. (4) can be well approximated by a simpler formula

$$d_0 = 1.24 \sqrt[3]{L_N - 15} - 1.8 \,, \quad (5)$$

which drops, for example, from 6.4 to 2.3 Å when L_N changes from 300 to 50 residues. As shown in Figure 3, the TM-score of d_0 defined as in eq. (5) has an approximately constant value of ≈0.17, independent of protein size for the random structure pairs.

Search Engine

To find the spatially optimal superposition of the template and the native structure that has the maximum (or close to the maximum) TM-score according to eq. (1), we use an iterative search algorithm, similar to that used by Zemla and coworkers,[1] Siew and coworkers,[2] Ortiz and coworkers[19] and Kihara and Skolnick.[25]

Starting with an initial fragment of the template that consists of L_{int} neighboring aligned residues, we superposed the fragment to the corresponding residues of the native structure according to Kabsch's rotation matrix.[11,12] Then, we collected all of the residues of the template with distance to native of less than d_0 and superposed this set of residues onto the native structure again. The process was repeated till the rotation matrix converged.

Since the converged superposition is usually sensitive to the initial selection of the fragment L_{int}, we ran an iterative process with $L_{\text{int}} = L_T$, $L_T/2$, $L_T/4$, L, 4, respectively. When $L_{\text{int}} < L_T$, we ran all the iterations with the location of initial fragments shifting continuously from the N- to the C-terminus. The rotation matrix with the highest TM-score was selected.

As confirmation of the optimization procedure, we ran the search engine about three times longer with additional randomly selected initial fragments for the abovementioned 1489 targets. There were only 92 (≈6%) cases with different TM-scores, all with a difference of less than 0.002. Based on the high convergence of the rotation matrix, we feel quite safe concluding our search engine is optimal or close to optimal for score maximization.

Z-Score of Relative RMSD

Bearing in mind the size and radius of gyration dependences of the RMSD as presented in eqs. (2 and 3), Betancourt and Skolnick[20] defined a relative RMSD (rRMSD) to eliminate the dependencies:

$$\text{rRMSD} = \frac{\text{RMSD}}{\sqrt{R_A^2 + R_B^2 - 2cR_AR_B}} \,. \quad (6)$$

which is just the ratio of the RMSD values of the structures of interest to the average RMSD of a pair of randomly related structures of the same radii of gyration. The standard deviation of the rRMSD for the random structure pairs is[20]

$$\delta = 0.09 + 1.16 e^{-L_N/1.6} + 0.25 e^{-L_N/36} \quad (7)$$

In our calculations, a Z-score like deviation of rRMSD to mean (which has a value of 1) is defined as

$$Z\text{-rRMSD} = \frac{\text{rRMSD} - 1}{\delta} \,. \quad (8)$$

RESULTS AND DISCUSSION

Benchmark Set of Targets and Templates

For a reliable evaluation of the scoring functions, we constructed a comprehensive benchmark protein set, which includes 1489 test proteins and covers the PDB library with lengths from 41 to 200 residues at 35% sequence identity. A list of the 1489 proteins is available at our website: http://www.bioinformatics.buffalo.edu/abinitio/1489.

For each target, the template structures were obtained using our threading program PROSPECTOR_3,[6] which was designed to match the target sequence to a non-homologous solved structure library culled from the PDB.[17] Template proteins whose sequence identity to the targets is >30% were excluded from the library. The highest scoring template from the PROSPECTOR_3 alignment was selected.

TM-Score Distribution of Templates

In PROSPECTOR_3,[6] if a template alignment has a Z-score, an energy in standard deviation units relative-to-mean above 15, or if the alignment has a Z-score above 7 and is structurally in consensus with other template alignments of similar Z-score, the alignment usually has a high possibility of being correct. It is therefore designated as an 'Easy' target according to PROSPECTOR_3,[6] since it should be easier for the modeling refinement than those targets that have lower Z-scores (typically with shorter alignment length and poorer alignment quality), which are designated as 'Medium/Hard' targets. In Figure 4(a), we show the TM-score distribution of threading templates in two different categories. As expected, the majority of the 'Easy' targets have TM-scores above 0.4, while most 'Medium/Hard' targets have TM-scores below 0.4, with an obvious gap between the 'Easy' and the 'Medium/Hard' target distributions.

As a control, we also present in Figure 4(a) the distribution of the templates identified from the structure alignment program SAL[25] that structurally aligns the native structures to the same template library as PROSPECTOR_3 and returns the structures with the highest Z-rRMSD scores to native as defined in eq. (8). In the 'Easy' set, where the 'gold standard' structure alignment has

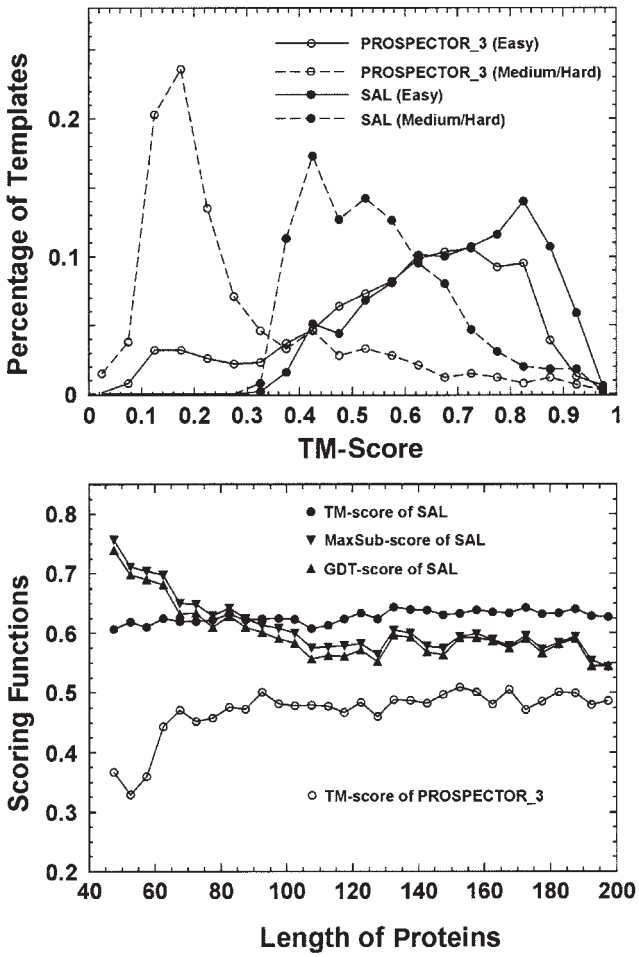

Fig. 4. (a) TM-score distributions for templates obtained from the threading program PROSPECTOR_3[6] and the structure alignment program SAL.[25] The templates are categorized into 'Easy' (877 cases) or 'Medium/Hard' (612 cases) sets according to their alignment confidence by PROSPECTOR_3. (b) Average scoring function versus protein size. The different templates evaluated by the different scoring functions are labeled by different types of points. The lines connecting the points serve to guide the eye.

sequence identity around 17%, the TM-score distribution of PROSPECTOR_3 is quite similar as that of SAL alignments, except for the tail-region where PROSPECTOR_3 has slightly more instances in which the TM-score is less than 0.4. For 'Medium/Hard' cases, where the 'gold standard' structure alignment has a sequence identity around 9%, far below the 'twilight zone' of sequence identity, the current version of PROSPECTOR_3 fails in most cases: 77% of the cases have a TM-score less than 0.4; 32% of the cases have a TM-score less than 0.17, which means the threading alignment does not provide more information than a random selection for those targets. Of course, this result is consistent with the scoring system of PROSPECTOR_3, since the category is defined according to its alignment confidence.

The alignment difference between the real threading and the 'gold standard' structure alignment also manifests itself in their sensitivities to the target protein size. As shown in Figure 4(b), the TM-score of the SAL alignment has almost no size dependence after introduction of the size-dependent scale as in eq. (5). On the other hand, the PROSPECTOR_3 alignment has obviously lower TM-scores for small targets than for larger ones, which highlights the difficulty of the current alignment method in dealing with small proteins. On the other hand, because of using the constant cutoff/scale for the match differences, the SAL alignment shows significant size-dependence when evaluated by MaxSub- or GDT-scoring functions [see Fig. 4(b)].

Correlations Between Scoring Functions and Final Full-Length Models

To evaluate how the scoring functions are related to the ability to construct full-length models from the initial templates, we built full-length models using the PROSPECTOR_3 templates as the only input. For the purpose of generality and clearness of the evaluation, we employed one of the most widely used modeling software programs, MODELLER, which is designed to build full-length models by optimally satisfying the spatial restraints extracted from the input templates.[6,10] This algorithm is representative of a general set of approaches exploited by different authors.[7–9,26–29]

Figure 5(a–c) show, respectively, the TM-score, the MaxSub-score and the GDT_TS score of the threading templates as the function of the Z-rRMSD of the final full-length models built by MODELLER. Here, we only present those targets (1048 of 1489) for which MODELLER could generate a model whose Z-rRMSD to native was below -1, since for structure pairs of too weak similarity the scoring assessment becomes less relevant. There is an obviously stronger correlation between the TM-score and the quality of final full-length models than between either of the other two scoring functions and the final models. To be more precise, we define the correlation coefficient as

$$C = \frac{\langle SZ\rangle - \langle S\rangle\langle Z\rangle}{\sqrt{(\langle S^2\rangle - \langle S\rangle^2)(\langle Z^2\rangle - \langle Z\rangle^2)}} \qquad (9)$$

where S and Z respectively represent the scoring functions of the initial templates and the Z-rRMSD to native of the final full-length models. The average of $\langle\ldots\rangle$ only includes those targets whose final model have Z-rRMSD values less than -1. Based on Figure 5(a–c), the TM-score has the highest correlation coefficient (-0.891) to the Z-rRMSD of final models among all three scoring functions, with the GDT_TS-score (-0.751) slightly better than the MaxSub-score (-0.746).

In Figure 5(d), we divide the Z-rRMSD space into 20 bins and calculate the fluctuations of the scoring functions of the initial templates for each Z-rRMSD bin. In general, the standard deviation of the template scores tends to increase with the Z-rRMSD values of the final models. Consistent with the correlation coefficients, the TM-score has the smallest dispersion for a given Z-rRMSD value of the three scores. The dispersions are similar for the GDT-score and MaxSub-score for very good Z-rRMSD

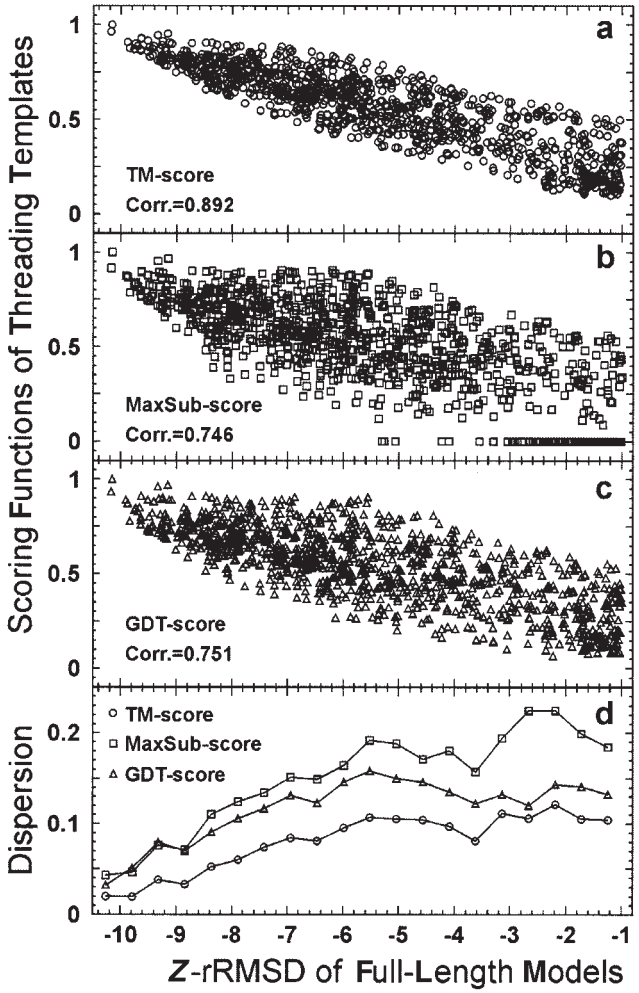

Fig. 5. (a–c) Template scoring functions versus the *Z*-rRMSD values of the final full-length models built by MODELLER.[7,10] (d) Standard deviation of the template scoring functions as a function of *Z*-rRMSD of the final full-length models.

models. When the absolute value of *Z*-rRMSD decreases (or the quality of templates becomes worse in general), in comparison with the MaxSub-score, the GDT-score, because of its multiple threshold cutoff resolutions, is more indicative of final model quality. Starting with a *Z*-rRMSD greater than −5.3, the MaxSub-score is 0 for some of the templates, showing no sensitivity to the quality of final models for those targets.

Among all 1489 targets, 50% have threading templates with a TM-score above 0.48. If we consider a successfully predicted final model as one with a *Z*-rRMSD value less than −5, then the false positive rate for templates whose TM-scores are greater than 0.48 is 16.5%, and the false negative rate for the templates whose TM-scores are less than 0.48 is 4.1%. Similarly, 50% of the targets have threading templates with MaxSub-scores above 0.42. Using the same maximum *Z*-rRMSD of −5 as the threshold for successful predictions produces a false positive rate and a false negative rate for the MaxSub-score of 23.5% and 9.2%, respectively. For the GDT-score, 50% of the targets have threading templates with a GDT-score greater than 0.475, and the false positive and false negative rates are 23.0% and 7.9%, respectively. Thus, on the basis of its ability to select models whose quality is better with the fewest false positives and false negatives, the TM-score is best, with essentially identical performance exhibited by both the Maxsub and the GDT scores.

Gap Density of Template Alignments

One of the differences between the TM-score and the original LG-[18] and S-score[14] formulas is that a gap penalty of $-10 \times N_{\mathrm{gap}}$ is included in the latter scores, where N_{gap} is the number of alignment gaps. This inclusion of the gap penalty can result in negative scores in some alignments.

For the purpose of further examining the TM-score, we checked the average effects of the number of gaps found in the template alignments on the final modeling results. We defined a normalized gap number (or gap density) as the ratio of N_{gap} to the number of aligned residues, where N_{gap} is calculated as $N_{\mathrm{fra}} - 1$. N_{fra} denotes the number of continuous fragments consisting of more than two residues. The data show that there is essentially no correlation between the gap density and the *Z*-rRMSD of final model (with a correlation coefficient of ≈0.004).

We also examined different ways to combine the gap penalty into eq. (1); none improved the correlation between the TM-scores and the *Z*-rRMSD of the final full-length models. The reason could be that in most model building programs,[7–9,26–29] including MODELLER, the quality of the final models is dictated by the tertiary spatial restraints extracted from the templates to give global fold information; such information is insensitive to the number of gaps in the alignment.

Automated Assessment of Protein Structure Prediction

As shown in the recent CASP experiments (Critical Assessment of Structure Prediction),[16,30,31] the accurate and automatic evaluation of the predicted tertiary models is not a trivial problem, because in lower quality models different metrics are sensitive to different features. In CASP5, the assessors developed various combined scoring functions that were efficiently used in the automatic assessment of comparative modeling[32] and fold recognition target predictions.[33] However, for difficult targets, especially those 'New Fold' targets, human visual assessment is usually necessary.[34]

Although the major purpose of the development of the TM-score in this work is to evaluate the relationship between partially aligned templates and the resulting final full-length models, we also explore the assessment power of the TM-score based on the quality of full-length models, especially for the hard New Fold targets. Because of the weak similarity of those models to the experimental structures and the lack of biological insight, here we use the human-expert evaluation in CASP5 as the 'gold standard.'

In Figure 6, we rank the first predicted model submitted by different groups of all five new fold targets, i.e. T0129,

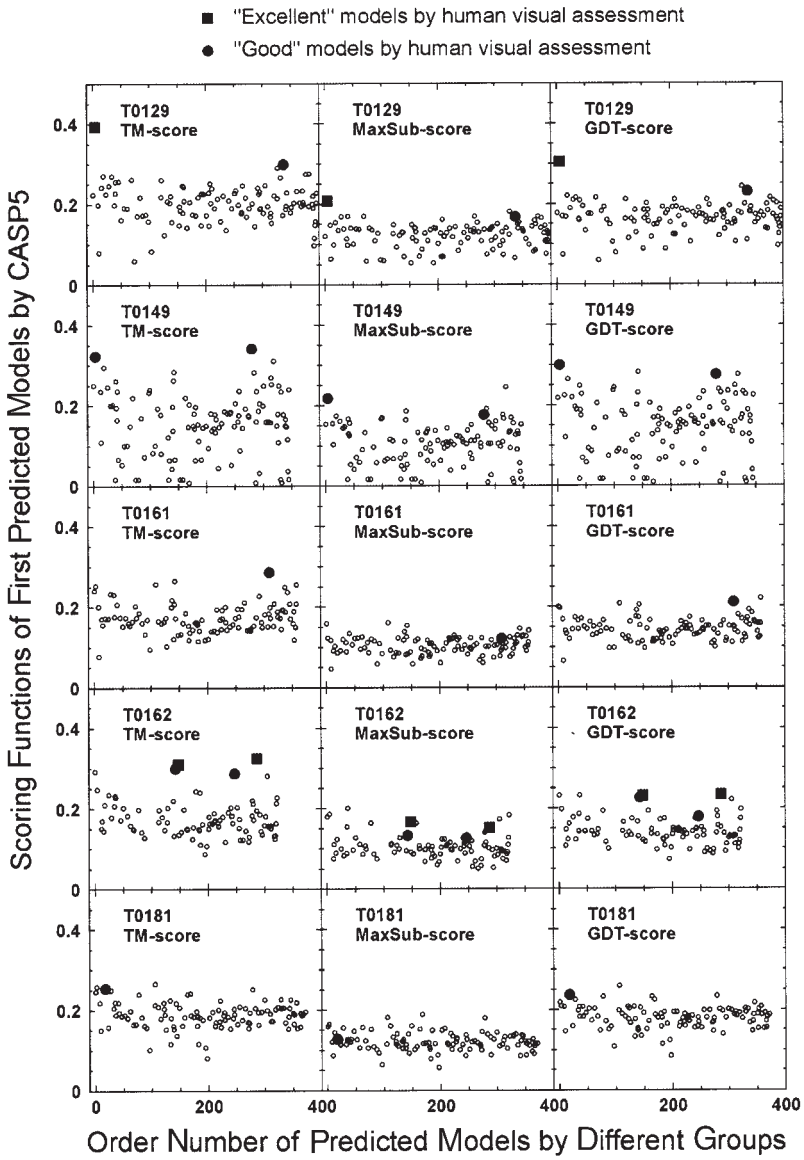

Fig. 6. Scoring functions of the first predicted model of five 'New Fold' targets in CASP5.[16] The horizontal axis is the order number of the predictions, which differentiates the predictions from different groups. The solid squares and solid circles denote, respectively, those predictions assigned by human visual assessment as 'excellent' and 'good.'[34]

T0149_2 (203–318), T0161, T0162_3 (114–281), and T0181, downloaded from the CASP5 webpage: http://prediction-center.llnl.gov/casp5/pubResultS/CASP_BROWSER/ and calculate the scoring functions of these models on the basis of TM-score (Column I), MaxSub-score (Column II), and GDT-score (Column III). According to the human visual evaluations by Aloy and coworkers[34] (data from http://www.russell.embl.de/casp5/NF/Table2_1st.html), the models were given 2 points ('excellent') when the visual assessment found that the overall fold was correct, or 1 point ('good') when the models were deemed to be partway to the correct fold and all others were given a score of 0. Among all the submitted models for the five new fold targets, there were three models in Aloy and coworkers' evaluation that received an 'excellent' score and seven models that received 'good.' These models are highlighted in Figure 6 as solid squares (excellent) and solid circles (good).

As shown in Column I of Figure 6, almost all of these successfully predicted models according to human visual assessments have obviously higher TM-scores than other predictions that were assigned by human-expert as incorrect. The TM-score of models labeled as 'excellent' are also scored somewhat higher than those labeled as 'good.' In general, using either the TM-score or the GDT-score can discriminate the excellent/good models from incorrect models; this is not the case when the MaxSub-score is used. However, there are still some examples, such as T0162, for which the GDT-score of the good model is lower than the GDT-scores of many incorrect models (see Row 4, Column III in Fig. 6).

TABLE I. Assessment Rank of Ten Successfully Predicted New Fold Models in CASP5[a]

Target_Group	Human Visual[b]	TM-Score[c]	GDT-Score[c]	MaxSub-Score[c]
T0129_002	Excellent	1	1	1
T0162_132	Excellent	2	4	6
T0162_373	Excellent	1	1	10
T0129_349	Good	2	3	9
T0149_002	Good	2	1	2
T0149_314	Good	1	4	6
T0161_349	Good	1	2	27
T0162_112	Good	3	5	14
T0162_271	Good	5	13	19
T0181_012	Good	3	2	37
Average		2.1	3.6	13.1

[a]Only first predicted model from each group is considered
[b]Human visual assessment result by Aloy *et al.*[34]
[c]Automated rank by the various scoring functions.

Table I lists the automated rank of all ten successfully predicted models according to TM-score, GDT-score and MaxSub score respectively. Consistent with the data in Figure 6, TM-score assessments have a slightly better average rank (2.1) than do GDT-scores (3.6); both are much better than the average rank of MaxSub-score assessments (13.1).

CONCLUSIONS

We have developed a new scoring function for the automated evaluation of protein structure predictions. First, the modeling errors are normalized by a protein size dependent scale so that the average TM-score of random protein pairs has no bias to the target protein's length, which sets up a minimum threshold, i.e. a TM-score$>$0.17, for any meaningful threading alignments or final model predictions. Second, rather than using specific distance cutoff and focusing only on the fractions of structures as what was done in the MaxSub- or GDT-scoring function, all the residues of the modeled proteins are evaluated in the TM-score.

For the purpose of an objective evaluation, we construct structure templates for a large-scale benchmark set of proteins from our threading program PROSPECTOR_3,[6] and build the full-length models using the widely used program MODELLER.[7,10] The reason we selected MODELLER is that its methodology forms the basis of many comparative modeling tools[7–9,26–29] and it is extensively used by structural biologists.[35] Based on the modeling results, the TM-score shows a stronger correlation to the quality of the final full-length models than the MaxSub-score[2] or the GDT_TS-score.[1,13] We also randomly selected 200 proteins in the benchmark set and rebuilt them using a different modeling algorithm, TASSER,[36] which is designed to assemble full-length models by rearranging the continuous fragments from initial template alignments and which has the ability to refine the aligned residues to make them closer to the native structure. The results are qualitatively similar to those obtained here using MODELLER, although the average correlations of all the three scores are slightly weaker in the TASSER refinements. (This is because TASSER generates better final models of lower RMSD to native on average.) This demonstrates that the features of templates described by the TM-score are not sensitive to the specific refinement methods. Obviously, the TM-score is much faster and more convenient to use in the judgment of the template quality than any real full-length modeling refinement program.

The reason that the TM-score shows a closer correlation between the initial template alignments and the final models than does the MaxSub-score is that the TM-score counts the template information of both high accuracy aligned regions and low accuracy aligned regions, while the MaxSub score neglects the alignment information included in the low accuracy aligned regions that could be of assistance in global modeling. On the other hand, unlike the RMSD in which the prediction errors are averaged with equal weights for all residues, the TM-score uses the LG-factor that weights the low and high accuracy regions differently. This also allows the TM-score to provide a more sensitive measure than the GDT-score. This assessment is validated by the lower false positive and negative rates for full molecule assembly when TM is used to identify good template alignments compared to either the Maxsub or the GDT scores. .

We further exploit the TM-score to the evaluation of predicted full-length models of the five available 'New Fold' targets in CASP5.[16] The rank of the TM-score of the first predicted models closely coincides with the rank by the human-expert visual analysis of Aloy and coworkers.[34] The average rank of all ten successfully predicted models is 2.1 by the TM-score (3.6 by GDT-score, 13.1 by the MaxSub-score). This result suggests that the TM-score may also be used as a useful complement to the automated assessment of protein full-length structure predictions.

The TM-score program is freely downloadable at http://bioinformatics.buffalo.edu/TM-score. For the users' convenience, the program also provides options for the output of MaxSub-scores and GDT_TS-scores, based on the same search engine as the TM-scores. For structures of around 200 residues, the calculation takes less than 1 s on a 2.4 GHz Pentium-4 processor.

ACKNOWLEDGMENTS

The authors thank Drs. D. Fischer, A. Szilagyi and H. J. Zhou for stimulating and useful discussions. This research was supported in part by NIH grant GM-48835 of the Division of General Medical Sciences.

REFERENCES

1. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
2. Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 2000;16(9):776–785.
3. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
4. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358(6381):86–89.

5. Bryant SH, Altschul SF. Statistics of sequence-structure threading. Curr Opin Struct Biol 1995;5(2):236–244.
6. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Proteins 2004;56:502–518.
7. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234(3):779–815.
8. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. Proteins 1998;32(4):475–494.
9. Zhang Y, Kolinski A, Skolnick J. TOUCHTONE II: A new approach to ab initio protein structure prediction. Biophysics J 2003;85:1145–1164.
10. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9(9):1753–1773.
11. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Cryst 1976;A 32:922-923.
12. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst 1978;A 34:827-828.
13. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31(13):3370–3374.
14. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A. A study of quality measures for protein threading models. BMC Bioinformatics 2001;2(1):5.
15. Rychlewski L, Fischer D, Elofsson A. LiveBench-6: large-scale automated evaluation of protein structure prediction servers. Proteins 2003;53 Suppl 6:542–547.
16. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 2003;53 Suppl 6:334–339.
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–242.
18. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95(11):5913–5920.
19. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11(11):2606–2621.
20. Betancourt MR, Skolnick J. Universal similarity measure for comparing protein structures. Biopolymers 2001;59(5):305–309.
21. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci 2001; 10(7):1470–1473.
22. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20(11):478–480.
23. Reva BA, Finkelstein AV, Skolnick J. What is the probability of a chance prediction of a protein structure with an RMSD of 6 Å? Fold Des 1998;3(2):141–147.
24. Gerstein M, Levitt M. Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. Proc Int Conf Intell Syst Mol Biol 1996;4:59–67.
25. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol 2003;334(4):793–802.
26. Aszodi A, Taylor WR. Homology modelling by distance geometry. Fold Des 1996;1(5):325–334.
27. Brocklehurst SM, Perham RN. Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipoylated H-protein from the pea leaf glycine cleavage system: a new automated method for the prediction of protein tertiary structure. Protein Sci 1993;2(4):626–639.
28. Havel TF, Snow ME. A new method for building protein conformations from sequence alignments with homologues of known structure. J Mol Biol 1991;217(1):1–7.
29. Srinivasan N, Blundell TL. An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. Protein Eng 1993;6(5):501–512.
30. Moult J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins 1999;Suppl 3:2–6.
31. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins 2001;Suppl 5:2–7.
32. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53 Suppl 6:352–368.
33. Kinch LN, Wrabl JO, Krishna SS, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV. CASP5 assessment of fold recognition target predictions. Proteins 2003;53 Suppl 6:395–409.
34. Aloy P, Stark A, Hadley C, Russell RB. Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins 2003;53 Suppl 6:436–456.
35. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
36. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.