Published online April 22, 2005

2302–2309 Nucleic Acids Research, 2005, Vol. 33, No. 7
doi:10.1093/nar/gki524

TM-align: a protein structure alignment algorithm based on the TM-score

Yang Zhang and Jeffrey Skolnick*

Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington Street, Buffalo, NY 14203, USA

Received March 5, 2005; Revised and Accepted April 1, 2005

ABSTRACT

We have developed TM-align, a new algorithm to identify the best structural alignment between protein pairs that combines the TM-score rotation matrix and Dynamic Programming (DP). The algorithm is ~4 times faster than CE and 20 times faster than DALI and SAL. On average, the resulting structure alignments have higher accuracy and coverage than those provided by these most often-used methods. TM-align is applied to an all-against-all structure comparison of 10 515 representative protein chains from the Protein Data Bank (PDB) with a sequence identity cutoff <95%: 1996 distinct folds are found when a TM-score threshold of 0.5 is used. We also use TM-align to match the models predicted by TASSER for solved non-homologous proteins in PDB. For both folded and misfolded models, TM-align can almost always find close structural analogs, with an average root mean square deviation, RMSD, of 3 Å and 87% alignment coverage. Nevertheless, there exists a significant correlation between the correctness of the predicted structure and the structural similarity of the model to the other proteins in the PDB. This correlation could be used to assist in model selection in blind protein structure predictions. The TM-align program is freely downloadable at http://bioinformatics.buffalo.edu/TM-align.

INTRODUCTION

Protein structure comparisons are employed in almost all branches of contemporary structural biology, ranging from protein fold classification (1,2), protein structure modeling (3) to structure-based protein function annotation (4,5). With the rapid increase in the number of solved protein structures in the Protein Data Bank (PDB) (6) and the progress on proteome-scale protein structure modeling (7–9) and functional annotation (10), the need for fast and accurate structure comparison algorithms has become more and more crucial. In general, there are two types of comparisons for protein tertiary structures. The first is to compare protein structures/models with an a priori specified equivalence between pairs of residues (such an equivalence can be provided by sequence or threading algorithms, for example). The most commonly used metric in this category is the root-mean-square deviation, RMSD, in which the root-mean-square distance between corresponding residues is calculated after an optimal rotation of one structure to another (11). Since the RMSD weights the distances between all residue pairs equally, a small number of local structural deviations could result in a high RMSD, even when the global topologies of the compared structures are similar. Furthermore, the average RMSD of randomly related proteins depends on the length of compared structures, which renders the absolute magnitude of RMSD meaningless (12). The recently proposed TM-score (13) overcomes these problems by exploiting a variation of Levitt–Gerstein (LG) weight factor (14) that weights the residue pairs at smaller distances relatively stronger than those at larger distances. Therefore, the TM-score is more sensitive to the global topology than to the local structural variations. Moreover, the value of the TM-score is normalized in a way that the score magnitude relative to random structures is not dependent on the protein's size, with a value of 0.17 for an average pair of randomly related structures (13).

The second type of structure comparison compares a pair of structures where the alignment between equivalent residues is not a priori given. Therefore, an optimal alignment needs to be identified, which is in principle an NP-hard problem with no exact solution (15). A variety of different structure alignment approaches have been proposed to search for the best structure alignment. These differ mainly in the score matrix used to assess the alignments and the search algorithm employed to identify the defined best alignment. For example, in DALI (16), the equivalency score is defined as the difference between the intra-structural residue–residue distances in the compared structures, and a Monte Carlo procedure is exploited to search for the minimum in the cumulative equivalency score. In CE (17), the score is measured by the intra-structural distance of

*To whom correspondence should be addressed. Tel: +1 716 849 6712; Fax: +1 716 849 6747; Email: skolnick@buffalo.edu

eight-residue fragments, and the alignment is built by gradually adding new eight-residue fragments to the existing alignment path. In STRUCTAL (14) and SAL (18), the authors used the inter-structural residue–residue distance based LG-score matrix and maximized the cumulative LG-score (14) or relative RMSD (12) by a heuristic iteration of Needleman–Wunsch dynamic program (19). However, since the LG-score is calculated based on the Kabsch rotation matrix (11) that was defined for minimizing the RMSD rather than maximizing the LG-score, this mismatch in the alignment optimization can slow down the convergence of the iteration procedure and reduce the efficiency of these algorithms.

In this work, we will extend the approaches of Levitt and Gerstein (14) and Kihara and Skolnick (18), but use the TM-score rotation matrix to speed up the process of identifying the best structure alignments. We at first test the developed algorithm on a small set of 200 non-homologous proteins and compare the results to existing methods. Then, we will apply the algorithm for large-scale native-to-native and model-to-native structure comparisons.

METHODS

TM-align only employs the backbone C_α coordinates of the given protein structures; however, the methodology is readily generalized to any type of atom.

Initial structural alignment

Three kinds of quickly identified initial alignments are exploited. The first type of initial alignment is obtained by aligning the secondary structures (SSs) of two proteins using dynamic programming (DP) (19). The element of the score matrix is assigned to be 1 or 0 depending on whether or not the SS elements of aligned residues are identical. Here, a penalty of -1 for gap-opening works the best. For a given residue, an SS state (α, β or coil) is assigned based on the C_α coordinates of five neighboring residues, i.e. ith residue is assigned as $\alpha(\beta)$ when

$$|d_{j,j+k} - \lambda_k^{\alpha(\beta)}| < \delta^{\alpha(\beta)}, \quad (j = i-2, i-1, i;\ k = 2,3,4) \qquad \mathbf{1}$$

is satisfied for all $d_{j,j+k}$ that denotes the C_α distance between the jth and $(j+k)$th residues; otherwise, it is assigned to be a coil. The final assignment is further smoothed by merging and removing singlet SS states. We note that the set of eight parameters are optimized based on 100 non-homologous training proteins by maximizing the SS assignment similarity to the DSSP definition (20), which defines protein SS elements on the basis of hydrogen bond patterns and requires the full set of backbone atomic coordinates. The optimized parameters are $\lambda_2^\alpha = 5.45$ Å, $\lambda_3^\alpha = 5.18$ Å, $\lambda_4^\alpha = 6.37$ Å, $\delta^\alpha = 2.1$ Å, $\lambda_2^\beta = 6.1$ Å, $\lambda_3^\beta = 10.4$ Å, $\lambda_4^\beta = 13$ Å, $\delta^\beta = 1.42$ Å. Using Equation 1, we achieve an average Q_3 accuracy of 85% with respect to the DSSP assignment for the representative 1489 non-homologous test protein set used in Ref. (8).

The second type of initial alignment is based on the gapless matching of two structures. As in SAL (18), for the smaller of the two compared proteins, we perform gapless threading against the larger structure, but rather than use RMSD as the comparison metric as was done in SAL, now the alignment with the best TM-score is selected.

The third initial alignment is also obtained by DP using a gap-opening penalty of -1, but the score matrix is a half/half combination of the SS score matrix and the distance score matrix selected in the second initial alignment.

Heuristic iteration

The above-obtained initial alignments are submitted to a heuristic iterative algorithm, which has been extensively used in refining NP-hard structure-based alignments (14,18,21). In this procedure, we first rotate the structures by the TM-score rotation matrix (13) based on the aligned residues in the initial alignments. The score similarity matrix is defined as

$$S(i,j) = \frac{1}{1 + d_{ij}^2 / d_0(L_{\min})^2}, \qquad \mathbf{2}$$

where d_{ij} is the distance of the ith residue in structure 1 and the jth residue in structure 2 under the TM-score superposition; $d_0(L_{\min}) = 1.24\sqrt[3]{L_{\min} - 15} - 1.8$ with $L_{\min}$ being the length of the smaller protein. A new alignment can be obtained by implementing DP on the matrix S(i, j) with an optimal gap-opening penalty of -0.6. We then again superimpose the structures by the TM-score rotation matrix according to the new alignment and obtain a newer alignment by implementing DP with the new score matrix. The procedure is repeated until the alignment becomes stable and the alignment with the highest TM-score is returned. Because of the consistency of the TM-score based rotation matrix and the DP similarity score, the alignments usually converge very fast, and typically 2–3 iterations are enough for the identification of the best alignment.

Here, in both the initial alignment identification and the heuristic iterations, we only exploit gap penalties for gap opening but not for gap extension. Another option is to eliminate the gap penalties entirely and consider a local cooperativity term to avoid overfragmentation within helices and strands.

RESULTS

Benchmark test

To test the performance of the algorithm, we collect a set of 200 non-homologous protein chains from the PDB, which range in size from 46 to 1058 residues and whose pairwise sequence identity is <30%. A list of the proteins as well as the full-atom structures is available at http://www.bioinformatics.buffalo.edu/TM-align/benchmark.

In Table 1, we present a summary of the structural alignments of 200 × 199 non-homologous protein pairs by TM-align, compared with three other most often-used structural alignment tools, i.e. CE (17), DALI (DaliLite 2.3) (16) and SAL (18). Here, for some algorithms (e.g. DALI), changing the order of the compared structures can result in different alignments. Moreover, the definition of TM-score (see Equation 3 below) depends on the target we select for normalization. We therefore count all comparisons with respect to both partners in Table 1.

We at first take averages for CE, SAL and TM-align over all the 39 800 structure pairs (upper half of the table). Since DALI only reports those alignments of significant Z-score and

Table 1. Structural alignments by different algorithms for 200 non-homologous PDB proteins

	Average over all pairs[a]				Average over pairs with TM_M[b]				$\langle t \rangle$[c]
	$\langle R \rangle$	$\langle L \rangle$	$\langle cov \rangle$	$\langle TM \rangle$	$\langle R_M \rangle$	$\langle L_M \rangle$	$\langle cov_M \rangle$	$\langle TM_M \rangle$	
Test set of all 39 800 structure pairs									
CE	6.52	64.3	34.7%	0.169	3.95	128.8	61.4%	0.441	2.25
SAL	7.33	95.3	47.3%	0.229	5.84	164.8	72.8%	0.474	10.00
TM-align	4.99	87.4	42.0%	0.253	4.45	166.2	73.1%	0.510	0.51
Test set of 17 086 pairs where DALI has an output									
CE	6.36	73.0	34.7%	0.185	3.95	129.2	61.2%	0.440	2.28
DALI	14.25	123.2	53.5%	0.223	9.40	175.2	76.8%	0.471	12.22
SAL	7.53	108.4	47.5%	0.241	5.83	164.4	71.7%	0.471	10.13
TM-align	5.18	101.9	43.4%	0.271	4.44	165.8	71.9%	0.506	0.52

[a]Results are averaged over all structure pairs. R, L, cov and TM denote, respectively, the RMSD (in the unit of Å), number of aligned residues, coverage of aligned regions over the target sequence and TM-score as defined in Equation 3.
[b]For each protein, only the pair with the maximum TM-score is considered, on which the averages are taken.
[c]Average CPU time (in the unit of second) per structure alignment on a 1.26 GHz PIII processor.

17 086/39 800 pairs have DALI outputs, we take averages for all four methods over these 17 086 structural pairs in the lower part of Table 1.

Columns 2–4 list the alignment accuracy and the coverage (fraction of aligned residues within the target protein). In general, algorithms with larger coverage tend to have lower accuracy. For example, DALI has the largest coverage of 53.5%, but the average RMSD of the corresponding aligned residues is 14.25 Å. CE has a higher accuracy than both DALI and SAL, but the alignment coverage is the lowest. TM-align has the highest accuracy and rank three coverage in the table. To have a single scoring function that can reasonably assess the alignment quality and balance the coverage and accuracy, we use the TM-score, which is defined as (13)

$$\text{TM-score} = \text{Max}\left[\frac{1}{L_{\text{Target}}}\sum_{i}^{L_{\text{ali}}}\frac{1}{1+\left(\frac{d_i}{d_0\left(L_{\text{Target}}\right)}\right)^2}\right]. \qquad \mathbf{3}$$

Here, L_{Target} is the length of target protein that other PDB structures are aligned to; L_{ali} is the number of aligned residues; d_i is the distance between the ith pair of aligned residues. $d_0\left(L_{\text{Target}}\right) = 1.24\sqrt[3]{L_{\text{Target}}-15}-1.8$ is a distance parameter that normalizes the distance so that the average TM-score is not dependent on the protein size for random structure pairs and can be thought of as the average distance between an aligned pair of residues in a randomly related pair of structures whose target structure is length L_{Target}.

Based on the average TM-score, TM-align is ranked best, followed by SAL, DALI and CE. Certainly, the rank of performance of the algorithms could be different when different evaluation criteria are used. For example, if one simply considers alignment coverage, DALI and SAL will rank better than TM-align. In fact, a variety of different evaluation methods have been considered in the literature (22). For example, many authors use SCOP (1) or CATH (2) as the gold standard and assess the structural alignments based on the fold classifications in these databases (23–25). Because the CATH and SCOP classifications are discrete, a drawback of this kind of evaluation is that the detailed alignment quality is not taken into account. Moreover, the creators of some databases such as CATH use information from other structural alignment algorithms. Recent studies (18,26) have shown that significant structural similarity exists in the proteins belonging to different fold families in the CATH and SCOP classifications. Here, the criteria we adopt is purely geometric; i.e. for the same set of proteins, the winners are those who find the more matched residues (coverage) with higher accuracy (low RMSD), where the TM-score represents an appropriate combined quality measure (13). It has been demonstrated elsewhere (13) that the TM-score has the strongest correlation with the foldability of alignments by the generally used modeling tool, MODELLER (27), in comparison with other similarity scores. Moreover, the rank of CASP5 models by TM-score is highly consistent with that of human-expert visual evaluations (28). Here, we do not include the alignment gap penalty in the TM-score evaluation because there is no obvious correlation between the gap density and the foldability of a given alignment (13).

While the data in columns 2–5 are the average over all structure pairs where the majority of them have different folds and low TM-score, a more practical question is to check the method's abilities to fish out the most significant structural match to a given target structure. In columns 6–9, we choose the match of the highest TM-score for each target protein and do the average for all 200 target proteins (or 198, when considering those proteins with DALI outputs). On the basis of either coverage or accuracy, TM-align is ranked second in this average. The average TM-score of TM-align again ranks the best.

In the last column of Table 1, we list the average CPU time per structure pair, where all the alignments are done on a 1.26 GHz Pentium III processor. The average CPU time per pair by TM-align is ~0.5 s, which is ~4 times faster than CE, ~20 times faster than DALI and SAL.

In Figure 1, we show a typical example of a structural comparison between 1atzA and 1auoA, which have a sequence identity of 16% and share a similar αβα-sandwich fold. While 1atzA has five β-strands and three α-helices on each side, 1auoA has seven β-strands in the middle and two α-helices in the left and four α-helices in the right side. The latter has also a unique long β-turn on the right side. An ideal structure alignment, therefore, should match two α-helices on the left, five β-strands in the middle and three α-helices on the right side of the two structures.

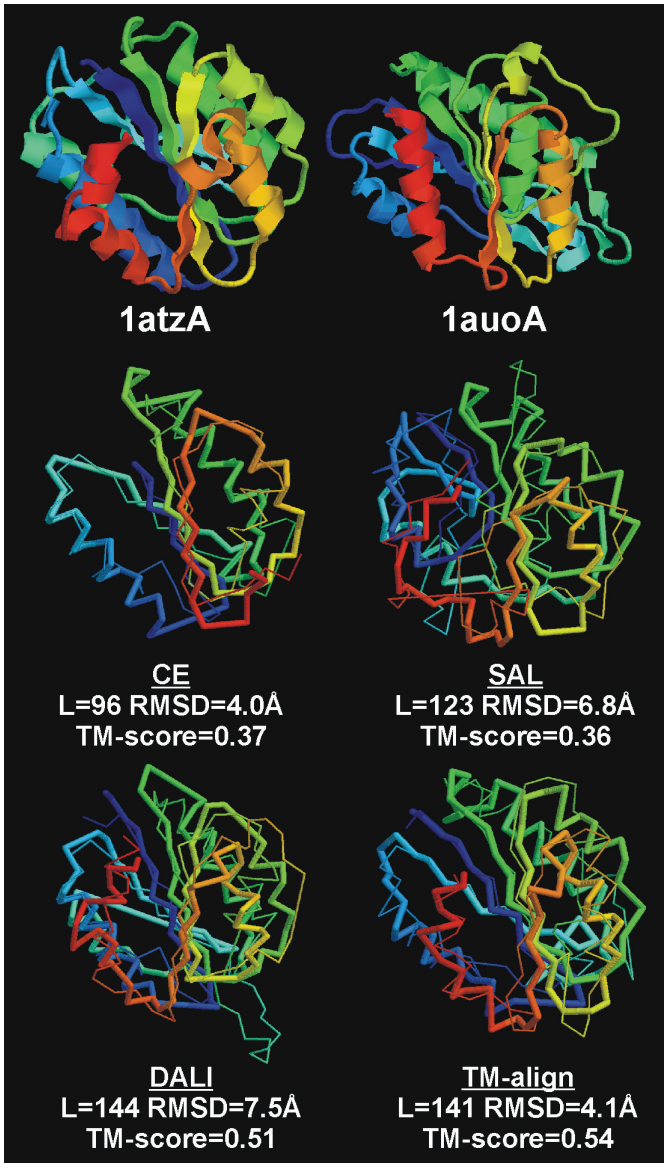

Figure 1. Illustrative example of structure alignments by different alignment methods for 1atzA and 1auoA. The first row is the ribbon diagram of the native structures of 1atzA (184 residues) and 1auoA (218 residues), which have a sequence identity 16% and adopt the common αβα-sandwich topology. The second and third rows are the structure superposition between the aligned residues by CE (17) and SAL (18), DALI (38) and TM-align algorithms, respectively. The thick and thin backbones denote the aligned residues from 1atzA and 1auoA, respectively. The indicated numbers are the length of aligned residues, the RMSD between the aligned residues, and the TM-score normalized by the length of 1atzA. All the pictures are generated by RASMOL (http://www.umass.edu/microbio/rasmol) with blue to red running from the N- to C-terminus.

As shown in Figure 1, CE aligns one α-helix in the left-hand side, four β-strands in the center and two α-helices in the right-hand side. The overall RMSD is 4.0 Å with 96 aligned residues, which results in a TM-score of 0.37. SAL aligns three α-helices in the left side, four β-strands in the middle, three α-helices on the right; but the third left-hand side α-helix in 1atzA is misaligned to the β-turn of 1auoA, which results in a high RMSD of 6.8 Å over 123 residues and a TM-score = 0.36. DALI has the longest alignment coverage. It aligns correctly two α-helices in the left, five β-strands in the middle and two α-helices in the right. But the third α-helix on the left is misaligned to the β-turn of another structure and the third α-helix on the right has large errors, which results in a high overall RMSD of 7.5 Å on 144 residues and a TM-score = 0.51. Only TM-align correctly aligns all the SSs, which have an RMSD of 4.1 Å over 141 residues and a TM-score = 0.54.

How many folds are there in the PDB?

The answer to this question obviously depends on how we define a protein fold, a definition that is usually subjective. In SCOP (1), for example, a fold is defined by the arrangement of assigned secondary structure elements on the basis of human visual inspection. This resulted in the identification of 800-folds in the last version of SCOP (August 1, 2003). In CATH (2), a fold at the Topology-level is defined by the structure alignment score of SSAP combined with some alignment coverage cutoff. Based on the last version of CATH 2.5.1 (January 28, 2004), there are 3300-fold families in the CATH database. Here, we try to use the TM-score to define the number of protein folds, partially because the TM-score is normalized so that a value of TM-score = 0.17 is that between two randomly related pairs of structures independent of target length (13). A TM-score cutoff = 0.5 is adopted; the criterion is somewhat empirical and protein structure modeling oriented. When an alignment has a TM-score >0.5, common protein structure modeling tools such as MODELER (27) could build reasonable full-length models with an RMSD <6.5 Å in most cases (29). This is also the threshold of structural similarity that the fold-recognition algorithms can identify with confidence. In a recent fold-recognition study of 1489 non-homologous targets, the majority of alignments identified by our threading program PROSPECTOR_3 (30) with a high confidence score (Easy Set) have a TM-score >0.5 while most of the low confidence threading alignments (Medium/Hard Set) have a TM-score <0.5. Here, in order to define the fold clusters in a definitive way so that it does not depend on the comparison order of the structures, we normalize the TM-score by the average length of two compared structures. This normalization also makes the structural similarity defined by TM-score transitive. Above this TM-score cutoff, on average, 87.3% of the residues in the smaller proteins and 74.6% of the residues in the larger proteins are aligned by TM-align. The average RMSD is 2.9 Å.

As of January of 2005, there are >30 000 entries (actually 30 123 on January 18, 2005) deposited in the PDB (6). After removing theoretical models and obsolete entries, we obtained 56 096 chains that have a length longer than 40 residues, where many of the chains are redundant. Figure 2 shows the number of folds in representative sets of proteins versus the sequence identity cutoff used for collecting non-redundant samples. The number of folds is calculated based on a star-like structure clustering algorithm (31) using a TM-score of 0.5 as the cutoff. As expected, the number of folds keeps increasing, when we use higher sequence identity cutoffs. However, the increase gradually saturates when the sequence identity cutoff increases. When we use the sequence identity cutoff of 95%, there are 10 515 protein chains and 1996 different folds. The data in Figure 2 also raises the old question of what sequence identity

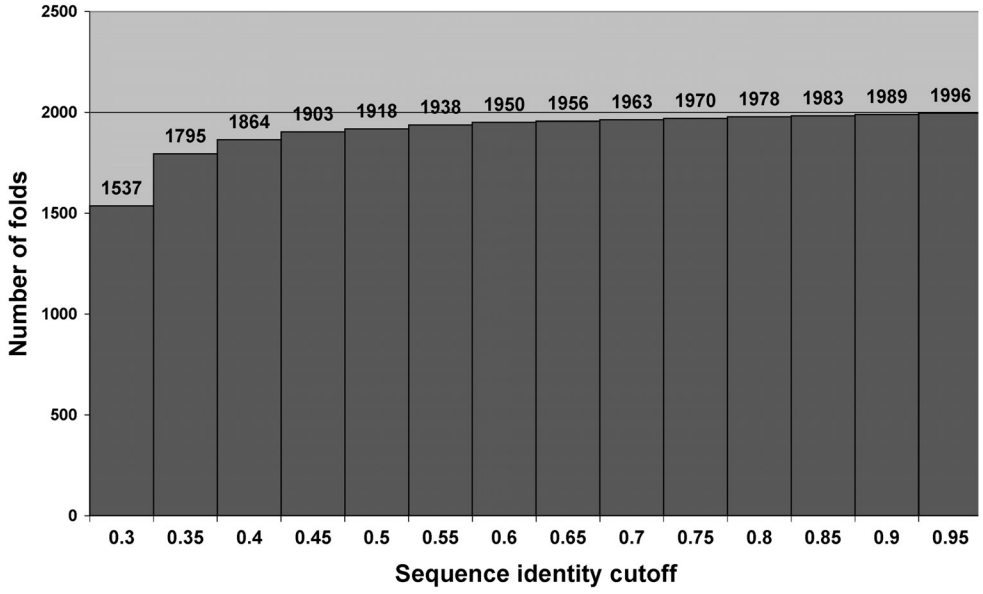

Figure 2. Number of folds included in the representative protein sets collected from the PDB library on January 28, 2005 using different sequence identity cutoffs. A fold is defined using a TM-score threshold of 0.5.

cutoff should be used when we construct a template library for the fold-recognition programs. Based on Figure 2, if a 30% sequence identity cutoff is used, we include only 77% of the protein folds in the PDB. Even if we use a 40% sequence identity cutoff, 7% of the folds will still be missed in the fold library. A similar dependence of the number of folds on the sequence identity cutoff can also be obtained if we use, e.g. SCOP or DALI. However, the absolute value of fold numbers in SCOP definition is lower than that shown in Figure 2. Based on our data, a similar number of folds can be obtained if we use a *Z*-score cutoff of ~7.3 in DALI.

The data in Figure 2 demonstrate that highly homologous proteins can adopt very different folds. In the set of 10 515 protein chains, TM-align found 52 protein pairs that have sequence identities >70% but have a TM-score <0.4. More cases should be found if we consider the whole set of 56 096 protein chains in the PDB. Most of these cases are NMR structures, except for four pairs that were solved by X-ray diffraction. One reason for the structurally divergent homologies is that the mutations of a few key residues could sometimes induce shifts of the minimum of the free-energy landscape and therefore trigger dramatic conformation changes (32). Structure differences can also be caused by changes in solvent conditions and ligand binding (33). Two typical examples are shown in Figure 3 where 1hngB, 1a64A and 1g4yB are X-ray structures and 1kkdA from NMR.

Comparison of misfolded proteins to PDB structures

For a given sequence, structure prediction algorithms generate a variety of structure decoys, which often include both correct and wrong (i.e. a structure different from that actually adopted by the sequence of interest) folds (34). An interesting question is how different are these structural decoys from the native structures of other proteins in the PDB library. Since the accessible conformational space of a medium size protein is astronomical (35), while the number of folds of solved proteins in the PDB (or even the folds existing in nature) is limited, it is conceivable that the possibility of a misfolded structure being close to any native structures should be low. On the other hand, as demonstrated in Figure 2 and by other authors (1,2,18,26), the space of solved protein structures is very dense, and the correctly folded decoys should be relatively easier to find similar folds in PDB. Therefore, it is tempting to consider the distance of models to the closest solved protein as an indicator of the correctness of the predicted structure. Having the developed structure alignment method, we can systematically check this possibility.

We first collect a set of 300 non-homologous proteins whose length ranges from 41 to 300 residues, including 250 single domain proteins and 50 multi-domain proteins. To make the benchmark representative, the 300 target proteins are equally taken from three categories of Easy, Medium, Hard sets, defined according to the score significance of template alignments by our fold-recognition program PROSPECTOR_3 (30) and approximately reflecting the difficulties of structure modeling of the targets. TASSER (36) is exploited to assemble full-length models using continuous threading template fragments by Monte Carlo simulations. The five lowest free-energy models are selected by SPICKER clustering (31).

We then search for the closest structural analogues of these five models by TM-align that are found in the PDB library (6). To guarantee that our exercise does not exploit homologous information, we exclude the proteins >30% sequence identity to the target structures from the PDB library. In Figure 4A(C), we show the TM-score (RMSD) of the closest templates identified by TM-align to native versus the TM-score (RMSD) of the TASSER models to native. As expected, there is a very strong correlation between the distances of the template to native and the model to native, although the models have on average a higher TM-score to native. The average RMSD of the template to native (8.5 Å) is slightly lower than that of the models (8.9 Å), which is partially because only a part of the total number of residues counted in the RMSD calculation for

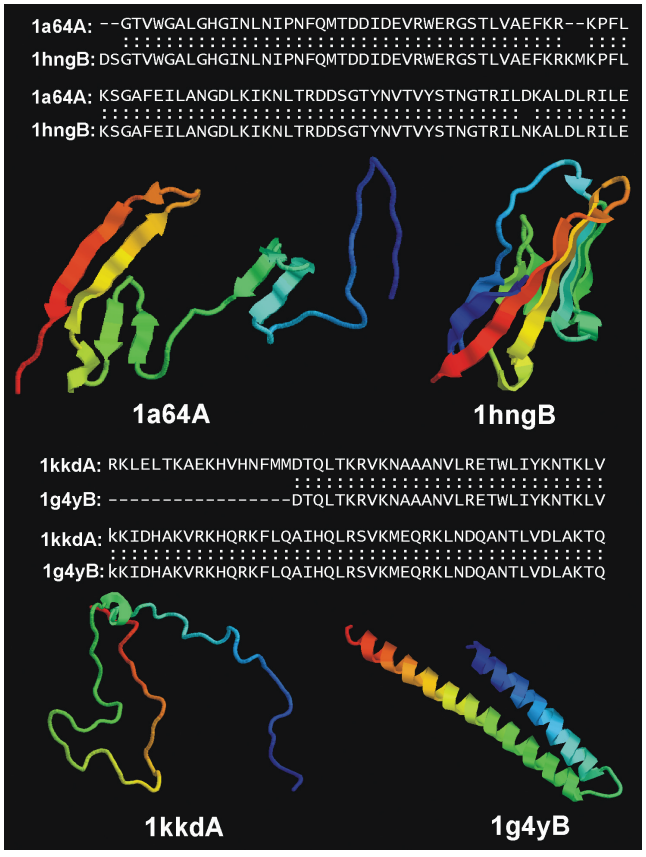

Figure 3. Two examples of protein pairs that have high sequence identities but adopt entirely different folds. In both examples, the upper parts show the sequence alignments of the proteins and ':' denotes the residues with identical amino acids; the lower parts are the cartoon structures of the proteins with blue to red running from N- to C-terminus. The proteins in the first example are from 1a64A (32) and the N-terminal domain of 1hngB (39). The deletion mutation of two key residues (K44 and M45) induces a domain swapping of two proteins. The proteins in the second example are from the calmodulin binding domain (CaMBD), where 1g4yB is the crystal structure from Ca^{2+}-loaded CaMBD in complex with calmodulin (40) and 1kkdA is the NMR structure from Ca^{2+}-free CaMBD in complex with calmodulin (33). Ca^{2+}-binding is responsible for the conformational changes of the two structures.

the templates (on average, 87% residues are aligned). It is noticeable that some points at the bottom of Figure 4C indicate a significant RMSD difference between the models and templates. A closer check shows that these TASSER models have some errors in the tail and loop regions while the TM-align automatically aligns only the core regions in the template. Figure 5 shows a typical example of this situation for 1c0fS, where the TASSER model has two tails and one loop incorrectly predicted. This results in an overall RMSD of 10.5 Å although the core region of the model is correct. In contrast, TM-align matches the core region of the model to 1kcqA with an RMSD = 1.9 Å over 76% of the residues (it is noted that 1kcqA was not used as the input template in the TASSER modeling of 1c0fS). However, the TM-score of the template is still slightly smaller than the model because of the missed regions (see Figure 5). This example also highlights the insensitivity of RMSD to the global topology of protein structures.

In Figure 4B and D, using the TM-score and RMSD, respectively, we plot the structure distances between TM-align selected templates and TASSER models versus the distance

Table 2. Comparison of the first model selected by different ranking methods

	Free-energy[a]	TM-align[b]	Random[c]	Combination[d]
⟨TM-score⟩	0.551	0.544	0.5042	0.559
⟨RMSD⟩ (Å)	8.89	9.19	10.13	8.71

[a]Ranked by the cluster size from SPICKER (31).
[b]The models are ranked on the basis of their distances to the closest non-homologous PDB structures found by TM-align.
[c]The first model is randomly selected from the five largest size clusters.
[d]Combined rank of free-energy and TM-align structural alignment. Here, for each model, a target function is defined as $C = Rank_1 + Rank_2/2$, where $Rank_1$ and $Rank_2$ are the ranks of the considered model on the basis of free-energy and TM-align, respectively. The first model is selected as the one having the lowest C.

between TASSER models and native structures. First of all, for almost all the models including both folded and misfolded ones, TM-align can find fairly close structure alignments in the PDB library, which is consistent with our earlier conclusion that the current PDB library is nearly a complete fold set (18,29). For example, even for models that have an RMSD from native of >20 Å, TM-align still finds alignments to other PDB structures <5 Å with >75% of the residues aligned in the majority of cases (Figure 4D). Because by design, structural alignments explore a large set of compact, protein-like structures, the number of which increases exponentially with protein size, the library of solved PDB structures might provide an essentially complete source of compact, protein-like structures detectable by structure alignment algorithms. In practice, this appears to be the case where even non-native decoys almost always have a reasonably close representative in the PDB.

Despite the fact all decoys are 'protein-like', it is interesting to note that there still exists a strong correlation (with correlation coefficient = 0.87) between the TM-score of the model and template, and that of model and native. This seems to suggest that the distance of models to the closest PDB structures may be considered as an indicator of model quality (37). To examine this idea, we list in Table 2 a quantitative comparison of the first models selected using different ranking methods. The rank based on the distance of TM-align structural alignment to the closest PDB structure is obviously better than random, which is consistent with the correlation data shown in Figure 4C. However, the rank of structural alignment still does not work as well as that by the free energy as shown in column 2. If we combine both ranks from the free energy and the TM-align structural alignments, we can obtain some gain in ranking the best model although it is quite marginal.

DISCUSSION

We have developed a protein structural alignment approach, called TM-align, which is an extension of ideas used in STRUCTAL by Levitt and Gerstein (14) and SAL by Kihara and Skolnick (18). The main difference between the TM-align algorithm and these previous methods is that the TM-score rotation matrix instead of Kabsch RMSD rotation matrix has been exploited in both heuristic DP iterations and final alignment selection. Because of the inherent consistency of the TM-score rotation matrix and the structural similarity scoring function, convergence is much faster. This also helps the algorithm identify more accurate alignments since the

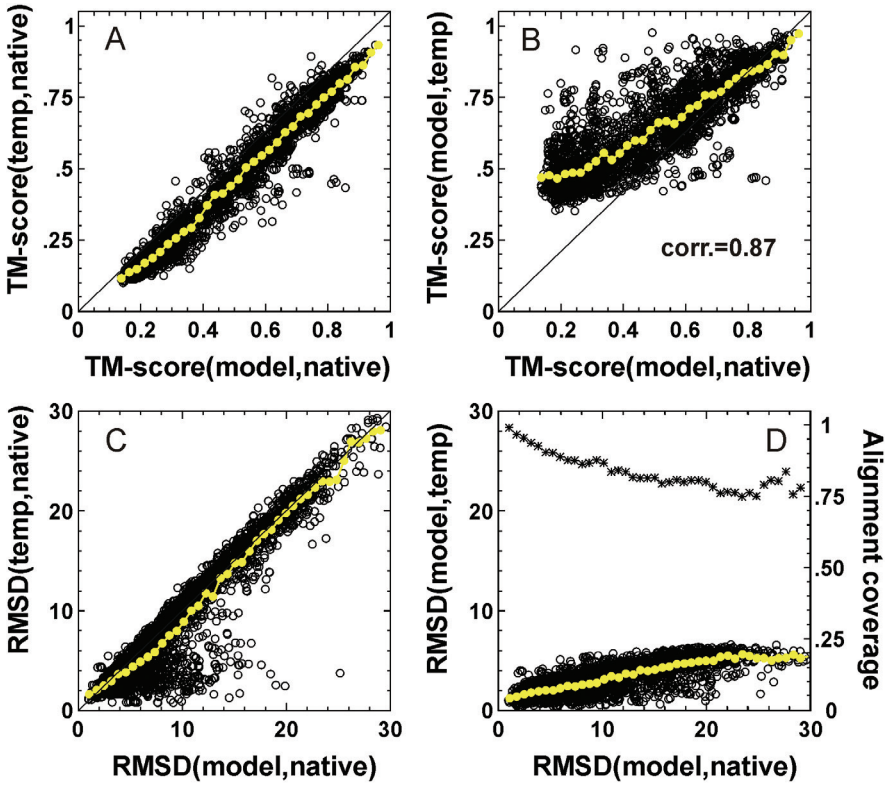

Figure 4. Structure alignments of the computer models by TASSER (8) to non-homologous proteins in the PDB library (6). (**A**) TM-score between the closest template to the native structure found by TM-align and the native structure versus the TM-score between the TASSER model and the native. (**B**) TM-score between the TASSER model and the closest found (highest TM-score) template versus the TM-score between the TASSER model and the native. (**C**) RMSD between the closest template to the native structure and the native structure versus RMSD between the model and the native. (**D**) RMSD between the model and the closest template versus the RMSD between the model and the native. The stars denote the alignment coverage of the closest templates found by TM-align. The yellow solid circles denote the average of the points fallen in the intervals of the horizontal axis in each picture. The black lines are to guide the eye.

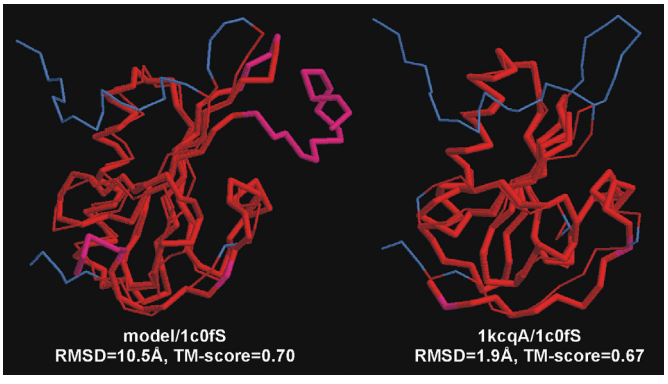

Figure 5. A comparison of a computer model generated by TASSER (8) and the closest PDB structure (template) found by TM-align. This is a typical example where the model has a much larger RMSD than the template because of the misoriented tails and loops. The thick backbones are the model or template and the thin ones the native structure of 1c0fS. The red residues are those residues where their distances are <5 Å in the TM-score rotation matrix.

TM-score rotation matrix weights smaller inter-structural distances stronger than larger inter-structural distances and is therefore more sensitive to the global structure topology than the RMSD rotation matrix. In a benchmark test of 200 × 199 non-homologous protein pairs, TM-align is ~20 times faster than SAL and yet generates more accurate alignments with comparable coverage. TM-align is also faster than two other often-used algorithms, CE (17) and DALI (38), and yet provides structure alignments of higher TM-score (which is not surprising since maximization of the TM-score is the goal of the TM-align objective function).

Because of its advantage in both speed and accuracy, TM-align is conveniently exploited in large-scale, sequence-independent structure comparisons. As an example, we used TM-align for an all-against-all structure alignment of 10 515 non-redundant protein chains in the PDB with pairwise sequence identity <95%. Approximately 2000 folds is obtained after clustering all the structures based on the threshold cutoff of TM-score = 0.5.

We find that, on many occasions, highly homologous proteins adopt very different folds. Consistent with this observation, the number of folds included by a representative set of proteins collected by sequence comparisons is sensitive to the sequence identity cutoff as shown in Figure 2. In general, there are always more or less lost folds when constructing a fold-recognition template library using a sequence identity threshold [typically 35–40% (30)]. A simple strategy to deal with this issue is to combine the sequence comparison procedure with a follow-up structural alignment search of the entire PDB to add missed folds.

We also used the TM-align algorithm to match the predicted structures to the solved non-homologous proteins in PDB. Consistent with the previous conclusion about the completeness of protein folds in PDB (18,29), both folded

and misfolded decoys are found to have close structure analogs with an average RMSD = 3 Å and 87% of the residues aligned. However, the correctly folded decoys tend to have a closer match to non-homologous PDB structures than that of misfolded ones. This finding indicates that some signal about fold correctness is carried by similarity of the models to PDB structures and the latter may be used as a complement to the free energy for the model selection in blind protein structure predictions. Presumably, this signal may be due to the fact that the models closer to PDB structures retain more protein-like packing of elementary secondary structure pieces and turns which eventually determine the global topology of the models.

A web-based server version of TM-align, as well as a freely downloadable program, is available at http://bioinformatics.buffalo.edu/TM-align.

ACKNOWLEDGEMENTS

We are grateful to Dr Andras Szilagyi for many stimulating discussions and very helpful suggestions. This research was supported in part by grant numbers GM-37408 and GM-48835 of the Division of General Sciences of the National Institutes of Health (NIH). Funding to pay the Open Access publication charges for this article was provided by NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
2. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
3. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53**, 334–339.
4. Skolnick,J., Fetrow,J.S. and Kolinski,A. (2000) Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.*, **18**, 283–287.
5. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
6. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
7. Chance,M.R., Fiser,A., Sali,A., Pieper,U., Eswar,N., Xu,G., Fajardo,J.E., Radhakannan,T. and Marinkovic,N. (2004) High-throughput computational and experimental techniques in structural genomics. *Genome Res.*, **14**, 2145–2154.
8. Zhang,Y. and Skolnick,J. (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl Acad. Sci. USA*, **101**, 7594–7599.
9. Kihara,D., Zhang,Y., Lu,H., Kolinski,A. and Skolnick,J. (2002) Ab initio protein structure prediction on a genomic scale: application to the Mycoplasma genitalim genome. *Proc. Natl Acad. Sci. USA*, **99**, 5993–5998.
10. Arakaki,A.K., Zhang,Y. and Skolnick,J. (2004) Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, **20**, 1087–1096.
11. Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* A, **34**, 827–828.
12. Betancourt,M.R. and Skolnick,J. (2001) Universal similarity measure for comparing protein structures. *Biopolymers*, **59**, 305–309.
13. Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
14. Levitt,M. and Gerstein,M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.
15. Lathrop,R.H. (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.*, **7**, 1059–1068.
16. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
17. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
18. Kihara,D. and Skolnick,J. (2003) The PDB is a covering set of small protein structures. *J. Mol. Biol.*, **334**, 793–802.
19. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
20. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
21. Hubbard,T.J. (1999) RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. *Proteins*, **37**, (Suppl. 3), 15–21.
22. Kolodny,R., Koehl,P. and Levitt,M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.
23. Leplae,R. and Hubbard,T.J. (2002) MaxBench: evaluation of sequence and structure comparison methods. *Bioinformatics*, **18**, 494–495.
24. Sierk,M.L. and Pearson,W.R. (2004) Sensitivity and selectivity in protein structure comparison. *Protein Sci.*, **13**, 773–785.
25. Novotny,M., Madsen,D. and Kleywegt,G.J. (2004) Evaluation of protein fold comparison servers. *Proteins*, **54**, 260–270.
26. Yang,A.S. and Honig,B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.*, **301**, 665–678.
27. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
28. Aloy,P., Stark,A., Hadley,C. and Russell,R.B. (2003) Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins*, **53**, 436–456.
29. Zhang,Y. and Skolnick,J. (2005) The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl Acad. Sci. USA*, **102**, 1029–1034.
30. Skolnick,J., Kihara,D. and Zhang,Y. (2004) The protein structure prediction problem could be solved using the current PDB library. *Proteins*, **56**, 502–518.
31. Zhang,Y. and Skolnick,J. (2004) SPICKER: a clustering approach to identify near-native protein folds. *J. Comput. Chem.*, **25**, 865–871.
32. Murray,A.J., Head,J.G., Barker,J.J. and Brady,R.L. (1998) Engineering an intertwined form of CD2 for stability and assembly. *Nature Struct. Biol.*, **5**, 778–782.
33. Wissmann,R., Bildl,W., Neumann,H., Rivard,A.F., Klocker,N., Weitz,D., Schulte,U., Adelman,J.P., Bentrop,D. and Fakler,B. (2002) A helical region in the C terminus of small-conductance Ca^{2+}-activated K^+ channels controls assembly with apo-calmodulin. *J. Biol. Chem.*, **277**, 4558–4564.
34. Zhang,Y., Kolinski,A. and Skolnick,J. (2003) TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.*, **85**, 1145–1164.
35. Levinthal,C. (1969) How to fold graciously. In Debrunner,P., Tsibris,J. and Munck,E. (eds), *Mossbauer Spectroscopy in Biological Systems.* University of Illinois Press, Illinois, pp. 22–24.
36. Zhang,Y. and Skolnick,J. (2004) Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophys. J.*, **87**, 2647–2655.
37. Bradley,P., Chivian,D., Meiler,J., Misura,K.M., Rohl,C.A., Schief,W.R., Wedemeyer,W.J., Schueler-Furman,O., Murphy,P., Schonbrun,J. *et al.* (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins*, **53**, 457–468.
38. Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.
39. Jones,E.Y., Davis,S.J., Williams,A.F., Harlos,K. and Stuart,D.I. (1992) Crystal structure at 2.8 Å resolution of a soluble form of the cell adhesion molecule CD2. *Nature*, **360**, 232–239.
40. Schumacher,M.A., Rivard,A.F., Bachinger,H.P. and Adelman,J.P. (2001) Structure of the gating domain of a Ca^{2+}-activated K^+ channel complexed with Ca^{2+}/calmodulin. *Nature*, **410**, 1120–1124.