Published in final edited form as:
Proteins. 2012 August ; 80(8): 2080–2088. doi:10.1002/prot.24100.

A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction

Yuedong Yang, **Jian Zhan**, **Huiying Zhao**, and **Yaoqi Zhou***
Indiana University School of Informatics, Indiana University-Purdue University and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 719 Indiana Ave., Walker Plaza Building Suite 319, Indianapolis, IN 46202, USA

Abstract

A structure alignment program aligns two structures by optimizing a scoring function that measures structural similarity. It is highly desirable that such scoring function is independent of the sizes of proteins in comparison so that the significance of alignment across different sizes of the protein regions aligned is comparable. Here, we developed a new score called SP-score that fixes the cutoff distance at 4Å and removes the size dependence by using a normalization prefactor. We further build a program called SPalign that optimizes SP-score for structure alignment. SPalign was applied to recognize proteins within the same structure fold and having the same function of protein-DNA or protein-RNA binding. For fold discrimination, SPalign improves sensitivity over TMalign for the chain-level comparison by 12% and over DALI for the domain-level comparison by 13% at the same specificity of 99.6%. The difference between TMalign and SPalign at the chain level is due to the inability of TMalign to detect single domain similarity between multi-domain proteins. For recognizing nucleic acid binding proteins, SPalign consistently improves over TMalign by 12% and DALI by 31% in average value of Mathews correlation coefficients for four datasets. SPalign with default setting is 14% faster than TMalign. SPalign is expected to be useful for function prediction and comparing structures with or without domains defined. The source code for SPalign and the server are available at http://sparks.informatics.iupui.edu.

Structure comparison of proteins allows detection of remote homologs that are otherwise not obvious from sequence comparison and aids in binding-site prediction for protein-ligand[1, 2], protein-protein[3, 4], protein-DNA[5, 6] and protein-RNA[7, 8, 9] interactions. Such comparison is productive because protein structures are often more conserved than their sequences[10]. Moreover, automatic structural comparison is complementary to manual protein structure classification[11, 12] that lags far behind the pace of newly determined structures due to structural genomics projects[13]. As a result, protein structure alignment has been an active area of research for more than thirty years with more than 50 computational methods published [14, 15, 16]. Representative examples in early studies include SSAP[17], DALI[18], and CE[19].

Central to a structure alignment method is the scoring function that measures structural similarity. One common measure is root mean square deviation (RMSD) between two

*To whom correspondence should be addressed: Phone: (317) 278-7674, Fax: (317) 278-9201, yqzhou@iupui.edu.

Supplement Materials: The supplement materials contain 14 lists of proteins (SCOP, SCOPc, rSCOP, rSCOPc, DNA-pos, DNA-neg, RNA-pos, RNA-neg, RNA-template, RNAc-pos, RNAc-neg, and RNAc-template). The lists for DNAc-pos and DNAc-neg are the same as those for DNA-pos and DNA-neg. All files along with pdb strucures are also available on http://sparks.informatics.iupui.edu.

structures. However, RMSD is heavily dependent on protein size and radius of gyration and very sensitive to the badly aligned regions and local structural changes, as all atoms in the structures are weighted equally [20]. Other alternative scoring functions have been proposed [14, 15, 16]. For example, Zemla proposed a global distance test score (GDT-score) by averaging the number of Cα atoms at several cutoff distances [21]. Levitt and Gerstein measured similarity by summing aligned pairs with a weight gradually decayed from 1 at distance of 0Å to 0 at infinity (LG-score) [22]. $\text{LG-score}=\sum(1+d_{ij}^2/d_0^2)^{-1}/L$ where d_{ij} is the distance between two aligned residues, L is a measure of protein size, and d_0 is a constant value, originally set to 5Å. Later, MaxSub was proposed with a d_0 of 3.5Å [23]. However, as RMSD, the magnitude of these scores depends on protein sizes [24].

To remove the dependency of structure similarity score on protein sizes, Zhang and Skolnick developed TM-score [24], and later applied to structure alignment[25]. The score is based on LG-score with an empirical size-dependent d_0 [$= 1.24(L - 13)^{1/3} - 1.8$] so that the score becomes size independent and makes structure comparison between proteins of different sizes possible. The equation for d_0, however, was obtained from the assumption that proteins are globular proteins and aligned in a predetermined size L. In other words, TM-score implicitly assumes that two structures align in a length of L (shorter length or average length). In reality, it is quite possible that only a part of a protein of unknown size aligns with a part of another protein, particularly for multi-domain proteins. Moreover, d_0 in TM-score is less than 0 for small fragments and can be as large as 10Å for a protein size of 1000 residues. A distance of 10Å, however, is not meaningful alignment for a pair of residues.

Here, we propose to remove the size dependence not by size-dependent d_0 but by a size-dependent normalization factor. This allows us to introduce an effective alignment length that removes the need to specify a length for normalization. The new score with its alignment program SPalign is tested in structure classification and prediction of nucleic-acid binding proteins and compared to DALI[26], CE[19], TMalign[25], and FrTMalign[27]. For recognizing structures in the same SCOP fold, SPalign is significantly more sensitive (>10%) in fold recognition than TMalign for chain-chain comparison and DALI for domain-domain comparison at the same specificity and similar in performance to TMalign for domain-domain comparison and DALI for chain-chain comparison. For predicting DNA/RNA-binding proteins, SPalign consistently improves over DALI and TM-score at both chain and domain levels.

METHODS

Datasets

SCOP: SCOP domain dataset—

We employed the dataset SCOP-20 that was utilized as a benchmark for testing the fold recognition program SPARKS X [28]. The dataset was built by using domains of sequence identity less than 20% and chain lengths greater than 60 from SCOP 1.75 [12]. After removing domains with C$_\alpha$ atoms only, we obtained 6367 domains.

SCOPc: SCOP chain dataset—

To further test our scoring function with multi-domain proteins, non-redundant chains for all domains contained in the SCOP-20 dataset are collected. There are a total of 5300 chains. We define that two chains are considered to be similar in structure if a domain in one chain belongs to the same fold of another domain in the other chain. This chain-level comparison is a real-world test because domains are often not defined for most newly solved structures.

rSCOP and rSCOPc datasets—In order to compare with slower structure alignment methods, we randomly chose 1058 and 1060 proteins from SCOP (rSCOP) and SCOPc (rSCOPc) datasets, respectively.

DNA and DNAc: Protein-DNA complex and non-binding datasets—We employed the dataset compiled by Gao and Skolnick [5] that contains 179 DNA binding protein domains and 3797 non-DNA binding protein domains (DB179 and NDB3787). The two sets were based on 35% sequence identity cutoff, a resolution of 3Å or better, a minimum length of 40 residues for proteins, 6 base pairs for DNA and 5 residues interacting with DNA (within 4.5Å of the DNA molecule). We also built two corresponding sets at the chain level by mapping domains to chains and removing redundant chains. Chains containing DNA binding domains are considered DNA binding chains. All other chains are non-DNA binding chains. Each chain was employed only once in the dataset. Numbers of DB and NDB chains are the same as the corresponding numbers of DB and NDB domains. The domain and chain datasets are referred as DNA and DNAc, respectively.

RNA and RNAc: Protein-RNA complex and non-binding datasets—The binding and non-binding datasets were compiled by Zhao et al [8]. The dataset at the domain level contains 212 RNA binding protein domains as the positive set, 6761 non-RNA binding protein domains as the negative set, and 250 RNA binding protein domains as templates. The two sets (denoted as RNA) were built by sequence identity of 30%, while the template library are built from 95% sequence identity cutoff. To mimic realistic situation of RNA-binding protein prediction, the proteins in RNA sets are structurally aligned to proteins in the template set. When compared to one query protein in the positive set, the templates with sequence identity higher than 30% are excluded. Similarly, we constructed the datasets at the chain level with 192 RNA-binding chains and 5785 non-RNA binding chains (RNAc).

SP-score

The SP-score is defined by

$$\text{SP-score}=\frac{1}{3L^{1-\alpha}}\left[\sum_{d_{ij}<2d_0}\left(\frac{1}{1+d_{ij}^2/d_0^2}-0.2\right)\right] \qquad (1)$$

where d_{ij} is the distance between C$_\alpha$ atoms of two aligned residues, d_0 was chosen 4.0Å somewhat in between 3.5Å in MaxSub and 5Å in LG score, α is a to-be-determined parameter for removing the dependence on protein length L, a constant of 0.2 is employed for a smooth cutoff for SP-score at $d_{ij} = 2d_0$, and a factor of 1/3 is employed to scale the threshold for fold discrimination to around 0.5. Note that the summation is for residues within $2d_0$ so that only meaningfully aligned pairs contribute to the score. Three L values are examined: the shorter length of two proteins in comparison (L_b), the average length (L_a) of the two proteins, and the effective alignment length (L_e). L_e is the number of core aligned residues (distance ≤ $2d_0$) plus the average number of surrounding residues in two proteins that are within $3d_0$ from any core residues. Defining an effective alignment length as a prefactor for normalization is possible because d_0 in SP-score, unlike TM-score, is length-independent. The corresponding scores are called SPb, SPa, and SPe, respectively. There is only one parameter α in our scoring function employed for removing the size dependence of the average score for proteins within the same fold as shown in Fig. 1 (α = 0.3).

SPalign Program

The program is written by C++. The alignment method employed similar heuristics as the TMalign program[25]. The seeds for alignments are initialized by gapless threading, secondary structure fitting, and fragment matching. The default value for the initial seed fragment size is 20. We also have an option called the “fast” mode by using similar fragment size as in TMalign. In this mode, the fragment will be set to 1/5 of the shorter length of two proteins in comparison and between 20 and 50. We improve the computational efficiency by stopping optimization if the RMSD between two fragments of size n is less than 1.2 * $n^{1/3}$. Another important difference from TMalign is that we minimize a weighed RMSD ($\sqrt{\sum (d_{ij}^{(k)})^2/(1+(d_{ij}^{(k-1)})^2/d_0^2)}$ where k is the iteration number and $d_{ij}^{(k-1)}=d_0$ at $k = 1$. This weighted RMSD is more consistent with TM-score and SP-score in search for the optimized rotation[29]. An equal weight is used in TMalign. As we shall see, this weight factor leads to 4% higher alignment score, compared to equal weight.

Performance Assessment

For two-state discrimination (same fold, RNA or DNA binding), we will mainly assess the method performance by Mathews correlation coefficient (MCC).

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}} \qquad (2)$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. Here positives can be two protein domains (by themselves or within chains) in the same fold (SCOP), binding to DNA or binding to RNA. MCC values range between −1 and 1 where 1 represents a perfect prediction and −1 a case where all predictions are incorrect. Higher MCC values indicate better prediction performance. To compare different methods, the highest MCC values are obtained by optimizing the thresholds. In addition, we calculated sensitivity [TP/(TP+FN)] and specificity [TN/(TN+FP)].

Other Methods

TMalign with the version updated on 01/30/2011 was downloaded from http://zhanglab.ccmb.med.umich.edu/TM-align/. DaliLite version 3.3 was downloaded from http://ekhidna.biocenter.helsinki.fi/dali_lite/downloads/v3/. frTMalign 1.0 was downloaded from http://cssb.biology.gatech.edu/fr-tm-align. CE was downloaded from http://cl.sdsc.edu/.

RESULTS

SCOP Datasets

In Fig. 1, we obtained the average alignment score for protein pairs of similar sizes in the same SCOP folds (the difference between the lengths of two proteins is less than 10%). SP-score at $\alpha = 0$ is a monotonically decreasing function of the size of proteins in comparison. This size dependence is mostly removed at $\alpha = 0.3$. The removal of the size dependence leads to a significant increase in MCC values in all our fold-recognition tests through all-against-all pairwise structure alignment. For example, SPalign with SP-score at $\alpha = 0$ with a shorter length for normalization leads to the MCC values of 0.35 and 0.24 at a cutoff value of 0.520 and 0.505 for recognizing the same fold for SCOP and SCOPc, respectively. By comparison, the corresponding MCC values for SPalign with SP-score at $\alpha = 0.3$ are 0.58 and 0.45 at the cutoff values of 0.546 and 0.541, respectively. Thus, removing size dependence leads a significant increase in MCC values (66% and 88% increase for SCOP and SCOPc sets, respectively). Here and hereafter, we employ SP-score at α = 0.3 only.

We first compare our SP-score by SPalign to TM-score by TMalign for fold recognition through pairwise all-against-all structure alignment. TM-score can be obtained by normalization based on average chain length (L_a, TMa for short notation) or the short chain length (L_b, TMb). We have corresponding scores for SP-score (SPa and SPb). In addition, we also have a version with an effective alignment length (L_e, SPe).

At both chain and domain levels, the scores based on the average chain length always lead to higher MCC values than the scores based on the shorter chain length. This is true for TM-score and SP-score. Although SPalign based on SP-score has a MCC value comparable to TMalign at the domain level, it achieves 24% higher MCC value (0.51 by SPe compared to 0.41 by TMa) at the chain level —the real world situation where single and multi-domain proteins are compared with each other because domains are not defined for many newly solved proteins. The corresponding absolute improvement in sensitivity is 12% from 32% to 44% with the same specificity of 99.6% for the SCOPc dataset. The effective alignment length contributes an increase in the MCC value from 0.45 by SPa to 0.51 by SPe. Thus, the effective alignment length is one of the main factors for large improvement of SP-score over TM-score for structure comparison between multi-domain proteins in SCOP fold discrimination.

We further employ reduced SCOP datasets (rSCOP and rSCOPc) to compare SPalign based on SP-score to DALI[26], CE[19] and frTMalign[27]. frTMalign improves the search algorithm to achieve higher TM-score at the expense of computational efficiency. As shown in Table 1 and Fig. 2, the MCC values by SPe are 0.60 for rSCOP and 0.50 for rSCOPc, which are close to 0.57 and 0.51, respectively, for the whole dataset. Interestingly, frTMalign[27] does not improve fold discrimination over TMb despite of its 6% higher TM-score. This is because frTMalign indistinguishably increases the TM-score of positive and negative pairs. As a result, the cutoff value for the best MCC value increases from 0.540 in TMalign to 0.555 in frTMalign. (frTMalign does not have an option as TMalign for using the average chain length for normalization). Another interesting observation is that the performance of DALI at the domain level is similar to that at the chain level while both TM-score and SP-score do not perform at the chain level as well as at the domain level. This leads to a slightly improved performance of DALI (MCC=0.505) over SPalign (MCC=0.497) at the chain level but significantly worse performance of DALI at the domain level (MCC=0.50 vs. 0.60 by SPe). The corresponding absolute improvement in sensitivity is 13% from 47% to 60% with the same specificity of 99.6% for the rSCOP dataset. As shown in Table 1, SPalign based on L_e is the only method that achieves consistently high MCC values for both chain and domain levels.

We demonstrate the difference between TMalign and SPalign using the pairwise alignment results of two chains 1z45A and 1lf6A (Fig. 3). Both chains are two domain proteins with one matching domain in the same SCOP fold (SCOP ID: b.30). TMalign gives a TM-score of 0.26, and their alignment length is 295 with RMSD of 8.7Å. frTMalign generates essentially the same alignment as TMalign despite of a slightly higher TM-score of 0.29 and 23 more aligned residues for RMSD of 8.3Å. By comparison, SPe aligned 178 residues with an RMSD of 2.9Å and the SP-score of 0.53. This mis-alignment by TMalign is not caused by the sub-optimal searching algorithm because TM-score for the alignment given by SPe is 0.25, even lower than 0.26 for the alignment by TMalign. If SPalign is employed to optimize TM-score, it will yield a TM-score of 0.31 with essentially the same overall alignment as that by TMalign or frTMalign. This pair of proteins are misaligned because the actual aligned length is significantly shorter than the sizes of both proteins. The smaller size of the two proteins is used in calculating d_0 in TM-score while d_0 in SPe is size-independent. It is noted that both DALI and CE can align these two chains. DALI aligns 175 residues with an RMSD of 3.5Å and CE aligns 184 residues with an RMSD of 3.7Å, compared to 178 residues with an RMSD of 2.9Å by SPalign.

Why does DALI perform poorly for the SCOP domain set, relative to SPe? We found that DALI often produces reasonable alignments but with low Z-scores for two domains in the same fold. For example, for SCOP domain pairs of d1c4qa (68 residues) and d2gy9q1 (87 residues), DALI aligns 52 residues with an RMSD of 2.7Å and a Z-score of 2.2. These two SCOP domains are not classified as the same fold by DALI because the Z-score is much lower than the cutoff of 5.0 optimized for the MCC value of the rSCOP set. By comparison, SPe aligns 54 residues with an RMSD of 2.4Å and a SP-score of 0.542. These two SCOP domains are classified as the same fold because of the cutoff value is 0.523 for the same dataset. This result suggests that DALI underestimates the significance of alignment for small proteins. Much lower Z-score for smaller proteins in DALI is confirmed in Fig. 1. Thus, the size dependence of DALI's Z-score leads to its relative poorer performance for the SCOP domain set. Fig. 1 also shows that CE's Z-score is surprisingly size-independent, as compared to DALI's Z-score. Yet, CE's MCC values are much lower than DALI's for both rSCOP and rSCOPc sets. This indicates that the size-dependence is not the only factor that determines the accuracy of structure alignment. It is the actual scoring function that distinguishes one method from another.

It is of interest to know how different methods perform at different size ratios of two proteins in comparison. Fig. 4 compares the maximum MCC values achieved by different methods at different size ratios for the rSCOP and rSCOPc datasets. Consistent with Fig. 2, SPalign is similar to TMalign for the domain set and similar to DALI for the chain set. In other words, SPalign has a consistent top performance across two benchmark sets. What is interesting is that the MCC values for all methods are much smaller when two protein sizes differ significantly. That is, it remains challenging for aligning two proteins in significantly different sizes.

The above results are for two-state fold discrimination. To evaluate residue-level accuracy, we calculated the fraction of residues in a protein (coverage) within different RMSD cutoffs. For the rSCOPc dataset, the coverage given by SPalign is 2% higher at 2Å and 4Å, respectively, than that by TMalign and 1% lower at 6Å. This result is expected because SP-score focuses on the contribution from 0–8Å only. Similar result is observed for the rSCOP dataset.

DNA/RNA-binding proteins

Table 2 shows the MCC values for two-state prediction of RNA/DNA binding proteins based on TM-score from TMalign and SP-score from SPalign for both DNA and RNA datasets. Unlike fold classifications, normalization based on a shorter chain length can lead to a higher or lower MCC value than based on an average chain length for TM-score and SP-score depending on the dataset. For example, MCC is 0.50 for TMb and 0.46 for TMa for the DNA set while MCC is 0.35 for TMb and 0.38 for TMa for the RNAc set. Thus, there is no clear trend for choosing shorter or average length for function prediction by TMalign. The effective length SPe removes this uncertainty because it yields the highest MCC values for DNA, DNAc, and RNAc datasets and its MCC value for the RNA set is only slightly lower than SPa's (0.41 versus 0.40). In average over four datasets, SPe improves MCC values by 12% over TMa or TMb. The largest improvement is for the RNAc dataset. The MCC of SPe (0.47) is 23% higher than that of TMa (0.38), and 34% higher than that of TMb (0.35). The corresponding sensitivity at specificity of 98% are 42%, 30%, and 33% for SPe, TMa and TMb respectively. That is, there is more than 9% improvement in sensitivity. For DNAc, the corresponding sensitivity at specificity of 98% are 46%, 39%, and 35% for SPe, TMa and TMb, respectively.

It is of interest to note that for the RNA set, the cutoff values for the highest MCC values for SPa, SPb and SPe are 0.709, 0.805 and 0.733, respectively. The corresponding cutoffs for TMa and TMb are 0.718 and 0.803, respectively. These cutoff values are essentially unchanged for the RNAc dataset (the difference is less than 0.01). Thus, higher global structural similarity is needed for binding discrimination than for fold discrimination, as found earlier[6, 8].

For all four datasets (DNA, DNAc, RNA, and RNAc), the performance of DALI is consistently poorer than that of SPalign and TMalign. The average MCC value for the four dataset is 17% lower than TMa, and 31% lower than SPe.

Fig. 5 compares the ROC curves for RNA and RNAc datasets. The true positive rate given by SPalign is consistently higher than those from TMalign and DALI at the same false positive rate. For the RNA dataset, the area under the curve (AUC) by SPalign (SPe) is 0.739 about 8% higher than the best performing TMalign (TMb) (0.682) and DALI (0.686). For the RNAc dataset, the area under the curve (AUC) by SPalign (SPe) is 0.768 about 11% higher than the best performing TMalign (TMa) (0.693) and 7% higher than DALI (0.716).

Employing DNA and RNA binding proteins for testing structure alignment allows us to compare TMalign and SPalign on the residue level. This can be done by predicting DNA/RNA binding residues according to the distances between the residues of the aligned query protein structure and DNA/RNA bases in the target template. We define a residue as a DNA/RNA-binding residue if the residue has one or more atoms with distance less than 4.5Å to any atoms in DNA/RNA.

MCC values for binding residue prediction according to structure alignment is shown in Table 3. These results are based on all DNA/RNA binding proteins, regardless if they are recognized by TM-score or SP-score. Interestingly, TM-score normalized by average length (TMa) performs consistently better than TM-score normalized by the shorter length (TMb) for predicting DNA/RNA binding residues although such consistency is not observed in predicting DNA/RNA binding proteins. SPe is consistently better than SPa except for the RNAc dataset. While SPe continues to have consistently higher MCC values than TMa in all the four datasets, the difference is smaller than predicting RNA/DNA binding proteins.

Alignment Significance

We employ the SCOP-20 dataset in all-against-all pairwise comparisons to calculate the probability of the two structures belonging to the same fold at a given score. This probability curve for SP-score based on L_e (SPe) can be fitted by the function $P = 1/(1 + \exp(-(\text{SP-score} - b)/a))$ with $b = 0.523$, $a = 0.044$ and reduced $\chi^2 = 0.0001$. In other words, p-value is 0.5 for SP-score=0.523. The higher SP-score is, the lower the p-value will be.

Computational Efficiency

Computational efficiency for alignment programs is important because time-consuming all-against-all pairwise comparison is often needed. We run pairwise calculation for the rSCOPc dataset. For removing the potential advantage of our SP-score function, we first employed SPalign to optimize TM-score. We found that the time for SPalign and for TMalign are 5870 and 5930 minutes, respectively on AMD 1800MHz Opteron CPU. Thus, the computational efficiencies of the searching algorithms for SPalign and TMalign are similar despite that we have employed a smaller fragment size as the seed and achieved 4.6% higher TM-score in average than TMalign. By comparison, the “fast” mode of our program leads to 3.5% higher

TM-score than TMalign with two times faster convergence. Although frTMalign can produce another 1.5% higher TM-score than SPalign, it is about 10 times slower than TMalign and SPalign. Using SPalign to optimize SP-score reduced the CPU time to 5100 minutes. Thus, overall efficiency of SPalign is about 14% improvement over that of TMalign. This efficiency gain is due to 8Å cutoff in calculating SP-score.

DISCUSSION

In this paper, we developed a new size-independent score called SP-score for protein structure alignment. The score was tested on SCOP structure classification and identification of DNA/RNA binding proteins. For structure classification, SP-score with an effective alignment length has a similar performance as TM-score for domain comparison but make 12% improvement over TM-score in sensitivity at constant specificity of 99.6% for multi-domain comparison. Large improvements over CE for both domain and chain-level comparisons, and over DALI for domain-level comparison are also observed. For function prediction, SP-score improves over TMalign and DALI at both domain and multi-domain levels. For example, SPalign (SPe) improves over the best performing TMalign by 4%, 9%, 5%, and 24% in MCC values for DNA, DNAc, RNA, and RNAc, respectively. Improvements on DNA/RNA binding site predictions and alignment coverage at the small RMSD cutoffs are also observed. An effective length and a fractional exponent are the key factors of the improvement in aligning multi-domain proteins from TMalign to SPalign. Only several selected methods are compared here because they are shown to be among state-of-the-art methods according to several comparative studies[30, 31, 32].

A good alignment score should be size-independent because it will allow a consistent comparison among proteins in different sizes. TM-score[24] relies on a predetermined size-dependent cutoff distance (d_0) for aligned residues to remove size dependence. The cutoff distance, however, becomes artificially small for small proteins or too large for large proteins. Moreover, using a predetermined size to determine d_0 is problematic because the size of aligned regions is unknown prior to alignment. Here, like MaxSub and LG-score, we set d_0 to a constant value. We further introduce a size-dependent prefactor to remove the size dependence. In addition, the constant values of d_0 and the distance cutoff $2d_0$ enable us to define an effective alignment size and remove contributions from meaningless alignment at long distance. The effective alignment size is employed for two proteins in comparison so that it removes the uncertainty in TM-score regarding whether or not the shorter protein length or the average protein length should be used in measuring structural similarity. However, size-independence alone is not enough to produce a good alignment method. For example, CE is less size-dependent than DALI but has a poorer performance. The actual functional form of a scoring function ultimately determines the overall performance of a structural alignment method.

The prefactor utilized for removing size dependence is $1/L^{0.7}$. A factor of $L^{0.3}$ in addition to normalization by L in LG-score has physical meaning because $L^{1/3}$ is the scaling exponent of radius of gyration for a polymer in poor solvent in the Flory theory[33]. The use of fractional exponent was inspired by its success in deriving the distance-scaled finite ideal-gas reference for generating knowledge-based energy function ($d^{0.39}$)[34], normalizing size-dependent domain-domain interactions for domain parsing ($L^{0.43}$)[35, 36], and rescaling raw alignment score for improving fold recognition ($L^{0.75}$)[28]. Recently, Gao and Skolnick also proposed to employ $0.18 - 0.35/L^{0.3}$ to remove the size dependence for interface alignment[3]. These studies highlight the usefulness of fractional exponent for normalization.

A fractional exponent was also employed to normalize RMSD in a structural diversity score that was defined as RMSD/$(L_{aligned}/L_{ave})^{1.5}$[37, 38]. Here $L_{aligned}$ and L_{ave} are the aligned length and the average protein size, respectively. It is difficult to know the size dependence of this score because it depends on the size dependence of RMSD and the residual size dependent of $L_{aligned}/L_{ave}$. One issue involving RMSD is that all atoms in the structures are weighted equally and it is very sensitive to badly aligned regions and local structural changes[20]. Moreover, the global minimum for RMSD is 0 with a short fragment of 1 to 3 residues and the diversity score will work only with elaborate restraints on alignments. These issues involving RMSD lead to introduction of LG-score[22] and the variants such as MaxSub, TM-Score and SPalign.

Our program SPalign has a “fast” mode in which the seed fragment length will be set to 1/5 of the shorter length of two proteins and between 20 and 50, rather than a constant value of 20 in the default mode. This fast-mode SPalign reduces the computational time by a factor of 2 without losing much discriminatory ability (<1% difference in MCC values) for all datasets (SCOP, RNA, RNAc, DNA, DNAc) except for SCOPc. We found that SPalign improves over SPalign-fast by 5% from 0.48 to 0.51 in SCOPc because of improved search for global minimum by starting with smaller fragments. We are currently working on improving the search for the global minimum because local minimum is the main reason behind significantly smaller MCC value for the SCOPc set (0.51) than the SCOP set (0.57). By contrast, similar performance in chain and domain levels is observed for DALI. The discrepancy between the chain and domain-level performances (MCC=0.41 vs. 0.57) is even larger for TMalign. The improvement in searching for higher TM-score is unlikely to help as illustrated by frTMalign (Table 1) because its implicit assumption that two chains are aligned in a predetermined size. Indeed, as demonstrated in Fig. 3, correctly aligned structures even have a lower TM-score than incorrectly aligned structures.

To further improve fold discrimination through structure comparison, one may have to go beyond sequential alignment. This is because many studies[2, 38, 39, 40, 41, 42, 43, 44] have found that some structures share the same overall shape but with different sequential order of secondary structure elements and such alignment has successfully applied to protein structure alignment[40, 42, 43], protein-protein[44] and protein-ligand binding interface alignment[2, 45]. Moreover, SCOP classifies some circular permutated proteins in the same fold. Work in this direction is in progress.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Nick Grishin and Bong-Hyun Kim for helpful comments. This work was supported by the National Institutes of Health grants GM R01 085003 and GM R01 067168 (Co-PI) to Y.Z and by the National Natural Science Foundation of China (grant number: 31000324) to Y.Y.

References

1. Gold ND, Jackson RM. SitesBase: a database for structure-based protein-ligand binding site comparisons. Nucleic acids research. 2006; 34:D231–234. [PubMed: 16381853]

2. Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. Proc Natl Acad Sci U S A. 2008; 105(14):5441–6. [PubMed: 18385384]

3. Gao M, Skolnick J. iAlign: a method for the structural comparison of protein-protein interfaces. Bioinformatics. 2010; 26(18):2259–2265. [PubMed: 20624782]

4. Mukherjee S, Zhang Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic acids research. 2009; 37(11)

5. Gao M, Skolnick J. Dbd-hunter: a knowledge-based method for the prediction of dna-protein interactions. Nucleic Acids Res. 2008; 36(12):3978–3992. [PubMed: 18515839]
6. Zhao H, Yang Y, Zhou Y. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected dfire-based energy function. Bioinformatics. 2010; 26(15):1857–1863. [PubMed: 20525822]
7. Draper DE. Themes in RNA-protein recognition. J Mol Biol. 1999; 293(2):255–270. [PubMed: 10550207]
8. Zhao H, Yang Y, Zhou Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucleic Acids Res. 2011; 39(8):3017–3025. [PubMed: 21183467]
9. Zhao H, Yang Y, Zhou Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. RNA Biology. 2011; 8:988–996. [PubMed: 21955494]
10. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. The EMBO journal. 1986; 5(4):823–826. [PubMed: 3709526]
11. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 2007; 35(Database issue):D291–D297. [PubMed: 17135200]
12. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008; 36(Database issue):D419–D425. [PubMed: 18000004]
13. Burley SK. An overview of structural genomics. Nat Struct Biol. 2000; 7 (Suppl):932–934. [PubMed: 11103991]
14. Koehl P. Protein structure similarities. Curr Opinion Struc Biol. 2001; 11:348–353.
15. Kolodny R, Petrey D, Honig B. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Current Opinion in Structural Biology. 2006; 16(3): 393–398. [PubMed: 16678402]
16. Hasegawa H, Holm L. Advances and pitfalls of protein structural alignment. Current opinion in structural biology. 2009; 19(3):341–348. [PubMed: 19481444]
17. Taylor W, Orengo CA. Protein structure alignment. J Molec Biol. 1989; 208:1–22. [PubMed: 2769748]
18. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. Journal of molecular biology. 1993; 233(1):123–138. [PubMed: 8377180]
19. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 1998; 11(9):739–747. [PubMed: 9796821]
20. Mizuguchi K, Go N. Seeking significance in three-dimensional protein structure comparisons. Curr Opin Struc Biol. 1995; 5(3):377–382.
21. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003; 31(13):3370–3374. [PubMed: 12824330]
22. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci U S A. 1998; 95(11):5913–5920. [PubMed: 9600892]
23. Siew N, Elofsson A, Rychlewski L, Fischer D. Maxsub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics. 2000; 16:776–785. [PubMed: 11108700]
24. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004; 57:702–710. [PubMed: 15476259]
25. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005; 33(7):2302–2309. [PubMed: 15849316]
26. Holm L, Rosenstrm P. Dali server: conservation mapping in 3D. Nucleic Acids Res. 2010; 38(Web Server issue):W545–W549. [PubMed: 20457744]
27. Pandit SB, Skolnick J. Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics. 2008; 9:531. [PubMed: 19077267]

28. Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics. 2011; 27(15):2076–2082. [PubMed: 21666270]

29. Kearsley SK. On the orthogonal transformation used for structural comparisons. Acta Crystallographica Section A. 1989; 45(2):208–210.

30. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures. Journal of Molecular Biology. 2005; 346(4):1173–1188. [PubMed: 15701525]

31. Ye Y, Godzik A. FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acid Research. 2004; 32:W582–W585.

32. Poleksic A. Algorithms for optimal protein structure alignment. Bioinformatics. 2009; 25(21): 2751–2756. [PubMed: 19734152]

33. Flory, P. Principles of polymer chemistry. Cornell University Press; 1953.

34. 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 Science. 2002; 11:2714–2726. [PubMed: 12381853]

35. Alexandrov N, Shindyalov I. PDP: protein domain parser. Bioinformatics. 2003; 19:429–430. [PubMed: 12584135]

36. Zhou H, Xue B, Zhou Y. DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile. Protein Sci. 2007; 16:947–955. [PubMed: 17456745]

37. Lu G. *TOP*: a new method for protein structure comparisons and similarity searches. J Appl Crystal. 2000; 33:176–183.

38. Lo W-C, Lyu P-C. CPSARST: an efficient circular permutation search tool applied to the detection of novel protein structural relationships. Genome Biology. 2008; 9:R11. [PubMed: 18201387]

39. Dai L, Zhou Y. Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations. J Mol Biol. 2011; 408:585–595. [PubMed: 21376059]

40. Xin Y, Christopher B. Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. Bioinformatics. 2005; 21(7):1010–1019. [PubMed: 15531601]

41. Abagyan RA, Maiorov VN. An automatic search for similar spatial arrangements of alpha-helices and beta-strands in globular proteins. J Biomol Struct Dyn. 1989; 6:1045–1060. [PubMed: 2818856]

42. Dundas J, Binkowski TA, DasGupta B, Liang J. Topology independent protein structural alignment. BMC bioinformatics. 2007; 8:388. [PubMed: 17937816]

43. Guerler A, Knapp E-WW. Novel protein folds and their nonsequential structural analogs Protein science: a publication of the Protein. Society. 2008; 17(8):1374–1382.

44. Gao M, Skolnick J. Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. Proc Natl Acad Sci USA. 2010; 107(52):22517–22522. [PubMed: 21149688]

45. Dundas J, Adamian L, Liang J. Structural signatures of enzyme binding pockets from order-independent surface alignment: A study of metalloendopeptidase and NAD binding proteins. Journal of Molecular Biology. 2011; 406(5):713–729. [PubMed: 21145898]

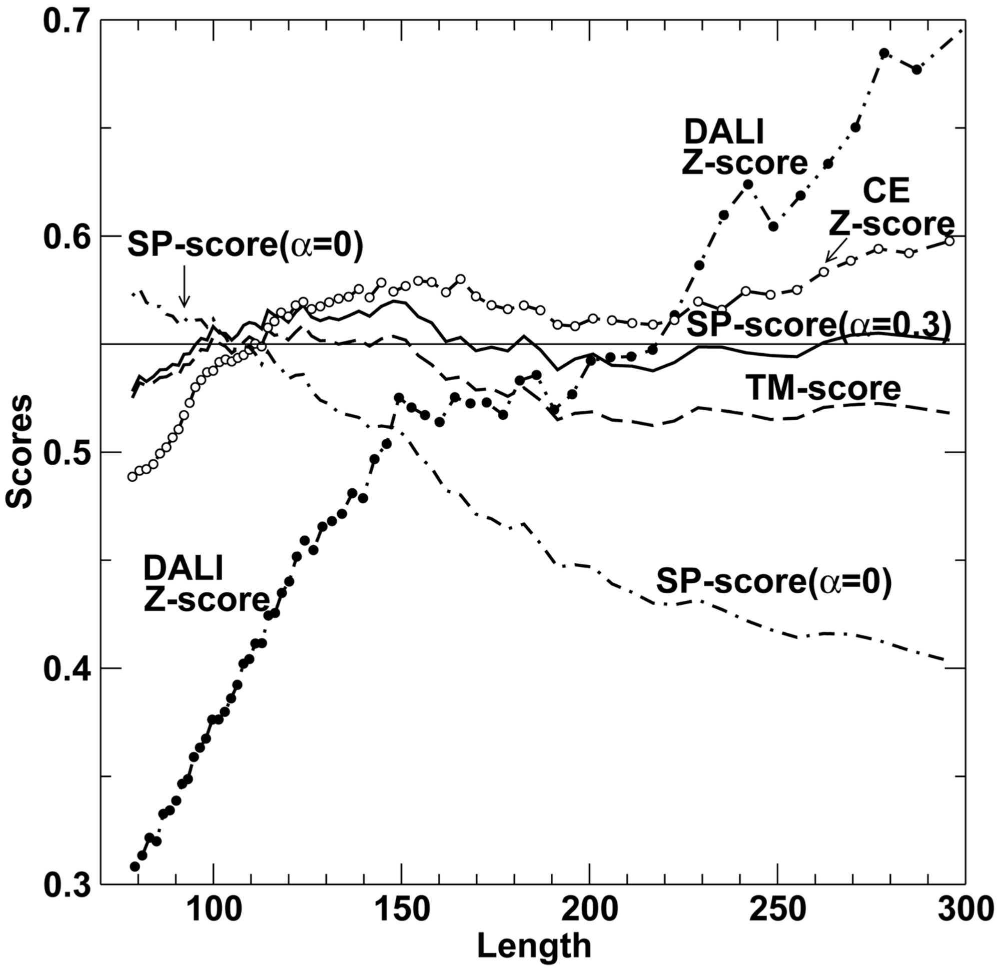

Figure 1.

The average alignment score for protein pairs in the same SCOP fold whose length difference is smaller than 10%*L* for the dataset of rSCOP. To ensure statistics, each bin contains 100 proteins. For clarity, SP-scores with the shorter length for normalization (SPb) at two different α values are shown and compared to TM-score (TMb), DALI Z-score and CE Score. SP-scores at $\alpha = 0$, DALI Z-scores, and CE Z-scores are multiplied by 4, divided by 15, and divided by 7, respectively, to bring these values in similar scale to other scoring functions to facilitate comparison.

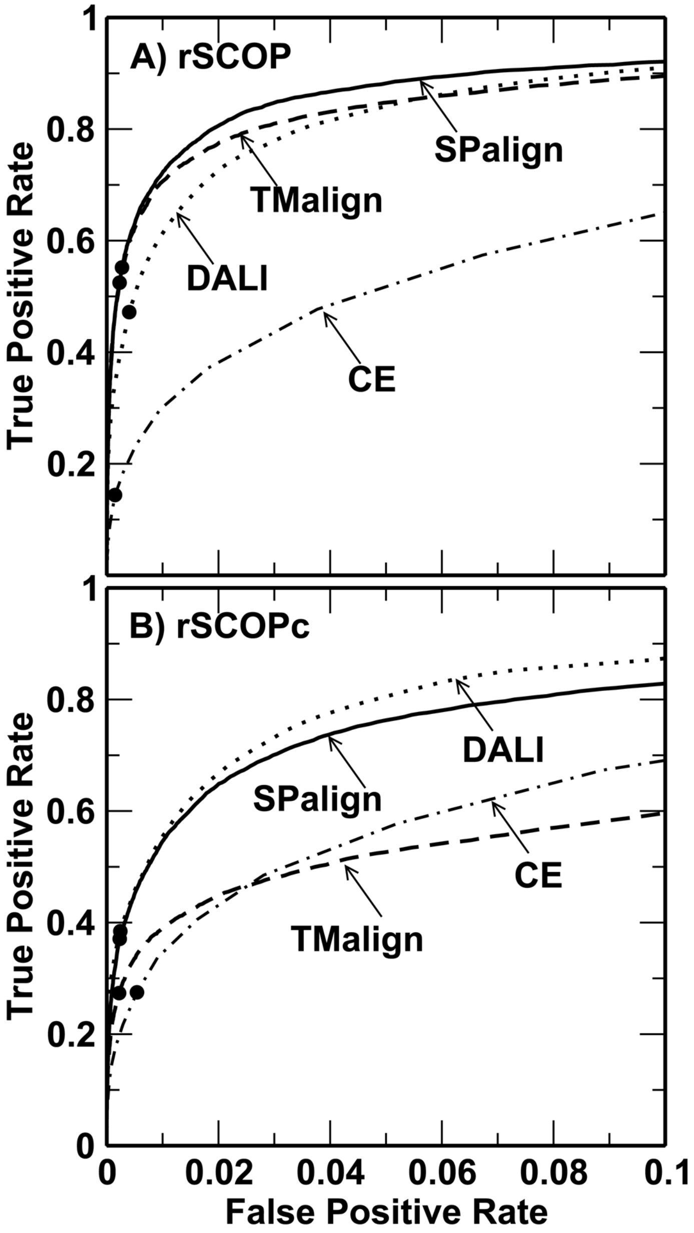

Figure 2.

The receiver operating characteristic (ROC) curve (true positive rate versus false positive rate) for the reduced SCOP (A) and SCOPc (B) sets by four methods (SPalign, TMalign, DALI, and CE) as labeled. SPalign and TMalign are based on the effective alignment and the average protein length, respectively. A filled circle indicates the location for the optimized MCC value for a given method. For clarity, we did not show the results from frTMalign whose performance is similar to TMalign based on the shorter chain length but is poorer than TMalign based on the average length.

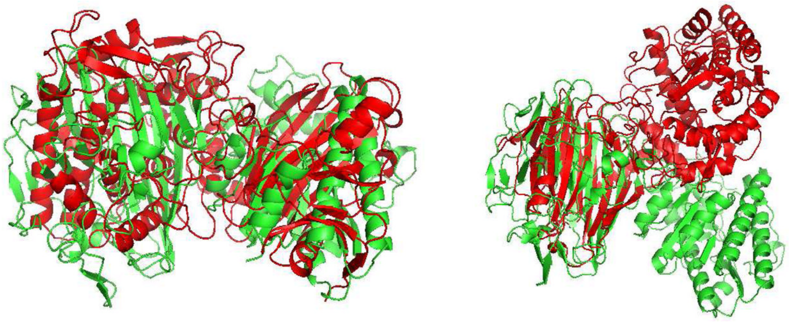

Figure 3.

Aligned structures of chain 1z45A and 1lf6A by TMalign (Left) and SPalign (Right), respectively. Unlike SPalign, TMalign is unable to detect the similarity of single domain between two two-domain proteins. This failure is not caused by sub-optimal search algorithm.

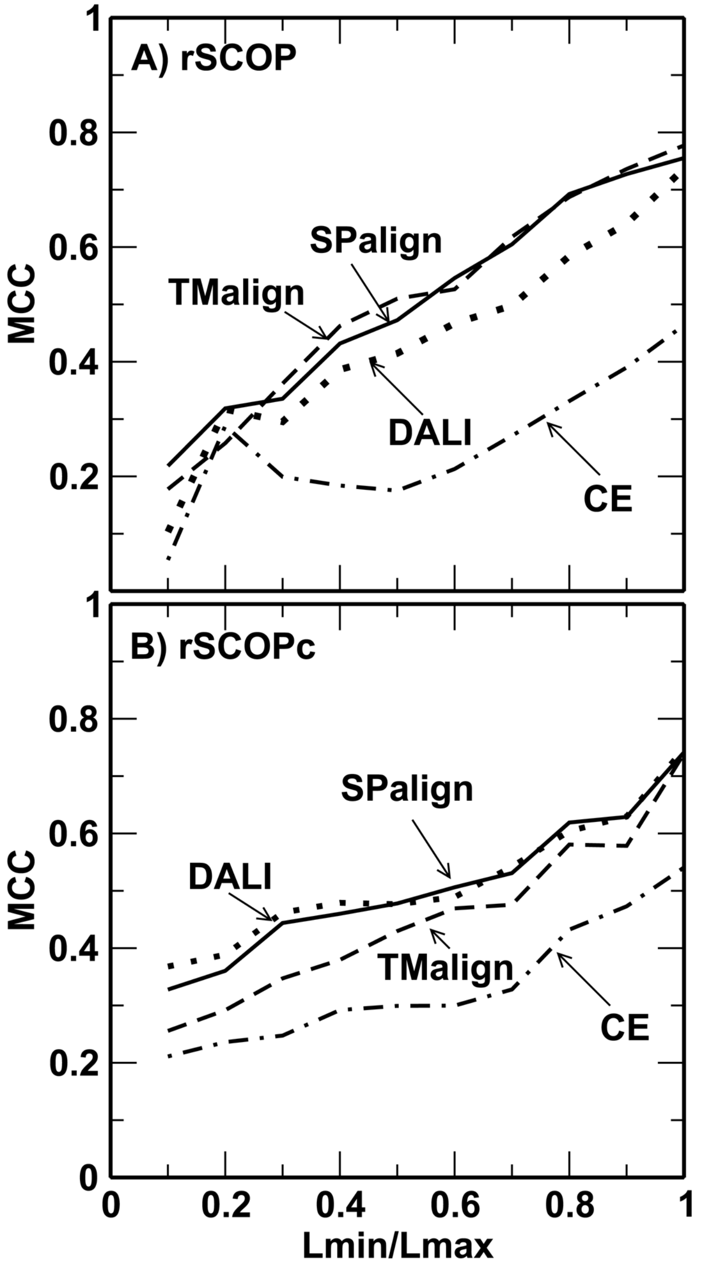

Figure 4.
The maximum MCC values at different protein size ratios (L_{min}/L_{max}) for the reduced SCOP (A) and SCOPc (B) sets by four methods (SPalign, TMalign, DALI, and CE) as labeled. Ten bins for L_{min}/L_{max} are [0.1,0.2), [0.2,0.3), …,[0.9,1) and 1.

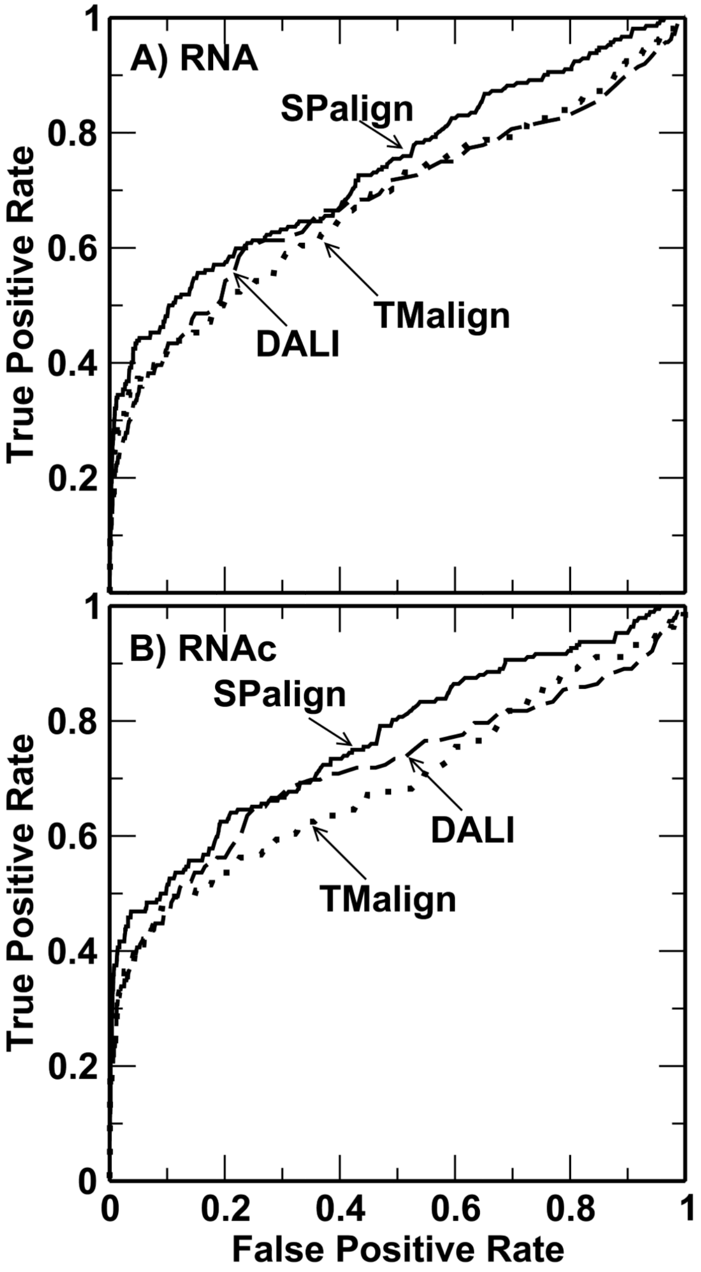

Figure 5.

ROC curves for the RNA (A) and RNAc (B) datasets given by DALI, TMalign based on TM-score and SPalign based on SP-score as labeled. The best performing TM-align based on the shorter chain length (RNA) or the average length (RNAc) is compared to SPalign based on the effective alignment length.

Table 1

Mathews correlation coefficients given by several methods for four SCOP datasets.

			TMalign[a]			SPalign[b]		
Set[c]	**CE**	**DALI**	**fr(L_b)**	L_b	L_a	L_b	L_a	L_e
SCOP				0.48	0.57	0.48	**0.58**	0.57
SCOPc				0.40	0.41	0.41	0.45	**0.51**
rSCOP	0.26	0.50	0.53	0.53	0.60	0.52	**0.61**	0.60
rSCOPc	0.33	**0.51**	0.40	0.41	0.41	0.42	0.45	0.50

[a] TM-score by a deeper search (frTMalign), normalized using shorter length (L_b, TMalign with option “-b”), and average length (L_a, TMalign with option “-a”).

[b] SP-score using shorter, average and effective lengths, respectively.

[c] rSCOP and rSCOPc are smaller, reduced sets for SCOP and SCOPc, respectively. SCOP and SCOPc are domain and chain-level datasets, respectively.

Table 2

Mathews correlation coefficients given by several methods for DNA and RNA datasets.

		TMalign[a]		**SPalign**[b]		
Set	**DALI**	L_b	L_a	L_b	L_a	L_e
DNA	0.39	0.50	0.46	0.50	0.48	**0.52**
DNAc	0.39	0.45	0.45	0.45	0.47	**0.49**
RNA	0.28	0.38	0.36	0.40	**0.41**	0.40
RNAc	0.36	0.35	0.38	0.42	0.45	**0.47**
Ave[c]	0.36	0.41	0.42	0.44	0.45	**0.47**

[a] TM-score normalized using shorter length (L_b, TMalign with option “-b”) and average length (L_a, TMalign with option “-a”).

[b] SP-score using shorter, average and effective lengths, respectively.

[c] The average of four datasets (DNA/DNAc and RNA/RNAc)

Table 3

The median MCC values on predicted DNA/RNA binding residues.

Dataset	TMb	TMa	SPb	SPa	SPe
DNA	0.58	0.62	0.61	0.63	**0.64**
DNAc	0.58	0.61	0.57	0.62	**0.63**
RNA	0.31	0.35	0.29	0.34	**0.37**
RNAc	0.31	0.32	0.33	**0.38**	0.36
Ave	0.45	0.48	0.45	0.49	**0.50**