SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PROTEIN ANALYSIS
SOFTWARE
STRUCTURAL BIOLOGY
BIOINFORMATICS

Received
15 November 2012

Accepted
25 February 2013

Published
14 March 2013

Correspondence and requests for materials should be addressed to J.X. (jinboxu@gmail.com)

Protein structure alignment beyond spatial proximity

Sheng Wang, Jianzhu Ma, Jian Peng & Jinbo Xu

Toyota Technological Institute at Chicago, USA 60637.

Protein structure alignment is a fundamental problem in computational structure biology. Many programs have been developed for automatic protein structure alignment, but most of them align two protein structures purely based upon geometric similarity without considering evolutionary and functional relationship. As such, these programs may generate structure alignments which are not very biologically meaningful from the evolutionary perspective. This paper presents a novel method DeepAlign for automatic pairwise protein structure alignment. DeepAlign aligns two protein structures using not only spatial proximity of equivalent residues (after rigid-body superposition), but also evolutionary relationship and hydrogen-bonding similarity. Experimental results show that DeepAlign can generate structure alignments much more consistent with manually-curated alignments than other automatic tools especially when proteins under consideration are remote homologs. These results imply that in addition to geometric similarity, evolutionary information and hydrogen-bonding similarity are essential to aligning two protein structures.

Protein structure alignment is a fundamental problem in computational structure biology and has been widely applied to protein sequence, structure and functional study[1]. In the past two decades many computer programs have been developed for automatic pairwise structure alignment[2–10] and multiple structure alignment[11–17]. However, the alignment accuracy of these programs is still low when judged by manually-curated structure alignments, especially on distantly but functionally related proteins[18]. Further, these programs may generate alignments which are not very biologically meaningful from the evolutionary perspective.

A protein structure alignment method consists of two major components: a scoring function measuring protein similarity and a search algorithm optimizing the scoring function. It is very challenging to design a scoring function to exactly capture all the (implicit and explicit) rules used by human experts, who align and classify protein structures using not only geometric similarity, but also evolutionary and functional information. A variety of popular structure alignment tools such as DALI[2], CE[4] and TMalign[19] use only 3D geometric similarity and may produce bad alignments even when proteins under consideration are close homologs[20].

It is observed that despite proteins in a family share a similar overall shape, their structures exhibit very high local flexibility due to evolutionary events (i.e., mutation, insertion and deletion) at the sequence or local substructure level. This kind of local conformation change due to evolutionary events cannot be accurately quantified by spatial proximity of aligned residues (after rigid-body superposition). Instead, evolutionary distance shall be a better measure. Inspired by this observation, DeepAlign uses amino acid and local substructure substitution matrices, which are derived from evolutionarily-related protein pairs, to align protein local structures.

Amino acid mutation information is useful when proteins under consideration are close homologs. Very few structure alignment programs such as Formatt[21] make use of amino acid substitution matrices in aligning protein structures. Our work differs from Formatt in that the latter uses only the BLOSUM[22] mutation matrix while we also use local substructure mutation matrices as well as hydrogen-bonding similarity to measure the similarity of two proteins. Note that our work is very different from those protein alignment methods such as PROMAL3D[23] and 3Dcoffee[24], which mainly focus on protein sequence alignment leveraging structural information. By contrast, we focus on protein structure alignment using sequence and evolutionary information.

BLOSUM is derived from close homologs, so it is not very sensitive for remote homologs. To deal with this, DeepAlign uses a substructure mutation matrix to measure the evolutionary distance of two proteins at the local substructure level. That is, we parse each protein backbone conformation to a sequence of local substructure alphabets and then measure the evolutionary distance of two proteins based upon the substructure mutation potential. In particular, we use the local conformation letter substitution matrix CLESUM described in[25,26] for such a purpose. CLESUM is derived from a set of non-redundant alignments of evolutionarily-related proteins.

CLESUM is better than spatial proximity in that it favors the alignment of conserved loop regions especially those appearing at the two ends of a regular secondary structure segment and that it also disfavors the alignment of two evolutionarily-unrelated alpha helices. In addition to evolutionary distance, we also make use of hydrogen-bonding similarity to improve alignments for hydrogen bonds, which greatly helps generate biologically more meaningful alignments especially for beta-strands.

It is challenging to optimize a scoring function to find the optimal structure alignment. Most search algorithms can only solve scoring functions to suboptimal and thus, generate a suboptimal alignment, which may impact alignment accuracy[1]. Along with the enlargement of the Protein Data Bank (PDB), it is now time-consuming to retrieve all the remotely-related protein structures in the PDB for a query protein structure. Therefore, a structure alignment tool shall also be computationally efficient. Different from other structure alignment methods, DeepAlign reduces its running time by carefully identifying a set of initial alignments. In particular, DeepAlign identifies evolutionarily-related structural fragment pairs using both amino acid and local substructure mutation scores and then build initial alignments based upon these fragment pairs. Starting from the initial alignments, DeepAlign iteratively optimizes the scoring function and finally produces high-quality alignments. By using the evolutionary information, DeepAlign can accurately identify a very small set of evolutionarily-related structural fragment pairs which are very likely contained in the optimal alignment. This greatly helps reduce the running time of DeepAlign since it avoids starting from many initial alignments that cannot lead to the optimal alignments.

We evaluate DeepAlign using several metrics including reference-independent and dependent alignment accuracy, evolutionary scores and superfamily/fold discrimination, based upon three manually-curated benchmarks and also SABmark[27]. Experimental results indicate that DeepAlign outperforms several popular tools on all the datasets when judged by human-curated alignments. DeepAlign works particularly well when proteins under consideration are not very close. The DeepAlign alignments also have better mutation scores and thus, are more biologically meaningful from the evolutionary perspective. Even evaluated by the pure geometric measures such as TMscore[28] and uGDT (un-normalized Global Distance Test[29]), DeepAlign still compares favorably to other popular tools such as DALI[2], MATT[14] and TMalign[19], all of which align protein structures based upon only geometric similarity.

Results

We evaluate our program DeepAlign using three manually-curated alignment databases and a few metrics: LALI (length of alignment), RefAcc (reference-dependent alignment accuracy), RMSD (root mean squared deviation), TMscore and mutation scores (i.e., BLOSUM and CLESUM). RefAcc is calculated as the percentage of correctly-aligned positions as judged by the gold standard (i.e., manual alignments), measuring consistency between automatic alignments and human-curated alignments. RMSD and TMscore, both of which are well-established, measure geometric similarity of protein structures, but TMscore is better than RMSD because TMscore is length-independent and not biased by few badly-aligned residue pairs. The evolutionary scores measure if one alignment is favorable or not from the evolutionary perspective. We calculate the evolutionary scores of a structure alignment using both the BLOSUM and CLESUM substitution matrices. Ideally, a good alignment shall have preferable performance regardless of the metrics.

The programs to be compared. We compare DeepAlign with several popular structure alignment tools such as DALI[2], TMalign[19], MATT[14], Formatt[21], which represents four very different methods. DALI is a distance matrix based approach, aligning two structures by matching their distance matrices. TMalign aligns two protein structures by maximizing the TMscore, using a similar search algorithm as STRUCTAL[28,30]. MATT aligns protein structures by concatenating the alignments of some short structural fragments. MATT also calculates a p-value to indicate the degree to which two proteins are structurally similar. Formatt is an extension of MATT by taking into consideration primary sequence similarity in aligning two structures.

The benchmarks. We use three manually-curated benchmarks: (i) A subset of CDD (Conserved Domain Database)[31] used in[20]; (ii) MALIDUP[32]; and (iii) MALISAM[33]. The CDD set contains 3591 manually-curated pairwise structure alignments. The human-curated alignments for CDD contain only the alignments of core residues. The CDD set has already been used to evaluate a bunch of pairwise structure alignment algorithms[34], including CE[4], FAST[8], LOCK2[35], MATRAS[36], VAST[10] and SHEBA[9]. MALIDUP has 241 manually-curated pairwise structure alignments for homologous domains originated from internal duplication within the same polypeptide chain. About half of the pairs in MALIDUP are remote homologs. MALISAM contains 130 protein pairs and the two proteins in any pair are structural analogs with different SCOP[37] folds. There is strong evidence indicating that proteins in a MALIDUP pair are not homologs[38]. Therefore, MALIDUP are the most challenging benchmark among these three. The alignments in these three databases are manually-curated, taking into consideration not only geometric similarity, but also evolutionary and functional relationship. Therefore, the manually-curated alignments make more biological sense and it is reasonable to use them as reference to judge automatically-generated alignments.

Performance on CDD. DeepAlign obtains the highest reference-dependent alignment accuracy of 93.8% among the five automatic structure alignment methods (in Table 1). DeepAlign also outperforms the methods evaluated in[34] in terms of ref-dependent alignment accuracy. That is, DeepAlign is more consistent with human experts than the other programs. In terms of TMscore and RMSD, the TMalign alignments are slightly better than the DeepAlign alignments, but the former are less consistent with manual alignments than the latter. This implies that the geometric similarity score used by TMalign (i.e., TMscore) does not accurately reflect the alignment criteria used by human experts. The DeepAlign alignments also have much better evolutionary scores than the other three programs no matter how the mutation scores are calculated. As a control, we also calculate the evolutionary scores of the manual alignments. The manual alignments have much lower mutation score per alignment because only core residues are aligned. However, the manual alignments have the best average mutation scores per aligned position. Note that the manual alignments are not explicitly driven by a specific mutation score. This confirms that human experts indeed take into consideration evolutionary relationship in aligning two protein structures and that TMalign may align many more evolutionarily-unrelated residues together than DeepAlign. Formatt has a similar mutation (i.e., BLOSUM/CLESUM) score per alignment as DeepAlign, but Formatt has a better average mutation score per aligned position than DeepAlign because Formatt has a smaller LALI. In terms of reference-independent or dependent alignment accuracy, DeepAlign is much better than Formatt partially because the latter has a much smaller LALI.

Performance on MALIDUP. DeepAlign obtains a reference-dependent alignment accuracy of 92%, greatly exceeding the other three tools (in Table 1). DeepAlign is 6% better than the second best algorithm DALI. Although the TMalign alignments have a longer alignment length and the MATT alignments have a smaller RMSD, both TMalign and MATT have much lower reference-dependent alignment accuracy. This again implies that the TMalign and MATT scoring functions greatly deviates from what are implicitly

Table 1 | Performance of five pairwise structure alignment tools on three benchmarks CDD, MALIDUP and MALISAM. See text for the explanation of LALI, RMSD, TMscore and RefAcc. "Blosum1 (Clesum1)" is the average mutation score per aligned position while "Blosum2 (Clesum2)" is the average mutation score per alignment. As a control, the performance of manually-curated alignments is also shown in the table

Method	LALI	RMSD	TMscore	RefAcc	Blosum1	Clesum1	Blosum2	Clesum2
CCD (3591)								
DeepAlign	134.8	2.86	0.667	93.8	0.261	1.782	43.45	243.71
DALI	130.8	2.75	0.663	92.8	0.165	1.684	28.78	225.15
MATT	128.6	2.53	0.655	91.4	0.152	1.728	30.19	229.59
Formatt	112.3	2.32	0.566	86.4	0.343	1.983	44.11	235.64
TMalign	138.4	2.84	0.686	85.6	0.047	1.531	15.25	211.88
Manual	62.6	1.66	0.345	100.0	0.677	2.499	43.89	157.67
MALIDUP (241)								
DeepAlign	85.5	2.61	0.622	92.0	0.314	1.872	29.31	158.28
DALI	83.5	2.65	0.600	86.4	0.172	1.700	18.63	147.53
MATT	82.3	2.47	0.608	79.8	0.178	1.824	18.84	150.00
Formatt	70.6	2.19	0.542	86.2	0.344	2.196	28.62	154.66
TMalign	87.0	2.62	0.631	81.0	0.110	1.600	12.50	137.64
Manual	77.9	2.49	0.587	100.0	0.294	1.853	27.67	154.81
MALISAM (130)								
DeepAlign	61.3	2.96	0.521	77.5	−0.601	1.108	−36.48	67.66
DALI	61.0	3.11	0.515	67.7	−0.595	0.925	−35.52	56.28
MATT	56.2	2.74	0.486	51.7	−0.625	1.013	−34.05	56.98
Formatt	44.9	2.42	0.411	56.3	−0.486	1.489	−21.1	65.69
TMalign	61.1	3.06	0.517	53.7	−0.684	0.739	−40.04	45.65
Manual	56.7	2.92	0.488	100.0	−0.556	1.240	−31.58	70.75

used by human experts. In terms of TMscore, DeepAlign is only slightly second to TMalign, but better than the others. However, the DeepAlign alignments have much better evolutionary scores, only second to the Formatt alignments in terms of the average mutation score per aligned position. Since the TMalign alignments on average are longer, this again confirms that TMalign may align many more evolutionarily unfavorable residues than DeepAlign. Formatt performs similarly on this dataset as on CDD.

Performance on MALISAM. DeepAlign obtains the highest ref-dependent alignment accuracy of 77.5% among all the five computer programs (in Table 1). DeepAlign is 10% better than the second best algorithm DALI. MALISAM is much more challenging than CDD and MALIDUP. In MALISAM, 80 pairs (i.e., 61.5% of the total) contain proteins with different SCOP folds[33]. The DALI and TMalign alignments have similar average alignment lengths as the DeepAlign alignments, but slightly higher RMSD. Furthermore, the DALI, MATT and TMalign alignments deviate significantly from the manual alignments. In terms of the BLOSUM scores, the difference between the DeepAlign alignments and others is not very significant. This is not unexpected because the proteins in this dataset are only weakly similar at sequence level and BLOSUM is not sensitive enough. However, the DeepAlign alignments have much better CLESUM score per alignment than the others, only slightly second to the manual alignments. That is, the DeepAlign alignments are more evolutionarily favorable than others at the local substructure level.

We also evaluate DeepAlign, TMalign, DALI, MATT and Formatt using another geometric similarity measure uGDT (unnormalized Global Distance Test[29]), an official metric used by CASP (Critical Assessment of Structure Prediction[39]) to evaluate the quality of a protein model and observe the same trend as TMscore. In addition, DeepAlign opens much fewer gaps in an alignment than the other three programs because DeepAlign uses evolutionary information to generate alignments. See Supporting Information for more detailed explanation.

In summary, our method DeepAlign performs significantly better than other popular tools when judged by manually-curated alignments and from the evolutionary perspective. Even if only geometric similarity measures such as TMscore and uGDT are considered, DeepAlign still compares favorably to other tools, only slightly second to TMalign, which aligns two protein structures by optimizing only TMscore. The manual alignments tend to have better mutation scores but much lower geometric similarity score (e.g., TMscore) partially because human experts tend to align only those evolutionarily conserved residues instead of spatially close residues (after rigid-body transformation). By contrast, DeepAlign achieves a good balance between geometric similarity and evolutionary conservation.

Discrimination of distant homologs and structural analogs. We use SABmark to test the performance of DeepAlign in identifying distant homologs and structural analogs[27]. SABmark-sup is the superfamily set in SABmark (version 1.65), containing 425 protein groups with low to intermediate sequence identity. SABmark-twi is the twilight set in SABmark, containing 209 groups with low sequence identity. Each SABmark-sup (-twi) group contains at most 25 structures sharing a SCOP superfamily (fold). It is believed that if two proteins are in the same SCOP superfamily, it is likely these two proteins are remote homologs. If two proteins share only the same SCOP fold, it is very likely that they are structural analogs instead of remote homologs. Given a protein structure, we align it to all the proteins in the benchmarks and then rank all the alignments by certain criteria. We examine if the top-ranked protein structures are in the same group as the query protein or not. DeepAlign uses its scoring function to rank the proteins. Similarly, TMalign, MATT and DALI use TMscore, P-value and Z-score to rank the alignments, respectively. The ranking results are evaluated by ROC (receiver operator curve) and AUC (area under curve). Formatt has a very similar result as MATT.

As shown in Figure 1(A), tested on SABmark-sup, DeepAlign has the best ROC curve, especially at the high specificity area. For

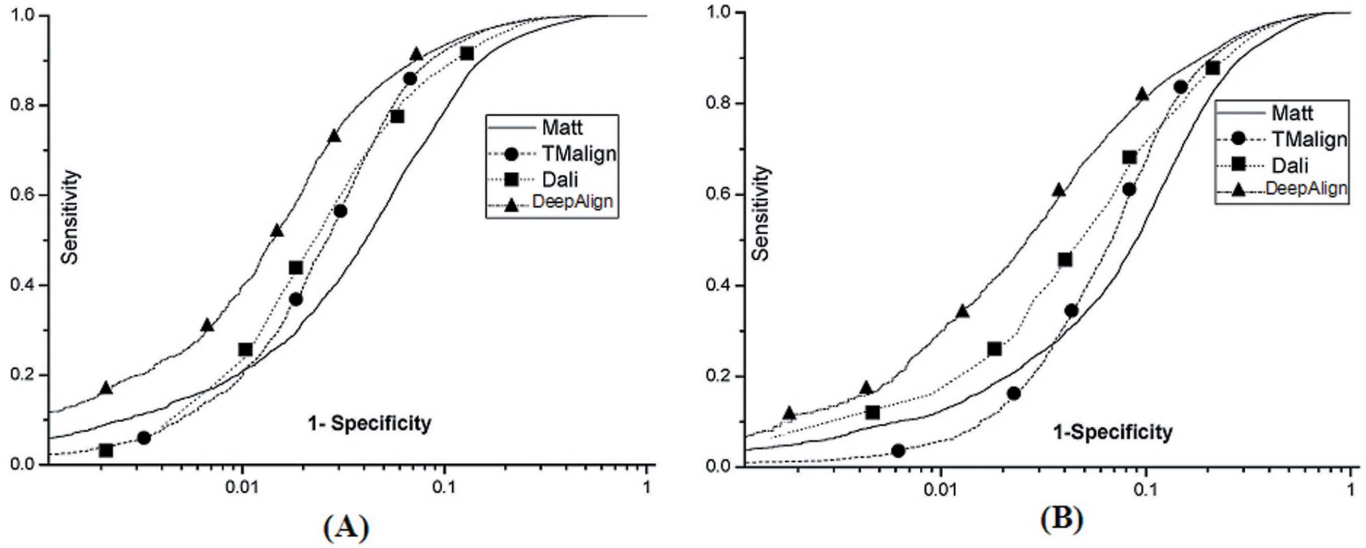

Figure 1 | The ROC curves by DeepAlign, DALI, MATT and TMalign on (A) SABmark-sup; and (B) SABmark-twi. The ROC curves for Formatt are not in the figure since they are very similar to those for MATT.

example, at the specificity level 0.99, DeepAlign has sensitivity around 0.4 while the other three have sensitivity only around 0.2. As shown in Table 2, DeepAlign, DALI, MATT and TMalign have AUCs 0.970, 0.956, 0.933 and 0.960, respectively. We also observe the same trend on the SABmark-twi set. SABmark-twi is more challenging because each group in this set consists of proteins similar at only SCOP fold level. However, DeepAlign outperforms others by an even larger margin. As shown in Figure 1(B), DeepAlign has sensitivity 0.6 at specificity 0.96 while the second best algorithm DALI has sensitivity only 0.4 at the same specificity level. MATT and TMalign have only sensitivity around 0.2 at this specificity level. As shown in Table 2, DeepAlign, DALI, MATT and TMalign have AUCs 0.933, 0.908, 0.873 and 0.903 on SABmark-twi, respectively. These results imply that DeepAlign scoring function is better than DALI's Z-score, MATT's P-value and TMscore in detecting the superfamily relationship of proteins.

Relative importance of the four score items. Our scoring function consists of four items: BLOSUM, CLESUM, the hydrogen-bonding score and TMscore. In order to estimate the impact of each item on structure alignment, we exclude each of them from the DeepAlign scoring function and regenerate the alignments for the protein groups in the three manually-curated benchmarks using the new scoring functions. Table 3 illustrates the ref-dependent alignment accuracy (RefAcc) that can be obtained by the four new scoring functions.

As shown in Table 3, when BLOSUM is excluded, the RefAcc on relatively easy benchmarks (i.e., CDD and MALIDUP) is decreased while that on the more challenging benchmark MALISAM is improved. This implies that evolutionary information at the amino acid level is useful for the alignment of homologous proteins, but may slightly hurt that of structurally analogous proteins. When CLESUM is excluded, the performance on CDD is slightly improved while on the two more challenging benchmarks MALIDUP and MALISAM dramatically decreased. That is, CLESUM is particularly useful for the alignment of structurally analogous proteins and remote homologs. Table 3 also shows that the TMscore and the hydrogen bonding score are useful across all similarity levels.

The hydrogen-bonding score has large impact on beta-containing proteins. For example, tested on the 94 beta-containing protein pairs in MAILSAM each of which contains a large portion of beta-strands, the overall accuracy is only 73.7% when the hydrogen-bonding score is not used. This is significantly worse than what can be obtained (79.3%) when the hydrogen-bonding score is used. Further, among the 334 CDD beta-containing protein pairs, using the hydrogen-bonding score can improve the alignment for 212 protein pairs by at least 5%.

In summary, the 3D geometric similarity (e.g., TMscore) is the major factor determining if two proteins are similar or not. However, in order to generate biologically meaningful structure alignments, other factors are also indispensable. The local substructure similarity is especially important for distantly-related proteins while the hydrogen-bonding score is important for beta-containing proteins.

Specific examples. *Case study 1: d1h99a1 and d1h99a2.* Both domains are in the same SCOP family (a.142.1.1) and they are mutants. Table S2 shows the evaluation of the structural alignments of these two domains generated by DeepAlign, DALI, MATT, Formatt and TMalign. DeepAlign, MATT and Formatt generate almost the same alignment as the human-curated and the alignments also have very good mutation scores. Although the DALI and TMalign alignments have slightly higher TMscore, but they have much worse mutation scores, which indicate that these alignments contain some

Table 2 | The AUCs obtained by DeepAlign, DALI, MATT/Formatt and TMalign on SABmark-sup and SABmark-twi

Method	SABmark-sup	SABmark-twi
DeepAlign	0.970	0.933
DALI	0.956	0.908
Formatt	0.934	0.874
MATT	0.933	0.873
TMalign	0.960	0.903

Table 3 | Impact of the four score items on alignment. The numbers in the table are the reference-dependent alignment accuracy (RefAcc). ''All'' indicates all score items are used. ''-'' indicates one score item is excluded

Dataset	All	-BLOSUM	-CLESUM	-Hydrogen	-TMscore
CDD	93.8	92.1	94.2	92.2	75.7
MALIDUP	92.0	90.7	87.3	89.5	78.3
MALISAM	77.5	78.5	65.4	74.4	48.1

evolutionarily unfavorable residue pairs. DALI misaligns one helix, as shown in Figure S1. TMalign misaligns one helix either. Figure S1 displays the DeepAlign and DALI alignments, but not the TMalign alignment due to space limit.

Case study 2: d1nekc_ and d1nekd_. These two domains are taken from the CDD database (ID: cd03493). Table S3 shows the evaluation of the structural alignments generated by different alignment tools for these two domains. The DeepAlign and DALI alignments are highly consistent with the human-curated alignment. MATT (as well as Formatt) mistakenly aligns the 1st and 2nd helices and the linker loop region, as shown in Figure S2. The DeepAlign and DALI alignments have very good TMscore, only slightly second the TMalign alignment which is generated by optimizing only TMscore. The DeepAlign alignment also has much better mutation scores than the MATT alignment, which even has a negative BLOSUM score. Overall, the MATT alignment has bad values in three performance metrics including TMscore, RefAcc and mutation scores.

Case study 3: d1ef5a_ and d1ndda_ in the CDD database (ID: cd00196). DALI, MATT, Formatt and TMalign misalign the loop region between the 1st alpha helix and the 3rd beta strand. MATT and TMalign also mistakenly align the 3rd beta strand (see Figure S3). Table S4 shows the evaluation of the structural alignments for these two domains. The DeepAlign alignment has the best ref-dependent alignment accuracy (RefAcc), better than DALI and much better than MATT and TMalign. In terms of TMscore, DeepAlign is slightly worse than DALI and TMalign, but much better than MATT. In terms of the CLESUM scores, DeepAlign significantly outperforms the other three programs. According to the BLOSUM scores, DeepAlign is not very different from the others. This is because that the two domains are not close homologs and BLOSUM is not suitable to measure their evolutionary relationship. Since BLOSUM does not work for this case, we calculate the sequence profile score at each position for the sub-alignments between the 38-th and the 53rd residues of d1ef5a_, which corresponds to the region between the 1st alpha helix and the 3rd beta strand (inclusive), as shown in Table S5. The protein sequence profile and the corresponding profile score is generated using the HHpred[40] package. The profile score usually ranges between -1.0 and 1.0. The higher the profile score, the better. Note that profile score is not used in the DeepAlign scoring function, but the DeepAlign alignment has positive profile scores at almost all the aligned positions. By contrast, the other three alignments have negative profile scores at some positions. This further confirms that DeepAlign aligns evolutionarily-related residues together even if the proteins are not close homologs.

Running time analysis. We measure the running time on an Ubuntu Linux PC with 2 GB RAM and Intel®Core™2 Quad CPU T5600 @ 1.83 GHz. The performance and running time of DeepAlign depend on three parameters *TopK*, *TopJ* and *M*. DeepAlign is run using default parameters (i.e., $TopK = 100$, $TopJ = 20$, $M = 10$), which are also used to generate the alignments discussed in section RESULTS. Tested on the 23074 protein pairs in SABmark-twi, DeepAlign, TMalign, MATT, Formatt and DALI have running times of 1878, 1073, 29192, 35138 and 54297 seconds, respectively. That is, DeepAlign is much faster than MATT, Formatt and DALI, but (not much) slower than TMalign. DeepAlign is slower than TMalign partially because that DeepAlign uses a scoring function of four items while TMalign uses only one of them (i.e., TMscore). That is, it takes a much longer time for DeepAlign to calculate its scoring function than TMalign.

Discussion

This paper has presented a novel method DeepAlign for automatic protein structure alignment, which can generate alignments highly consistent with manually-curated alignments. Manually-curated alignments usually make much more biological sense since they are built by human experts taking into consideration evolutionary and functional relationship, in addition to geometric similarity. The novelty of DeepAlign lies in its scoring function, which considering not only 3D geometric similarity, but also evolutionary information at the sequence and local substructure levels as well as hydrogen-bonding similarity. Note that the DeepAlign scoring function is the natural combination of four different items and there are no parameters to be fine-tuned. Therefore, we do not bias DeepAlign towards a specific performance metric.

We have tested DeepAlign with four widely-used structure alignment tools TMalign, MATT, Formatt and DALI on three manually-curated benchmarks. These benchmarks contain many distantly-related protein structures (which are remote homologs or structural analogs) and are very challenging for any automatic alignment tools. Many proteins in the most challenging benchmark MALISAM are not even in the same SCOP fold although they share structurally analogous motifs. Nevertheless, DeepAlign can still align two protein structures consistently with human experts. DeepAlign also tends to align many more evolutionarily-related residues together than the other tools and open much fewer gaps in an alignment. In addition, DeepAlign compares favorably to other tools when evaluated by purely geometric similarity measures such as TMscore and uGDT. DeepAlign is also better than the other tools in discriminating two proteins similar at the superfamily or fold level.

Currently DeepAlign uses only the local conformation and the amino acid substitution matrices to measure equivalence of two residues. We can also use other information such as sequence profile, represented by a position-specific scoring matrix[41] or Hidden Markov Model[42], to measure evolutionary distance, which shall be more sensitive than BLOSUM62. DeepAlign uses dynamic programming to generate a complete alignment and thus, cannot generate sequence order-independent alignments[43]. We will extend our algorithm to deal with non-sequential alignments. We will also extend DeepAlign to some other applications such as binding-site recognition[44] and protein-interface alignment[45].

Methods

The scoring function. The scoring function is used to determine how likely two residues, of the two proteins under consideration, shall be aligned. Our scoring function is composed of amino acid mutation score, local substructure substitution potential, hydrogen-bonding similarity and geometric similarity[46]. In particular, the equivalence of two residues i and j is estimated by the following scoring function.

$$Score(i,j) = (\max(0, BLOSUM(i,j)) + CLESUM(i,j)) \times v(i,j) \times d(i,j) \qquad (1)$$

Meanwhile, *BLOSUM* and *CLESUM* measure the evolutionary distance of two proteins at the sequence and local substructure levels, respectively. *BLOSUM* is the widely-used amino acid substitution matrix BLOSUM62[22], *CLESUM* is the local structure substitution matrix[25,26], $v(i,j)$ measures the hydrogen-bonding similarity

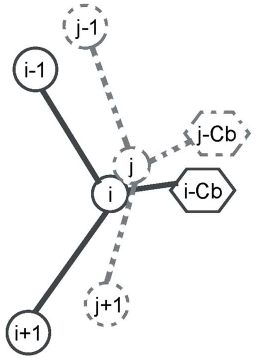

Figure 2 | This picture shows the three vectors of two proteins used in the hydrogen-bonding score. One protein is represented by solid lines and the other by dashed lines.

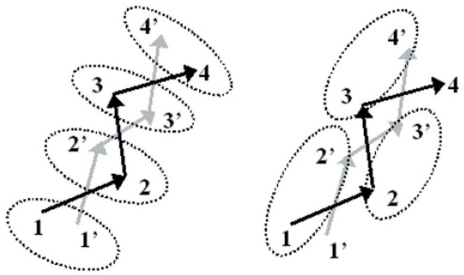

Figure 3 | An illustration of one wrong alignment between two beta-strands. Residues 1, 2, 3, and 4 in dark belong to protein *A* and residues 1′, 2′, 3′, and 4′ in gray belong to protein *B*. The aligned residue pairs are in dotted circle. (A) The wrong alignment generated by optimizing only TMscore. (B) The correct alignment optimizing the product of hydrogen-bonding score and TMscore.

and $d(i,j)$ measures the spatial proximity of two aligned residues (after rigid-body superposition).

In addition to the BLOSUM62 substitution matrix, other matrices (e.g., PAM250[47]) can also be used to measure the evolutionary distance of two proteins at the sequence level[41]. The *max()* function in Eq. (1) is used to handle the situation where two proteins to be aligned are distantly-related. In this case, we will only rely on CLESUM to measure the evolutionary distance. In the future we may use sequence profile similarity to measure evolutionary distance, which usually is more sensitive than BLOSUM matrices. One issue of using sequence profiles lies in that it is time-consuming to generate sequence profiles.

CLESUM is the substitution matrix for the 17 local conformation letter defined in[25,26]. Each letter represents one typical C_α conformation of a 4-mer protein segment. The 17 representative local conformations are generated by clustering all the 4-mer structural fragments in a subset of non-redundant protein structures. CLESUM is derived from a representative set of pairwise structure alignments in the FSSP database[48]. CLESUM disfavors the match of two unrelated helices but favors the alignment of two evolutionarily related loop regions. Loop regions are usually harder to align than alpha helices and beta strands if only spatial proximity is used in the scoring function.

In Eq. (1), $d(i,j)$ is the spatial proximity of two aligned (or equivalent) residues after the rigid-body superimposition. Here we use TMscore[28] to calculate $d(i,j)$, which is defined as follows.

$$d(i,j) = \frac{1}{1 + \left(|A_i - B_j|/d_0\right)^2} \quad (2)$$

Where A_i and B_j are the (transformed) 3D coordinates of the two C_α atoms and $d_0 = 1.24 \times \sqrt[3]{L_s - 15} - 1.8$ is a length-dependent normalization factor, which is used to offset the impact of protein length. TMscore is a widely-used measure in the field of protein structure prediction that has demonstrated excellent performance in identifying structurally similar proteins[49]. Generally speaking, when two protein structures have a TMscore larger than 0.6, it is highly likely they have similar folds. Otherwise, if the TMscore is less than 0.4, it is very likely they have different folds.

In Eq. (1), $v(i,j)$ is used to quantify hydrogen-bonding similarity through measuring the difference of the three vectors formed by the C_α and C_β atoms (see Figure 2). This term is defined as follows.

$$v(i,j) = \frac{1}{3}\left(\sum_{x=\{-1,1\}} \frac{(A_i - A_{i-x}) \bullet (B_j - B_{j-x})}{|A_i - A_{i-x}||B_j - B_{j-x}|} + \frac{(A_i - A_{i\_cb}) \bullet (B_j - B_{j\_cb})}{|A_i - A_{i\_cb}||B_j - B_{j\_cb}|}\right) \quad (3)$$

where A_{i_cb} denotes the corresponding C_β atom of A_i. This score helps align hydrogen bonds more accurately. As shown in Figure 3(A), the method that optimizes only spatial proximity (e.g., TMscore) leads to a wrong alignment, which can be corrected by incorporating v(i,j) to the scoring function, as shown in Figure 3(B).

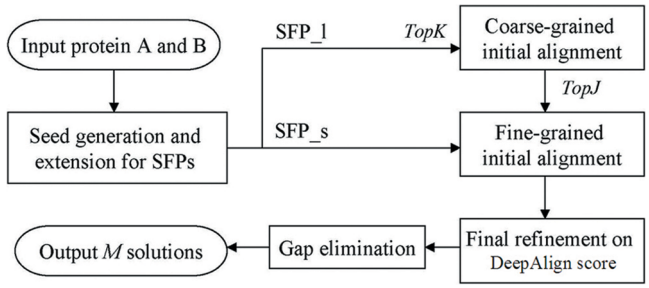

Figure 4 | The DeepAlign algorithm flowchart.

The search algorithm. *Overview*. The DeepAlign algorithm flowchart is illustrated in Figure 4. It consists of the following steps: (a) identifying similar fragment pairs (SFPs) using amino acid and local substructure mutation matrices; (b) generating an initial alignment from one SFP; and (c) refining alignments by dynamic programming and gap elimination.

Similar fragment pairs (SFPs). DeepAlign measures the equivalence of two residues *i* and *j* using amino acid and local substructure substitution matrices as follows.

$$Similarity(i,j) = \max(0, BLOSUM(i,j)) + CLESUM(i,j)$$

Where CLESUM is the local conformation substitution matrix and BLOSUM is the amino acid substitution matrix as described before. Using this score, we can identify two evolutionary-related instead of only geometric similar fragments and thus, generate better initial alignments. We use two types of SFPs: short SFP with 6-8 residues and long SFP with 9-18 residues. A short SFP, denoted as SFP_s, shall have a similarity score at least 0 while a long SFP, denoted as SFP_l, shall have a similarity score at least 10. It is obvious that each SFP_l must contain at least one SFP_s. We use SFP_l and SFP_s to build coarse-grained and fine-grained initial alignments, respectively. SFP_l is slightly longer than the average length of a helix while SFP_s has a similar length as a typical beta strand. By combining long and short SFPs, we can speed up our algorithm without losing accuracy. The higher score one SFP has, the more likely it is contained in the best alignment. Therefore, we sort all SFPs and only keep those top-ranked SFPs.

Generating initial alignments using SFPs. We select *TopK* long SFPs (i.e., SFP_l) and from each of them generate one coarse-grained initial alignment. In particular, we first calculate the rotation matrix using the Kabsch method[50] to minimize the RMSD of the two fragments in a SFP. Then we use this rotation matrix to transform one protein and generate an initial alignment using dynamic programming (DP) to maximize the scoring function, subject to the restriction that the distance deviation of two aligned residues shall be less than $3 \times$ d0. All these *TopK* coarse-grained initial alignments are sorted by the alignment score and only top *TopJ* ($\leq$*TopK*) are kept for further refinement. Starting from a coarse-grained initial alignment, we recalculate the rotation matrix using the SFP_s contained in the SFP_l and then realign the two proteins using dynamic programming to maximize the scoring function, which results in a better initial alignment.

Iterative refinement of alignment. Starting from an initial alignment, an iterative dynamic programming refinement procedure is applied to improving the alignment, with the goal to maximizing the scoring function. This procedure is very similar to that in many structure alignment methods such as CE[4], ProSup[51] and TMalign[19].

Gap elimination. As shown in[2], an aligned fragment pair (AFP) shall not be too short (say less than 4 residues). However, since our scoring function does not explicitly penalize gap openings, the resultant alignment may have more gap openings than desirable. To deal with this, we use some heuristics to merge one very short AFP (less than 4 residues) to its neighboring AFPs to reduce the number of gap openings.

Availability. DeepAlign is available at http://ttic.uchicago.edu/~jinbo/software.htm.

1. Hasegawa, H. & Holm, L. Advances and pitfalls of protein structural alignment. *Current opinion in structural biology* **19**, 341–348 (2009).
2. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology* **233**, 123–123 (1993).
3. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography* **60**, 2256–2268 (2004).
4. Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein engineering* **11**, 739 (1998).
5. Taylor, W. R. & Orengo, C. A. Protein structure alignment. *Journal of Molecular Biology* **208**, 1–22 (1989).
6. Wang, S. & Zheng, W. CLePAPS: fast pair alignment of protein structures based on conformational letters. *Journal of bioinformatics and computational biology* **6**, 347 (2008).
7. Ye, Y. & Godzik, A. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic acids research* **32**, W582 (2004).
8. Zhu, J. & Weng, Z. FAST: a novel protein structure alignment algorithm. *Proteins: Structure, Function, and Bioinformatics* **58**, 618–627 (2005).
9. Jung, J. & Lee, B. Protein structure alignment using environmental profiles. *Protein Engineering* **13**, 535–543 (2000).
10. Gibrat, J., Madej, T., Spouge, J. & Bryant, S. The VAST protein structure comparison method. *Biophys J* **72**, 298 (1997).
11. Guda, C., Lu, S., Scheeff, E. D., Bourne, P. E. & Shindyalov, I. N. CE-MC: a multiple protein structure alignment server. *Nucleic acids research* **32**, W100 (2004).

12. Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. & Lesk, A. M. MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics* **64**, 559–574 (2006).
13. Lupyan, D., Leo-Macias, A. & Ortiz, A. R. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* **21**, 3255 (2005).
14. Menke, M., Berger, B. & Cowen, L. Matt: local flexibility aids protein multiple structure alignment. *PLoS computational biology* **4**, e10 (2008).
15. Wang, S., Peng, J. & Xu, J. Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. *Bioinformatics* **27**, 2537–2545 (2011).
16. Wang, S. & Zheng, W. M. Fast Multiple Alignment of Protein Structures Using Conformational Letter Blocks. *Open Bioinformatics Journal* **3**, 69–83 (2009).
17. Ye, Y. & Godzik, A. Multiple flexible structure alignment using partial order graphs. *Bioinformatics* **21**, 2362 (2005).
18. Mayr, G., Domingues, F. & Lackner, P. Comparative analysis of protein structure alignments. *BMC Structural Biology* **7**, 50 (2007).
19. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302 (2005).
20. Kim, C. & Lee, B. Accuracy of structure-based sequence alignment of automatic methods. *BMC bioinformatics* **8**, 355 (2007).
21. Nadimpalli, S., Daniels, N. & Cowen, L. Formatt: Formatt: Correcting Protein Multiple Structural Alignments by Incorporating Sequence Alignment. *BMC Bioinformatics* **13**, 259 (2012).
22. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915 (1992).
23. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research* **36**, 2295 (2008).
24. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G. & Notredame, C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *Journal of molecular biology* **340**, 385–395 (2004).
25. Zheng, W. M. The use of a conformational alphabet for fast alignment of protein structures. *Bioinformatics Research and Applications*, 331–342 (2008).
26. Zheng, W. M. & Liu, X. A protein structural alphabet and its substitution matrix CLESUM. *Transactions on Computational Systems Biology II*, 59–67 (2005).
27. Van Walle, I., Lasters, I. & Wyns, L. SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics* **21**, 1267 (2005).
28. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**, 702–710 (2004).
29. Zemla, A. LGA: a method for finding 3D similarities in protein structures. *Nucleic acids research* **31**, 3370–3374 (2003).
30. Levitt, M. & Gerstein, M. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of sciences* **95**, 5913 (1998).
31. Marchler-Bauer, A. *et al*. CDD: a Conserved Domain Database for protein classification. *Nucleic acids research* **33**, D192 (2005).
32. Cheng, H., Kim, B. H. & Grishin, N. V. MALIDUP: a database of manually constructed structure alignments for duplicated domain pairs. *Proteins: Structure, Function, and Bioinformatics* **70**, 1162–1166 (2008).
33. Cheng, H., Kim, B. H. & Grishin, N. V. MALISAM: a database of structurally analogous motifs in proteins. *Nucleic acids research* **36**, D211 (2008).
34. Kim, C., Tai, C. H. & Lee, B. Iterative refinement of structure-based sequence alignments by Seed Extension. *BMC bioinformatics* **10**, 210 (2009).
35. Shapiro, J. & Brutlag, D. FoldMiner and LOCK 2: protein structure comparison and motif discovery on the web. *Nucleic acids research* **32**, W536 (2004).
36. Kawabata, T. MATRAS: a program for protein 3D structure comparison. *Nucleic acids research* **31**, 3367 (2003).
37. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* **247**, 536–540 (1995).
38. Cheng, H., Kim, B. H. & Grishin, N. V. Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. *Journal of molecular biology* **377**, 1265–1278 (2008).
39. Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology* **15**, 285–289 (2005).
40. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **33**, W244–W248 (2005).
41. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389 (1997).
42. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951 (2005).
43. Xie, L. & Bourne, P. E. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. *Proceedings of the National Academy of sciences* **105**, 5441 (2008).
44. Brylinski, M. & Skolnick, J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of sciences* **105**, 129 (2008).
45. Gao, M. & Skolnick, J. iAlign: a method for the structural comparison of protein–protein interfaces. *Bioinformatics* **26**, 2259 (2010).
46. Koehl, P. Protein Structure Classification. *Reviews in Computational Chemistry* **22**, 1 (2006).
47. Dayhoff, M. O. & Schwartz, R. M. *In Atlas of protein sequence and structure.* (1978).
48. Holm, L. & Sander, C. The FSSP database of structurally aligned protein fold families. *Nucleic acids research* **22**, 3600 (1994).
49. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics* **26**, 889–895 (2010).
50. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **32**, 922–923 (1976).
51. Lackner, P., Koppensteiner, W. A., Sippl, M. J. & Domingues, F. S. ProSup: a refined tool for protein structure alignment. *Protein Engineering* **13**, 745 (2000).

Acknowledgements

Funding: This work is financially supported by the National Institutes of Health grant R01GM089753 (to JX), the National Science Foundation grant DBI-0960390 (to JX) and the National Science Foundation CAREER award CCF-1149811 (to JX).

Author contributions

S.W. conceived, designed and implemented the algorithm. J.M., J.P. and J.X. helped design the algorithm and interpret the results. S.W., J.M. and J.X. wrote the manuscript. All authors read and approved the final manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Wang, S., Ma, J., Peng, J. & Xu, J. Protein structure alignment beyond spatial proximity. *Sci. Rep.* **3**, 1448; DOI:10.1038/srep01448 (2013).