Bioinformatics, 31(19), 2015, 3139–3146
doi: 10.1093/bioinformatics/btv354
Advance Access Publication Date: 9 June 2015
Original Paper

OXFORD

Structural bioinformatics

UniAlign: protein structure alignment meets evolution

Chunyu Zhao and Ahmet Sacan*

Center for Integrated Bioinformatics, School of Biomedical Engineering, Science and Health System, Drexel University, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on February 11, 2015; revised on May 26, 2015; accepted on June 2, 2015

Abstract

Motivation: During the evolution, functional sites on the surface of the protein as well as the hydrophobic core maintaining the structural integrity are well-conserved. However, available protein structure alignment methods align protein structures based solely on the 3D geometric similarity, limiting their ability to detect functionally relevant correspondences between the residues of the proteins, especially for distantly related homologous proteins.

Results: In this article, we propose a new protein pairwise structure alignment algorithm (UniAlign) that incorporates additional evolutionary information captured in the form of sequence similarity, sequence profiles and residue conservation. We define a per-residue score (UniScore) as a weighted sum of these and other features and develop an iterative optimization procedure to search for an alignment with the best overall UniScore. Our extensive experiments on CDD, HOMSTRAD and BAliBASE benchmark datasets show that UniAlign outperforms commonly used structure alignment methods. We further demonstrate UniAlign's ability to develop family-specific models to drastically improve the quality of the alignments.

Availability and implementation: UniAlign is available as a web service at: http://sacan.biomed.drexel.edu/unialign

Contact: ahmet.sacan@drexel.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein structure alignment can reveal distant evolutionary relationships between proteins, that sequence alignment alone is incapable to capture, and thus plays an essential role in understanding protein function (Hasegawa and Holm, 2009). Supposing the evolutionary continuity of structure and function, identification of structure similarities could elucidate the possible function of newly discovered proteins. Structure alignment identifies functionally equivalent residues in proteins and is often used as the gold standard for improving multiple sequence alignments (Orengo *et al.*, 1997). Despite the importance of the problem and the recent advances in the field, there is yet no widely accepted method for structural alignment. Large scale comparisons of existing methods have concluded that there is no single best method that works well for all proteins (Kolodny *et al.*, 2005).

Various pairwise protein structure alignment programs have been developed, differing in their representation of protein structure, the scoring function used to evaluate the 'goodness' of an alignment, and the optimization algorithm used to search for the best alignment with respective to the scoring function (Jung and Lee, 2000; Shindyalov and Bourne, 2001; Ye and Godzik, 2003). Generally, there are two different strategies to generate a structural alignment: directly searching for the optimal alignment by piecing together small aligned substructures, e.g. DaliLite (Holm *et al.*, 2008) and CE (Levitt and Gerstein, 1998); and iteratively optimizing a rough initial alignment using cycles of geometric superposition and residue pair collection steps, e.g. TMalign (Zhang and Skolnick, 2005) and Deepalign (Wang *et al.*, 2013).

The scoring function is used to evaluate how good an alignment is and recognize the optimal alignment among all the candidates. Common scoring functions utilize the root mean square distance (RMSD) between the aligned residues, taking into account the length of the proteins. Traditionally, protein structure alignment has been described as a geometric optimization problem, prudently optimizing the geometric superposition of proteins. This limits the usefulness of the resulting alignments, requiring researchers to seek additional validation from the amino acid types or catalytic activity of the aligned residues. This is evident from the presence of alternative alignments that are equally good in terms of the quality of their geometric superposition, but that vary widely in terms of their accuracy in identifying functionally and evolutionarily equivalent residues (Kim and Lee, 2007). For example, it is possible to achieve a good geometric superposition of immunoglobulin, despite misaligning a conserved disulfide bridge (Gerstein and Levitt, 1998). Furthermore, pure geometric information based structure alignment programs are found highly sensitive to conformational changes (Pirovano *et al.*, 2008).

It has been suggested that structural alignment can benefit from the evolutionary information provided by the alignment of homologous proteins (Hasegawa and Holm, 2009; Kim and Lee, 2007). This is not surprising since information extracted from homologous proteins represent general features of the protein family and allow the identification of similarity to a remote sequence or family, even when the similarity to each individual aligned sequence is not significant (Gerstein and Levitt, 1998).

In recent years, researchers have begun to integrate sequence similarity information into the scoring function. For example, Formatt (Daniels *et al.*, 2012) incorporates amino acid substitution matrices derived from evolutionarily-related protein pairs when constructing the alignment. Deepalign (Wang *et al.*, 2013) incorporates the BLOSUM mutation matrix, a local substructure mutation matrix, and hydrogen-bonding similarity into its scoring function. While these methods utilize the amino acid similarity as described by the BLOSUM substitution matrix, they do not utilize evolutionary information available from the history of the proteins being aligned.

It is well established that evolutionary profiles have a dominant advantage over sequence-based alignments, and provide greater accuracy in fold recognition (Yan *et al.*, 2013) and protein classification tasks (Rost and Sander, 1994). Motivated by this, we introduce UniScore, a new similarity score for structure alignment, which integrates evolutionary information in the form of sequence and conservation profiles; as well as amino acid, secondary structure and geometric similarity measures. Using this new similarity score, we implement UniAlign, a new protein pairwise structure alignment algorithm, which focuses on identifying not only structurally equivalent, but also evolutionarily favorable residue alignments. We demonstrate that compared to other methods, the alignments generated by UniAlign are in better agreement with hand-curated reference alignments. Furthermore, for difficult cases when UniAlign and other methods fail to generate good alignments, we propose family-specific alignment models to drastically improve the alignments.

2 Methods

UniAlign is aimed to recognize the maximal number of evolutionarily important residues as being structurally equivalent with minimal spatial deviations after the optimal rotation and translation. Two goals need to be accomplished for this task; a scoring function that can differentiate the optimal alignment from several candidate alignments must be first defined, and then applied in a heuristic search algorithm.

2.1 UniScore

An objective scoring function of a protein structure alignment program should reflect how likely two residues shall be aligned such that the program will be able to align those functionally important residue pairs more accurately. Based on the observation that important residues are likely to result in a loss of function were they to mutate into other residues during the evolution, we recover the functional importance of residues by analyzing residue conservation among homologous proteins. While sequence similarity has previously been investigated, to the best of our knowledge, this is the first study to systematically utilize evolutionary information, including conservation score and sequence profiles, in structure alignments.

While we utilize evolutionary information in structure alignment, we do not abandon other types of information that can help determine residue equivalences. We define UniScore as the weighted average of various sources of protein similarity measures. Specifically, for a residue pair i and j from two proteins being aligned, UniScore is defined as:

$$\begin{aligned} Uniscore(i,j) &= w_{\mathrm{geo}} \times Uni_{\mathrm{geo}}(i,j) + w_{\mathrm{pro}} \times Uni_{\mathrm{pro}}(i,j) \\ &\quad + w_{\mathrm{con}} \times Uni_{\mathrm{con}}(i,j) + w_{\mathrm{seq}} \times Uni_{\mathrm{seq}}(i,j) + w_{\mathrm{sse}} \times Uni_{\mathrm{sse}}(i,j) \end{aligned} \tag{1}$$

where Uni_{xxx} represents different types similarity measures, including geometric, profile, conservation, sequence and secondary structure similarity. The UniScore of an alignment is the sum of the UniScores of the aligned residue pairs. The weights w_{xxx} adjust the contributions of different similarity measures and are normalized to a sum of one. For the sequence similarity component Uni_{seq}, we use the BLOSUM62 amino acid substitution matrix. We describe each of the other similarity measures in more detail below.

2.1.1 Geometric similarity

We adopt the TMscore as the geometric component of UniScore, as TMscore has been proven to perform excellently as a geometric similarity measure (Zhang and Skolnick, 2004). For an alignment of length L_T, Uni_{geo} is defined as:

$$Uni_{\mathrm{geo}} = \mathrm{MAX}\left[\frac{1}{L_T}\sum_{i=1}^{L_T}\frac{1}{1+\left({}^{d_{i,j}}\big/_{d_0(L_{\min})}\right)^2}\right] \tag{2}$$

where $d_{i,j}$ is the Euclidean distance between the i th and j th aligned residues from the two proteins, and $d_0 = 1.24\sqrt{L_{\min} - 15} - 1.8$ is an empirical scaling factor to normalize the distance based on the length of the smaller protein (Zhang and Skolnick, 2004). It is worth noting that 'MAX' here refers to a heuristic search to find the rigid superposition with the maximum TMscore; and the summation is over all the aligned residues.

2.1.2 Evolutionary information as sequence profiles

The story of mutation, substitution, and genetic drift of one protein can be told from the amino acid patterns observed at each position of the multiple sequence alignments (MSA) of homologous proteins. A sequence profile represents the propensity of each amino acid to occur at each position of that protein. For each protein being aligned, we construct the MSA from the HSSP database of curated homologous proteins (Sander and Schneider, 1991). When HSSP does not contain an entry for a protein, we collect homologous proteins from a PSI-BLAST query (Altschul *et al.*, 1997) against NCBI's non-redundant sequence database (with E-value cutoff = 0.005, 3 iterations, and a maximum of 1000 search results).

From the constructed MSA of homologous proteins, we generate the position specific score matrix (PSSM), following sequence weight and pseudo-count calculations as used in (Altschul *et al.*, 1997). When comparing residues *i* and *j*, we calculate the score of aligning their PSSM columns as the sum-of-pairs of substitutions looked up from the BLOSUM matrix and weighted by the amino acid frequencies in the PSSM. Uni_{pro} of an alignment is then the sum of these scores for all aligned residue pairs. Although there are other methods for comparing two profiles, there is no statistically significant difference between these methods (Edgar and Sjolander, 2004).

2.1.3 Evolutionary information as conservation similarity scores

Whereas the sequence profile describes the amino acid composition at each position of the protein, the conservation score describes the variability at each position. With conservation, we aim to capture equivalent residues that share a similar conservation level, but that may or may not share a similar amino acid composition. Generally speaking, the conservation value should be able to normalize against redundancy and bias in the MSA without loss of evolutionary information (Valdar, 2002). We calculate the conservation scores of a single protein from MSA using a sequence-weighted sum-of-pairs scheme (Shatsky *et al.*, 2006).

While the conservation score is routinely used to evaluate functional importance of residues, we are not aware of any study that aligns proteins based on conservation levels of the residues. While it is expected that highly conserved residues are more likely to align, no quantification of conservation-based similarity is available. Here, in order to systematically quantify likelihood of aligning residues with different conservation values, we generate a conservation similarity score table as illustrated in Figure 1, similar to the generation of the BLOSUM substitution matrix (Henikoff and Henikoff, 1992). We generate the conservation values of the aligned residues for all proteins in the CDD reference alignment database and discretize these conservation values into 20 conservation categories by equal frequency binning. This essentially gives us a set of alignments where each residue is now encoded by a discrete conservation level. A conservation similarity scoring table tabulating the log-odds ratios of observing the alignment of any two conservation levels is calculated from these reference alignments. Uni_{con} between two residues can then be looked up from this substitution table. Notice in the heatmap shown in Figure 1 that alignment of not only the highly conserved residues, but also of highly variable residues is favorable. The conservation similarity score table is provided in Supplementary documents.

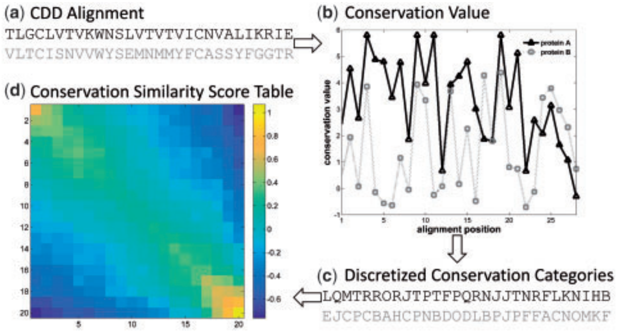


Fig. 1. Calculation of the conservation similarity score table. For each protein in the (**a**) CDD reference alignments, (**b**) the conservation values of the residues are calculated and (**c**) converted into one of 20 discrete conservation levels, encoded here by letters A (least conserved) through T (most conserved). (**d**) A conservation similarity score table shown as a heatmap here, is calculated using log-odds ratios of observing different conservation levels aligned in the encoded alignments

2.1.4 Secondary structure similarity

We obtain the secondary structure assignments from DSSP (Kabsch and Sander, 1983), or calculate it from the alpha carbon distances (Zhang and Skolnick, 2005) when DSSP entry is not available. We then calculate a secondary structure scoring table (available in the Supplementary documents) from the CDD reference alignments as log-odds ratio of observed substitutions. Uni_{sse} of aligned residues is then looked up from this substitution table.

2.2 UniAlign

Equally important as the scoring function of a protein structure alignment method is the heuristic search algorithm used to find an alignment with optimal score. The UniAlign algorithm consists of 4 main steps (Algorithm 1). First, an initial alignment is constructed as a set of residues pairs from the two proteins, using a fragment-based search. Second, the proteins are geometrically superposed based on this initial alignment. Third, the spatial proximity of the residues in the superposition, along with the other components of the UniScore are used to collect a new set of residue correspondences. The second and third steps are repeated until the UniScore of the alignment converges. We describe each of these steps in more detail below.

Algorithm 1. UniAlign Algorithm

```
Input: Two protein structures, A and B of length $L_A$ and $L_B$.
Output: Aligned residue pairs and rotation/translation matrices.
1. Construct the initial alignment (Algorithm 2)
2. while pairs have not converged do:
3.     Geometrically superpose the structures.
4.     Calculate the UniScore similarity matrix.
5.     Collect residue pairs using dynamic programming.
6. Return aligned residue pairs and the final UniScore.
```

2.2.1 Initial alignment

We consider structure alignments resulting from gapless alignment of all pairs of fragments of length L_f from the two proteins. Similar to (Pandit and Skolnick, 2008), we use $L_f = 8$ if the smaller protein has less than 100 residues, and $L_f = 12$ otherwise. Proteins shorter than 8 residues are used as a single fragment. Unlike TMalign and Fr-TM-align (Pandit and Skolnick, 2008), which use different and possibly conflicting criteria for initializing and optimizing the alignment, we use the same UniScore evaluation for each of these steps. For each fragment pair, we use their alignment to obtain a 3D transformation and use this transformation to superpose the entire proteins (Algorithm 2). The calculation of UniScore and collection of residue pairs is done as in the main algorithm, except without the iterative optimization loop. The UniScore of the alignment is assigned into the corresponding alignment path in T_{init} initial alignment score table. If a residue pair [*i*, *j*] is part of alignments resulting from multiple fragment pair alignments, we keep the largest UniScore in $T_{\text{init}}[i, j]$. Once all fragments are assessed, we use dynamic programming with free end gaps, on the T_{init} table to find an alignment path that produces the largest sum of UniScores. This alignment path is used as the initial set of residue correspondences, to be optimized by the rest of the UniAlign algorithm.

Algorithm 2. Generate the initial alignment.

```
Input: Two protein structures, A and B of length $L_A$ and $L_B$.
Out: Aligned residue pairs from the two proteins.
$T_{\text{init}} \leftarrow L_A$ by $L_B$ initial alignment score table.
foreach fragment $F_A$ of length $L_f$ from A do:
    foreach fragment $F_B$ of length $L_f$ from B do:
        Geometrically superpose $F_A$ and $F_B$
        Use this 3D transformation to superpose A and B
        Collect residue pairs [$i$, $j$] from the UniScore similarity matrix
        Replace all smaller $T_{\text{init}}$ [$i$, $j$] values with UniScore of this alignment.
Return residue pairs from dynamic programming on $T_{\text{init}}$.
```

2.2.2 UniScore-enriched Gaussian-weighted RMSD superposition

Given a set of residue correspondences from two proteins, Kabsch least square method (Kabsch, 1978) has been the standard method for generating the rotation translation matrices (R/T) that superpose one structure onto the other with optimal RMSD. However, Kabsch's method is sensitive to outliers and the resulting superposition can be skewed by flexible regions, such as loop and hinge regions. In order to overcome this problem, we utilize a weighted RMSD superposition, where the distance penalty of each aligned residue pair is weighted by a Gaussian function of the distance, effectively reducing the contribution of large distances and focusing the superposition on residues that can be aligned well.

We observe that the original Gaussian-weighted RMSD (Damm and Carlson, 2006) tends to yield poor results for significantly different structures (low sequence identity $\leq 20\%$). Several actions were taken in UniAlign to refine the performance of Gaussian-weighted RMSD. First, in order to avoid over-fitting to very few pairs, we resort to the standard RMSD if there are less than 10 pairs of residues aligned closer than $sqrt(RMSD)$. Second, the UniScore of the aligned residue pairs is used as the convergence criteria during iterative superposition, ensuring the superposition step improves the same criteria as the rest of UniAlign algorithm. Third, whereas the original method uses standard RMSD superposition in its first step, we use the geometric superposition available from the previous iteration (of the iterative optimization in Algorithm 1) to speed up convergence and ensure a smooth exploration of the search space.

2.2.3 Collecting residue pairs from the geometric superposition

Geometric superposition, while bringing some of the residue pairs close to each other in space, may make or break other residue pairs. Thus, we collect new residue pairs that are consistent with the geometric superposition, along with the rest of the components measured by UniScore. We consider a score table T_{align} where each entry is the individual UniScores of the pairs of residues from the two proteins. Residues that have similar evolutionary, sequence and secondary structure features, as well as that are close in space, will have high UniScore values. We use dynamic programming (with affine gap penalty) on T_{align} to collect a new set of correspondences, such that the sum of their UniScores is optimal. These new correspondences are then used for another round of superposition and residue collection; iteratively optimizing the alignment until the total UniScore cannot be improved further.

2.2.4 Parameter optimization

The weights in UniScore controlling the contributions of different features and the gap penalties used in dynamic programming are optimized using grid search, on a small subset of the CDD database (Marchler-Bauer *et al.*, 2013), with the objective of maximizing the fraction of correctly aligned residues. The training dataset and the optimized parameters can be found in the Supplementary File.

3 Results and discussion

3.1 Experiments

3.1.1 Benchmark datasets

We evaluated UniAlign on three large-scale datasets that are commonly used to assess sequence and structure alignment methods: CDD, HOMSTRAD and BAliBASE. The subset of NCBI's human-curated Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2013) used in (Kim and Lee, 2007) contained a total of 3591 pairwise alignments with the corresponding ASTRAL SCOP domains (Chandonia *et al.*, 2004). HOMSTRAD (Stebbings and Mizuguchi, 2004) is a curated database of structure-based alignments for 3454 homologous structures from 1032 protein families; giving a total of 9536 pairwise alignments of protein structures. BAliBASE (Thompson *et al.*, 2005) contains 162 multiple alignments, involving 1944 pairwise sequence alignments from five different reference sets indicating various divergence levels.

A global sequence alignment with the PDB sequences was used to identify start and end positions of the sequences listed in these benchmark datasets and to correct for discrepancies such as missing residues. Unlike other assessment studies that use pre-segmented domains, we use the full length chains as the input structures. We denote the alignment generated by different structure alignment methods as the *test alignment*. Only the homologous regions marked in the reference datasets were used to evaluate the test alignments.

3.1.2 Comparison with other methods

Three widely used structure alignment methods were chosen for comparison: DaliLite, TMalign and Deepalign. DaliLite is a classical geometry-based alignment method that uses a Monte-Carlo procedure to minimize the intra-structural distances of aligned substructures and generates the final alignment by gradually adding new eight-residue fragments to the existing alignments (Holm *et al.*, 2008). TM-align is another method that has been shown to perform well as a purely geometric information based structure alignment method (Zhang and Skolnick, 2005). TM-align utilizes the TMscore, a length-normalized geometric scoring measure that, compared to RMSD, attenuates the contribution of large distances. Deepalign is a recently developed method that integrates the BLOSUM mutation matrix, local substructure mutation matrices and hydrogen-bonding similarity in its scoring function (Wang *et al.*, 2013). We note that the so-called evolutionary distance in Deepalign is only a simple transformation of the BLOSUM mutation matrix and does not utilize the evolutionary history of the proteins being aligned.

3.1.3 Accuracy of the alignment

Following other studies that assess the performance of structure alignment methods, we use the fraction of correctly aligned residues *fcar* as the primary criteria to assess the test alignments, in comparison with the manually curated reference alignments (Kim and Lee, 2007; Nayeem *et al.*, 2006; Sauder *et al.*, 2000). *fcar*$_\delta$ is defined as the fraction of residues that have a shift error of at most δ. When the reference dataset specifies a core region (e.g. for the CDD database), we calculate *fcar* only for the core region.

3.2 Performance on benchmarks

The results from running UniAlign and other structure alignment methods on the three benchmark datasets are summarized in Table 1. For all three datasets, UniAlign aligned a higher fraction of the residues correctly, achieving *fcar*$_0$ of 93.7, 91.4 and 73.5% for CDD, HOMSTRAD and BAliBASE datasets, respectively. Since UniAlign optimizes for the UniScore, it is no surprise that it generates alignments with the best UniScore. Geometric quality of the UniAlign alignments, as measured by TMscore, was also comparable to or better than those generated by other alignments. This indicates that incorporating evolutionary information did not deteriorate the performance of UniAlign as a structure alignment method.

Secondary structure states of the aligned residues were best matched by Deepalign, likely due to the consideration of hydrogen bonding in its scoring function. Residue pairs in the UniAlign alignments had the highest scores for their sequence, profile and conservation scores. Note that Deepalign utilizes sequence information, whereas DaliLite and TMalign make use of only the geometric information. Consequently, the sequence and evolutionary scores of the alignments from DaliLite and TMalign are significantly lower than those of UniAlign and Deepalign.

BAliBASE reference alignments were more difficult to reproduce by the structure alignments, even though these sequences were not more remotely related to each other than those in the other databases, as measured by sequence identity. CDD database contained alignments with poorer geometric similarity than the other databases, as measured by TMscore. On the other hand, CDD had a higher secondary structure score, indicating that its human curators may have paid special attention to the secondary structure elements when constructing the alignments and determining the core regions. Compared to the other databases, HOMSTRAD alignments contained proteins that were more similar to each other in both sequence and structure.

Table 1. Comparison of the performance of four structure alignment methods on three benchmark datasets

Method	*fcar*$_0$	UNI	GEO	SSE	SEQ	PRO	CON
CDD core regions (3591 pairs with sequence identity 21.7%)							
UniAlign	**93.7%**	**2.09**	0.682	0.141	**0.282**	**0.163**	**0.054**
Deepalign	91.5%	1.99	0.654	**0.151**	0.265	0.123	0.045
DaliLite	92.1%	2.00	0.662	0.141	0.095	0.041	0.041
TMalign	85.1%	2.05	**0.684**	0.143	0.047	-0.002	0.038
CDD core	**100.0%**	1.09	0.341	**0.232**	**0.688**	**0.566**	**0.071**
HOMSTRAD (9536 pairs with sequence identity 35.7%)							
UniAlign	**91.4%**	**2.74**	**0.802**	0.134	**1.696**	**1.236**	**0.168**
Deepalign	90.3%	2.69	0.789	**0.136**	1.685	1.215	0.163
DaliLite	83.1%	2.68	0.797	0.134	1.526	1.108	0.158
TMalign	87.0%	2.68	0.793	0.134	1.559	1.135	0.159
HOMSTRAD	**100.0%**	2.67	0.762	**0.135**	1.658	1.210	0.164
BAliBASE (1944 pairs with sequence identity 23.4%)							
UniAlign	**73.5%**	**2.36**	**0.733**	0.121	**0.504**	**0.626**	**0.114**
Deepalign	71.6%	2.26	0.706	**0.126**	0.487	0.585	0.104
DaliLite	68.9%	2.26	0.712	0.119	0.283	0.478	0.097
TMalign	68.3%	2.30	0.729	0.120	0.275	0.461	0.096
BAliBASE	**100.0%**	2.00	0.601	**0.126**	0.414	0.558	0.101

Note: For each dataset, the best performance values are shown in bold. The scores of the reference database alignments are also shown in bold, when it is better than the best value from the structure alignment methods.

Note that DaliLite failed to report any result for 14 pairs from CDD, 646 pairs from HOMSTRAD, and 51 pairs from BAliBASE. These missing alignments were excluded when calculating the average scores, inflating the reported statistics for DaliLite. Deepalign failed to produce an alignment for one CDD pair. UniAlign and TMalign generated an alignment in all cases.

Whereas *fcar*$_0$ evaluates the exact agreement of the residue correspondences with respect to the reference alignment, it is suggested that an approximate alignment that superposes the correct regions of the proteins may be sufficient in certain applications, such as fold recognition. Figure 2 shows the accuracy of the CDD alignments under different allowed shift errors δ.

UniAlign outperforms other methods for all shift error tolerance levels. The accuracy of UniAlign, Deepalign and DaliLite increased by 2–3% when a single shift error was allowed, whereas the accuracy of TM-align increased by 8% for $\delta = 1$. This suggests a substantial room for improvement of existing TM-align alignments by considering single-residue shifts. Whereas an additional 3% of the residues from Deepalign, DaliLite and TM-align had a shift error of $\delta = 2$, UniAlign alignments contained fewer residues with two-residue shift error. The fraction of residues with a shift-error of 3–8 were small for all methods, indicating a deficiency in generation of initial alignments, such that the remaining cases misaligned by each method cannot be corrected by small adjustments of their existing alignments.

3.3 Dependence of performance on sequence and sequence similarity

A successful structure alignment method should be able to generate accurate alignments for different types of proteins it is applied to. Here, we characterize the performance of UniAlign and other structure alignment methods with respect to the level of sequence and structure similarity levels of the aligned proteins.

Figure 3 illustrates the accuracy of aligning proteins with different sequence similarity levels, where similarity is measured as the fraction of identical amino acid residues in the reference alignment. UniAlign is robust with respect to the homology level of the proteins

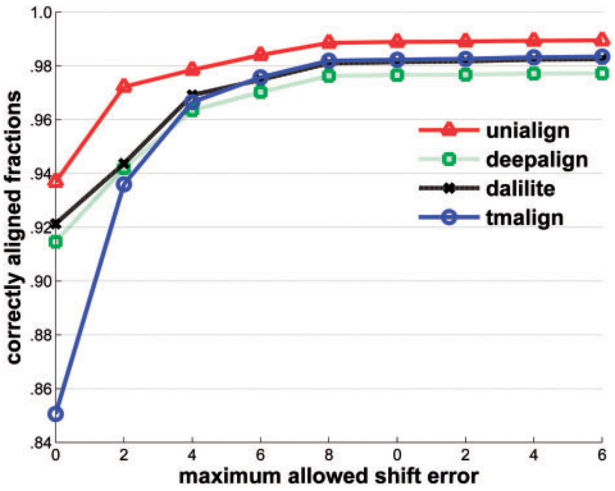


Fig. 2. Average *fcar* of CDD alignments of different methods as a function of the shift error tolerance level δ. The y-axis starts from 0.84 to 1.0

it is applied to and consistently produces good alignments at all sequence identity levels. The other methods perform poorer on more remotely homologous proteins, consistent with the commonly accepted notion that closely related proteins are easier to align. Surprisingly however, the performance of other methods decreases for proteins with 45+% sequence identity compared to those with 40–45% identity. We attribute this to the presence of equally good alignments when only geometric similarity is considered—UniAlign is able to distinguish among these alternatives by utilizing additional non-geometric information. TMalign was the most sensitive to the sequence similarity level of the proteins and performed its best when the proteins were 35–40% identical.

In order to characterize the geometric similarity of the dataset proteins, we used the TMscore measure of the superpositions produced from the residue correspondences of the reference alignments (using the entire CDD alignments, not just the core regions). For structurally highly similar proteins (TMscore > 0.5), the performance of all the methods were similar (Fig. 4).

At lower structural similarity levels Deepalign, DaliLite and TM-align produced significantly less accurate alignments. On the other hand, UniAlign was robust with respect to structural similarity level of the proteins it was applied to, consistently producing highly accurate alignments. We attribute this to the fact that at lower structural similarity levels, geometric information alone is not sufficient for identifying functionally equivalent residues and there is a greater benefit from incorporating evolutionary information. Note that although Deepalign utilizes sequence information, its behavior and performance at different sequence and structure similarity levels were similar to those of DaliLite, which uses geometric information alone.

3.4 Case study

We demonstrate the advantage UniAlign has over other structure alignment methods using a case study of proteins from the immunoglobulin superfamily: a heterogeneous group of proteins built on a common fold comprised of a sandwich of two beta sheets, listed in CDD with the identifier CD00096 (Marchler-Bauer *et al.*, 2013). The residue correspondences of the reference alignment and of the test alignments from structure alignment methods are shown in Figure 5a. Here, the accuracy of a test alignment is determined by the agreement of the core residue correspondences with those in the reference alignment (shown with capital letters).

UniAlign aligns all of the core residues correctly, whereas TM-align produces one-residue shifted alignments and DaliLite and Deepalign produces two-residue shifted alignments. From the geometric similarity point of view, all of these shifted alignments are as good as the reference alignment, yet they misalign functionally equivalent residues, including the highly conserved cysteine bridge and tryptophan residues. This demonstrates that geometric information alone is insufficient in recognizing biologically relevant alignments and additional sequence and evolutionary features need to be considered in order to obtain accurate alignments.

Figure 5b and 5c show the geometric superposition of the two proteins resulting from DaliLite and UniAlign. UniAlign aligns all of the beta strands correctly, whereas DaliLite misaligns them as can be observed by focusing on the ends of these beta strands. The regularity of the beta strand elements in general is an important factor for DaliLite (and other pure geometry based structure alignment methods) to produce such incorrect alignments of residues that otherwise superpose tightly in 3D space.

3.5 Family-specific optimization

There were several reference alignments for which none of the structure alignment methods (including UniAlign) was able to produce the correct alignment. Among these were alignments of the proteins from the calmodulin-like (CBP) protein family in HOMSTRAD database. The CBP family contained 8 all-alpha protein structures, with an average pairwise sequence identity of 38%. Calmodulin has a flexible linker connecting two globular calcium-binding domains, which throws a wrench into the structure alignments, because of the difficulty of simultaneously superposing the two domains with a rigid alignment. The accuracy of the alignments obtained by Deepalign, TM-align, DaliLite and UniAlign were 45.8, 56.6, 65.9 and 66.2%, respectively. Although flexible structure alignment may be the natural solution to align these proteins, Fr-TM-align (Pandit and Skolnick, 2008), which is popular for its support of flexible

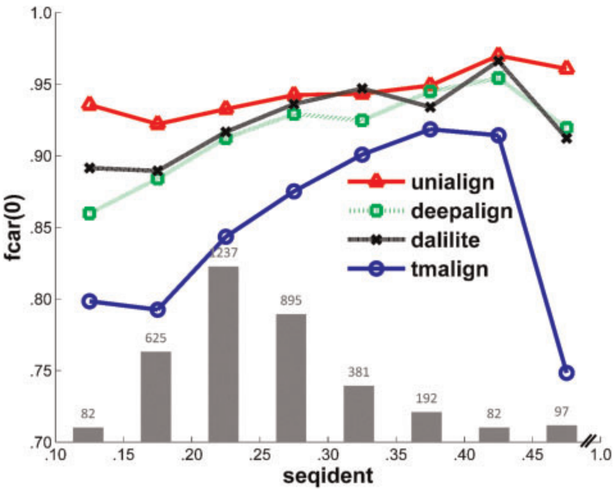


Fig. 3. Dependence of alignment accuracy on the level of homology of the proteins from the CDD dataset. Alignments were grouped into sequence identity bins of 5% width. Line plots show the average *fcar₀* values of various methods, whereas the histogram shows the number of alignments in each bin

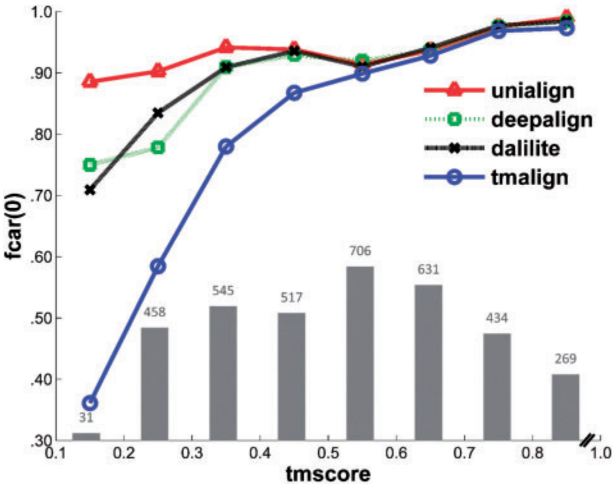
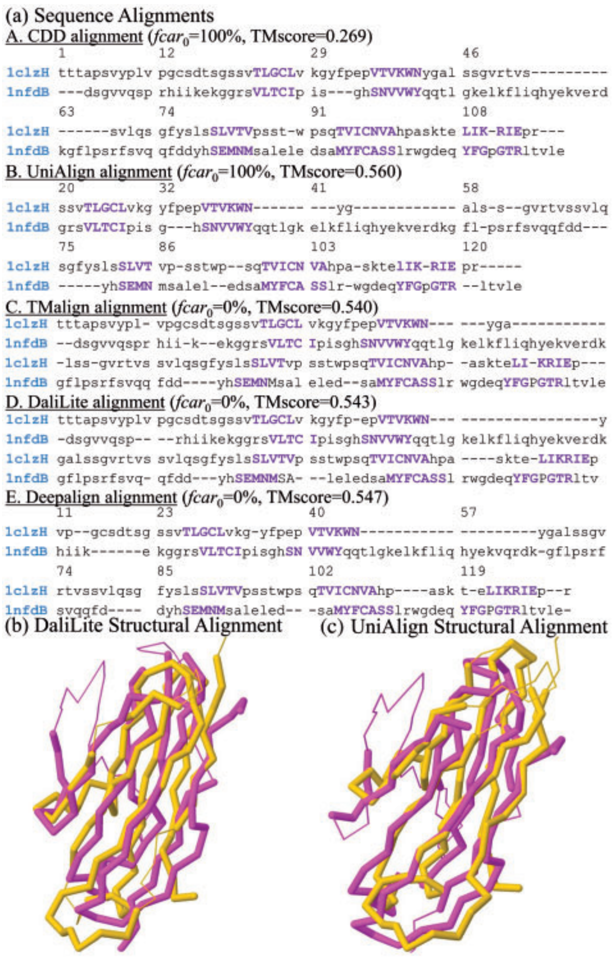


Fig. 4. Dependence of alignment accuracy on the level of structural similarity of the proteins from the CDD dataset. Structural similarity is measured by the TMscore of the superposition generated from the reference alignments. Proteins are grouped into structural similarity bins of size 0.1. Line plots show the average *fcar₀* values of different methods, whereas the histogram shows the number of alignments in each bin



(a) Sequence Alignments

A. CDD alignment (*fcar*$_0$=100%, TMscore=0.269)

```
      1          12              29              46
1clzH tttapsvyplv pgcsdtsgssvTLGCLv kgyfpepVTVKWNygal ssgvrtvs---------
1nfdB ---dsgvvqsp rhiikekggrsVLTCIp is---ghSNVVWYqqtl gkelkfliqhyekverd
      63         74              91              108
1clzH ------svlqs gfyslsSLVTVpsst-w psqTVICNVAhpaskte LIK-RIEpr---
1nfdB kgflpsrfsvq qfddyhSEMNMsalele dsaMYFCASSlrwgdeq YFGpGTRltvle
```

B. UniAlign alignment (*fcar*$_0$=100%, TMscore=0.560)

```
      20         32              41              58
1clzH ssvTLGCLvkg yfpepVTVKWN------ ---yg------------ als-s--gvrtvssvlq
1nfdB grsVLTCIpis g---hSNVVWYqqtlgk elkfliqhyekverdkg fl-psrfsvqqfdd---
      75         86              103             120
1clzH sgfyslsSLVT vp-sstwp--sqTVICN VAhpa-sktelIK-RIE pr-----
1nfdB -----yhSEMN msalel--edsaMYFCA SSlr-wgdeqYFGpGTR --ltvle
```

C. TMalign alignment (*fcar*$_0$=0%, TMscore=0.540)

```
1clzH tttapsvypl- vpgcsdtsgssvTLGCL vkgyfpepVTVKWN--- ---yga-----------
1nfdB --dsgvvqspr hii-k--ekggrsVLTC IpisghSNVVWYqqtlg kelkfliqhyekverdk
1clzH -lss-gvrtvs svlqsgfyslsSLVTvp sstwpsqTVICNVAhp- -askteLI-KRIEp---
1nfdB gflpsrfsvqq fdd----yhSEMNMsal eled--saMYFCASSlr wgdeqYFGPGTRltvle
```

D. DaliLite alignment (*fcar*$_0$=0%, TMscore=0.543)

```
1clzH tttapsvyplv pgcsdtsgssvTLGCLv kgyfp-epVTVKWN--- ----------------y
1nfdB -dsgvvqsp-- --rhiikekggrsVLTC IpisghSNVVWYqqtlg kelkfliqhyekverdk
1clzH galssgvrtvs svlqsgfyslsSLVTVp sstwpsqTVICNVAhpa ----skte-LIKRIEp
1nfdB gflpsrfsvq- qfdd---yhSEMNMSA- --leledsaMYFCASSl rwgdeqYFGPGTRltv
```

E. Deepalign alignment (*fcar*$_0$=0%, TMscore=0.547)

```
      11         23              40              57
1clzH vp--gcsdtsg ssvTLGCLvkg-yfpep VTVKWN----------- ---------ygalssgv
1nfdB hiik------e kggrsVLTCIpisghSN VVWYqqtlgkelkfliq hyekvqrdk-gflpsrf
      74         85              102             119
1clzH rtvssvlqsg  fyslsSLVTVpsstwps qTVICNVAhp----ask t-eLIKRIEp--r
1nfdB svqqfd----  dyhSEMNMsaleled-- -saMYFCASSlrwgdeq YFGPGTRltvle-
```

(b) DaliLite Structural Alignment

(c) UniAlign Structural Alignment

Fig. 5. Case study: comparison of structure alignments of proteins 1clz H:115-231 and 1nfd B:1-117 from the cd00096 family of all-beta immunoglobulin proteins. (**a**) Residue correspondences from the reference CDD alignment and structure alignments. The core residues are shown in capital letters and marked in purple. (**b**) and (**c**) display the 3D geometric superposition of the DaliLite and UniAlign alignments. Structures are drawn in Jmol (Hanson, 2010), with the aligned residues shown in thicker backbone

Table 2. Comparison of performance on CBP family before (lower triangle) and after (upper triangle) the family-specific optimization

fcar$_0$	1aj4	1br1	1tn4	2sas	2scp	3cln	4cln	5tnc
1aj4		**0.952**	**0.975**	**0.946**	**0.785**	**0.965**	**0.966**	**0.975**
1br1	0.425		**0.959**	**0.688**	**0.837**	**0.972**	**0.972**	**0.959**
1tn4	**0.975**	0.338		**0.952**	**0.859**	**1.00**	**1.00**	**1.00**
2sas	0.390	0.382	0.418		**0.966**	**0.957**	**0.963**	**0.946**
2scp	0.444	0.433	0.454	0.931		**0.842**	**0.843**	**0.854**
3cln	0.865	0.958	0.539	0.454	0.453		**1.00**	**1.00**
4cln	**0.966**	0.958	0.898	0.472	0.450	1.00		**1.00**
5tnc	0.968	0.476	1.00	0.432	0.473	1.00	0.898	

Note: For each protein pair, better of the two alignments is shown in bold.

alignments, also failed to align these globular domains, with *fcar*$_0$ = 41.49%.

Since the weights in our UniScore formulation control the contributions from different types of information, we can adjust these parameters to better align protein families with unique requirements. Setting aside a single test protein from the CBP family, we optimized the parameters using the rest of the CBP proteins. The test protein is then aligned with each of the CBP proteins and the accuracy is recorded. This optimization and testing procedure is repeated with each CBP protein set aside as the test case. The accuracy of the structure alignments obtained before and after this optimization is shown in Table 2. UniAlign achieves a boost of 27.4% on the average, when the family-specific optimization is performed.

We observed that optimization of the parameters for the CBP family reduced the weight of the geometric component from 0.29 to 0.06, while increasing the weights of the other components. This again illustrates the benefit of incorporating non-geometric features into a structural alignment method to detect functionally equivalent residues even under big conformational differences in the structures.

4 Conclusion

In this study, we have introduced UniAlign, a new structural alignment method that integrates different sources of information in order to achieve a more accurate alignment. Compared to classical methods that utilize only the geometry of the proteins and the recently developed methods that incorporate sequence information; UniAlign produces alignments that are in better agreement with expert-curated datasets. UniAlign is robust with respect to the sequence homology or the geometric similarity levels of the proteins being aligned. Furthermore, adjustment of UniAlign's parameters allows for development of family-specific models that highlight the features most relevant to the proteins in that family.

The increased accuracy achieved by UniAlign is at the cost of increased demands in computing time. For an average sized pair of proteins, it can take up to 15 min to calculate a structure alignment, with most of this time spent on the homology search to construct a multiple sequence alignment. The running times can be significantly reduced by caching and re-using the evolutionary information calculated for each protein in their alignments with different proteins. A detailed running time analysis is provided in the Supplemental Data.

We expect a number of downstream applications to benefit from the additional accuracy provided by UniAlign. Ability to develop family-specific alignment models will find use in structure classification problem. Integration of evolutionary information is likely to improve the protein-protein interaction prediction protocols that rely on structural alignment.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Chandonia,J.M. *et al.* (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.

Damm,K.L. and Carlson,H.A. (2006) Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys. J.*, **90**, 4558–4573.

Daniels,N.M. *et al.* (2012) Formatt: Correcting protein multiple structural alignments by incorporating sequence alignment. *BMC Bioinformatics*, **13**, 259.

Edgar,R.C. and Sjolander,K. (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, **20**, 1301–1308.

Gerstein,M. and Levitt,M. (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci.*, **7**, 445–456.

Hanson,R.M. (2010) Jmol - a paradigm shift in crystallographic visualization. *J. Appl. Crystallogr.*, **43**, 1250–1260.

Hasegawa,H. and Holm,L. (2009) Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*, **19**, 341–348.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.

Holm,L. *et al.* (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.

Jung,J. and Lee,B. (2000) Protein structure alignment using environmental profiles. *Protein Eng.*, **13**, 535–543.

Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A*, **34**, 827–828.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kim,C. and Lee,B. (2007) Accuracy of structure-based sequence alignment of automatic methods. *BMC Bioinformatics*, **8**, 355.

Kolodny,R. *et al.* (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.

Levitt,M. and Gerstein,M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci. USA*, **95**, 5913–5920.

Marchler-Bauer,A. *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.*, **41**, D348–D352.

Nayeem,A. *et al.* (2006) A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models. *Protein Sci.*, **15**, 808–824.

Orengo,C.A. *et al.* (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pandit,S.B. and Skolnick,J. (2008) Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics*, **9**, 531.

Pirovano,W. *et al.* (2008) The meaning of alignment: lessons from structural diversity. *BMC Bioinformatics*, **9**, 556.

Rost,B. and Sander,C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Sauder,J.M. *et al.* (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, **40**, 6–22.

Shatsky,M. *et al.* (2006) Optimization of multiple-sequence alignment based on multiple-structure alignment. *Proteins*, **62**, 209–217.

Shindyalov,I.N. and Bourne,P.E. (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic Acids Res.*, **29**, 228–229.

Stebbings,L.A. and Mizuguchi,K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res.*, **32**, D203–D207.

Thompson,J.D. *et al.* (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.

Valdar,W.S.J. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.

Wang,S. *et al.* (2013) Protein structure alignment beyond spatial proximity. *Sci. Rep.*, **3**, 1448.

Yan,R. *et al.* (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci. Rep.*, **3**, 2619.

Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**, ii246–ii255.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.