BMC Bioinformatics

Research article

Topology independent protein structural alignment

Joe Dundas[1], TA Binkowski[1], Bhaskar DasGupta*[2] and Jie Liang[1]

Address: [1]Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607-7052, USA and [2]Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053, USA

Email: Joe Dundas - jdunda1@uic.edu; TA Binkowski - abinkowski@anl.gov; Bhaskar DasGupta* - dasgupta@cs.uic.edu; Jie Liang - jliang@uic.edu

* Corresponding author

Published: 15 October 2007

BMC Bioinformatics 2007, **8**:388 doi:10.1186/1471-2105-8-388

Received: 2 July 2007
Accepted: 15 October 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/388

Abstract

Background: Identifying structurally similar proteins with different chain topologies can aid studies in homology modeling, protein folding, protein design, and protein evolution. These include circular permuted protein structures, and the more general cases of non-cyclic permutations between similar structures, which are related by non-topological rearrangement beyond circular permutation. We present a method based on an approximation algorithm that finds sequence-order independent structural alignments that are close to optimal. We formulate the structural alignment problem as a special case of the maximum-weight independent set problem, and solve this computationally intensive problem approximately by iteratively solving relaxations of a corresponding integer programming problem. The resulting structural alignment is sequence order independent. Our method is also insensitive to insertions, deletions, and gaps.

Results: Using a novel similarity score and a statistical model for significance *p*-value, we are able to discover previously unknown circular permuted proteins between nucleoplasmin-core protein and auxin binding protein, between aspartate rasemase and 3-dehydrogenate dehydralase, as well as between migration inhibition factor and arginine repressor which involves an additional strand-swapping. We also report the finding of non-cyclic permuted protein structures existing in nature between AML1/core binding factor and ribofalvin synthase. Our method can be used for large scale alignment of protein structures regardless of the topology.

Conclusion: The approximation algorithm introduced in this work can find good solutions for the problem of protein structure alignment. Furthermore, this algorithm can detect topological differences between two spatially similar protein structures. The alignment between MIF and the arginine repressor demonstrates our algorithm's ability to detect structural similarities even when spatial rearrangement of structural units has occurred. The effectiveness of our method is also demonstrated by the discovery of previously unknown circular permutations. In addition, we report in this study the finding of a naturally occurring non-cyclic permuted protein between AML1/Core Binding Factor chain F and riboflavin synthase chain A.

Background

The classification of protein structures often rely on the topology of secondary structural elements. For example, the Structural Classification of Proteins (SCOP) system

classifies proteins structures into common folds using the topological arrangement of secondary structural units [1]. Most protein structural alignment methods can reliably classify proteins into similar folds given the structural units from each protein are in the same sequential order. However, the evolutionary possibility of proteins with different structural topology but with similar spatial arrangement of their secondary structures pose a problem. One such possibility is the circular permutation.

A circular permutation is an evolutionary event that results in the N and C terminus transferring to a different position on a protein. Figure 1[2] shows a simplified example of circular permutation. There are three proteins, all consist of three domains (A, B, and C). Although the spatial arrangement of the three domains are very similar, the ordering of the domains in the primary sequence has been circularly permuted. Lindqvist *et. al*. observed the first natural occurrence of a circular permutation between jackbean concanavalin A and favin [3]. Although the jackbean-favin permutation was the result of post-translational ligation of the N and C terminus and cleavage elsewhere in the chain, a circular permutation can arise from events at the gene level through gene duplication and exon shuffling.

Permutation by duplication [4,5] is a widely accepted model where a gene first duplicates and fuses. After fusion, a new start codon is inserted into one gene copy while a new stop codon is inserted into the second copy. Peisajovich *et al*. demonstrated the evolutionary feasibility of permutation via duplication by creating functional intermediates at each step of the *permutation by duplication model* for DNA methyltransferases [6]. Identifying structurally similar proteins with different chain topologies, including circular permutation, can aid studies in homology modeling, protein folding, and protein design. An algorithm that can structurally align two proteins independent of their backbone topologies would be an important tool.

The biological implications of thermodynamically stable and biologically functional circular permutations, both natural and artificial, has resulted in much interest in detecting circular permutations in proteins [3,7-11]. The more general problem of detecting non-topological structural similarities beyond circular permutation has received less attention. We refer to these as *non-cyclic permutations* from now on. Tabtiang *et al*. were able to create a thermodynamically stable and biologically functional non-cyclic permutation, indicating that non-cyclic permutations may be as important as circular permutations [12]. In this study, we present a novel method that detects spatially similar structures that can identify structures related by circular and more complex non-cyclic permutations. Detection of non-cyclic permutation is possible by our algorithm by virtue of a recursive combination of a local-ratio approach with a global linear-programming formulation. This paper is organized as follows. We first show that our algorithm is capable of finding known circular permutations with sensitivity and specificity. We then report the discovery of three new circular permutations and one example of a non-cyclic permutation that to our knowledge have not been reported in literature. We conclude with remarks and discussions.

Results and discussion

For availability of our alignment software please see [13].

Detection of known circular permutations

We first demonstrate the ability of our algorithm to detect circular permutations by examining known examples of circular permutations. The results are summarized in Table 1 and Table 2.

In Table 1 we compare results against DaliLite and K2. As expected, DaliLite returned the largest sequential alignment. K2 did not find circular permutations even when the option to ignore sequence order constraints was selected.

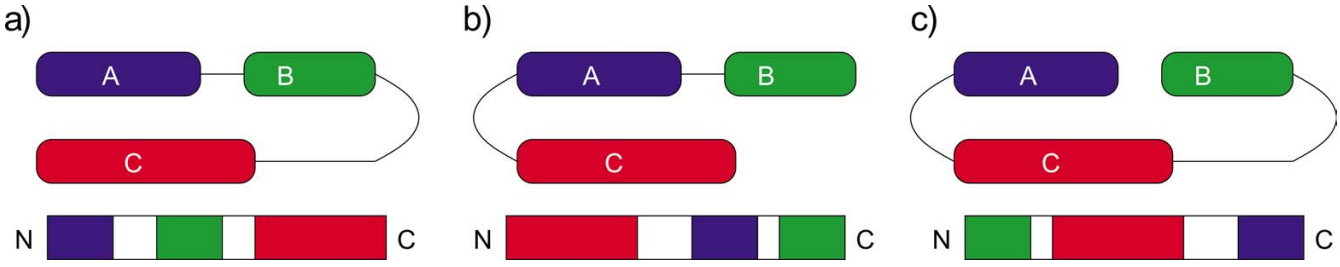

Figure 1

Circular permutation example. The cartoon illustration of three protein structures whose domains are similarly arranged in space but appear in different order in primary sequences. The location of domains A, B, C in primary sequences are shown in a layout below each structure. Their orderings are related by circular permutation [2].

Table 1: Known circular permutation results

Protein 1	Protein 2	Us			DaliLite		K2	
PDB(Length)	PDB(Length)	*N*	RMSD	*p – value*	*N*	RMSD	*N*	RMSD
1rinA(180)	2cna_(237)	**152***	**0.875**	10^{-6}	106	1.7	60	0.92
1glh_(214)	1cpn_(208)	**192***	**1.163**	10^{-5}	156	0.4	156	0.41
1exg_(110)	1tul_(102)	**74***	**1.485**	10^{-4}	63	4.0	34	2.26
1rhgA(145)	1bcfA(158)	**118***	**1.500**	10^{-4}	94	2.3	81	1.51
1ihwA(52)	1sso_(62)	**46***	**0.502**	10^{-3}	45	2.9	28	1.93

Comparison of results against DaliLite and K2. DaliLite is not expected to find sequence order independent alignments. K2 did not find the circular permutation even when the sequence order independent options was selected.

In Table 2 we compare our alignment results to the methods of MASS [14], OPAAS [15], SAMO [7], and Topofit [16]. Each method is able to detect circular permutations. However, Table 2 shows that our method normally finds more equivalent residues with a lower RMSD. Compared with SAMO our method found less aligned residues in 4 of the 5 shown alignments. However, our *cRMSD* values are considerably better. At the time of this writing, SAMO only outputs the *cRMSD* and the number of equivalent residues (N) of the alignment, without specifying the residue equivalence relationships between the two aligned protein structures. This makes it difficult to compare the quality of the alignments. Table 2 shows that our method finds better alignments in terms of cRMSD than other structural alignment methods when the two proteins are related by a circular permutation.

The GANSTA method by Kolbeck et al [17] can also align similar structures independent of the connectivity. The approach is somewhat similar to the Blast method in sequence alignment, where a set of seeds of high-similarity pairs of secondary structural elements (SSE) are first identified, and are then aligned through a genetic algorithm, regardless of the connectivity.

The SCALI method by Yuan and Bystroff [18] assembles from a library of gapless alignment of fragments of local sequence-structure hierarchically, enforcing compactness and conserved contacts, but disregard the sequence ordering of the fragments. The aligned local fragments are then incremented by adding a new fragment pair. This process is organized as a tree, where nodes corresponds to the addition of new fragments. A breadth-first tree search method was then carried out, with a number of heuristic conditions to limit the search space.

Instead of only aligning regular SSE fragments, our method differs from GANSTA and has no restriction on spatial patterns belonging to a regular SSE, and therefore is also applicable to loop regions. Our method differs from SCALI in that our fragments are not prebuilt, but are exhaustive fragments ranging from size 4–7. Compared to both methods, our method provides a guaranteed optimal ratio of aligned structures, whereas the heuristics employed by GANSTA and SCALI cannot guarantee that a good alignment can be found, and when an alignment is found, there is no guarantee that it will be within a certain ratio of the best possible alignment. In practice, we find that GANSTA often requires 3–5 hours for aligning a pair of proteins, and sometimes no results are returned. In contrast, our method usually terminates between 30 seconds – 5 minutes. The SCALI website consists of pre-computed results of aligned structures and does not allow user input for a customized alignment, therefore it is difficult to compare performance of our method with SCALI on the examples reported in Table 2.

Discovery of novel circular permutations and a novel non-cyclic permutation

The effectiveness of our method is also demonstrated by the discovery of previously unknown circular permutations. In an attempt to test our algorithm's ability to discover new circular permutations, we structurally aligned a subset of 3,336 structures from PDBSELECT 90% [19]. We first selected proteins from PDBSELECT90 (sequences have less than 90% identities) whose N and C termini were no further than 30Å apart. From this subset of 3,336 proteins, we aligned two proteins if they met the following conditions: the difference in their lengths was no more than 75 residues, and they had approximately the same secondary structure content. To compare secondary structure content, we determined the percentage of the residues labeled as helix, strand, and other for each structure. Two structures were considered to have the same secondary structure content if the difference between each secondary structure label was less than 10%. Within the approximately 200,000 alignments, we found 426 candidate circular permutations. Of these circular permutations, 312 were symmetric proteins that can be aligned with or without a circular permutation. Of the 114 non-symmetric circular permutations, 112 were already known in literature, and 3 are novel. We describe these three novel circular permutations as well as a novel non-cyclic permutation in some details.

Table 2: Known circular permutation results

Protein 1	Protein 2	Us			MASS		OPAAS		SAMO		Topofit	
PDB(Length)	PDB(Length)	*N*	*R*	*p*	*N*	*R*	*N*	*R*	*N*	*R*	*N*	*R*
1rinA(180)	2cna_(237)	**152***	**0.875**	10^{-6}	164*	1.2	167*	1.48	174*	1.581	152*	1.09
1glh_(214)	1cpn_(208)	**192***	**1.163**	10^{-5}	206*	0.49	No	solution	170*	3.283	206*	0.49
1exg_(110)	1tul_(102)	**74***	**1.485**	10^{-4}	60*	1.9	No	solution	93*	2.88	52*	1.79
1rhgA(145)	1bcfA(158)	**118***	**1.500**	10^{-4}	106*	1.7	63*	2.12	126*	2.309	109*	1.4
1ihwA(52)	1sso_(62)	**46***	**0.502**	10^{-3}	39*	1.7	No	solution	48*	2.713	35*	1.47

Comparison of our alignment results with that of MASS, OPAAS, SAMO, and Topofit for known circular permutations. Each method detected the circular permutations. Our method normally returns more equivalent residues at a lower RMSD. *N* indicates the number of aligned residues. An * next to the number of aligned residues indicates that a circular permutation was found. *R* indicates the cRMSD of the alignment. *p* indicates the *p*-value of our alignment.

Nucleoplasmin-core and auxin binding protein

The first novel circular permutation we found was between the nucleoplasmin-core protein in *Xenopu laevis* (PDB ID 1k5j, chain E) [20] and the auxin binding protein in maize (PDB ID 1lrh, chain A, residues 37 through 127) [21]. The overall structural alignment between 1k5jE (Figure 2a, top) and 1lrhA (Figure 2a, bottom) has an RMSD value of 1.36Å with an alignment length of 68 residues and a significant *p*-value of 2.7×10^{-5} after Bonferroni correction. These proteins are related by a circular permutation. The short loop connecting two antiparallel strands in nucleoplasmin-core protein (in ellipse, top of Fig 2b) becomes disconnected in auxin binding protein 1 (in ellipse, bottom of Fig 2b), and the N- and C- termini of the nucleoplasmin-core protein (in square, top of Fig 2b) are connected in auxin binding protein 1 (square, bottom of Fig 2b).

Aspartate racemase and type II 3-dehydrogenate dehyrdalase

Another circular permutation we found is between the aspartate racemase (PDB ID 1iu9, chain A) [22] and type II 3-dehydrogenate dehydralase (PDB ID 1h0r, chain A) [23]. The overall structural alignment between 1iu9A (Figure 3a, top) and 1h0rA (Figure 3a, bottom) has an RMSD value of 1.49Å with an alignment length of 59 residues and a significant *p*-value of 4.7×10^{-4} after Bonferroni correction. These proteins are related by a circular permutation. The loop connecting the first helix with the first strand in aspartate racemase (in rectangle, top of Figure 3b) becomes disconnected in 3-dehydrogenate dehydralase (in rectangle, bottom), while the N- and C-termini of the aspartate racemase (in ellipse, top) are connected in the dehydralase by an insertion (shown in green) (Figure 3b, bottom). Figure 3c depicts the topology of these two proteins.

Migration inhibition factor and arginine repressor

The majority of circular permutations maintain their overall three dimensional structures. However, it is possible that additional structural changes may occur beyond circular permutation. We have discovered a novel circular permutation between the microphage migration inhibition factor (MIF, PDB ID 1uiz, chain A, from *Xenopus laevis*) and the C-terminal domain of arginine repressor (AR, 1xxa, chain C, from *Escherichia coli*) [24,25], which contains in addition to circular permutation a spatial swapping of two antiparallel strands, and a change in the orientation of a helix. The overall folds of these two protein are different by the SCOP definition. The MIF factor belongs to the tautomerase fold, and the C-terminal domain of arginine repressor belongs to the DCoH-like fold. The overall structural alignment between 1uiz chain A (Figure 4a, top) and 1xxa chain C (Figure 4a, bottom) has an RMSD value of 1.74Å between 24 residues, with a *p*-value of 1.3×10^{-2} after Bonferroni correction. They are related by a circular permutation. The short loop of MIF (Figure 4b, top, in rectangle) connecting the first helix and the second strand from the N-terminus becomes disconnected in arginine repressor (AR, Figure 4a, bottom, in rectangle). The relaxing of spatial constraints imposed by the connection allows strand 1 of MIF to swap positions with strand 4 of MIF. This can also be clearly seen in Figure 4a, where a strand colored in red (strand 2' in AR, corresponding to strand 4 in MIF) swaps position with the strand colored in blue (strand 4' in AR, corresponding to strand 1 in MIF). Although the strands have changed positions spatially, their topology remains the same (Figure 4c and 4d). The circular permutation and strand swapping cause additional structural changes. In MIF, helix 1 was connected with a short loop to strand 2 (Figure 4b, top, in rectangle). With the creation of the new N- and C-termini replacing the original short loop (Figure 4b, bottom, rectangle), helix 1 loses the spatial constraints imposed by the connection, and was pulled over when strand 1 and strand 4 swap positions. The net results for helix 1 is that its orientation in arginine repressor (Figure 4b, bottom) is almost perpendicular to its original orientation in MIF (Figure 4b, top).

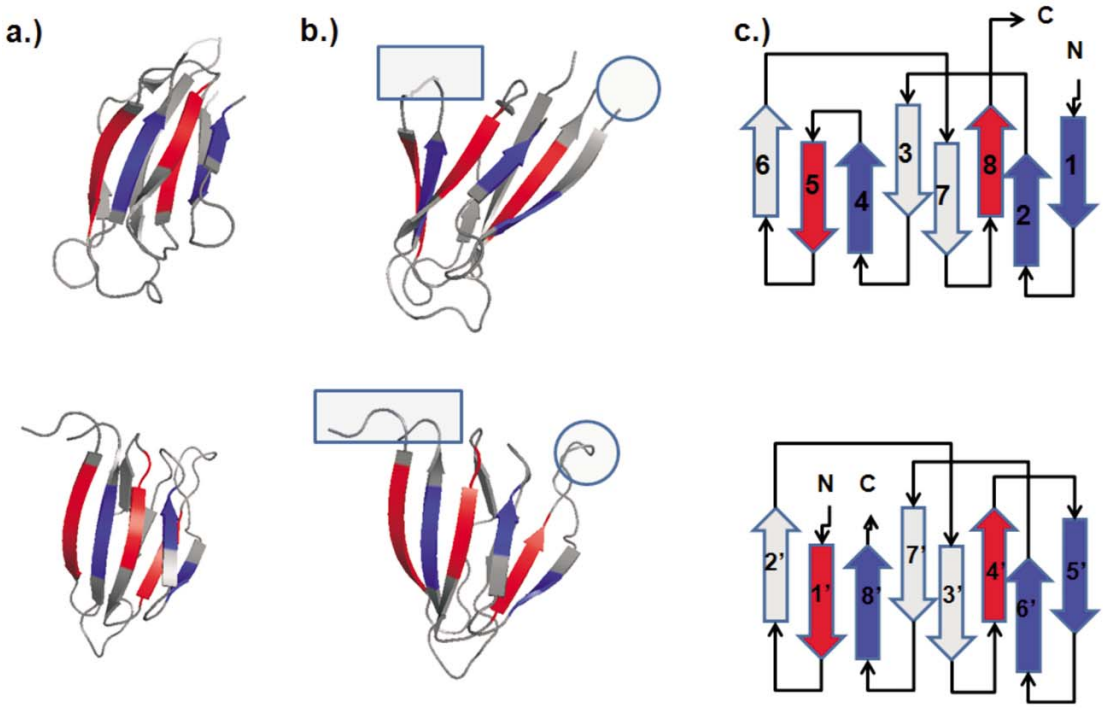

Figure 2

Nucleoplasmin-core and auxin binding protein 1. A new circular permutation discovered between nucleoplasmin-core (1k5j, chain E, top panel), and the fragment of residues 37–127 of auxin binding protein 1 (1lrh, chain A, bottom panel). a) These two proteins superimpose well spatially, with an RMSD value of 1.36Å for an alignment length of 68 residues and a significant *p*-value of 2.7×10^{-5} after Bonferroni correction. b) These proteins are related by a circular permutation. The short loop connecting strand 4 and strand 5 of nucleoplasmin-core (in rectangle, top) becomes disconnected in auxin binding protein 1. The N- and C- termini of nucleoplasmin-core (in ellipse, top) become connected in auxin binding protein 1 (in ellipse, bottom). For visualization, residues in the N-to-C direction before the cut in the nucleoplasmin-core protein are colored red, and residues after the cut are colored blue. c) The topology diagram of these two proteins. In the original structure of nucleoplasmin-core, the electron density of the loop connecting strand 4 and strand 5 is missing.

Beyond circular permutation

The information that naturally occurring circular permutations contain about the folding mechanism of proteins has led to a lot of interest in their detection. Another interesting class of permuted proteins is the non-cyclic permutation. Although there has been previous work on the detection of non-cyclic permutations [14-16,26], compared to cyclic-permutations there has been relatively little research of noncyclic-permutations. As an example of this important class of topologically permuted proteins, Tabtiang *et al* (2004) were able to artificially create a non-cyclic permutation of the Arc repressor that was thermodynamically stable, refolds on the sub-millisecond time scale, and binds operator DNA with nanomolar affinity [12]. This raises the question of whether or not these non-cyclic permutations can arise naturally. Here we report the discovery of a *possibly* naturally occurring non-cyclic permutation between chain F of AML1/Core Binding Factor (AML1/CBF, PDB ID 1e50, Figure 5, top) and chain A of riboflavin synthase (PDB ID 1pkv, Figure 5a, bottom) [27,28]. The two structures align well with a RMSD of 1.23 Å with an alignment length of 42 residues, and a significant *p*-value of 2.8×10^{-4} after Bonferroni correction. The topology diagram of AML1/CBF (Figure 5b) can be transformed into the topology diagram of riboflavin synthase (Figure 5f) by the following steps: Remove the the loops connecting strand 1 to helix 2, strand 4 to strand 5, and strand 5 to strand 6 (Figure 5c). Connect the C-terminal end of strand 4 to the original N-termini (Figure 5d). Connect the C-terminal end of strand 5 to the N-terminal end of helix 2 (Figure 5e). Connect the original C-termini to the N-terminal end of strand 5. The N-terminal end of strand 6 becomes the new N-termini and the C-terminal end of strand 1 becomes the new C-termini (Figure 5f).

Algorithm comparison

Zhu *et al* (2005) demonstrated the quality of their structural alignment algorithm (FAST [29]) by comparing their alignments with manually curated alignments in the HOMSTRAD database [30]. As of March 2007, HOM-

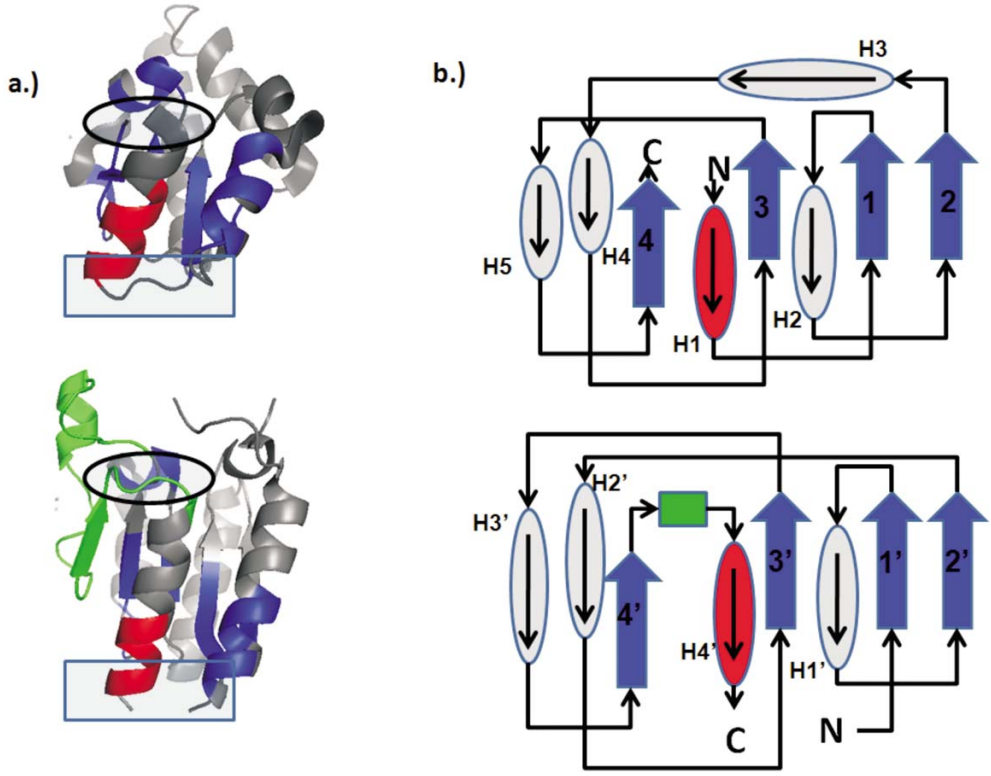

Figure 3

Aspartate racemase and type II 3-deydrogenate dehyralase. A new circular permutation discovered between a) aspartate racemase (1iu9, chain A, top) and type II 3-dehydrogenate dehydralase (1h0r, chain A, bottom) superimpose well spatially with an RMSD of 1.49Å between 59 residues, with a significant *p*-value of 4.7×10^{-4}. b) These proteins are related by a circular permutation. The loop connecting helix 1 with strand 1 in aspartate racemase (in rectangle, top) becomes disconnected in type II 3-dehydrogenate dehydralase (in rectangle, bottom), but the N- and C- termini of aspartate racemase (in ellipse, top) becomes connected in dehydrogenate dehydralase (in ellipse, bottom) with an insertion (shown in green). For visualization, residues of aspartate racemase in the N-to-C direction before the cut in the dehydrogenate dehydralase are colored red, and residues after the cut are colored blue. c) The topology diagram of these two proteins. Here an ellipse represents a helix and a block arrow represents a strand.

STRAD contains 3,454 proteins structures in 1,032 families. We randomly chose 10 structures from families that consisted of more than 20 protein structures. Within each family, we compared the structures using our alignment method to determine accuracy. Within alignments, our method's predicted equivalent residues agreed with HOMSTRAD 93% of the time. Discrepancies occur normally when our method would shift a fragment pair by one or two residues along the backbone. Zhu *et al*. chose 11 representatives from different structural classes as examples (Table IV of [29]). Table 3 is a representation of Table IV from [29] comparing our results with that of FAST [29] and DaliLite [31]. In all alignments, our method found sequentially ordered alignments, therefore, there is no bias in favor of our sequence order independent method. It can be seen from Table 3 that the equivalent residues that our method predicts are consistent with the manually determined residues of HOMSTRAD.

Conclusion

The approximation algorithm introduced in this work can find good solutions for the problem of protein structure alignment. Furthermore, this algorithm can detect topological differences between two spatially similar protein structures. The alignment between MIF and the arginine repressor demonstrates our algorithm's ability to detect structural similarities even when spatial rearrangement of structural units has occurred. In addition, we report in this study the finding of a naturally occurring non-cyclic permuted protein between AML1/Core Binding Factor chain F and riboflavin synthase chain A.

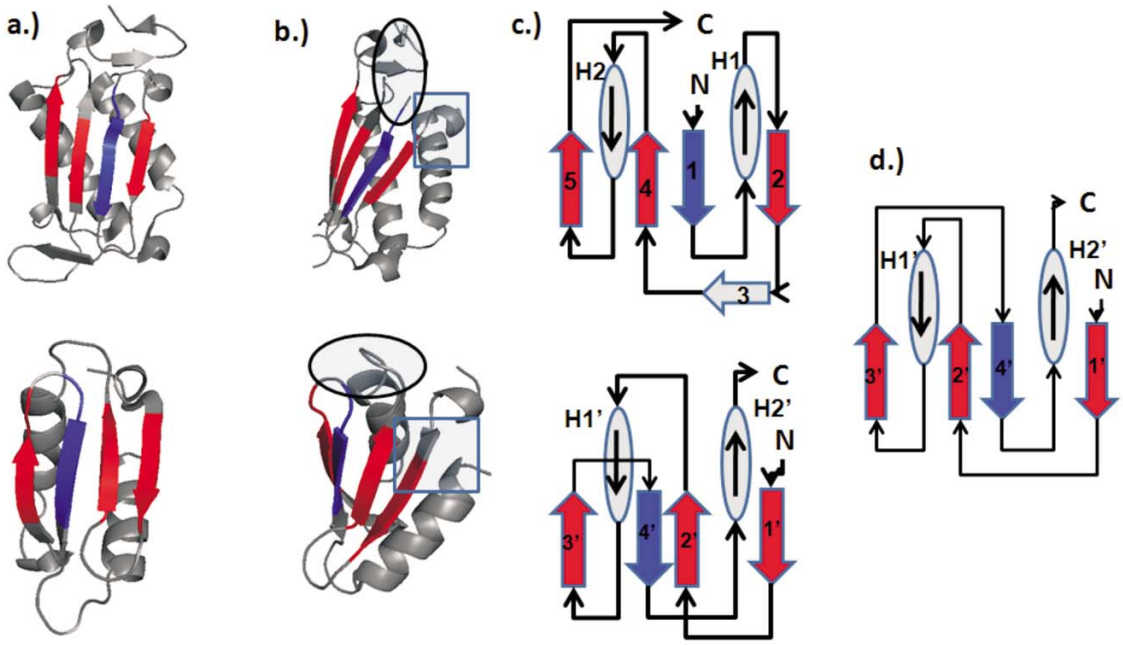

Figure 4

Microphage migration inhibition factor and C-terminal domain of arginine repressor. A new circular permutation discovered between a) the microphage migration inhibition factor (MIF, PDB ID 1uiz, chain A, top) and the C-terminal domain of arginine repressor (AR, 1xxa, chain C, bottom). a) These two proteins superimpose well spatially, with a RMSD of 1.74Å for an alignment length of 24 residues, and a *p*-value of 1.3×10^{-2}. b.) These proteins are related by a circular permutation. The loop connecting helix 1 with strand 2 of MIF (in rectangle, top) becomes disconnected in arginine repressor, the N- and C- termini of MIF (in ellipse, top) becomes connected in arginine repressor (in ellipse, bottom). The disconnection of helix 1 from strand 2 of MIF removes some spatial constraints, allowing strand 1' in AR to swap places with strand 4'. c) The topology diagram of these two proteins. d.) The artificial topology diagram for arginine repressor, where strand 2' and strand 4' are spatially swapped back. The diagram for AR in (c) has the same topology as the diagram in (d).

In our method, the scoring function plays a pivotal role in detecting substructure similarity of proteins. We expect future experimentation on optimizing the parameters used in our similarity scoring system can improve detection of topologically independent structural alignment. In this study, we were able to fit our scoring system to an Extreme Value Distribution (EVD), which allowed us to perform an automated search for circular permuted proteins. Although the *p*-value obtained from our EVD fit is sufficient for determining the biological significance of a structural alignment, the structural change between the microphage migration inhibition factor and the C-terminal domain of arginine repressor (Figure 3) indicates a need for a similarity score that does not bias heavily towards cRMSD measure for scoring circular permutations.

Whether naturally occurring circular permutations are frequent events in the evolution of protein genes is currently an open question. Lindqvist *et al*. (1997) pointed out that when the primary sequences have diverged beyond recognition, circular permutations may still be found using structural methods [3]. In this study, we discovered three examples of novel circularly permuted protein structures and a non-cyclic permutation among 200,000 protein structural alignments for a set of non-redundant 3,336 proteins. This is an incomplete study, as we restricted our studies to proteins whose N- and C- termini distance were less than 30Å. We plan to relax the N to C distance and include more proteins in future work to expand the scope of the investigation.

Methods

Approach

In this study, we describe a new algorithm that can align two protein structures or substructures independent of the connectivity of their secondary structure elements. We first exhaustively fragment the two proteins separately. An approximation algorithm based on a fractional version of the local-ratio approach for scheduling split-interval graphs [32] is then used to search for the combination of

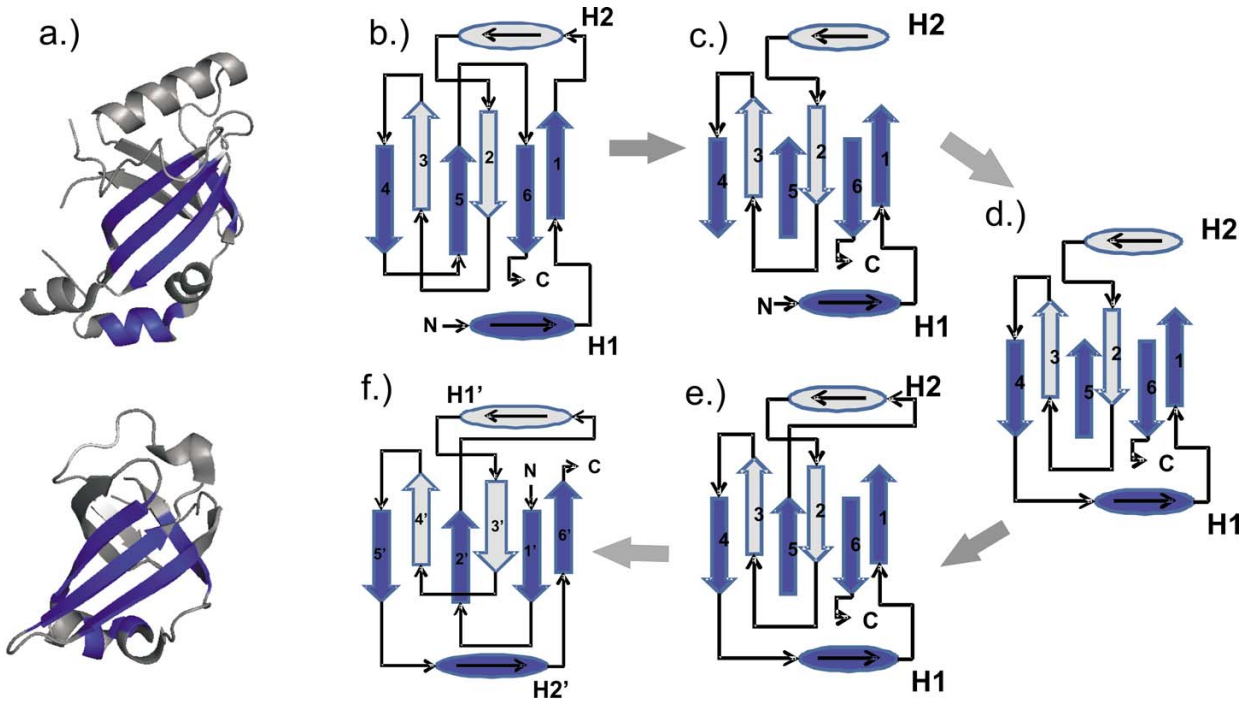

Figure 5

A non-cyclic permutation. A novel non-cyclic permutation discovered between AML1/Core Binding Factor (AML1/CBF, PDB ID 1e50, Chain F, top) and riboflavin synthase (PDBID 1pkv, chain A, bottom) a) These two proteins superimpose well spatially, with an RMSD of 1.23 Å and an alignment length of 42 residues, with a significant *p*-value of 2.8×10^{-4} after Bonferroni correction. Aligned residues are colored blue. b) These proteins are related by multiple permutations. The steps to transform the topology of AML1/CBF (top) to riboflavin (bottom) are as follows: c) Remove the the loops connecting strand 1 to helix 2, strand 4 to strand 5, and strand 5 to helix 6; d) Connect the C-terminal end of strand 4 to the original N-termini; e) Connect the C-terminal end of strand 5 to the N-terminal end of helix 2; f) Connect the original C-termini to the N-terminal end of strand 5. The N-terminal end of strand 6 becomes the new N-termini and the C-terminal end of strand 1 becomes the new C-termini. We now have the topology diagram of riboflavin synthase.

Table 3: Alignment quality

Proteins		HOMSTRAD		FAST			US		
PDB(PDB)	PDB(PDB)	N	RMSD	N	M%	RMSD	N	M%	RMSD
1dfaA	1qceA	57	2.5	55	55%	1.2	**45**	**72%**	**1.1**
1hx8A	1hg5A	258	1.1	255	99%	1.1	**247**	**98%**	**1.0**
2ahjA	1rieA	192	4.3	187	89%	2.0	**168**	**99%**	**1.3**
1h7sA	1b63A	105	2.2	98	99%	2.0	**96**	**100%**	**1.9**
1ed9A	1ew2A	403	5.6	343	98%	1.7	**252**	**100%**	**1.2**
1oyc_	2tmdA	330	3.6	284	97%	2.3	**193**	**94%**	**1.4**
1fjnA	1ica_	33	4.7	28	100%	1.9	**33**	**100%**	**1.8**
1tpn_	1fbr_	43	2.4	40	93%	2.2	**39**	**97%**	**2.2**
1e12A	1c3wA	220	1.7	214	97%	1.5	**170**	**100%**	**0.9**
1af6A	1a0tP	377	4.6	323	97%	1.8	**281**	**97%**	**1.5**
1hc1_	1lla_	582	2.3	546	97%	1.7	**380**	**100%**	**1.4**

Table IV from Zhu *et al*. (2005) with the addition of our alignment results. Zhu *et al*. chose the following alignment examples to cover a broad range of structural classes. For each alignment, our method returned sequence ordered alignments. N is the number of aligned residues corresponding to each method and M% is the number of aligned residues generated by the corresponding algorithm that are equivalent to HOMSTRAD's aligned residues.

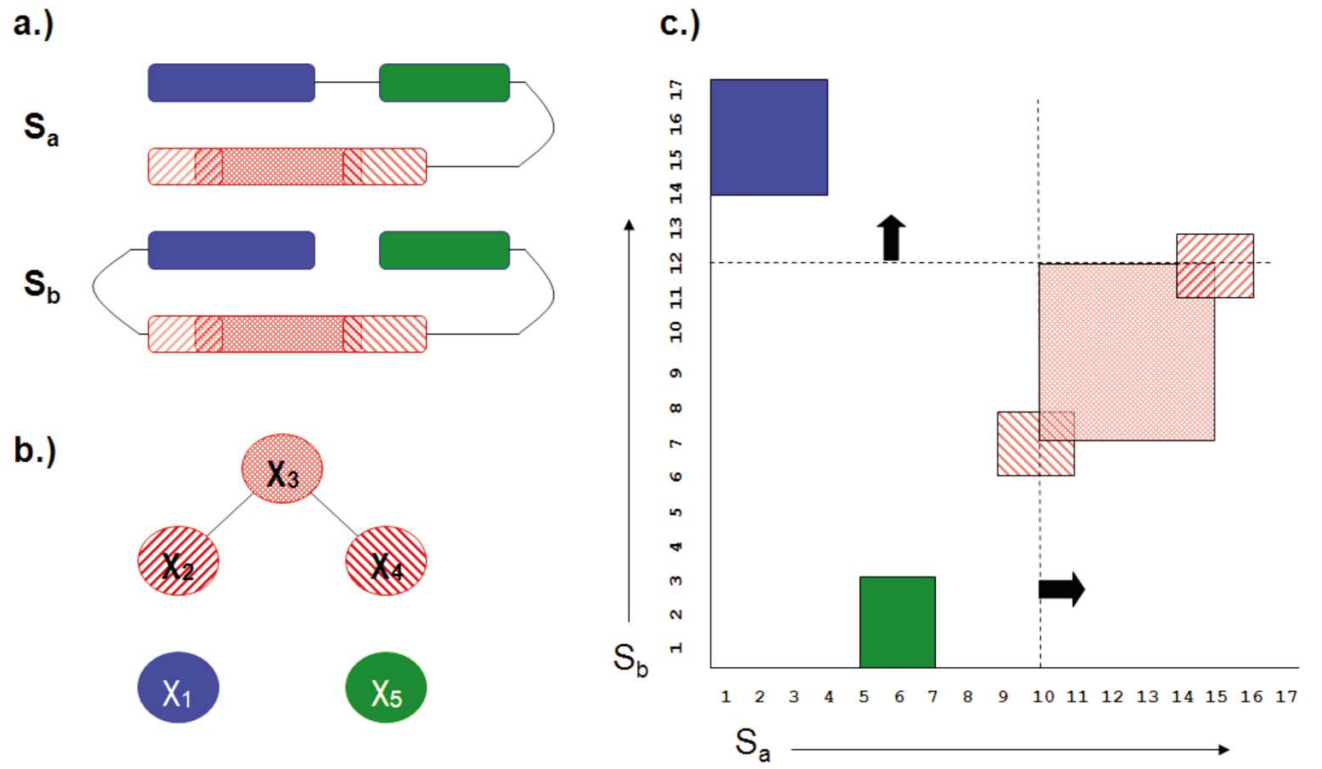

Figure 6

Implementation example with vertex sweep. An illustration of the first iteration of our algorithmic approaches for $BSSI_{\Lambda,\sigma}$: a) The cartoon representation of circularly permuted proteins S_a and S_b; b) The problem represented as a graph where each node $\chi_i \in \Lambda$ represents an aligned fragment pair and each edge represents two inconsistent pairs; c) An illustration how sweep lines (dashed) can identify inconsistent aligned pairs as required to generate the interval clique inequalities. A rectangle is an ordered fragment pair (e.g., the solid green rectangle is the pair $\chi_5 = (\lambda^a_{5,3}, \lambda^b_{1,3})$).

peptide fragments from both structures that will optimize the global alignment of the two structures.

The methods discussed here do resemble the methods in our previous conference paper [2]. However, they are similar because they both use the same approximation algorithm used for scheduling split interval graphs that appears in [32]. Beyond the approximation algorithm for scheduling split-interval graphs, the methods are different. Figure 1 does appear in our previous conference paper [2]. However, Figure 6 and Table 4 are different due to errors in the corresponding figure and table in that previous paper. Also, note that the previous conference version [2] had a recursive formulation of the algorithm as opposed to the *non-recursive* formulation as described in this paper. There are other differences too, including significant improvements/corrections of notations.

Basic Definitions and Notations

The following definitions/notations are used uniformly throughout the paper unless otherwise stated:

- Protein structures are denoted by S_a, S_b,....

- A substructure $\lambda^a_{i,k}$ of a protein structure S_a is a continuous fragment $\lambda^a_{i,k}$, where i is the residue index of the beginning of the substructure and k is the length (number of residues) of the substructure. We will denote such a substructure simply by λ^a if i and k are clear from the context or irrelevant.

- A residue $a_t \in S_a$ is a *part* of a substructure $\lambda^a_{i,k}$ if $i \le t \le i + k - 1$.

Table 4: Constraints

Interval clique inequalities:	**(2)**
	Line sweep at a_t = 1
$y_{\chi_{5,\lambda_a}} \leq 1$	
	Line sweep at a_t = 5
$y_{\chi_{1,\lambda_a}} \leq 1$	
	Line sweep at a_t = 9
$y_{\chi_{4,\lambda_a}} \leq 1$	
	Line sweep at a_t = 10
$y_{\chi_{3,\lambda_a}} + y_{\chi_{4,\lambda_a}} \leq 1$	
	Line sweep at a_t = 12
$y_{\chi_{3,\lambda_a}} \leq 1$	
	Line sweep at a_t = 14
$y_{\chi_{3,\lambda_a}} + y_{\chi_{2,\lambda_a}} \leq 1$	
	Line sweep at a_t = 16
$y_{\chi_{2,\lambda_a}} \leq 1$	
Interval clique inequalities:	**(3)**
	Line sweep at b_t = 1
$y_{\chi_{1,\lambda_b}} \leq 1$	
	Line sweep at b_t = 6
$y_{\chi_{4,\lambda_b}} \leq 1$	
	Line sweep at b_t = 7
$y_{\chi_{4,\lambda_b}} + y_{\chi_{3,\lambda_b}} \leq 1$	
	Line sweep at b_t = 9
$y_{\chi_{3,\lambda_b}} \leq 1$	
	Line sweep at b_t = 12
$y_{\chi_{2,\lambda_b}} + y_{\chi_{3,\lambda_b}} \leq 1$	
	Line sweep at b_t = 13
$y_{\chi_{2,\lambda_b}} \leq 1$	
	Line sweep at b_t = 14
$y_{\chi_{5,\lambda_b}} \leq 1$	
Consistency inequalities:	**(4,5)**
$y_{\chi_{1,\lambda_a}} - x_{\chi_1} \geq 0$	$y_{\chi_{1,\lambda_b}} - x_{\chi_1} \geq 0$
$y_{\chi_{2,\lambda_a}} - x_{\chi_2} \geq 0$	$y_{\chi_{2,\lambda_b}} - x_{\chi_2} \geq 0$
$y_{\chi_{3,\lambda_a}} - x_{\chi_3} \geq 0$	$y_{\chi_{3,\lambda_b}} - x_{\chi_3} \geq 0$
$y_{\chi_{4,\lambda_a}} - x_{\chi_4} \geq 0$	$y_{\chi_{4,\lambda_b}} - x_{\chi_4} \geq 0$
$y_{\chi_{5,\lambda_a}} - x_{\chi_5} \geq 0$	$y_{\chi_{5,\lambda_b}} - x_{\chi_5} \geq 0$

The constraints of the conflict graph for the set of fragments in Figure 6c.

• Λ_a is the set of all continuous substructures or fragments of protein structure S_a that is under consideration in our algorithm.

• $\chi_{i,\,j,\,k}$ (or simply χ when the other parameters are understood from the context) denotes an ordered pair $(\lambda^a_{i,k}, \lambda^b_{j,k})$ of equal length substructures of two protein structures S_a and S_b.

• Two ordered pairs of substructures $(\lambda^a_{i,k}, \lambda^b_{j,k})$ and $(\lambda^a_{i',k'}, \lambda^b_{j',k'})$ are called *inconsistent* if and only if at least one of the pairs of substructures $\{\lambda^a_{i,k}, \lambda^a_{i',k'}\}$ and $\{\lambda^a_{j,k}, \lambda^a_{j',k'}\}$ are not disjoint.

We are now ready to formalize our substructure similarity identification problem as below:

Problem name: Basic Substructure Similarity Identification ($\text{BSSI}_{\Lambda,\,\sigma}$).

Instance: a set $\Lambda = \{\chi_{i,\,j,\,k} | i,\, j,\, k \in \} \subset \Lambda_a \times \Lambda_b$ of ordered pairs of equal length substructures of S_a and S_b and a similarity function $\sigma : \Lambda \to \mathbb{R}^+$ mapping each pair of substructures to a positive similarity value.

Valid Solutions: a set of substructure pairs $\{\chi_{i_1,j_1,k_1}, \chi_{i_2,j_2,k_2}, \ldots \chi_{i_t,j_t,k_t}\}$ that are mutually consistent.

Objective: *maximize* the total similarity of the selection $\sum_{\ell=1}^{t} \sigma(\chi_{i_\ell,j_\ell,k_\ell})$.

An Algorithm Based on the Local-Ratio Approach

The $BSSI_{\Lambda,\,\sigma}$ problem is a special case of the well-known maximum weight independent set problem in graph theory [33]. In fact, $BSSI_{\Lambda,\,\sigma}$ itself is MAX-SNP-hard (i. e., there is a constant $0 < \varepsilon < 1$ such that no polynomial-time algorithm can return a solution with a value of the objective function that is within 1 - ε times the optimum [34] unless P = NP) even when all the substructures are restricted to have lengths at most 2 [32, Theorem 2.1]. Our approach is to adopt the approximation algorithm for scheduling split-interval graphs [32] which itself is based on a fractional version of the local-ratio approach. For ease in description of our algorithm, we introduce the following definitions.

Definition 1 *For any subset,* $\Delta \subseteq \Lambda$ *the conflict graph* $G_\Delta = (V_\Delta, E_\Delta)$ *is the graph in which* $V_\Delta = \{\chi | \chi \in \Delta\}$ *and* $E_\Delta = \{\{\chi, \chi'\} | \chi, \chi', \in \Delta$ *and the pair* $\{\chi, \chi'\}$ *is not consistent*$\}$

Definition 2 *The closed neighborhood* $Nbr_\Delta[\chi]$ *of a vertex* χ *of* G_Δ *is* $\{\chi' \mid \{\chi, \chi'\} \in E_\Delta\} \cup \{\chi\}$.

For an instance of $BSSI_{\Lambda,\sigma}$ with $\Delta \subseteq \Lambda$ we introduce three types of indicator variables as follows. For every $\chi = (\lambda_a, \lambda_b) \in \Delta$, we introduce three indicator variables $x_{\chi'}$ $\gamma_{\chi\lambda_a}$ and $\gamma_{\chi\lambda_b} \in \{0,1\}$. x_χ indicates whether the substructure pair should be used ($x_\chi = 1$) or not ($x_\chi = 0$) in the final alignment. $\gamma_{\chi\lambda_a}$ and $\gamma_{\chi\lambda_b}$ are artificial selection variables for λ_a and λ_b that allows us to encode consistency in the selected substructures in a way that guarantees good approximation bounds. Our algorithm for solving an instance of $BSSI_{\Lambda,\sigma}$ can now be described as follows. We initialize $\Delta = \Lambda$. Then, the following algorithm is executed:

1. Solve a linear programming (LP) formulation of $BSSI_{\Lambda,\sigma}$ by relaxing a corresponding integer programming version of the $BSSI_{\Lambda,\sigma}$ problem.

maximize

$$\sum_{\chi \in \Delta} \sigma(\chi) \cdot x_\chi \qquad (1)$$

Subject to

$$\sum_{a_t \in \lambda^a \in \Lambda_a} \gamma_{\chi\lambda_a} \leq 1 \quad \forall a_t \in S_a \qquad (2)$$

$$\sum_{a_t \in \lambda^b \in \Lambda_b} \gamma_{\chi\lambda_b} \leq 1 \quad \forall a_t \in S_b \qquad (3)$$

$$\gamma_{\chi\lambda_a} - x_\chi \geq 0 \quad \forall \chi \in \Delta \qquad (4)$$

$$\gamma_{\chi\lambda_b} - x_\chi \geq 0 \quad \forall \chi \in \Delta \qquad (5)$$

$$x_\chi, \gamma_{\chi\lambda_a}, \gamma_{\chi\lambda_b} \geq 0 \quad \forall \chi \in \Delta \qquad (6)$$

2. For every vertex $\chi \in V_\Delta$ of G_Δ, compute its *local conflict number* $\alpha_\chi = \sum_{\chi' \in \mathrm{Nbr}_\Delta[\chi]} x_{\chi'}$. Let χ_{min} be the vertex with the *minimum* local conflict number. Define a new similarity function σ_{new} from σ as follows:

$$\sigma_{new}(\chi) = \begin{cases} \sigma(\chi) & \text{if } \chi \notin \mathrm{Nbr}_\Delta[\chi_{min}] \\ \sigma(\chi) - \sigma(\chi_{min}) & \text{otherwise} \end{cases}$$

3. Create $\Delta_{new} \subseteq \Delta$ by removing from Δ every substructure pair χ such that $\sigma_{new}(\chi) \leq 0$. Push each removed substructure on to a stack in arbitrary order.

4. If $\Delta_{new} \neq \emptyset$ then repeat from step 1 setting $\Delta = \Delta_{new}$ and $\sigma = \sigma_{new}$. Otherwise, continue to step 5.

5. Repeatedly pop the stack, adding the substructure pair to the alignment as long as the following conditions are met:

(a) The substructure pair is consistent with all other substructure pairs that already exist in the selection.

(b) The *cRMSD* of the alignment does not change by a threshold. This condition bridges the gap between optimizing a local similarity between substructures and optimizing the tertiary similarity of the alignment by guaranteeing that each substructure from a substructure pair is in the same spatial arrangement in the global alignment.

A brief intuitive explanation of the various inequalities in the LP formulation as described above in terms of their original integer programming formulation is as follows:

• The "interval clique" inequalities in Equation (2) (resp. Equation (3)) ensure that the various substructures of S_a (resp. S_b) in the selected substructure pairs from Δ are mutually disjoint.

• Inequalities in Equation 4 and Equation 5 ensure consistencies between the indicator variable for each substructure pair χ and its two substructures λ_a and λ_b.

• Inequalities in Equation 6 relax the 0–1 values of the indicator variables to any fractional value between 0 and 1.

In implementation, the graph G_Δ is considered implicitly via intersecting intervals. The interval clique inequalities can be generated via a *sweepline* approach (see Figure 6c). The running time depends on the number of iterations needed to solve the LP formulations. Let LP(n, m) denote the time taken to solve a linear programming problem on n variables and m inequalities. Then the worst case running time of the above algorithm is $O(|\Lambda| \cdot \mathrm{LP}(3|\Lambda|, 5|\Lambda| + |\Lambda_a| + |\Lambda_b|))$. However, the worst-case time complexity happens under the excessive pessimistic assumption that

each iteration removes exactly one vertex of G_Λ, namely χ_{min} only, from consideration, which is unlikely to occur in practice as our computational results show. A theoretical pessimistic estimate of the performance ratio of our algorithm can be obtained as follows. Let α be the maximum of all the $\alpha_{\chi_{min}}$'s over all iterations. Proofs in [32] translate to the fact that the algorithm returns a solution whose total similarity is *at least* $\frac{1}{\alpha}$ times that of the optimum and, if Step 5(b) is omitted from the algorithm, then $\alpha \leq 4$. The value of α even with Step 5(b) is much smaller than 4 in practice (*e.g.* $\alpha = 2.89$).

Simple example

We present a simplified example to illustrate the first iteration of our algorithmic approach for two protein structures S_a and S_b (Figure 6a) selected for alignment. Here S_b is the structure to be aligned to the reference structure S_a. We systematically cut S_b into fragments of length 4–7 and exhaustively compute a similarity score of each fragment from S_b to all possible fragments of equal length in S_a. Each fragment pair can be thought of as a vertex in a graph (Figure 6b). *Abusing notations slightly for ease of understanding*, let the vertices be denoted by vertex corresponds to a rectangle in Figure 6c. Suppose we have the following similarity scores for aligned substructures: $\sigma(\chi_1) = 8$, $\sigma(\chi_2) = 5$, $\sigma(\chi_3) = 7$, $\sigma(\chi_4) = 3$ and $\sigma(\chi_5) = 6$. Then, our objective function is to maximize

$8x_{\chi_1} + 5x_{\chi_2} + 7x_{\chi_3} + 3x_{\chi_4} + 6x_{\chi_5}$. Figure 6b shows the conflict graph for the set of fragments. A sweep line (shown as dashed lines in Figure 6c) is implicitly constructed ($O(n)$ time after sorting) to determine which vertices of fragment pairs overlap. A conflict is shown in Figure 6b as edges between vertices. χ_1 and χ_5 do not conflict with any other substructure pairs, while χ_2 and χ_4 conflict with χ_3. For this graph, the constraints in the linear programming formulation are shown in Table 4. The linear programming problem is solved using the BPMPD package [35].

Similarity Score σ

The similarity score $\sigma(\chi_{i,\ j,\ k})$ between two aligned substructures $\lambda^a_{i,k}$ and $\lambda^b_{j,k}$ is a weighted sum of a shape similarity measure derived from the *cRMSD* value, which is then modified for the secondary structure content of the aligned substructure pairs, and a sequence composition score (*SCS*). Here cRMSD values are the *coordinate root mean square distance*, which are the square root of the mean of squares of Euclidean distances of coordinates of corresponding C_α atoms. Formally, for two sets of n points **v** and **w**, the cRMSD is defined as $\sqrt{\sum_1^n \|v_i - w_i\|^2}$.

cRMSD scaling by secondary structure content

We scale the *cRMSD* according to the secondary structure composition of the two substructures (λ^a and λ^b) that compose the substructure pair χ. We extracted 1,000 α-helices of length 4–7 (250 of each length) at random from protein structures contained in PDBSELECT 25% [19]. We exhaustively aligned helices of equal length and obtained the *cRMSD* distributions shown in Figure 7(a–d). We then exhaustively aligned equal length β-strands (length 4–7) from a set of 1,000 (250 of each length) strands randomly extracted from protein structures in PDBSELECT 25% [19] and obtained the distributions shown in Figure 7(e–h). For each length, the mean *cRMSD* value of the strands is approximately two times larger than the mean RMSD of the helices. Therefore, we introduce the following empirical scaling factor

$$s(\lambda_a, \lambda_b) = \frac{\sum_{i=1}^{N} \delta(A_{a,i}, A_{b,i})}{N}$$

, where

$$\sigma(A_{a,i}, A_{b,i}) = \begin{cases} 2, & \text{if residues } A_{a,i} \text{ and } A_{b,i} \text{ are both helix,} \\ 1, & \text{otherwise,} \end{cases}$$

to modify the *cRMSD* of the aligned substructure pairs to remove bias due to different secondary structure content. We use DSSP [36] to assign secondary structure to the residues of each protein.

Sequence composition

The score for sequence composition *SCS* is defined as:

$$SCS = \sum_{i=1}^{k} B(A_{a,i}, A_{b,i}),$$

where $A_{a,\ i}$ and $A_{b,\ i}$ are the amino acid residue types at aligned position i. $B(A_{a,\ i}, A_{b,\ i})$ is the similarity score between $A_{a,\ i}$ and $A_{b,\ i}$ based on a modified BLOSUM50 matrix, in which a constant is added to all entries such that the smallest entry is 1.0.

Combined similarity score

The combined similarity score $\sigma(\chi)$ of two aligned substructures is calculated as follows:

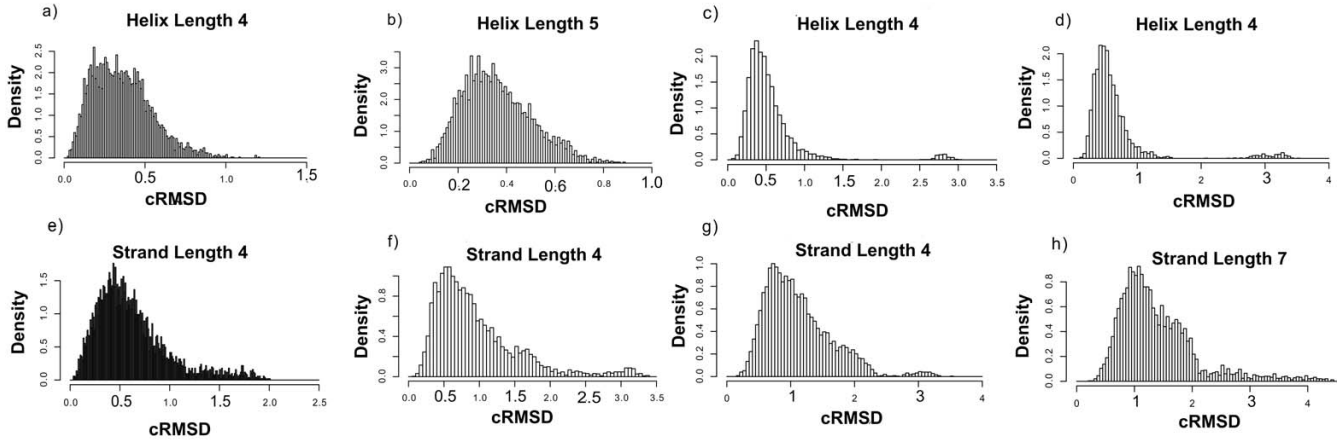

Figure 7

Secondary Structure cRMSD distributions. The cRMSD distributions of a) helices of length 4 b) helices of length 5 c) helices of length 6 d) helices of length 7 e) strands of length 4 f) strands of length 5 g) strands of length 6 and h) strands of length 7.

$$\sigma(\chi_{i,j,k}) = \alpha[C - s(\lambda_a, \lambda_b) \cdot \frac{cRMSD}{k^2}] + SCS, \qquad (7)$$

In current implementation, the values of α and C are empirically set to 100 and 2, respectively.

Similarity score for aligned molecules

The output of the above algorithm is a set of aligned substructure pairs $X = \{\chi_1, \chi_2, \ldots \chi_m\}$ that maximize Equation (1).

The alignment X of two structures is scored following Equation (7) by treating X as a single discontinuous fragment pair:

$$\sigma(X) = \alpha[C - s(X) \cdot \frac{cRMSD}{N_X^2}] + SCS. \qquad (8)$$

In this case $k = N_X$, where N_X is the total number of aligned residues.

To investigate the effect that the size of each the proteins being aligned has on our similarity score, we randomly

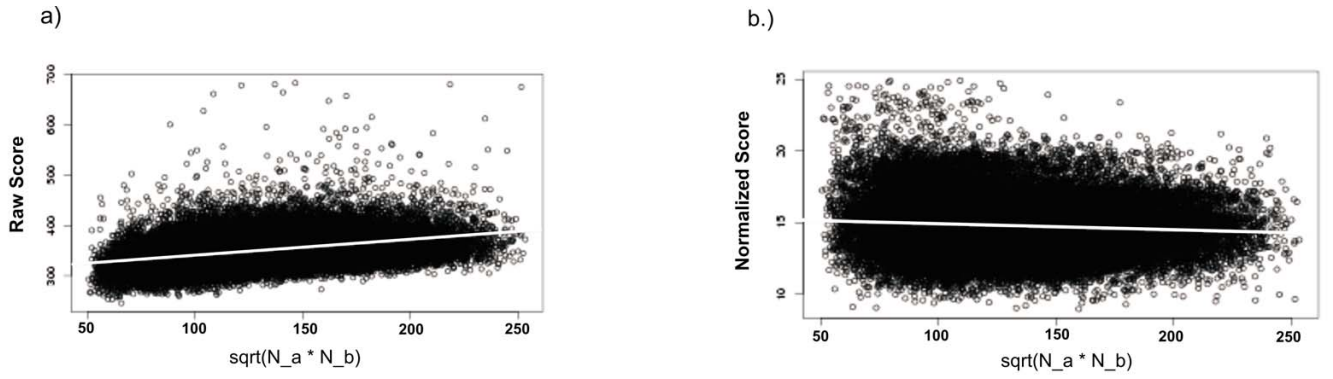

Figure 8

Similarity Score versus length. a) Linear fit between *raw similarity score* $\sigma(X)$ (equation 8) as a function of the geometric mean $\sqrt{N_a \cdot N_b}$ of the length of the two aligned proteins (N_a and N_b are the number of residues in the two protein structures S_a and S_b). The linear regression line (grey line) has a slope of 0.314. b) Linear fit of the normalized similarity score $\tilde{\sigma}(X)$ (equation 9) as a function of the geometric mean of the length of the two aligned proteins. The linear regression line (grey line) has a slope of -0.0004.

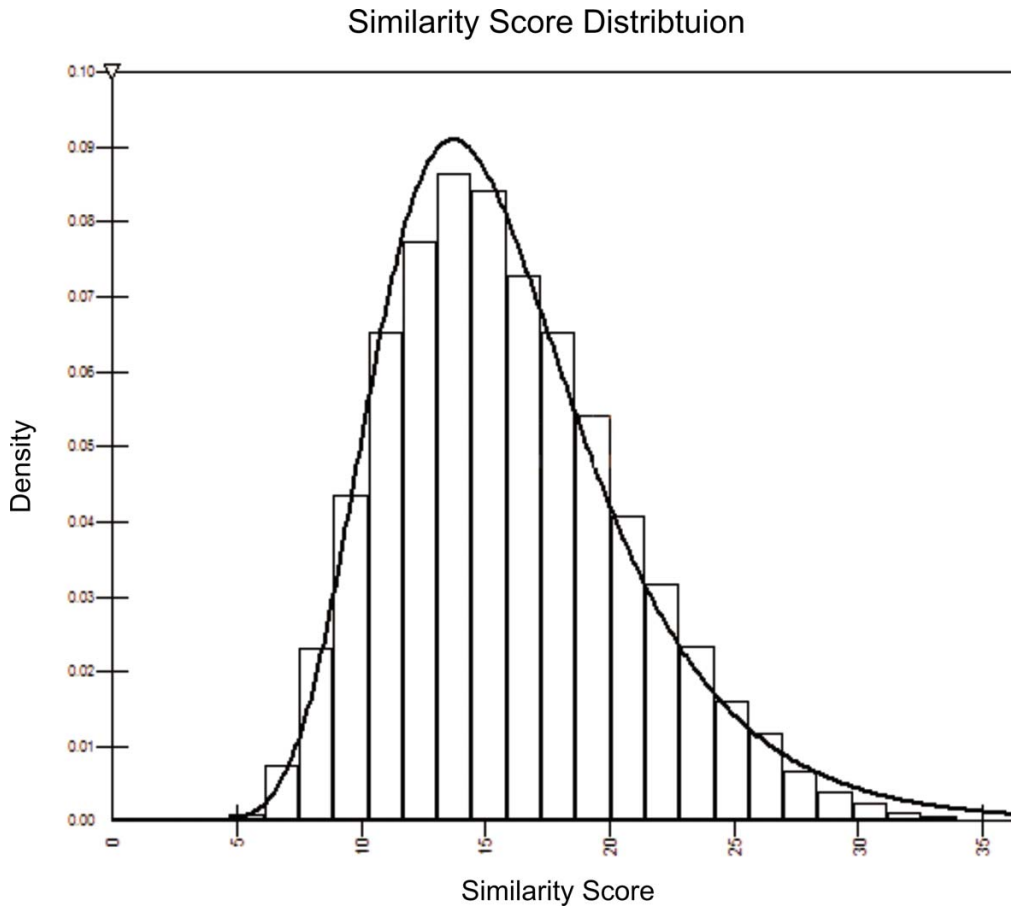

Figure 9
Similarity Score Distribution. The distribution of the normalized similarity scores obtained by aligning 200,000 pairs of proteins randomly selected from PDBSELECT 25% [19]. The distribution can be fit to an Extreme Value Distribution, with parameters α = 14.98 and β = 3.89.

aligned 200,000 protein pairs from PDBSELECT 25% [19]. Figure 8a shows the similarity scores $\sigma(X)$ (equation 8) as a function of the geometric mean of two aligned structure lengths $\sqrt{N_a \cdot N_b}$. Where N_a and N_b are the number of residues in S_a and S_b, respectively. The regression line (grey line) has a slope of 0.314, indicating that $\sigma(X)$ is not ideal for determining the significance of the alignment because larger proteins produce higher similarity scores. This is corrected by a simple normalization scheme:

$$\tilde{\sigma}(X) = \frac{\sigma(X)}{N_X}, \qquad (9)$$

where N is the number of equivalent residues in the alignment is used. Figure 8b shows the normalized similarity score as a function of the geometric mean of the aligned protein lengths. The regression line (grey line) has a negligible slope of -4.0×10^{-4}. In addition, the distribution of the normalized score $\tilde{\sigma}(X)$ can be approximated by an extreme value distribution (EVD) (Figure 9). This allows us to compute the statistical significance given the score of an alignment [37,38].

Authors' contributions

All authors contributed equally to this paper. All authors read and approved the final manuscript.

Acknowledgements

A shorter preliminary version of this paper was presented at the 7th Workshop on Algorithms in Bioinformatics (WABI) during September 2007. Bhaskar DasGupta was supported by NSF grants IIS-0346973, IIS-0612044 and DBI-0543365. Joe Dundas was partially supported by NSF grant IIS-0612044. Jie Liang was supported by NSF grant DBI-0133856 and NIH grants GM68958 and GM079804.

References

1. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structure.** *J Mol Biol* 1995, **247:**536-540.
2. Binkowski TA, DasGupta B, Liang J: **Order independent structural alignment of circularly permuted proteins.** *Conf Proc IEEE Eng Med Biol Soc* 2004, **4:**2781-2784.
3. Lindqvist Y, Schneider G: **Circular permutations of natural protein sequences: structural evidence.** *Curr Opinions Struct Biol* 1997, **7:**422-427.
4. Ponting CP, Russell RB: **Swaposins: circular permutations within genes encoding saposin homologues.** *Trends Biochem Sci* 1995, **20:**179-180.
5. Jeltsch A: **Circular permutations in the molecular evolution of DNA methyltransferase.** *J Mol Evol* 1999, **49:**161-164.
6. Peisajovich SG, Rockah L, Tawfik DS: **Evolution of new protein topologies through multistep gene rearrangements.** *Nature Genetics* 2006, **38:**168-173.
7. Chen L, Wu LY, Wang Y, Zhang S, Zhang XS: **Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison.** *BMC Struct Biol* 2006, **6:**18.
8. Chen L, Zhou T, Tang Y: **Protein structure alignment by deterministic annealing.** *Bioinformatics* 2005, **21:**51-62.
9. Szustakowski JD, Weng Z: **Protein structure alignment using a genetic algorithm.** *Proteins: Structure, Function, and Genetics* 2000, **38:**428-440.
10. Jung J, Lee B: **Protein structure alignment using environmental profiles.** *Prot Eng* 2000, **13:**535-543.
11. Uliel S, Fliess A, Amir A, Unger R: **A simple algorithm for detecting circular permutations in proteins.** *Bioinformatics* 1999, **15:**930-936.
12. Tabtiang RK, Cezairliyan BO, Grant RA, Cochrane JC, Sauer RT: **Consolidating critical binding determinants by noncyclic rearrangement of protein secondary structure.** *PNAS* 2004, **7:**2305-2309.
13. **CPalign: Software for topology independent protein structural alignment** [http://gila.bioengr.uic.edu/lab/]
14. Dror O, Benyamini H, Nussinov R, Wolfson HJ: **MASS: multiple structural alignment by secondary structures.** *Bioinformatics* 2003, **19:**i95-i104.
15. Shih ES, Hwang MJ: **Alternative alignments from comparison of protein structures.** *Proteins* 2004, **56:**519-527.
16. Ilyin VA, Abyzov A, Leslin CM: **Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point.** *Protein Science* 2004, **13:**1865-1874.
17. Kolbeck B, May P, Schmidt-Goenner T, Steinke T, Knapp EW: **Connectivity independent protein-structure alignment: a hierarchical approach.** *BMC Bioinformatics* 2006, **7:**510.
18. Yuan X, Bystroff C: **Non-sequential structure-based alignments reveal topology independent core packing arrangements in proteins.** *Bioinformatics* 2005, **21(7):**1010-1019.
19. Hobohm U, Sander C: **Enlarged representative set of protein structures.** *Protein Science* 1994, **3:**522.
20. Dutta S, Akey IV, Dingwall C, Hartman KL, Laue T, Nolte RT, Head JF, Akey CW: **The crystal structure of nucleoplasmin-core implication for histone binding and nucleosome assembly.** *Mol Cell* 2001, **8:**841-853.
21. Woo EJ, Marshall J, Bauly J, Chen JG, Venis M, Napier RM, Pickersgill RW: **Crystal structure of the auxin-binding protein 1 in complex with auxin.** *EMBO J* 2002, **21:**2877-2885.
22. Liu L, Iwata K, Yohda M, Miki K: **Structural insight into gene duplication, gene fusion and domain swapping in the evolution of PLP-independent amino acid racemases.** *FEBS LETT* 2002, **528:**114-118.
23. Hermoso JA, Monterroso B, Albert A, Galan B, Ahrazem O, Garcia P, Martinez-Ripoll M, Garcia JL, Menendez M: **Structural basis for selective recognition of pneumococcal cell wall by modular endolysin from phage Cp-1.** *Structure* 2003, **11:**1239.
24. Suzuki M, Takamura Y, Maeno M, Tochinai S, Iyaguchi D, Tanaka I, Nishihira J, Ishibashi T: **Xenopus laevis macrophage migration inhibitory factor is essential for axis formation and neural development.** *J Biol Chem* 2004, **279:**21406-21414.
25. Van Duyne GD, Ghosh G, Maas WK, Sigler PB: **Structure of the oligomerization and L-arginine binding domain fo the arginine repressor of Escherichia Coli.** *J Mol Biol* 1996, **256:**377-391.
26. Alexandrov NN, Fischer D: **Analysis of topological and nontopological structural similarities in the PDB: New examples with old structures.** *Proteins* 1996, **25:**354-365.
27. Warren AJ, Bravo J, Williams RL, Rabbitts TH: **Structural basis for the heterodimeric interaction betweeen the acute leukemia-associated transcription factors AML1 and CBFbeta.** *EMBO J* 2000, **19:**3004-3015.
28. Meining W, Eberhardt S, Bacher A, Ladenstein R: **The structure of the N-terminal domain of riboflavin synthase in complex with riboflavin at 2.6A resolution.** *J Mol Biol* 2003, **331:**1053-1063.
29. Zhu J, Weng Z: **FAST: A novel protein structure alignment algorithm.** *PROTEINS: Structure, Function, and Bioinformatics* 2005, **58:**618-627.
30. Mizuguchi K, Deane CM, Blundell TL, Overington JP: **HOMSTRAD: a database of protein structure alignments for homologous families.** *Protein Science* 1998, **7:**2469-2471.
31. Holm L, Park J: **DaliLite workbench for protein structure comparison.** *Bioinformatics* 2000, **16:**566-567.
32. Bar-Yehuda R, Halldorsson MM, Naor J, Shacknai H, Shapira I: **Scheduling split intervals.** *14th ACM-SIAM Symposium on Discrete Algorithms* 2002:732-741.
33. Alon N, Feige U, Wigderson A, Zuckerman D: **Derandomized graph products.** *Computational Complexity* 1995, **5:**60-75.
34. Arora S, Lund C, Motwani R, Sudan M, Szegedy M: **Proof verification and hardness of approximation problems.** *Journal of ACM* 1998, **45:**501-555.
35. Meszaros C: **Fast Cholesky factorization for interior point methods of linear programming.** *Comp Math Appl* 1996, **31:**49-51.
36. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22:**2577-2637.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
38. Binkowski TA, Adamian L, Liang J: **Inferring functional relationship of proteins from local sequence and spatial surface patterns.** *J Mol Biol* 2003, **332:**505-526.