Topology Independent Protein Structural Alignment

Joe Dundas[1,⋆], T.A. Binkowski[1], Bhaskar DasGupta[2,⋆⋆], and Jie Liang[1,⋆⋆⋆]

[1] Department of Bioengineering, University of Illinois at Chicago, Chicago,
IL 60607–7052

[2] Department of Computer Science, University of Illinois at Chicago, Chicago,
Illinois 60607-7053

dasgupta@cs.uic.edu
jliang@uic.edu

Abstract. Protein structural alignment is an indispensable tool used for many different studies in bioinformatics. Most structural alignment algorithms assume that the structural units of two similar proteins will align sequentially. This assumption may not be true for all similar proteins and as a result, proteins with similar structure but with permuted sequence arrangement are often missed. We present a solution to the problem based on an approximation algorithm that finds a sequence-order independent structural alignment that is close to optimal. We first exhaustively fragment two proteins and calculate a novel similarity score between all possible aligned fragment pairs. We treat each aligned fragment pair as a vertex on a graph. Vertices are connected by an edge if there are intra residue sequence conflicts. We regard the realignment of the fragment pairs as a special case of the maximum-weight independent set problem and solve this computationally intensive problem approximately by iteratively solving relaxations of an appropriate integer programming formulation. The resulting structural alignment is sequence order independent. Our method is insensitive to gaps, insertions/deletions, and circular permutations.

1 Introduction

The classification of protein structures often depend on the topology of secondary structural elements. For example, Structural Classification of Proteins (SCOP) classifies proteins structures into common folds using the topological arrangement of secondary structural units [16]. 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

[⋆] Partially supported by NSF grant IIS-0346973.

[⋆⋆] Corresponding author. Supported by NSF grants IIS-0346973, IIS-0612044 and DBI-0543365.

[⋆⋆⋆] Supported by NSF grant DBI-0133856 and NIH grants GM68958 and GM079804.

R. Giancarlo and S. Hannenhalli (Eds.): WABI 2007, LNBI 4645, pp. 171–182, 2007.

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 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.

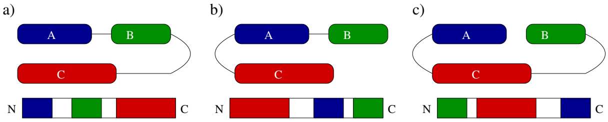

Fig. 1. The cartoon illustration of three protein structures whose domains are similarly arranged in space but appear in different order in primary sequences. The location of domains A,B,C in primary sequences are shown in a layout below each structure. Their orderings are related by circular permutation.

Lindqvist *et al.* (1997) observed the first natural occurence of a circular permutation between jackbean concanavalin A and favin. 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 [18] 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 [17]. 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 [6, 20, 10, 12]. 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 [21]. 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.

This work has incorporated several major improvements and new results over the short paper in [5]. First, the algorithm has been improved so the number of aligned residues in an alignment is significantly increased, without compromise in RMSD values. Second, we have developed a new similarity score for a pair of aligned structures. It incorporates correction of the alignment length, and gives more reliable results. Third, we have developed a method to estimate the statistical significance of a structural alignment by calculating the p-value of a similarity score. Finally, the overall running time is significantly improved and we are able to report results of a large scale exhaustive search of circularly permuted proteins in the PDB database. This includes the discovery of three previously unknown circularly permuted proteins. In addition, we also report the discovery of a new non-cyclicly permuted protein. To our knowledge, this is the first reported naturally occurring non-cyclic permutation between two structures.

The rest of the 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.

2 Method

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 seperately. An approximation algorithm based on a fractional version of the local-ratio approach for scheduling split-interval graphs [3] is then used to search for the combination of peptide fragments from both structures that will optimize the global alignment of the two structures.

2.1 Basic Definitions and Notations

The following definitions/notations are used uniformly throughout the paper unless otherwise stated. Protein structures are denoted by $S_a, S_b, \ldots$. 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 $i \leq t \leq i + k - 1$. 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 can now formalize our substructure similarity identification problem as follows. We call it the *Basic Substructure Similarity Identification* ($\mathrm{BSSI}_{\Lambda,\sigma}$) problem. An instance of the problem is a set $\Lambda = \{\chi_{i,j,k} \,|\, i,j,k \in \mathbb{N}\} \subset \Lambda_a \times \Lambda_b$ of ordered pairs of equal length substructures of S_a and S_b and a similarity function $\sigma : \Lambda \mapsto \mathbb{R}^+$ mapping each pair of substructures to a positive similarity value. The goal is to find a set of substructure pairs $\{\chi_{i_1,j_1,k_1}, \chi_{i_2,j_2,k_2}, ...\chi_{i_t,j_t,k_t}\}$ that are mutually consistent and *maximizes* the total similarity of the selection $\sum_{\ell=1}^{t} \sigma(\chi_{i_\ell,j_\ell,k_\ell})$.

2.2 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. In fact, $BSSI_{\Lambda,\sigma}$ itself is MAX-SNP-hardeven when all the substructures are restricted to have lengths at most 2 [3, Theorem 2.1]. Our approach is to adopt the approximation algorithm for scheduling split-interval graphs [3] which itself is based on a fractional version of the local-ratio approach.

Definition 1. *The closed neighborhood* $Nbr_\Delta[\chi]$ *of a vertex* χ *of* G_Δ *is* $\{\chi' \mid \{\chi,\chi'\} \in E_\Delta\} \bigcup \{\chi\}$. *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*$\}$

For an instance of $BSSI_{\Delta,\sigma}$ 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_χ, $y_{\chi_{\lambda_a}}$ and $y_{\chi_{\lambda_b}} \in \{0,1\}$. x_χ indicates whether the substructure pair should be used ($x_\chi = 1$) or not ($x_\chi = 0$) in the final alignment. $y_{\chi_{\lambda_a}}$ and $y_{\chi_{\lambda_b}}$ 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. We initialize $\Delta = \Lambda$. Then, the following algorithm is executed:

1. Solve the following LP relaxation of a corresponding integer programming formulation of $BSSI_{\Delta,\sigma}$:

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

$$subject\ to \sum_{a_t \in \lambda^a \in \Lambda_a} y_{\chi_{\lambda_a}} \leq 1 \qquad \forall a_t \in S_a \tag{2}$$

$$\sum_{a_t \in \lambda^b \in \Lambda_b} y_{\chi_{\lambda_b}} \leq 1 \qquad \forall a_t \in S_b \tag{3}$$

$$y_{\chi_{\lambda_a}} - x_\chi \geq 0 \qquad \forall \chi \in \Delta \tag{4}$$

$$y_{\chi_{\lambda_b}} - x_\chi \geq 0 \qquad \forall \chi \in \Delta \tag{5}$$

$$x_\chi, y_{\chi_{\lambda_a}}, y_{\chi_{\lambda_b}} \geq 0 \qquad \forall \chi \in \Delta \tag{6}$$

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
$\sigma_{new}(\chi) = \begin{cases} \sigma(\chi) & \text{if } \chi \notin \mathrm{Nbr}_\Delta[\chi_{min}] \\ \sigma(\chi) - \sigma(\chi_{min}) & \text{otherwise} \end{cases}$
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 to a stack in arbitrary order.
4. If $\Delta_{new} \neq \emptyset$ then set $\Delta = \Delta_{new}$, $\sigma = \sigma_{new}$ and go to Step 1. Otherwise, go to Step 5.
5. Repeatedly pop the stack, adding the substructure pair to the alignment as long as the following conditions are met:
   - The substructure pair is consistent with all other substructure pairs that already exist in the selection.
   - 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.

In implementation, the graph G_Δ is considered implicitly via intersecting intervals. The interval clique inequalities can be generated via a *sweepline* approach. The running time depends on the number of iterations needed to solve the LP formulations. Let $\mathrm{LP}(n, m)$ 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 [3] 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.* $\alpha = 2.89$).

Due to lack of space we provide the implementation details of our algorithmic approach in a full version of the paper. We just note here that the linear programming problem is solved using the BPMPD package [14] and to improve computational efficiency, only the top-scoring 1200 substructure pairs are initially used in our algorithm.

3 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.

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% [8]. We exhaustively aligned helices of equal length and obtained the $cRMSD$ distributions shown in Figure 2(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% [8] and obtained the distributions shown in Figure 2(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 $s(\lambda_a, \lambda_b) = \frac{\sum_{i=1}^{N} \delta(A_{a,i}, A_{b,i})}{N}$, to modify the $cRMSD$ of the aligned substructure pairs, where $\delta(A_{a,i}, A_{b,i}) = \begin{cases} 2, & \text{if residues } A_{a,i} \text{ and } A_{b,i} \text{ are both helix} \\ 1, & \text{otherwise} \end{cases}$, to remove bias due to different secondary structure content. We use DSSP [11] 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:

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

In current implementation, 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 $X = \{\chi_1, \chi_2, \dots \chi_m\}$ that maximize Equation (1). The alignment X of two structures is scored following Equation (7) by treating X as a single discontinuous fragment pair:

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

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

3.1 Statistical Significance

To investigate the effect that the size of each the proteins being aligned has on our similarity score, we randomly aligned 200,000 protein pairs from PDBSELECT

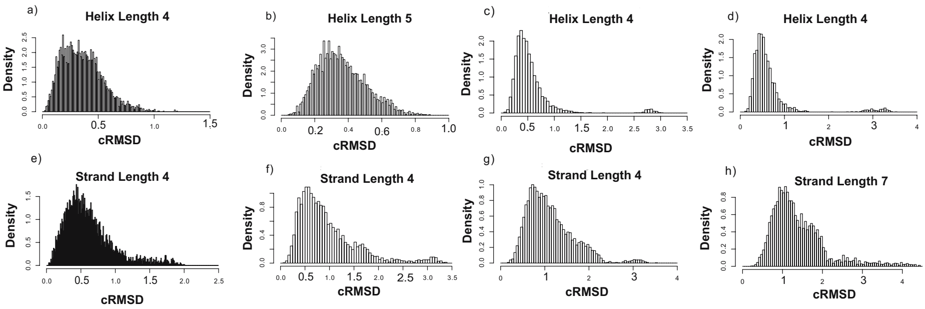

Fig. 2. The cRMSD distributions of a) helices of length 4 b) helices of length 5 c) helices of length 6 d) helices of length 7 e) strands of length 4 f) strands of length 5 g) strands of length 6 and h) strands of length 7

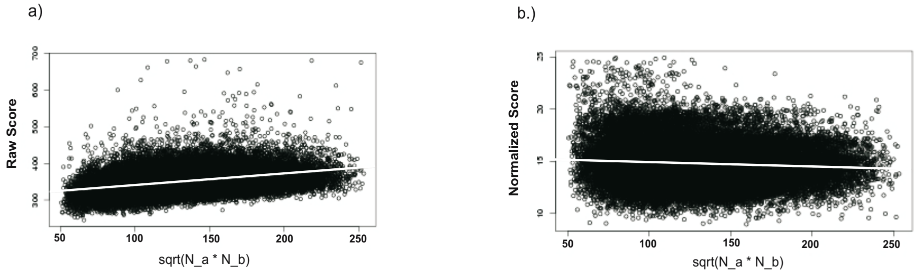

Fig. 3. a) Linear fit between *raw similarity score* $\sigma(X)$ (equation 8) as a function of the geometric mean $\sqrt{N_a \cdot N_b}$ of the length of the two aligned proteins (N_a and N_b are the number of residues in the two protein structures S_a and S_b). The linear regression line (grey line) 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.

25% [8]. Figure 3a 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:

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

where N is the number of equivalent residues in the alignment is used. Figure 3b 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 4). This allows us to compute the statistical significance given the score of an alignment [1, 4].

4 Results

4.1 Discovery of Novel Circular Permutation and Novel Non-cyclic Permutation

In Appendix, we demonstrate the ability of our algorithm to detect circular permutations by examining known examples of circular permutations. 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% [8]. We first selected proteins from PDBSELECT90 (sequences have less than 90% identities) whose N and C termini were no further than 30 Å 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 labelled 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 one novel circular permutations as well as one novel non-cyclic permutation in some details. The newly discovered circular permutation between migration inhibition factor and arginine repressor, which involves an additional strand-swappng is described in Appendix.

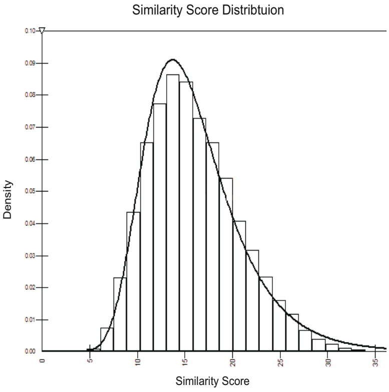

Fig. 4. The distribution of the normalized similarity scores obtained by aligning 200,000 pairs of proteins randomly selected from PDBSELECT 25% [8]. The distribution can be fit to an Extreme Value Distribution, with parameters $\alpha = 14.98$ and $\beta = 3.89$.

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) and the auxin binding protein in maize (PDB ID `1lrh`, chain A, residues 37 through 127). The overall structural alignment between `1k5jE` (Figure 5a, top) and `1lrhA`(Figure 5a, 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

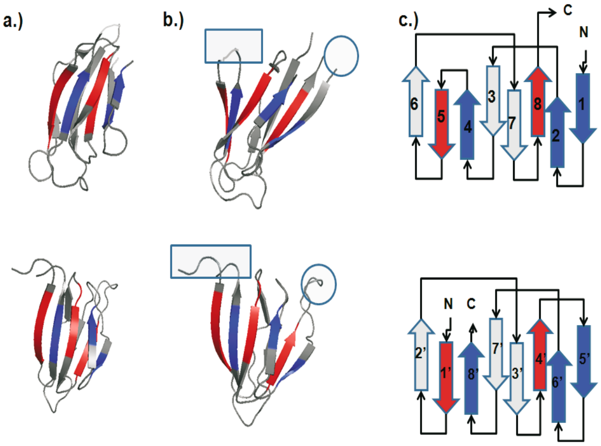

Fig. 5. A new circular permutation discovered between nucleoplasmin-core (`1k5j`, chain E, top panel), and the fragment of residues 37-127 of auxin binding protein 1 (`1lrh`, chain A, bottom panel). a) These two proteins superimpose well spatially, with an RMSD value of 1.36Å for an alignment length 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.

of Fig 5b) becomes disconnected in auxin binding protein 1 (in ellipse, bottom of Fig 5b), and the N- and C- termini of the nucleoplasmin-core protein (in square, top of Fig 5b) are connected in auxin binding protein 1 (square, bottom of Fig 5b). The novel circular permutation between aspartate racemase and type II 3-dehydrogenate dehyrdalase is described in detail in Appendix.

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. However, there has been little work on the detection of non-cyclic permuted proteins. As an example of this important class of topologically permuted proteins, Tabtiang *et al* (2004) were able to artificially create a noncyclic permutation of the Arc repressor that was thermodynamically stable, refolds on the sub-millisecond time scale, and binds operator DNA with nanomolar affinity [21]. This raises the question of whether or not these non-cyclic permutations can arise naturally.

Here we report the discovery of a naturally occurring non-cyclic permutation between chain F of AML1/Core Binding Factor (AML1/CBF, PDB ID `1e50`, Figure 6, top) and chain A of riboflavin synthase (PDB ID `1pkv`, Figure 6a,

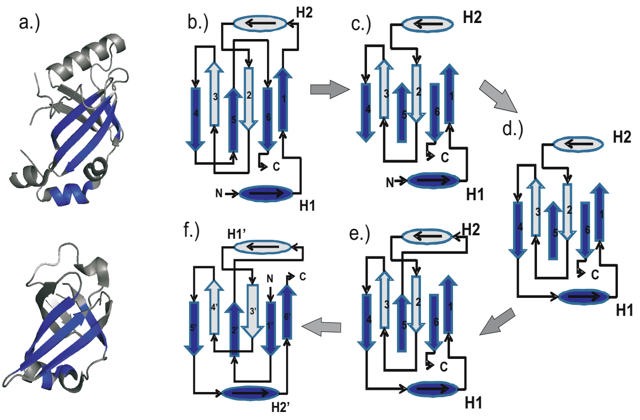

Fig. 6. A novel non-cyclic permutation discovered between AML1/Core Binding Factor (AML1/CBF, PDB ID `1e50`, Chain F, top) and riboflavin synthase (PDBID `1pkv`, chain A, bottom) a) These two proteins superimpose well spatially, with an RMSD of 1.23 Å and an alignment length of 42 residues, with a significant 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.

bottom). 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 6b) can be transformed into the topology diagram of riboflavin synthase (Figure 6f) 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 6c). Connect the C-terminal end of strand 4 to the original N-termini (Figure 6d). Connect the C-terminal end of strand 5 to the N-terminal end of helix 2 (Figure 6e). 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 6f).

5 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 first example of a naturally occurring non-cyclic permuted protein between AML1/Core Binding Factor chain F and riboflavin synthase chain A.

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 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 [12]. 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.

References

1. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997)
2. Arora, S., Lund, C., Motwani, R., Sudan, M., Szegedy, M.: Proof verification and hardness of approximation problems. Journal of the ACM 45(3), 501–555 (1998)
3. Bar-Yehuda, R., Halldorsson, M.M., Naor, J., Shacknai, H., Shapira, I.: Scheduling split intervals. In: 14th ACM-SIAM SODA, pp. 732–741. ACM Press, New York (2002)
4. Binkowski, T.A., Adamian, L., Liang, J.: Inferring functional relationship of proteins from local sequence and spatial surface patterns. J. Mol. Biol. 332, 505–526 (2003)
5. Binkowski, T.A., DasGupta, B., Liang, J.: Order independent structural alignment of circularly permutated proteins. In: EMBS 2004, pp. 2781–2784 (2004)
6. Chen, L., Wu, L., Wang, Y., Zhang, S., Zhang, X.: Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison. BMC Struct. Biol. 6, 18 (2006)
7. Hermoso, J.A., Monterroso, B., Albert, A., Galan, B., Ahrazem, O., Garcia, P., Martinez-Ripoll, M., Garcia, J.L., Menendez, M.: Structural Basis for Selective Recognition of Penumococcal Cell Wall by Modular Endolysin from Phage Cp-1. Structure v11, 1239 (2003)

8. Hobohm, U., Sander, C.: Enlarged representative set of protein structures. Protein Science 3, 522 (1994)
9. Holm, L., Park, J.: DaliLite workbench for protein structure comparison. Bioinformatics 16, 566–567 (2000)
10. Jung, J., Lee, B.: Protein structure alignment using enviromental profiles. Prot. Eng. 13(8), 535–543 (2000)
11. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983)
12. Lindqvist, Y., Schneider, G.: Circular permutations of natural protein sequences: structural evidence. Curr. Opinions Struct. Biol. 7, 422–427 (1997)
13. Liu, L., Iwata, K., Yohada, M., Miki, K.: Structural insight into gene duplication, gene fusion and domain swapping in the evolution of PLP-independent amino acid racemases. FEBS LETT v528, 114–118 (2002)
14. Meszaros, C.S.: Fast Cholesky factorization for interior point methods of linear programming. Comp. Math. Appl. 31, 49–51 (1996)
15. Mizuguchi, K., Deane, C.M., Blundell, T.L, Overington, J.P.: HOMSTRAD: a database of protein structur alignments for homologous families. Protein Sci. 7, 2469–2471 (1998)
16. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structure. J. Mol. Biol. 247, 536–540 (1995)
17. Peisajovich, S.G., Rockah, L., Tawfik, D.S.: Evolution of new protein topologies through multistep gene rearrangements. Nature Genetics 38, 168–173 (2006)
18. Ponting, R.B., Russell, R.B.: Swaposins: circular permutations within genes encoding saposin homologues. Trends Biochem Sci. 20, 179–180 (1995)
19. 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. 279, 21406–21414 (2004)
20. Szustakowski, J.D., Weng, Z.: Protein structure alignment using a genetic algorithm. Proteins: Structure, Function, and Genetics 38, 428–440 (2000)
21. Tabtiang, R.K., Cezairliyan, B.O., Grand, R.A., Cochrane, J.C., Sauer, R.T.: Consolidating critical binding determinants by noncyclic rearrangement of protein secondary structure. PNAS 7, 2305–2309 (2004)
22. Van Duyne, G.D., Ghosh, G., Maas, W.K., Sigler, P.B.: Structure of the oligomerization and L-arginine binding domain of the arginine repressor of Escherichia coli. J. Mol. Biol. 256, 377–391 (1996)
23. Zhu, J., Weng, Z.: FAST: A Novel Protein Structure Alignment Algorithm. PROTEINS: Structure, Function, and Bioinformatics 58, 618–627 (2005)