UC Santa Barbara
UC Santa Barbara Previously Published Works

Title
PADS: Protein structure Alignment using Directional shape Signatures

Permalink
https://escholarship.org/uc/item/6nn5g485

Journal
Database Systems For Advanced Applications, Proceedings, 3453

ISSN
0302-9743

Authors
Aghili, Alireza
Agrawal, D
El Abbadi, A

Publication Date
2005

Peer reviewed

PADS: Protein Structure Alignment using Directional Shape Signatures [*]

S. Alireza Aghili, Divyakant Agrawal, and Amr El Abbadi

Department of Computer Science,
University of California-Santa Barbara,
Santa Barbara, CA 93106
{aghili,agrawal,amr}@cs.ucsb.edu

Abstract. *A novel data mining approach for similarity search and knowledge discovery in protein structure databases is proposed. PADS (**P**rotein structure **A**lignment by **D**irectional shape **S**ignatures) incorporates the three dimensional coordinates of the main atoms of each amino acid and extracts a geometrical shape signature along with the direction of each amino acid. As a result, each protein structure is presented by a series of multidimensional feature vectors representing local geometry, shape, direction, and biological properties of its amino acid molecules. Furthermore, a distance matrix is calculated and is incorporated into a local alignment dynamic programming algorithm to find the similar portions of two given protein structures followed by a sequence alignment step for more efficient filtration. The optimal superimposition of the detected similar regions is used to assess the quality of the results. The proposed algorithm is fast and accurate and hence could be used for analysis and knowledge discovery in large protein structures. The method has been compared with the results from CE, DALI, and CTSS using a representative sample of PDB structures. Several new structures not detected by other methods are detected.*
Keywords. Shape Similarity, Protein Structure Comparison, Biological Data Mining, Bioinformatics.

1 Introduction

Protein structure similarity has been extensively used to highlight the similarities and differences among *homologous* three dimensional protein structures. The corresponding applications include *drug discovery, phylogenetic analysis,* and *protein classification* which have attracted tremendous attention and have been broadly studied within the past decade. The proteins have a primary sequence, which is an ordered sequence of amino acid molecules, e.g. AALHSIAISAJSH. However, they also appear to conform into a three dimensional shape (*fold*) which is highly conserved in the protein evolution. The fold of a protein strongly indicates its functionality and the potential interactions with other protein structures. Meanwhile, the protein sequences as well as

---

their structures may change over time due to mutations during evolution or natural selection. High sequence similarity implies descent from a common ancestral family, and the occurrence of many topologically superimposable substructures provides suggestive evidence of evolutionary relationship [8]. This is because the genetic mechanisms rarely produce topological permutations. For two given proteins, if the sequences are similar then the evolutionary relationship is apparent. However the three dimensional structure of proteins, due to their conformational and functional restraints, are much more resilient to mutations than the protein sequences. There exist functionally similar proteins which *sequence-level* similarity search fails to accurately depict the true similarity. Such cases introduce a big challenge and the necessity of incorporating *structure-level* similarity. Meanwhile, there are two main problems in protein structure similarity:

– *Complexity.* The problem of structure comparison is NP-hard and there is no exact solution to the protein structure alignment [9]. A handful of heuristics [4–6, 8, 12–15, 18] have been proposed in which, to achieve the best result the similarity might need to be evaluated using a series of techniques in conjunction. However, none of the proposed methods can guarantee optimality within any given precision! There are always cases where one heuristic fails to detect, while some of the others succeed.

– *Curse of Dimensionality.* The total number of discovered protein structures has been growing exponentially. Currently the Protein Data Bank (PDB)[1] contains 27,112 protein structures (as of September 8th, 2004.). The growth in the content of PDB demands faster and more accurate tools for structure similarity and the classification of the known structures.

Table 1. Notations.

TERM	DESCRIPTION
atom	Any of the *Nitrogen*(N), *Oxygen*(O), *Hydrogen*(H), or *Carbon*(C) atoms found in protein chains. Carbon atoms that are located on the *backbone* of the protein chains are called C_α, and those on the *side chains* of the protein are called C_β. The atoms that are located closer to the backbone are much more resilient to topological and mutational changes, compared to those atoms that are further away from the backbone. different atom combinations are approximated. For instance, the NH_3^+ and CO^- molecules may be approximated by just considering the considering the coordinates of their corresponding N and C atoms.
amino acid (*residue*)	There are 20 different amino acid molecules in nature (*Alanine, Glycine, Serine, . . .*) which are the alphabets of proteins. Each amino acid is labeled by a capital letter (A, B, F, T, ...) which is made of a number of atoms. All the amino acids have the main N, O, C, and C_α atoms, however this is not true of other atoms like C_β (e.g., Glycine does not have C_β). In this paper, the terms *amino acid* and *residue* are used interchangeably.
protein	A protein is an ordered sequence of amino acids (i.e. ALFHIASUHG. . .). Additionally, each amino acid and as a result each protein chain takes a three-dimensional shape in nature (i.e. in solvents, reactions, . . .). Given two proteins, they may be compared by just aligning their sequences or further inspecting their three-dimensional conformations. Each protein may be either represented by the sequence of its amino acid constituents or its three-dimensional conformation. The topological shape of a protein is one of the very main key factors in defining its functionalities.
SSE	Secondary Structure Element (SSE) is the ordered arrangement or conformation of amino acids in localized regions of a protein molecule. The two main secondary structures are the α-helix and β-sheets. A single protein may contain multiple secondary structures.

We first provide the basic definitions of terms used throughout the paper in Table 1. In this paper, we consider both the sequence and structure of protein chains for more efficient similarity comparison. The main goal of protein structure similarity is to superimpose two proteins over the maximum number of residues (amino acids) with a minimal distance among their corresponding matched atoms. These methods typically employ the three dimensional coordinates of the C_α atoms of the protein backbone and sometimes, in addition, the side chain comprising C_β atoms but exclude the other amino acid atoms when making global structural comparisons. When superimposing two protein structures, side chain conformations (coordinates of O, C, C_β, N, H atoms) may vary widely between the matched residues however the C_α atoms of the backbone trace and the corresponding SSEs are usually well conserved. However, there are situations where the local comparison of the side chain atoms can be of great significance, for instance, in the comparison of residues lining an active or binding sites especially when different *ligands*[1] are bound to the same or similar structures [10].

Distances between the atom coordinates or residual feature vectors or their corresponding biochemical properties are often used to compare protein structures. These features are considered either separately or in combination, as a basis for structural comparison. Some of these features include: *physical properties*, *local conformations*, *distance from gravity center*, *position in space*, *global/local direction in space*, *side chain orientation*, and *secondary structure type*. First, each amino acid of the target and query proteins are represented by a feature vector, and hence each protein is mapped into an ordered sequence of feature vectors. Comparison of the features of the query and target proteins is used as a basis to attribute the similarity. Dynamic programming [16, 20] may be used to discover the similarities between any two protein structures using any number and combination of features of individual residues or regional segments. As a result, a local alignment of the structural features may be deployed to give the best sequential alignment of the given protein structure pairs. Subsequently, the structures should be superimposed according to the results of the alignment. However, a single global alignment of the given protein structures might be meaningless while dissimilar regions (fragments) may affect the overall superimposition drastically. Hence, each fragment of the aligned protein structures should be superimposed individually and independently to explore local similarities. are superimposed on each other, independent of the other similar regions.

The rest of the paper is organized as follows: section 2, discusses the background and related work. Section 3 introduces the formulation and the proposed technique. Section 4 discusses the experimental results, and section 5 which concludes the work.

2 Background & Related Work

Given two protein chains $P = p_1 - p_2 - ... - p_m$ and $Q = q_1 - q_2 - ... - q_n$ (each p_i and q_j denote the feature vectors extracted from the ith and j^{th} amino acid molecule of P and Q, respectively), there are a variety of heuristics to find *optimal* structural similarities (global or local) among them. The techniques map the entire or the best matching regions of the given structures to each other. These algorithms may be classified into three main categories based on their choice of feature vectors and the detail level: *i*) algorithms incorporating only C_α atom coordinates as representatives of amino acid

---

[1] *ligand*: An atom, molecule, or ion that forms a complex around a central atom.

residues and inspecting their inter-atomic distances [12, 13, 18], *ii*) algorithms incorporating SSEs to find initial alignments and filter out non-desired segments [4, 13–15, 19], and *iii*) algorithms using geometric hashing as an indexing scheme to retrieve similar structural alignments [17].

The methods may also be classified based on their choice of heuristics used to align one structure against the other in order to determine the *equivalent pairs*. The term equivalent pairs is defined as the pairs of atoms (or fragments) from the given protein chains whose distance is less than a *threshold*. The threshold or cut-off value may either be a contextual characteristic of the employed method, or provided by the user, or directly learned from the input dataset. The context and the domain properties of the applied method determines the choice of the distance function and the cut-off thresholds, which explains why different structure similarity methods may return non-identical though mostly coherent results. There also exist methods[2] which employ a combination of the listed techniques, including *Dynamic programming* methods [5, 16, 18, 20], *Bipartite and Clique Detection* methods [6, 12, 13], *Match list* methods [4, 6, 12]. Different methods have different notions of similarity score or distance function. These differences make the alignment score not a tangible criterion for comparison. Some of the most frequently used indicators of the quality of a structural comparison include the Root Mean Square Deviation (RMSD) and the extent of the match which is the number of aligned residues. These factors along with the alignment score may be used to asses the quality of the alignment. PADS extends our earlier proposal [3], which considers both the sequence and structure of protein chains and constructs a rotation-invariant geometrical representation from each structure for more efficient similarity comparison. The following section introduces the theoretical aspects and formulation of the proposed protein structure similarity technique.

3   The PADS method

PADS is a novel method for fast and accurate protein structure similarity using directional shape signatures. The algorithm not only exploits the topological properties of the amino acid and protein structures, but also incorporates the biochemical properties (SSE assignments) of the protein chains into account. PADS starts by identifying the geometrical properties of each amino acid of the given proteins along with their directions and their SSE assignments. As a result, each protein structure is represented by a series of directional shape signature feature vectors, one for each amino acid. In the next step, a score matrix is constructed on the corresponding feature vectors. A local *structural alignment* [20] based on shape, direction and biological features detects the optimal local matching regions among the two proteins. For each of the locally matched regions (pertaining to length and score constraints), a *sequence alignment* is performed to facilitate a visualization of the sequence similarities. Thereafter, the best locally matched regions are topologically superimposed. The corresponding RMSD value, length of the aligned fragments, and sequence alignment score are reported for the assessment of the quality of the match. A *linear time* least-square solution to superimpose the ordered sets of protein feature vectors is applied (due to space limitations, the details are provided in [2]). We sort the results based on their extent(L) and RMSD value and report a list of top alignments with the best scores φ, where $\varphi = L/RMSD$.

---

[2] For a detailed survey and comparative study of these methods refer to [2].

3.1 Shape signature extraction

Consider a protein structure P made of an ordered set of amino acids $[a_1, ..., a_N]$, where each a_i is a vector of three-dimensional coordinates of atoms such as C_α, C, O, N, H or other side chain atoms. Hence each amino acid residue constitutes a 3D polyhedron in 3D Euclidean space. For instance, if 6 significant atoms (as in Figure 1-a) of a_i are considered, then a_i would be represented by a vector of 6 three-dimensional vectors, one for the position of each of its constituent atoms.
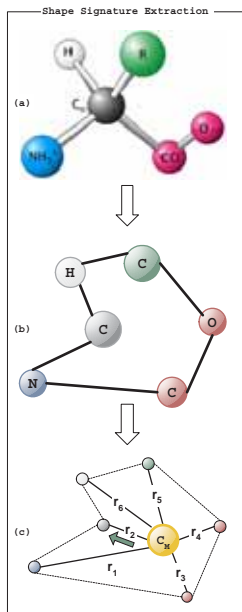

Fig. 1. Shape signature extraction process. (a) An amino acid molecule consisted of $N(NH_3^+)$, C_α, $C(CO^-)$, O, $R(C_\beta)$, and H atoms. (b) The same amino acid visualized as a three-dimensional polyhedron with its vertices being the coordinates of the corresponding atoms, after removing the bonds. (c) Directional Shape Signature Extraction: The distances between the center of mass C_M (or $C_\odot$) and all the atoms are calculated $(r_1, r_2, \ldots)$ along with the direction of the amino acid as:
$$\overrightarrow{C_M C_\alpha}.$$

Definition 1 *Let $S = (v_1, \ldots, v_n)$ be a polyhedron amino acid in 3D Euclidean space. Let v_i denote an atom of S positioned at $v_i = [v_{ix}, v_{iy}, v_{iz}]$ with molar mass μ_i. The* **Center of Mass**[3] *of S is a multidimensional point, $C_\odot(S)$, and is defined as*

$$C_\odot(S) = [C^S_{\odot x}, C^S_{\odot y}, C^S_{\odot z}], \quad where$$

$$C^S_{\odot x} = \frac{\sum_{i=1}^n \mu_i v_{ix}}{\sum_{i=1}^n \mu_i}, \; C^S_{\odot y} = \frac{\sum_{i=1}^n \mu_i v_{iy}}{\sum_{i=1}^n \mu_i}, \; and \; C^S_{\odot z} = \frac{\sum_{i=1}^n \mu_i v_{iz}}{\sum_{i=1}^n \mu_i}.$$

---

[3] The notations $C_\odot(S)$ and C_M are used interchangeably to denote the center of mass.

For instance, let $S = (N, C_\alpha)$ be an amino acid made of only two atoms, N (Nitrogen: molar mass 14.01 g/mol) and C_α (Carbon: : molar mass 12.01 g/mol) positioned at locations [10, 4, 12] and [2, 6, 1], respectively. The center of mass of S is a 3D point and is calculated as $C_\odot(S) = [\frac{(10 \times 12.01) + (2 \times 14.01)}{12.01 + 14.01}, \frac{(4 \times 12.01) + (6 \times 14.01)}{12.01 + 14.01}, \frac{(12 \times 12.01) + (1 \times 14.01)}{12.01 + 14.01}]$
$= [5.7, 5.08, 6.08]$.

Definition 2 *Let $S = (v_1, \ldots, v_n)$ be the polyhedron amino acid with center of mass $C_\odot(S)$.* ***Shape Signature*** *of S, $\sigma(S) = (r_1, \ldots, r_n)$, is defined as the distance between each of the atoms of S to $C_\odot(S)$:*

$$r_i = \sqrt{(v_{ix} - C_{\odot x}^S)^2 + (v_{iy} - C_{\odot y}^S)^2 + (v_{iz} - C_{\odot z}^S)^2}.$$

For instance, let S be the same amino acid as in the previous example with $C_\odot(S) = [5.7, 5.08, 6.08]$. The shape signature of S is $\sigma(S) = (r_1, r_2)$ where
$r_1 = \sqrt{(10 - 5.7)^2 + (4 - 5.08)^2 + (12 - 6.08)^2} = 7.4$ and
$r_2 = \sqrt{(2 - 5.7)^2 + (6 - 5.08)^2 + (1 - 6.08)^2} = 6.35.$

The localized shape signature as described above captures the general shape of each amino acid and is *invariant to rotation and displacement*. The invariance property facilitates the matching of the amino acids solely based on their shape and topological properties. This is a particularly helpful summarization since most protein structures in PDB belong to different coordinate systems. Being able to capture the local and global shape of the amino acids and proteins (invariant to rotation and displacement) facilitates the initial step of protein structure similarity. also be taken into account. The next definition captures the conformational property and orientation of the amino acid structures by augmenting the direction of each amino acid molecule onto its corresponding shape signature.

Definition 3 *Let $S = (v_1, \ldots, v_n)$ be a polyhedron amino acid with the center of mass $C_\odot$. Let v_α (for some $0 < \alpha \leq n$) denote the coordinates of C_α atom of S. The* ***Direction*** *of S, $\overrightarrow{D(S)}$, is defined as the direction of the vector connecting $C_\odot$ to v_α, or in other words $\overrightarrow{D(S)} = \overrightarrow{C_\odot v_\alpha}$.*

Figure 1 depicts the steps involved in extracting the directional shape signature. We excluded C_β from the shape signature because not all amino acids possess C_β (*Glycine*, GLY) and Hydrogen(H) side chain atoms, and due to their dramatic topological variances in different amino acids. On the other hand, a good shape signature should not only capture the topological and shape properties but also biologically motivated features. As a result, PADS incorporates the secondary structure assignment of each amino acid for a more meaningful and efficient structure comparison. Let P be a protein structure with amino acids $[p_1, \ldots, p_N]$ where each p_i is a vector of the three-dimensional coordinates of atoms of the ith residue. Different amino acids have different, though unique, number of atoms. For instance, *Serine* is an amino acid residue which has only 14 atoms while *Arginine* has 27 atoms. PADS also incorporates the distances from $C_\odot$ to the coordinates of C_α, Nitrogen(N) of the *amino group*, Carbon(C) and uncharged Oxygen(O) of the *carboxyl group*, which are common among all amino acids and are topologically more resilient than other side chain atoms.

Definition 4 *Let $P = [p_1, \ldots, p_N]$ be a protein structure where each p_i represents the list of coordinates of atoms that constitute the ith amino acid of P. The* ***Directional***

shape signature of P, P^ϑ, *is defined as the feature vector* $P^\vartheta = [p_1^\vartheta, \ldots, p_N^\vartheta]$ *where each* p_j^ϑ *is a feature vector*

$$(\ |\overrightarrow{C_\odot N}|, \ |\overrightarrow{C_\odot C_\alpha}|, \ |\overrightarrow{C_\odot C}|, \ |\overrightarrow{C_\odot O}|, \ \overrightarrow{C_\odot C_\alpha}, \ SSE_j \ ),$$

comprising the distances from the center of mass of the jth *amino acid to its N, C_α, C and O atoms(Def. 2) along with its corresponding direction(Def. 3), and its secondary structure assignment.*

3.2 Local alignment procedures

This section introduces the alignment procedures to be performed on the extracted directional shape signatures of the corresponding proteins. Structural local alignment starts by constructing a score matrix, S, on the directional shape signatures of the given proteins. This score matrix is used to structurally align the corresponding signatures in the alignment step.

Let P and Q be two protein structures with their corresponding directional shape signatures $P^\vartheta = [p_1^\vartheta, \ldots, p_N^\vartheta]$ and $Q^\vartheta = [q_1^\vartheta, \ldots, q_M^\vartheta]$, where p_i^ϑ and q_j^ϑ denote the feature vectors $[r_{i,1}^p, r_{i,2}^p, r_{i,3}^p, r_{i,4}^p, \overrightarrow{v_i^p}, SSE_i^p]$ and $[r_{j,1}^q, r_{j,2}^q, r_{j,3}^q, r_{j,4}^q, \overrightarrow{v_j^q}, SSE_j^q]$, respectively. The entry $S_{i,j}$, of the score matrix S, denotes the symmetric normalized[4] score of replacing p_i^ϑ by q_j^ϑ residue and is defined

$$S_{i,j} = \sum_{k=1}^{4} (r_{i,k}^p - r_{j,k}^q)^{-2} + \cos(\overrightarrow{v_i^p}, \overrightarrow{v_j^q})^{-1} + SSE_{i,j}^{PQ},$$

where $cos(\boldsymbol{U}, \boldsymbol{V})$ denotes the cosine of the angle between vectors $\boldsymbol{U}$ and $\boldsymbol{V}$, and

$$SSE_{i,j}^{PQ} = \begin{cases} +G & SSE_i^p = SSE_j^q \\ -G & SSE_i^p \neq SSE_j^q. \end{cases}$$

The value of the constant G is empirically chosen to be 10, which is equal to half of the range of the normalized score values. The constant G is used to favor the residue pairs that belong to similar SSEs, and to penalize those that belong to different SSEs. This constant is a tuning parameter of PADS and the user may choose to penalize the residues which have different SSE assignments with a different value for G as desired. Once the calculation of the score matrix is completed, a dynamic programming alignment algorithm is applied to align the given structures. We have deployed the local alignment algorithm [20] using the affine cost gap model with opening and extending gap penalty of -5 and -2, respectively.

Note that, PADS performs two consecutive alignment procedures, *structural alignment* and *sequence alignment*. Structural alignment aligns the corresponding proteins based on their directional shape signatures to find the best structurally-matched-regions. Thereafter, the sequence alignment [16] is performed on the amino acid sequences of the structurally-matched-regions for further refinement of the alignment. For each of the discovered locally matched regions satisfying length and score constraints[5],

---

[4] Scores are normalized on the range $[1 \ldots 20]$ for all i, j such that $0 < S_{i,j} \leq 20$ to be similar to that of PAM [7] score matrix and CTSS [5].

[5] *Length* longer than 10 and *Score* above the 60% of the overall average score.

a *sequence alignment* is performed to facilitate the visualization of the sequence similarities and further refinement. The aligned residue coordinates passed through structural and sequence alignment steps are then passed to the superimposition stage.

Why did we need to perform the superimposition? The detected best local alignment passed from the structural alignment step is not necessarily the most optimal alignment because the directional shape signatures do not include any information on the proximity/locality of the amino acids (i.e., Center of mass ($C_\odot$) was not taken as part of the directional shape signature). Including such locality features in the shape signature would not have been very meaningful because the proteins have different coordinate frames. Should the locality information be included in the shape signature, then two very similar proteins with different coordinate frames may be reported non-similar because of their location differences. Additionally, the detected patterns may have very poor RMSD if the gaps produced by the structural alignment are in turn and twist regions of the protein structures. The sequence alignment step aims at eliminating those regions from affecting the superimposition process. After the local regions are passed to the superimposition step, the given proteins are translocated to a common coordinate frame. Once the structures are in a common coordinate system, they are optimally superimposed on each other (with the necessary displacements and rotations) achieving the minimal RMSD. Finally, after performing the superimposition, the RMSD values and the length of the best matched regions are reported. Figure 2 provides a summary of PADS procedure.

Input: Protein chains $P = [p_1, \ldots, p_N]$ and $Q = [q_1, \ldots, q_M]$, where each p_i and q_j represent the list of coordinates of atoms that constitute the ith and j^{th} amino acids of P and Q, respectively.

Output: Pairs of aligned/matched fragments of P and Q, reported with their corresponding RMSD and fragment length.

---

1. Directional Shape Signature Extraction:
   - Calculate the center of mass of each amino acid molecule p_i and q_j, as $C_\odot(p_i)$ and $C_\odot(q_j)$, for $1 \leq i \leq N$ and $1 \leq j \leq M$.
   - Calculate the distances between each of the atoms of p_i and q_j molecules to their corresponding center of mass $C_\odot(p_i)$ and $C_\odot(q_j)$, respectively.
   - Extract the direction of each amino acid molecule p_i and q_j.
   - Inspect and include the SSE assignment of each p_i and q_j in the shape signature.
2. Structural local alignment
   - Calculate the score matrix for P and Q protein chains as described in section 3.2.
   - Run the dynamic programming on the calculated score matrix to find the best structurally-matched (aligned) fragment pairs of P and Q.
   - Report the fragment pairs to the next step.
3. Sequence alignment
   - Run the global sequence alignment on the sequences of the structurally-matched fragment pairs
   - Remove the gapped regions of the alignment from the fragments, and report the non-gapped subfragments of the alignment to the next step.
4. Optimal Superimposition
   - Find the best rotation and translation matrix to superimpose the matched non-gapped fragment pairs.
   - Report the RMSD and the length of the matched fragment pairs in the sorted order.

Fig. 2. PADS structure similarity procedure.

4   Experimental Results

We implemented our proposed technique using *Java 1.4.1* and ran our experiments on an *Intel Xeon 2.4 GHz* with *1GB* of main memory. Our experiments incorporated a representative of PDB database using the PDBSELECT[6] method [11] which does not contain any homologue protein pairs. The PDBSELECT database is an archive of 2216 *non-homologue* protein chains with a total number of 352855 residues (as of December 2003). Each of the protein pairs from the PDBSELECT protein database has less than 25% sequence identity (non-homologue). As a result, protein pairs with low sequence similarity may not be efficiently compared solely based on a sequence-level similarity procedure and therefore introduce a challenging problem where the combination of structure and sequence alignment is inevitable. As mentioned before, PADS incorporates a combination of structural and sequence alignment for efficient protein similarity comparison.

The performance comparison of PADS with other structural alignment methods is not always possible. One of the main challenges is the *running time* comparison of the proposed technique against current existing heuristics. This is mainly because most of the available techniques are provided as web services in which the results are notified back to the user through an e-mail. As a result, the time interval between submitting a query and obtaining the results does not truly reflect the running time of the applied method. There are many factors that may affect the running time. The servers may include pre-evaluated results for the known structures, and hence the results may be returned very fast. They may be using parallel clusters or various hardware setups for faster computation of the results. The DALI [12] interactive database search[7] may report the results back in 5 to 10 minutes or 1 to 2 hours depending on whether the query protein has a homologue in the database [5]. Meanwhile the most important obstacle is the fact that various structural alignment techniques may lead to non-identical results which makes the quality assessment an even harder problem. There are cases when the regions found very similar by one technique are not validated by other techniques[8]. Since there is no exact solution to the structural alignment problem, a combination of various techniques along with domain expert is needed to evaluate and ascertain all the similarities.

In the experiments, we discovered motifs not reported by other alignment tools such as CE [18], DALI [12], and CTSS [5]. The aligned fragment pairs are reported as a pair of fragments (r_1, r_2) where r_1 and r_2 denote the location of the matched fragments in the first and second protein chains, respectively. One such motif discovered by our technique was between 1AKT:_ (made of 147 residues and 1108 atoms) and 1CRP:_ (made of 166 residues and 2619 atoms) protein chains (having 8.9% sequence identity) with RMSD 0.58 Å. Figure 3 shows the results of structural alignments on 1AKT:_ and 1CRP:_ protein chains using CE[9] and PADS, respectively. These results are reported after finding the best similar regions (fragments) followed by the optimal superimposition of the structures of the corresponding matched fragments. However, the results are shown at the sequence level for the sake of visualization. In figure 3(b),

---

[6] For more information refer to *http://homepages.fh-giessen.de/ hg12640/pdbselect/*

[7] *http://www.embl-ebi.ac.uk/dali/*

[8] Please refer to Table VI in [18]

[9] The results of CE were obtained by submitting the corresponding protein chains to CE's interactive web server at *http://cl.sdsc.edu/ce.html*

the fragments reported by PADS are demonstrated using the output of CE as the base for better visual comparison of the results. The local fragments are identified by three numbers in the R(L,φ) format, where R, L and $\varphi = \frac{L}{R}$ denote *RMSD*, *length* and the *fragment score* of the aligned (matched) fragments, respectively. The fragment score denotes the *quality* of the matched fragments and the best aligned fragment is the one with the highest fragment score. PADS reports the aligned fragment pairs sorted by their corresponding fragment scores in decreasing order.

Table 2. Comparison of detected similar regions between 1AKT:_ and 1CRP: protein chains using PADS and DALI methods with alignment rank $\varphi = \frac{Fragment\ Length}{RMSD}$.

Rank	φ	Fragment size	RMSD (Å)	PADS		DALI	
				1AKT:_	1CRP:_	1AKT:_	1CRP:_
–						[1–8]	[4–11]
2	11.66	14	1.2	[10–23]	[12–35]	[12–15] [18–23]	[12–15] [16–21]
–						26-29	41–44
5	3.7	20	5.4	[35–54]	[51–70]	[30–36] [43–58]	[53–59] [69–84]
4	6.66	28	4.2	[75–101]	[98–125]	[65–81] [83–92] [93–100]	[88–104] [107–116] [118–125]
3	7.64	13	1.7	[108–121]	[130–142]	[104–112] [121–124]	[130–138] [140–143]
–						[129–133]	[146–150]
1	29.31	17	**0.58**	[131–147]	[149–165]	[135–147]	[151–163]

Table 2 shows a detailed comparison of PADS against DALI[10] [12] on the very same pair of protein chains. Each column pair (1AKT:_ , 1CRP:_) indicates the location of the aligned fragments in the corresponding protein chains. The correspondence of the detected aligned fragments of PADS and DALI are noted in rows and labeled with φ to indicate the quality of the aligned fragments and their corresponding ranks as reported by PADS technique. There are some matched fragments reported by PADS, which do not have counterparts in the results returned by DALI. However, it is interesting to note that, the fragments matched using PADS with higher φ tend to be those fragment pairs having a higher level of similarity to their corresponding aligned fragments as reported by DALI. As a result, highly-ranked matched fragment pairs reported by PADS, have very similar counterparts in the results reported by DALI. We use DALI to validate the quality of our results, while DALI is designed with very insightful domain expertise and is expected to return biologically meaningful results. PADS results are very similar, though not identical, to that of DALI and in some cases, the fragment pairs reported by PADS are a combination of some consecutive fragment pair outputs of DALI. Meanwhile, running PADS on 1AKT:_ and 1CRP:_ protein chains takes only 0.1 CPU seconds.

Similarly, the reported results on the very same pair of protein chains were compared against the CTSS [5] algorithm. CTSS reports the best aligned fragment pair between 1AKT:_ and 1CRP:_ protein chains to be ([89–113],[140–164]) with length 24 and

---

[10] The results of DALI were obtained by submitting the corresponding protein chains to DALI's interactive web server hosted by European Bioinformatics Institute at *http://www.ebi.ac.uk/dali/*
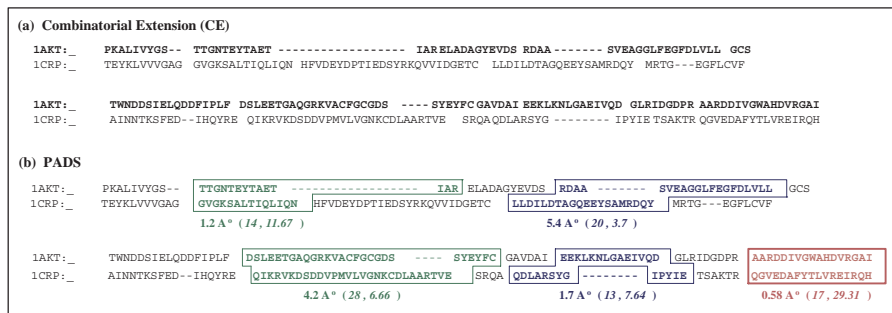
Fig. 3. (a) Structural alignment (shown at the sequence level) between 1AKT:_ and 1CRP:_ using CE. (b) The RMSD, extent and score of local fragments discovered by PADS structural alignment (shown at the sequence level) between 1AKT:_ and 1CRP:_ (The output of CE is also shown for comparison purposes).

RMSD 2.14 Å with a fragment score of φ=11.21. On a relative note, the best aligned fragment pair reported by PADS is ([131–147],[149–165]) of length 17, though with an RMSD of 0.58 Å and the fragment score of φ=29.31. Although the best fragment pair reported by PADS has smaller length however it is aligned with a substantially better RMSD value (by a factor of 3.6) and higher quality of the alignment (by a factor of 2.6) noted by φ. The calculation of the value of φ in our algorithm is identical with its counterpart in the CTSS method. The intuition behind PADS finding a better fragment pair compared with CTSS, is as follows. The CTSS method approximates each protein chain by a spline (curve), however PADS represents each chain as a series of directional shape signatures (a sequence of polyhedrons in multidimensional space). To give a better visual example, suppose we would like to represent a snake, then CTSS approximates its shape with a rope while PADS approximates the shape using a chain of polyhedral beads for a more precise approximation.

5 Conclusion and Future Work

In this paper, we introduced a novel data representation technique incorporating multi-dimensional shape similarity and data mining techniques for the problem of structural alignment of protein structure databases. We evaluated the quality of the results of PADS on a pair of protein chains and compared the corresponding results with the other methods. The results demonstrate highly *accurate* (the reported fragments have very high score with the RMSD value much better than all other methods), *consistent* (the fragment pairs reported similar by PADS had high overlap with regions reported similar by other methods) results compared with DALI, CE, and CTSS protein structure similarity methods, while running only in fractions of a second. PADS may be used in collaboration with other protein alignment methods such as DALI and CE for providing a larger number of fragment pairs. One could potentially use PADS to get an instant feedback of the location and quality of the matched regions, and thereafter run the time-consuming DALI method to achieve the most accurate results, if desired. We intend to perform database-against-database structure similarity search for protein classification and add a 3D visualization tool to PADS for better assessment of fragment pair discovery.

12

References

1. Protein data bank(pdb). http://www.rcsb.org/pdb/holdings.html, 2004.
2. S. A. Aghili, D. Agrawal, and A. E. Abbadi. Pads: Protein structure alignment using directional shape signatures. Technical Report 2004-12, UCSB, May 2004.
3. S. A. Aghili, D. Agrawal, and A. E. Abbadi. Similarity search of protein structures using geometrical features. In *Proceedings of Thirteenth Conference on Information and Knowledge Management (CIKM)*, pages 148–149, 2004.
4. P. Bradley, P. Kim, and B. Berger. Trilogy: Discovery of sequence-structure patterns across diverse proteins. *Proc. Natl. Academy of Science*, 99(13):8500–5, 2002.
5. T. Can and Y. Wang. Ctss: A robust and efficient method for protein structure alignment based on local geometrical and biological features. In *IEEE Computer Society Bioinformatics Conf.*, pages 169–179, 2003.
6. O. Çamoğlu, T. Kahveci, and A. Singh. Towards index-based similarity search for protein structure databases. In *IEEE Computer Society Bioinformatics Conf.*, pages 148–158, 2003.
7. M. Dayhoff and R. Schwartz. Atlas of protein sequence and structure. *Nat. Biomed. Res. Found.*, 1978. Washington.
8. J. Gibrat, T. Madej, and S. Bryant. Surprising similarities in structure comparison. *Current Opinion Structure Biology*, 6(3):377–85, 1996.
9. A. Godzik. The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, 5:1325–1338, 1996.
10. D. Higgins and W. Taylor. *Bioinformatics: Sequence, Structure and Databanks.* Oxford University Press, 2000.
11. U. Hobohm, M. Scharf, and R. Schneider. Selection of representative protein data sets. *Protein Science*, 1:409–417, 1993.
12. L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Molecular Biology*, 233(1):123–138, 1993.
13. L. Holm and C. Sander. 3-d lookup: Fast protein database structure searches at 90% reliability. In *ISMB*, pages 179–185, 1995.
14. G. Lua. Top: a new method for protein structure comparisons and similarity searches. *J. Applied Crystallography*, 33(1):176–183, 2000.
15. T. Madej, J. Gibrat, and S. Bryant. Threading a database of protein cores. *Proteins*, 23:356–369, 1995.
16. S. Needleman and C. Wunsch. General method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Molecular Biology*, 48:443–453, 1970.
17. X. Pennec and N. Ayache. A geometric algorithm to find small but highly similar 3d substructures in proteins. *Bioinformatics*, 14(6):516–522, 1998.
18. I. Shindyalov and P. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Engineering*, 11(9):739–747, 1998.
19. A. Singh and D. Brutlag. Hierarchical protein structure superposition using both secondary structure and atomic representations. In *Proc. Int. Conf. Intelligent System Mol. Bio.*, pages 284–93, 1997.
20. R. Smith and M. Waterman. Identification of common molecular subsequences. *J. Mol. Bio.*, 147(1):195–197, 1981.