PADS: Protein Structure Alignment using Directional Shape Signatures*

S. Alireza Aghili†
Department of Computer Science
University of California Santa Barbara
Santa Barbara, CA 93106
aghili@cs.ucsb.edu

Divyakant Agrawal
Department of Computer Science
University of California Santa Barbara
Santa Barbara, CA 93106
agrawal@cs.ucsb.edu

Amr El Abbadi
Department of Computer Science
University of California Santa Barbara
Santa Barbara, CA 93106
amr@cs.ucsb.edu

ABSTRACT

*A novel approach for similarity search on the protein structure databases is proposed. PADS (**P**rotein **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 the given amino acid. As a result, each protein is presented by a series of feature vectors representing local geometry, shape, direction, and secondary structure assignment of its amino acid constituents. Furthermore, a residue-to-residue distance matrix is calculated and is incorporated into a local alignment dynamic programming algorithm to find the similar portions of two given proteins and finally a sequence alignment step is used as the last filtration step. 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 the analysis of large protein structure similarity. The method has been tested and 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.*

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.5 [**Pattern Recognition**]: Models, Design Methodology; J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms

Design, Performance, Algorithms

*This research was supported by the NSF grants under EIA02-05675, EIA99-86057, and IIS02-09112.

†To whom all the correspondences should be forwarded.

Keywords

Shape Similarity, Protein Structure Alignment, Biological Databases, Biological Data Mining

1. INTRODUCTION

Protein structure similarity has been extensively used to highlight the similarities and differences among related/similar (*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. However, they appear to conform into a three dimensional shape, *fold*, which is highly conserved in the protein evolution. The fold of a protein highly 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, insertions, and deletions during evolution or natural selection. Extensive sequence similarity implies descent from a common ancestral gene, and the occurrence of many topologically superimposable substructures provides suggestive evidence of evolutionary relationship [6]. 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 protein structures, due to structural, conformational and functional restraints placed on them, are much more resilient to mutations than the protein sequences. As a result, the structural similarity among various protein fragments may be used to understand the differences in the observed functionalities and potentially their evolutionary relationships. 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 [7]. A handful of heuristics [2, 3, 4, 6, 10, 11, 13, 14, 19] have been proposed in which, for 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.

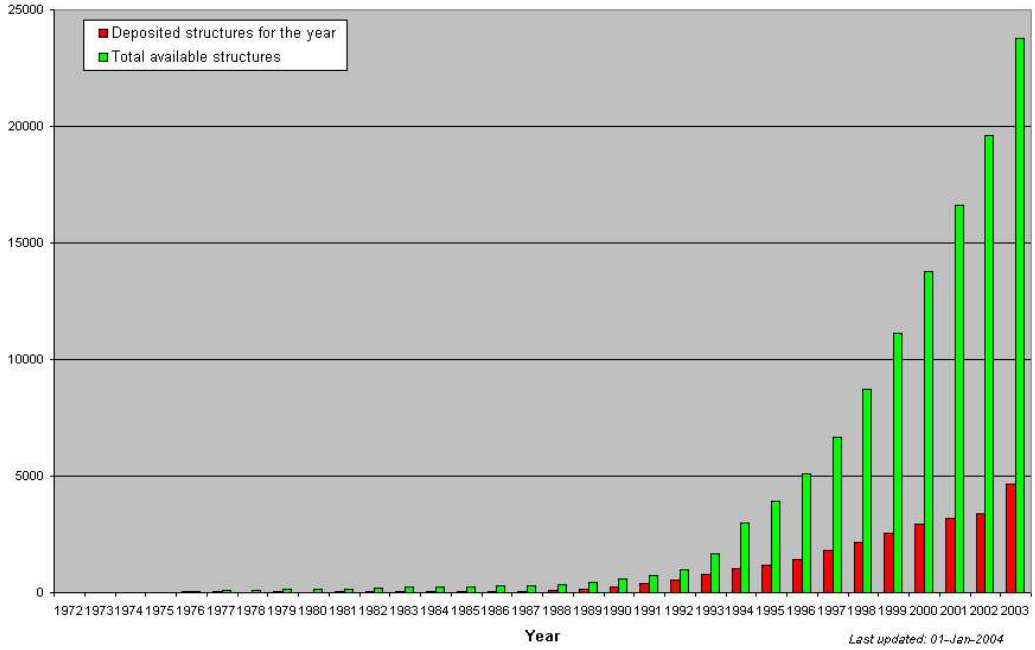

Figure 1: Protein Data Bank (PDB) content growth.

- *Curse of Dimensionality.* The number of discovered protein structures has been growing very rapidly. Currently[1] the Protein Data Bank (PDB)[1] contains 25,551 protein structures. The growth in the content of PDB demands faster and more accurate tools for structure similarity and the classification of the known structures. Figure 1 demonstrates the growth of PDB in the past decades.

Let's start our discussion by providing the definition of the terms used throughout the paper:

- *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 (core) 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. When approximating different atom combinations, some molecules may be replaced by an atom. For instance, the NH_3^+ and CO^- molecules may be approximated by just considering the coordinates of their corresponding N and C atoms.

- *amino acid (residue)*: There exist 20 different amino acid molecules in nature (*Alanine*, *Glycine*, *Serine*, ...) which are the alphabets of proteins. Each amino acid is labeled by a character (A, B, F, T, ...) and is made of a number of atoms. All the amino acids have the main N, O, C, and C_α atoms, however that 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 residues are used interchangeably.

[1]as of May 18^{th}, 2004.

- *protein*: A protein is an ordered sequence of amino acids. 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 looking through their amino acid consitituent sequences or further inspecting their three-dimensional conformations. Each protein may be either represented by the sequence of its amino acid constituents or the actual three-dimensional conformation that it takes. The topological shape (conformation) of the protein is one of the very main key factors in defining its functionalities. Different amino acids having different atoms may take similar conformations which facilitates the classification of amino acids (and as a result: proteins) into various classes based on their conformational and chemical properties.

In this paper, we consider both the topological shapes and the corresponding amino acid sequences of the 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 secondary structure elements, α-helices and β-sheets, 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 site especially when

different ligands are bound to the same or similar structures [8]. As a result, depending on what types of questions one wishes to answer, when comparing protein structures, the choice of which atom coordinates to consider can be extremely crucial.

Distances between the atom coordinates or residual feature vectors or molecular properties are often used to compare protein structures. These molecular properties or relationships among the individual residues are considered either separately or in combination with each other 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 protein and a target protein is used as a basis to represent the distance/similarity between their corresponding matched amino acid residues or regions of interest. Dynamic programming [15, 21] may be used to discover the similarities/differences between any two protein structures using any number and combination of features of individual residues or regional segments. Hence, a local alignment algorithm based on the structural features is deployed to give the best sequential alignment of the given protein structures. 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 may affect the overall superimposition drastically. Hence, each domain of the aligned protein structures should be superimposed individually and independently to explore local similarities. These domains are mainly identified from the output of the structural local alignment. As a result, each of the similar regions of the two proteins are superimposed on each other, independent of the other similar regions. The choice of the features, used for protein 3-D comparison, rely upon the type of questions that are to be answered.

The rest of the paper is organized as follows: Section 2, discusses the background and related work. Section 3 introduces the terminology and formulation of the problem and the proposed technique. Section 4 demonstrates an empirical performance analysis and the simulation results followed by Section 5 which concludes the work.

2. BACKGROUND & RELATED WORK

Given two protein chains $P = p_1 - p_2 - ... - p_m$ and $Q = q_1 - q_2 - ... - q_n$, 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 residues and inspecting their inter-atomic distances [10, 11, 19], *ii*) algorithms incorporating Secondary Structure Elements (SSE) to find initial alignments and filter out non-desired segments [2, 11, 13, 14, 20], and *iii*) algorithms using geometric hashing as an indexing scheme to retrieve similar structural alignments [18].

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 is either 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 which employ a combination of the listed techniques.

- *Dynamic programming* methods [3, 15, 19, 21], construct an $m \times n$ distance or score matrix M where each cell of the matrix, $M_{i,j}$, corresponds to the distance or score of matching residue p_i of protein A, with residue q_j of protein B. Should distance be used, cells are filled with the score between the corresponding feature vectors (consisting of C_α atom coordinates, side chain atom coordinates, secondary structure assignments, or a combination of other residual properties) of p_i and q_j. The score is inversely proportional to the designated distance value. Starting from the upper left corner of the matrix, $M_{1,1}$, the algorithm seeks the optimal segments of one protein against the other, where optimality is defined as the longest segments with the highest score, or lowest distance value. Further heuristics might be applied, *i*) to merge some of these segments, *ii*) to run some variations of dynamic programming [15, 21] as a refinement step, and/or *iii*) to prune the non-desired (below cut-off threshold) segments.

- *Bipartite and Clique Detection* methods [4, 10, 11], represent each structure as a graph with its nodes being C_α atom coordinates or secondary structure elements or a combination of some other molecular properties. Each edge of the graph is labeled with the distance between the corresponding nodes where distance is defined as the distance between their corresponding feature vectors. Furthermore, they find a maximal common subgraph whose nodes are very close to each other, using a tree search algorithm (e.g., depth-first or breath-first search). Each vertex in the subgraph corresponds to a unique vertex in each of the structures. Some of these methods include a further step, namely finding a maximal bipartite graph to map the vertices of one structure to the other. The nodes in the subgraph are considered equivalent among the structures. Finally the equivalent vertices or their corresponding atoms are directly superimposed. These superimposition may be further merged for longer matches.

- *Match list* methods [2, 4, 10], construct two lists, the first one includes all pairs of atom coordinates (e.g., C_α, side chain, or other properties) of protein A with their counterparts in protein B whose distance is below a certain threshold (e.g., 3.0 Å). Similarly the second list holds all the similar atom pairs of protein B to protein A with distance below a threshold. The method chooses the most similar pairs in the list, and merges

them to extend the match pairs to a contiguous sequence of existing match pairs of longer length, while filtering out non-significant matches. Furthermore, the matching pairs are sorted based on their distance and length, and all the matched pairs whose length is below a certain threshold T (e.g., $T = 3$ results in *triangular extension*) are pruned from further investigation. For instance (a_{i-1}, b_{j+1}), (a_i, b_j) and (a_{i+1}, b_{j-1}) may be merged to achieve an extended triangular pair $[(a_{i-1}, b_{j+1}), (a_{i+1}, b_{j-1})]$.

There are also a variety of optimization techniques to discover the relationship among the features of one structure against the other. These methods include, *Monte Carlo optimization*, *Double dynamic programming*, and *Genetic algorithms*. One important question yet to be answered is how to assess the quality of the discovered similar patterns. The final alignment distance (or alignment score) may be used to evaluate the quality of the matches within a family of proteins to infer evolutionary relationships. However 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. We now describe some of the most popular protein alignment methods in the literature namely DALI [10, 11], CE [19], VAST [6, 14], CTSS [3], and PSI [4]:

DALI [10, 11] calculates a distance matrix (D^A) for any given protein A. Each cell $D^A_{i,j}$ contains the intermolecular distance between the i^{th} and j^{th} C_α atom coordinates of A. Given two protein chains A and B, DALI seeks the similar regions, denoted by *contact maps*, between D^A and D^B distance matrices and finds the optimal clique on the contact maps obtained from the structures. DALI uses a Monte Carlo optimization to search the best 40,000 matches. The matches are further extended by combining those contact maps which are common in both distance matrices. Holm and Sander have further improved this technique by incorporating a preprocessing filtration step using secondary structure elements [11]. DALI interactive database search may respond to a query in 5-10 minutes or 1-2 hours depending on whether the query protein structure has a sequence homologue in the database or not [3]. Considering the extent and fast growth of the PDB protein database, response time plays an important role in providing an effective search.

VAST structure similarity method [14] performs a hierarchical alignment. Given two proteins A and B, it constructs a bipartite graph on the SSE pairs of protein A against the SSE pairs of protein B, and an edge is inserted between every two pairs of vertices (from A and B) having a similarity more than a cut-off value. An initial SSE alignment is found by applying a maximal clique algorithm on the bipartite graph, and is further extended to C_α atom coordinates incorporating Gibbs sampling. There are cases[2] where VAST produces a lower RMSD alignment with fewer matched residues while DALI produces a longer alignment for the cost of larger RMSD, although both of the techniques use Monte Carlo refinement. It is unclear whether either is better than the other. Driven by these cases, the VAST structure similarity technique might be suitable for the identification of highly conserved core elements of a protein family (e.g., for a threading experiment), while the DALI structure similarity technique might be useful for the identification of a larger set of similar sites (e.g., in homology modeling) [6].

[2]The alternative alignments between *T4 glutaredoxin* (PDB code 1ABA) and *Escherichia coli* disulfide bond formation protein (PDB code 1DSB-A).

PSI [4] extracts feature vectors corresponding to triplet SSEs and builds an R^*-tree index structure on the feature space using Minimum Bounding Rectangles (MBR). PSI builds a Triplet Pair Graph (TPG) on the similar triplet pairs of query and target proteins and runs Depth-First Search (DFS) on the TPG graph to find the Largest Weight Connected Component (LWCC). The LWCC corresponds to the most similar subset of SSEs of query and target proteins. PSI then constructs a bipartite graph on the subsets extracted from the LWCC, with the edges indicating the quality of an alignment of the corresponding SSE pairs.

CTSS [3] calculates a spline fitting to approximate the positions of the C_α atoms and computes, for each residue, the curvature and torsion values at the C_α positions along the spline. It further runs a dynamic programming local alignment algorithm [21] on the curvature-torsion feature vectors and superimposes the corresponding residues. However the detected best local alignment is not necessarily the most optimal structural alignment and the algorithm needs to perform locally sensitive superimposition to find the best regions. This is because the shape signatures used in CTSS do not capture the locality of the amino acids.

Finally, CE [19] performs a combinatorial extension of an alignment path defined by Aligned Fragment Pairs (AFP) instead of dynamic programming and Monte Carlo optimization. AFPs are based on local geometry rather than global features such as orientation of SSEs or overall topology. Continuous alignment paths which are made of the combinations of AFPs are selectively extended for a long optimal alignments.

The following section introduces the theoretical 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 amino acid directional shape signatures. The algorithm not only exploits the topological properties of the amino acid and protein structures, but also incorporates the SSE assignments of the amino acids 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* [21] 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 (Section 3.4). 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 $\varphi = L/RMSD$.

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 2-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.

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*

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

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

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 $C_\odot(S) = [\frac{(10\times12.01)+(2\times14.01)}{12.01+14.01}, \frac{(4\times12.01)+(6\times14.01)}{12.01+14.01}, \frac{(12\times12.01)+(1\times14.01)}{12.01+14.01}] = [5.7, 5.08, 6.08]$.

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 $C_\odot(S)$:*

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

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 shape of the amino acids and the global shape of a protein (invariant to rotation and displacement) facilitates the initial step of protein structure similarity. Meanwhile, in addition to shape similarity, the location sensitive conformation of the amino acids (and as a result protein shape structure) should 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 2 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). The Hydrogen(H) side chain atom was also discarded for the same reason, and due to its 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 list of the three-dimensional coordinates of atoms of the i^{th} 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. Meanwhile PADS 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 i^{th} 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 j^{th} 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 Structural local alignment

This section introduces the structural alignment procedure 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 [21].

Let P and Q be two protein structures with their corresponding directional shape signatures $P^\vartheta = [p_1^\vartheta, \ldots, p_N^\vartheta]$ and

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

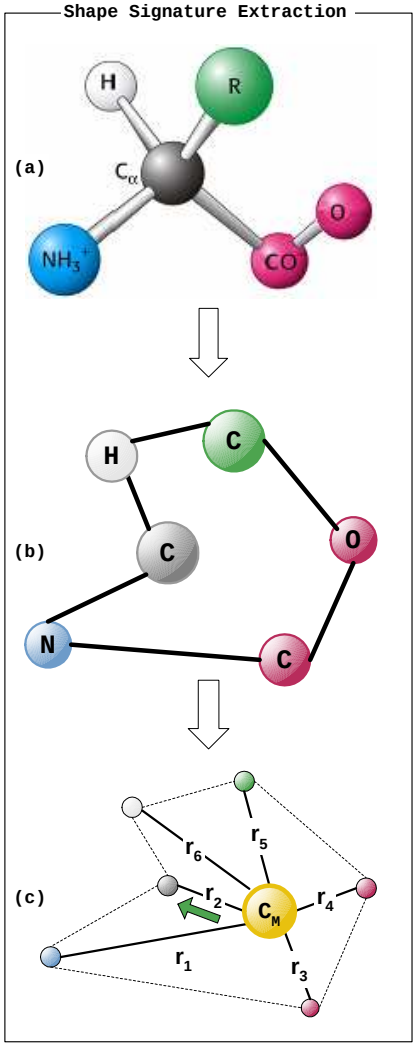

Figure 2: Shape signature extraction process. (a) An amino acid molecule consisted of N(NH_3^+), C_α, C(CO^-), O, R(C_β), and H atoms. (b) The same amino acid visualized as a three-dimensional polyhedron with its vertices being the coordinates of the corresponding atoms, after removing the bonds. (c) 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, ...) along with the direction of the amino acid as: $\overrightarrow{C_M C_\alpha}$.

$Q^\vartheta = [q_1^\vartheta, \ldots, q_M^\vartheta]$, where $p_i^\vartheta = [r_{i,1}^p, r_{i,2}^p, r_{i,3}^p, r_{i,4}^p, \overrightarrow{v_i^p}, SSE_i^p]$, and $q_j^\vartheta = [r_{j,1}^q, r_{j,2}^q, r_{j,3}^q, r_{j,4}^q, \overrightarrow{v_j^q}, SSE_j^q]$. The entry $S_{i,j}$, of the score matrix S, denotes the symmetric normalized[4] score of replacing the p_i^ϑ residue by the 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 $\cos(\vec{U}, \vec{V})$ denotes the cosine of the angle between vectors $\vec{U}$ and $\vec{V}$, 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 used to align the given structures. We have deployed the local alignment algorithm [21] 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 is performed on the amino acid sequences of the structurally-matched-regions for further refinement of the alignment, which is described in the next section.

3.3 Sequence alignment

For each of the discovered locally matched regions satisfying length and score constraints[5], a *sequence alignment* [15] is performed to facilitate the visualization of the sequence similarities and further refinement. We deployed the PAM250 [17] scoring matrix for the sequence alignment and incorporated gap penalty of -10 as is usually the case for the global sequence alignment. The structural alignment followed by sequence alignment provides a good picture of the local similarities. The aligned residue coordinates passed through structural and sequence alignment steps are then passed to the superimposition stage. The next section describes the details of the superimposition process.

3.4 Optimal Superimposition

The Root Mean Square Deviation (RMSD) is a frequently used measure to assess the goodness of a topological match among two sets of coordinates. The RMSD value indicates the average level of deviations among the matched or aligned residues. Given two ordered sets of residues, a smaller RMSD indicates a better topological alignment. After the proteins are superimposed on each other, only those

[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 [5] score matrix and CTSS.

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

topological matches which are within a certain threshold (e.g 3.0 Å) contribute to the overall RMSD value. Alternatively, all the topologically matched residues may contribute to the overall RMSD value. In that case, a longer match would be found for the cost of a larger RMSD.

Let P^M and Q^M denote the set of matched residues of protein chains P and Q, respectively, for $P^M = \{p_1, \ldots, p_N\}$ and $Q^M = \{q_1, \ldots, q_N\}$ where $|P| \geq N$ and $|Q| \geq N$. The RMSD value corresponds to *the total distance among the topologically equivalent residues once they have been optimally superimposed* (after the necessary translation/translocation and rotation) on each other, as follows [8, 12]:

- Calculating the *translation vector*:
  1. Find the *coordinate center*[6] for each set of matched residues P^M and Q^M from the two structures, as
  $$C(P) = \sum_{i=1}^{N} p_i^M \text{ and } C(Q) = \sum_{i=1}^{N} q_i^M.$$
  2. Translate each structure P and Q such that their coordinate centers are located at the origin of the coordinate system: $P^\tau = \tau(P) = P - C(P)$ and $Q^\tau = \tau(Q) = Q - C(Q)$. Hence P^τ and Q^τ represent the translated version of protein P and protein Q respectively.

- Calculating the *rotation matrix* (*Kearsley* method [12]):
  1. Add a selected combination of sums and differences[7] of the matched pair coordinates and generate a symmetric 4×4 matrix M.
  2. Extract the eigenvalues and eigenvectors of M through diagonalization. Select the lowest eigenvalue and use its corresponding eigenvector to calculate the 3×3 rotation matrix $\Re$.
  3. Multiply the translated version of the second structure by the rotation matrix to produce the superimposition of protein Q over protein P, as $Q^{\Re\tau} = \Re \times Q^\tau$.

Finally, calculate the square root of the sum of the Euclidean distance(ℓ_2) between each pair of matched residues of P^τ and $Q^{\Re\tau}$, divided by the number of matched pairs N, hence

$$RMSD = \sqrt{\sum_{i=1}^{N} \ell_2(p_i^\tau, q_i^{\Re\tau})}.$$

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 locality features (e.g., center of mass) in the shape signature would not have been very meaningful because the proteins have different coordinate frames. Should the location 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 3 provides a summary of PADS procedure.

[6]The coordinate center corresponds to the center of mass of each set of matched residues without taking the masses into account ($\mu_i = 1$, for all i).
[7]For more details refer to [8, 12].

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[8] method [9] 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 may be very helpful. 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 [10] interactive database search[9] 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 [3]. Meanwhile the most important obstacle is the fact that various structural alignment techniques may lead to non-identical

[8]For more information please refer to *http://homepages.fh-giessen.de/ hg12640/pdbselect/*
[9]*http://www.embl-ebi.ac.uk/dali/*

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 i^{th} and j^{th} amino acids of P and Q, respectively.

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

1. Directional Shape Signature Extraction:
   - Calculate the center of mass for each amino acid constituent 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 to their corresponding center of mass $C_\odot(p_i)$ and $C_\odot(q_j)$ respectively
   - Extract the direction of each amino acid molecule of p_i and q_j
   - Inspect and include the SSE assignment of each amino acid 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 to the next step.
4. Optimal Superimposition
   - Find the best rotation and translation matrix to superimpose the matched fragment pairs (non-gapped)
   - Report the RMSD and the length of the matched fragment pairs in the sorted order.

Figure 3: PADS structure similarity procedure.

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[10]. 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 [19], DALI [10], and CTSS [3]. 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 4 shows the results of structural alignments on 1AKT:_ and 1CRP:_ protein chains using CE[11] 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 4(b), 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 1 shows a detailed comparison of PADS against DALI[12] [10] 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

[10]Please refer to Table VI in [19]

[11]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*

[12]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/*

Table 1: Comparison of detected similar regions between 1AKT:_ and 1CRP: protein chains using PADS and DALI methods, where $\varphi = \frac{Fragment\ Length}{RMSD}$ ranks the aligned fragment pairs in PADS.

				PADS		DALI	
Rank	φ	Fragment size	RMSD (Å)	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]

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 [3] algorithm. CTSS reports the best aligned fragment pair between 1AKT:_ and 1CRP:_ protein chains to be ([89–113],[140–164]) with length 24 and 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.

Figure 5 (Appendix)[16] depicts the linear SSE structure of 1AKT:_ and 1CRP:_ protein chains without running any alignment, extracted from CATH structural classification database. A close inspection of Figure 4 and Table 1, for the best aligned fragment pair reported by PADS ([131–147],[149–165]), reveals the high similarity of the fragments on their SSE constitutes. Moreover, the second best aligned fragment pair ([10–23],[12–35]), not only preserves the secondary assignment similarity (as suggested by Figure 5), but also has the significance that both regions are being the longest regions (sites) which interact with *ligands*[13] and *metals*.

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

5. CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel data representation technique incorporating multidimensional 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 colaboration 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.

6. REFERENCES

[1] Protein data bank(pdb). http://www.rcsb.org/pdb/holdings.html, 2004.

[2] 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.

[3] 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.

[4] 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.

[5] M. Dayhoff and R. Schwartz. Atlas of protein sequence and structure. *Nat. Biomed. Res. Found.*, 1978. Washington.

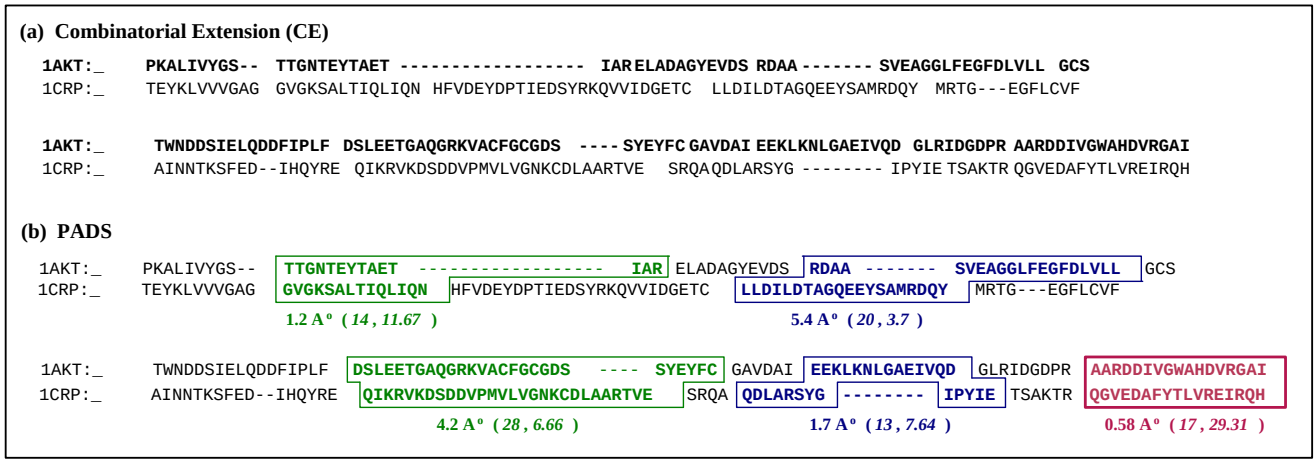

Figure 4: (a) Structural alignment (shown at the sequence level) between 1AKT:_ and 1CRP:_ using CE. (b) The RMSD, extent and score of local fragments discovered by PADS structural alignment (shown at the sequence level) between 1AKT:_ and 1CRP:_ (The output of CE is also shown for comparison purposes).

[6] J. Gibrat, T. Madej, and S. Bryant. Surprising similarities in structure comparison. *Current Opinion Structure Biology*, 6(3):377–85, 1996.

[7] A. Godzik. The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, 5:1325–1338, 1996.

[8] D. Higgins and W. Taylor. *Bioinformatics: Sequence, Structure and Databanks.* Oxford University Press, 2000.

[9] U. Hobohm, M. Scharf, and R. Schneider. Selection of representative protein data sets. *Protein Science*, 1:409–417, 1993.

[10] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Molecular Biology*, 233(1):123–138, 1993.

[11] L. Holm and C. Sander. 3-d lookup: Fast protein database structure searches at 90% reliability. In *ISMB*, pages 179–185, 1995.

[12] S. Kearsley. On the orthogonal transformation used for structural comparisons. *Acta Cryst.*, A45:208, 1989.

[13] G. Lua. Top: a new method for protein structure comparisons and similarity searches. *J. Applied Crystallography*, 33(1):176–183, 2000.

[14] T. Madej, J. Gibrat, and S. Bryant. Threading a database of protein cores. *Proteins*, 23:356–369, 1995.

[15] 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.

[16] C. Orengo and A. Michie. Cath:a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, 1997.

[17] W. Pearson. Rapid and sensitive sequence comparison with fastp and fasta. *Methods in Enzymology*, 183:63–98, 1990.

[18] X. Pennec and N. Ayache. A geometric algorithm to find small but highly similar 3d substructures in proteins. *Bioinformatics*, 14(6):516–522, 1998.

[19] I. Shindyalov and P. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Engineering*, 11(9):739–747, 1998.

[20] 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.

[21] R. Smith and M. Waterman. Identification of common molecular subsequences. *J. Mol. Bio.*, 147(1):195–197, 1981.

APPENDIX

Please find Figure 5 in the next page.

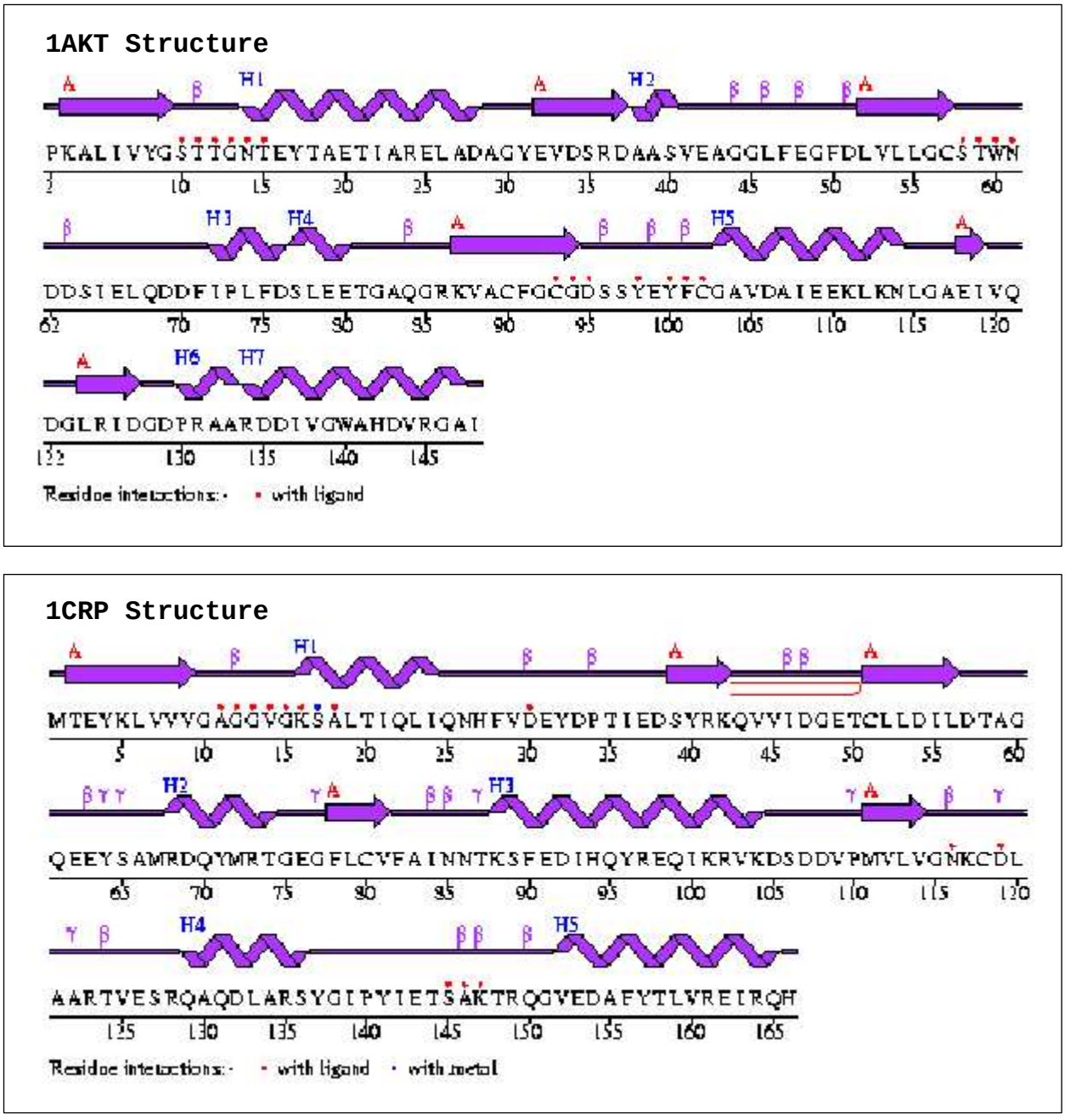

Figure 5: The linear secondary structure of 1AKT:_ and 1CRP:_ protein chains from CATH structural classification database. Residues marked with dots denote significant sites of ligand and metal interactions.