Identifying the Building Blocks of Protein Structures from Contact Maps Using Protein Sequence and Evolutionary Information

Hazem Radwan A. Ahmed
School of Computing and Information Science
Queen's University
Kingston, Ontario, Canada. K7L 3N6
Email: hazem@cs.queenu.ca

Janice I. Glasgow
School of Computing and Information Science
Queen's University
Kingston, Ontario, Canada. K7L 3N6
Email: janice@cs.queensu.ca

***Abstract*— 1D protein sequences, 2D contact maps and 3D structures are three different representational levels of detail for proteins. The problem of protein 3D structure prediction from 1D protein sequences remains one of the challenges of modern bioinformatics. The main issue here is that it is computationally complex to reliably predict the full 3D structure of a protein from its 1D sequence. A 2D contact map has, therefore, been used as an intermediate step in this problem. A contact map is a simpler, yet representative, alternative for the 3D protein structure. In this paper, we focus on the problem of identifying similar substructural patterns of protein contact maps (the building blocks of protein structures) using a structural pattern matching approach that incorporates protein sequence and evolutionary information. These substructural patterns are of particular interest to our research, because they could potentially be used as building blocks for a computational bottom-up approach towards the ultimate goal of protein structure prediction from contact maps. The results are benchmarked using a large standard protein dataset. We assess the consistency and the efficiency of identifying these similar substructural patterns by performing different statistical analyses (e.g., Harrell-Davis Quantiles and Bagplots) on different subsets of the experimental results. We further studied the effect of the local sequence information, global sequence information, and evolutionary information on the performance of the method. The results show that both local and global sequence information are more helpful in locating short-range contacts than long-range contacts. Moreover, incorporating evolutionary information has remarkably improved the performance of locating similar short-range contacts between contact map pairs.**

Keywords - protein structure prediction; contact map; case-based reasoning; evolutionary information; sequence information.

I. INTRODUCTION

Since the human genome sequence was revealed in April 2003, the need to predict protein structures from protein sequences has dramatically increased [2]. Proteins are complex macromolecules that are associated with several vital biological functions for any living cell. Such as, transporting oxygen, ions, and hormones; protecting the body from foreign invaders; and catalyzing almost all chemical reactions in the cell. Proteins are made of long sequences of amino acids that fold into three-dimensional structures. Because protein folding is not easily observable experimentally [3], protein structure prediction has been an active research field in bioinformatics as it can ultimately broaden our understanding of the structural and functional properties of proteins. Moreover, predicted structures can be used in structure-based drug design, which attempts to use the structure of proteins as a basis for designing new ligands by applying principles of molecular recognition [4].

In recent decades, many approaches have been proposed for understanding the structural and functional properties of proteins. These approaches vary from time-consuming and relatively expensive experimental determination methods (e.g., X-ray crystallography [5] and NMR spectroscopy [6]) to less-expensive computational protein modeling methods for protein structure prediction (e.g., ab-initio protein modeling [7], comparative protein modeling [8], and side-chain geometry prediction [9]).

Although the computational methods attempt to circumvent the complexity of the experimental methods with an approximation to the solution (predicted protein structures versus experimentally-determined structures), analyzing the three-dimensional structure of proteins computationally is not a straightforward task. Hence, two-dimensional representations of protein structures, such as distance and contact maps, have been widely used as a promising alternative that offers a good way to analyze the 3D structure using a 2D feature map [19]. This is because they are readily amenable to machine learning algorithms and can potentially be used to predict the three-dimensional structure, achieving a good compromise between simplicity and competency [45].

Despite the exhaustive research done in an effort to reliably predict the structure of proteins from their sequences, the gap between known protein sequences and computationally predicted protein structures is continuously growing because of the computational complexity associated with the problem [10]. The "Divide and Conquer" principle is applied in our research in an attempt to handle such a complex problem, by dividing it into two separate yet dependent subproblems, using a Case-Based Reasoning (CBR) approach [18]. Firstly, a contact map representing the contacts between amino acids is predicted using protein sequence information. Secondly, protein structure is predicted using its predicted contact map [19].

Since contact map prediction offers a possible shortcut to predict protein tertiary structure, researchers have considered various approaches with encouraging results for predicting protein contact maps from sequence information and structural features. Approaches of protein contact map prediction vary from those that apply neural networks [11][12], to those that consider support vector machines [13] and association rules [14]. Various statistical approaches have also been attempted, including correlated mutations [15][16] and hidden Markov models [17].

Our research focuses on the problem of protein structure prediction from contact maps. In particular, we apply a CBR framework [18] to determine the alignment of secondary structures based on previous experiences stored in a case base, along with detailed knowledge of the chemical and physical properties of proteins [19].

The proposed CBR framework is based on the premise that similar problems have similar solutions. CBR solves new problems (e.g., protein structure prediction) by adapting the most similar retrieved solutions of previously solved problems. Several challenges arise here: firstly, how to retrieve the contact maps from the case-base with the most similar solutions (substructural patterns); secondly, how to adapt the new problem "query protein" to the retrieved solutions "template proteins"; thirdly, how to evaluate the adapted solution in an attempt to have a close solution to the native structure of the query protein, which will be saved as a new case in the repository of the CBR system for a later use. These three challenges correspond to the three main phases of our CBR system: *Retrieval*, *Adaptation*, and *Evaluation*, as shown in Figure 1.

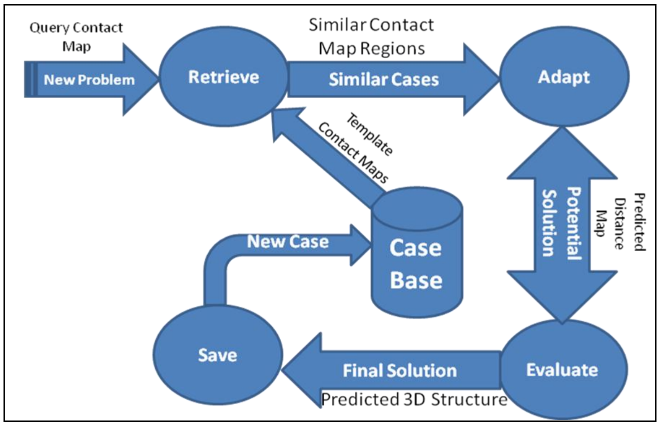


Figure 1. CBR framerwork for determining protein structure from contact maps.

The main challenge of the *Retrieval* phase for our problem domain is to find the template contact maps with the most similar solutions to the query contact map. Thus, it is necessary to have a robust similarity measure for contact maps to reliably compare each new contact map (i.e., a new case), with the contact maps in the case-base (i.e., previously solved cases). This measure is used so that the retrieved contact maps from the case-base have substructural patterns (e.g., secondary structures and structure motifs) that are in common with the new contact map.

Proposing an approach to determine and locate regions of similarity between contact map pairs help to identify these common or similar substructural patterns between the query contact map and every similar retrieved template contact map (known structure) from the case-base, which is crucial for the *Adaptation* phase of the CBR system. The main objective of this phase is to adapt the retrieved solutions to the new problem in the following way. All amino acid residues in the common substructural patterns for the current query contact map that have corresponding residues in the template structure are given the coordinate information from these residues [13]. This paper, an expansion of the work in [1], primarily focuses on the *Adaptation* phase, with the goal of identifying similar substructural patterns between protein contact map pairs using both sequence and evolutionary information.

The paper is organized as follows: Section II provides the reader with the background material required to understand the concepts used in our study. It describes distance and contact maps, gives examples of structural patterns of contact maps, and discusses protein similarity relationships at different representational levels of detail, as well as the structural classification of protein domains. Section III presents the experimental setup and the details of the multi-regional analysis of the contact map method used in our experiments. Section IV discusses the experimental benchmark dataset used in the study and shows the performance of the proposed method using statistical analyses, including a quantile-based analysis and a correlation analysis. The final section highlights the contributions, summarizes the main results of the study, and presents a set of potential directions for future research.

II. BACKGROUND MATERIAL

Contact and distance maps provide a compact 2D representation of the 3D conformation of a protein, and capture useful interaction information about the native structure of proteins. Contact maps can ideally be calculated from a given structure, or predicted from protein sequence. The predicted contact maps have received special attention in the problem of protein structure prediction, because they are rotation and translation invariant (unlike 3D structures). While it is not simple to transfer contact maps back to the 3D structure (unlike distance maps), it has shown some potential to reconstruct the 3D conformation of a protein from accurate and even predicted (noisy) contact maps [20].

A. Distance and Contact Maps

A distance map, D, for a protein of n amino acids is a two-dimensional n x n matrix that represents the distance between each pair of atoms of the protein. The distance may be that between alpha-carbon atoms (Cα) [21], beta-carbon atoms (Cβ), or it may be the minimum distance between any pair of atoms belonging to the side chain or to the backbone of the two residues [22][23]. While the best definition for inter-residue distance is the minimum distance between side-chain or alpha-carbon atoms with a cut-off distance of around 1.0 Ångstrom (0.1 nm) [24], backbone atom-based definitions (e.g., Cα or Cβ distances) with longer distance cut-offs are more readily projected into three dimensions [26]. As shown in Figure 2(a), the darker the distance map region is, the closer the distance of its corresponding atom pairs is. The distance information can be further used to infer the interactions among residues of proteins by constructing another same-sized matrix called a contact map.

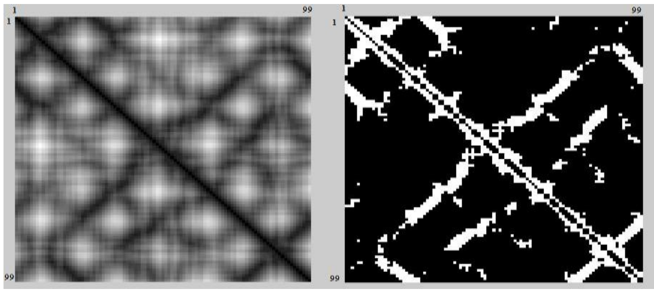
A contact map, *C*, is a two-dimensional binary symmetric matrix that represents pairs of amino acids that are in contact. A pair of residues is considered to be in contact if the distance between their alpha-carbon atoms is less than or equal a predefined threshold, i.e., their positions in the three-dimensional structure of the protein are within a given distance threshold (usually measured in Ångstroms), as shown in Figure 2(b).

An element of the ith and jth residues of a contact map, C*(i,j)*, can be defined as follows [19]:

$$\mathbf{C}\boldsymbol{(i,j)} = \begin{cases} \boldsymbol{1;} \quad \boldsymbol{if\ \ D(i,j)} \leq \boldsymbol{Threshold} \\ \mathbf{0;} \quad \boldsymbol{otherwise} \end{cases}$$

Where *D(i,j)* is the distance between amino acids i and j, ***1*** denotes *contacts* (white), and **0** denotes *no contacts* (black).

According to extensive experimental results presented in [25], contact map thresholds, ranging from 10 to 18 Å allow the reconstruction of 3D models from contact maps to be similar to the protein's native structure. In this paper, different contact map thresholds were applied in a series of experiments in order to find the contact map threshold that best suits our experiments.



(a) (b)

Figure 2. (a) Distance map for a protein of 99 amino acid residues. (b) contact map for the same protein of 99 amino acids after applying a distance threshold of 10 Ångstrom on its distance map. (local contacts < 3.8 Å are ignored – refer to Section III-C for details.)

As shown in Figure 3, a threshold of 8 Å missed some topological information about the protein, whereas a threshold of 12 Å added many contacts that are irrelevant to the topology of the protein. This suggests that a threshold of 10 Å is a good compromise; therefore, it was adopted in our experiments.

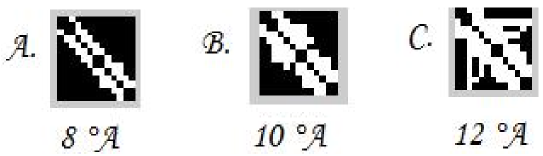


Figure 3. Applying different thresholds for a substructural pattern of contact map.

B. *Structural patterns of Contact Maps*

Different secondary structures of proteins have distinctive structural patterns in contact maps. In particular, an α-helix appears as an unbroken row of contacts between i, i ± 4 pairs along the main diagonal, while beta-sheets appear as an unbroken row of contacts in the off-diagonal areas. A row of contacts that is parallel to the main diagonal represents a pair of parallel β-sheets, while a row of contacts that is perpendicular to the main diagonal represents a pair of anti-parallel β-sheets [26]. Furthermore, contacts between secondary structure elements could also be recognized through a contact map. In general, contacts between α-helices and other secondary structure elements appear as broken rows or "tire tracks". If the two contacting elements are both helices, then the contacts appear in every 3 or 4 residues in both directions, following the periodicity of the helix. If one of the elements is a β-sheet, a periodicity of 2 in the contacts will appear, since the side chains in strands alternate between the two sides of the β-sheet [26].

C. *The Classification of Protein Domains*

The Structural Classification of Proteins (SCOP) database was designed by G. Murzin et al. [32] to provide an easy way to access and understand the information available for protein structures. The database contains a detailed and comprehensive description of the structural and evolutionary relationships of the proteins of known structure. Structurally and evolutionarily related proteins are classified into similar levels in the database hierarchy. Evolutionarily-related proteins are those that have similar functions and structures because of a common descent or ancestor. The main levels in the classification hierarchy of the SCOP database are as follows: 1) *Family* level that implies clear evolutionary relationship, 2) *Superfamily* level that implies probable common evolutionary origin, and 3) *Fold* level that implies major structural similarity.

D. *Protein Similarity Relationships*

Understanding protein similarity relationships is vital for the further understanding of protein functional similarity and evolutionary relationships. Although a protein with a given sequence may exist in different conformations, the chances that two highly-similar sequences will fold into distinctly-different structures are so small that they are often neglected in research practice [30]. This suggests that sequence similarity could generally indicate structure similarity. Furthermore, a pair of proteins with similar structure has similar contact maps [31]. Therefore, as shown in Figure 4, by the transitivity relationship, a logical inference could be drawn regarding the association between sequence similarity and contact map similarity. The premise of the method of multi-regional analysis of contact maps in this paper is based on this transitive similarity relationship between contact map and protein sequence (via structure).

That being said, the counter relationships between contact map and sequence similarity, as well as structure and sequence similarity, are still questionable. This is due to the fact that protein structures are evolutionarily conserved better than protein sequences [33], since the protein sequences evolve rapidly compared to protein structures.

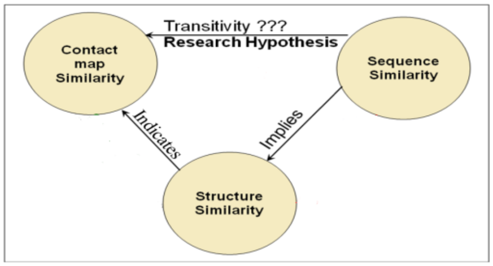


Figure 4. Protein similarity relationships at different levels of detail.

III. Method and experimental setup

This section describes the multi-regional analysis of the contact maps method used in the experiments. The method examines whether sequence similarity information helps in a pattern matching approach to locate regions of similarity in contact maps (the target substructural patterns) that correspond to local similarities in protein structures. The first stage of this method aims to align pairs of protein sequences for each combination pair of contact maps to find the most local similar subsequences. The next stage aims to quantify the similarity of contact maps regions that correspond to these similar subsequences found in the first stage. Finally, different statistical analyses were considered to evaluate the performance of the method, and to determine how well local protein sequence similarity leads to corresponding local contact map similarity.

A. *Experimental Dataset*

The benchmark Skolnick dataset is adopted for our experiments. The Skolnick dataset is a standard benchmark dataset of 40 large protein domains, divided into four categories as shown in Table I. It was originally suggested by J. Skolnick and described in [34]. The dataset has been used in several recent studies related to structural comparison of proteins [34][35][36]. The 40 protein domains are: 1b00A (1), 1dbwA (2), 1nat (3), 1ntr (4), 1qmpA (5), 1qmpB (6), 1qmpC (7), 1qmpD (8), 1rn1A (9), 1rn1B (10), 1rn1C (11), 3chy (12), 4tmyA (13), 4tmyB (14), 1bawA (15), 1byoA (16), 1byoB (17), 1kdi (18), 1nin (19), 1pla (20), 2b3iA (21), 2pcy (22), 2plt (23), 1amk (24), 1aw2A (25), 1b9bA (26), 1btmA (27), 1htiA (28), 1tmhA (29), 1treA (30), 1tri (31), 3ypiA (32), 8timA (33), 1ydvA (34), 1b71A (35), 1bcfA (36), 1dpsA (37), 1fha (38), 1ier (39), and 1rcd (40).

Each domain entry contains a name and its assigned index in parentheses. The domain name is the PDB code for the protein containing it. If multiple protein chains exist, the chain index is appended to the PDB code.

TABLE I. Protein domains in Skolnick dataset

Categories	Global sequence similarity	Sequence length (residues)	Domain indices
1	15-30% (low)	124	1-14
2	7-70% (Med)	170	35-40
3	35-90% (High)	99 (Short)	15-23
4	30-90% (High)	250 (Long)	24-34

B. *Sequence Analysis*

For the sequence analysis stage, we align every combination pair of sequences. The SIM algorithm [37] is used for this purpose. This algorithm employs a dynamic programming technique to find user-defined, non-intersecting alignments that are the best (i.e., with the highest similarity score) between pairs of sequences. The results from the alignments are sorted descendingly according to their similarity score [38].

In this method, we are only interested in alignments of subsequence of at least 10 residues, and at most 20 residues. We are not interested in alignments of length less than 10 residues because these alignments would not form a complete substructural pattern (for example, the lengths of alpha helices vary from 4 or 5 residues to over 40 residues, with an average length of about 10 residues [39]). We are also not interested in long alignments because most methods for contact maps analysis are known to be far more accurate on local contacts (those contacts that are clustered around the main diagonal), than nonlocal (long-range) contacts [40]. Thus, to eliminate one source of uncertainty of the long-range contacts, alignments of a length greater than 20 residues are not considered.

In this experiment, a large penalty for opening a gap is used, since it is evident that affine gap scores [27][28] with a large penalty for opening a gap and a much smaller one for extending it, have generally proven to be effective. Opening gap penalty is a penalty for the first residue in a gap, and extended gap penalty is a penalty for every additional residue in a gap. Therefore, in our experiments, the open and extended gap penalties are set to 10 and 1 respectively. In an effort to analyze pairs of protein sequences, the best 100 local sequence alignments are generated from every pair of sequences. Then, a selection strategy is used to select the two alignments of 10-20 residues with the most and least similarity score (to check the performance in case of low and high similarity).

As for the substitution matrix, BLOSUM62 was adopted to score sequence alignment. The BLOSUM substitution matrix was developed by Henikoff [41] as a new approach for the Percent Accepted Mutation (PAM) scoring matrix that was developed earlier by Margaret Dayhoff who pioneered this approach in the 1970's [29]. Unlike PAM, the BLOSUM62 matrix did an excellent job in detecting similarities even for distant protein sequences.

C. Contact Map Analysis

The second stage of the method is to locate contact map regions that correspond to the most and least similar protein subsequences. In order to unbiasedly analyze the diagonal contact map regions, we ignored local contacts between each residue and itself on the main diagonal. Comparing the main diagonal of contact maps (protein backbone) will neither add meaningful information for their similarity nor dissimilarity (for example, even too distant contact maps will share a similar main diagonal). Furthermore, local contacts with distance less than 3.8 Å are ignored, based on the fact that the minimum distance between any pair of different residues cannot be less than 3.8 Å [40].

Two common similarity measures for binary data that can be used to measure contact map similarity are Simple Matching Coefficient (SMC) [43] and Jaccard's Coefficient (J) [44]. SMC is based on the Hamming distance. If two contact map regions have the same size, we can use SMC to count the number of elements in positions where they have similar values. SMC is useful when binary values hold equal information (i.e., symmetry).

$$SMC = \frac{C_{11} + C_{00}}{S} \tag{1}$$

where C_{11} is the count of nonzero elements (contacts) of both contact maps, C_{00} is the count of zero elements (no-contacts) of both contact maps, and S is the contact map size (i.e., the square of the sequence length for the contact map).

In contact maps, however, binary values do not hold equal information because of the fact that zero values hold no information (they mean there is no contact between protein residues) as opposed to non-zero values where contacts between protein residues occurred. Another drawback of SMC is that it considers counting zero values for both contact maps (C_{00}). These regions represent the "double absence" where there are no contacts for both contact maps, making them of less interest in this study.

On the other hand, Jaccard's Coefficient (J) is widely used in information retrieval as a measure of similarity. It is suitable for asymmetric information on binary (and non-binary) variables where binary values do not have to carry equal information, resolving the first issue of SMC. Furthermore, Jaccard's Coefficient (J) does not consider counting zero elements in the matrix (no contacts), removing the effect of the "double absence" that has neither meaningful contribution to the similarity, nor the dissimilarity, of contact maps. Therefore, Jaccard's Coefficient was chosen as the contact map similarity metric that best suits our experiments.

$$J = \frac{C_{11}}{S - C_{00}} \tag{2}$$

Where C_{11} is the count of nonzero elements (contacts) of both contact maps, C_{00} is the count of zero elements (no-contacts) of both contact maps, and S is the contact map size (i.e., the square of the sequence length for the contact map).

D. Sequence Gap and Region Displacement Problem

The displacement problem happens when a pair of aligned subsequences is very similar (greater than 70%), but their corresponding diagonal contact map regions are not as similar (less than 50-60%). This is noticed to occur as a result of a slight shift in the aligned subsequence pair either because of a gap in the alignment, or because of a slightly shifted alignment. In this case, if the right displacement is considered for one of the aligned subsequence in the correct direction with the correct number of residues, their corresponding diagonal contact map regions will perfectly overlay one another and their similarity can go up to 90%, as shown in Figure 5. The current experimental setup, however, (e.g., open gap penalty, extended gap penalty, etc.) are optimized to minimize the displacement problem. As shown in Figure 7, the proposed method was successful in locating the exact correct boundaries of contact map regions that perfectly overlay one another, in an effort to maximize their similarity. That is, if any boundary is shifted only by one or two residues, the local contact map similarity will be significantly dropped, as shown in Figure 5 and Figure 6.

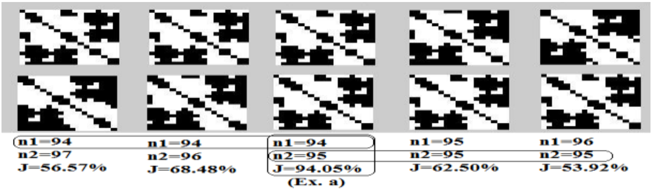


Figure 5. One example of the calculated region boundaries (n1 = 94 & n2 = 95) shows that the selected boundaries have the maximum Jaccard's coefficient (J = 94%) as opposed of 68% and 56% if the lower boundary is shifted by only one residue at a time, or 62% and 53% if the upper boundary is shifted by one residue at a time, instead.

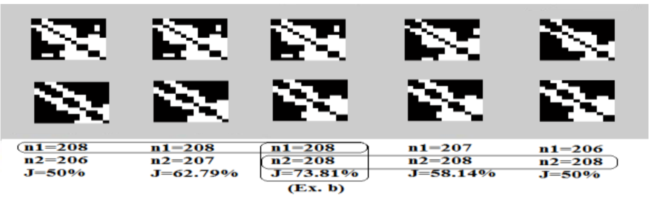
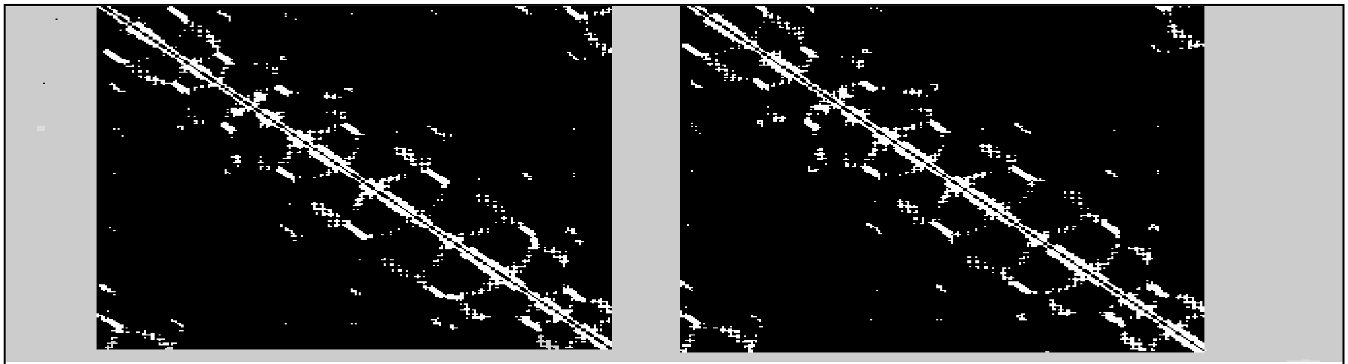


Figure 6. Another example of the calculated region boundaries of (Ex. b) also shows that the selected boundaries have the maximum Jaccard's coefficient (J = 73%) as opposed of 62% and 50% if the lower boundary is shifted by only one residue at a time, or 58% and 50% if the upper boundary is shifted by one residue at a time, instead.



```
        1AMK: 250 residues                1AW2A: 255 residues
39.7% identity in 237 residues overlap; Score: 421.0; Gap frequency: 2.1%

1AMK,       6 PIAAANWKCNGTTASIEKLVQVFNEHTIS-HDVQCVVAPTFVHIPLVQAKLRNP--KYVI
1AW2A,      3 PVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDVAVAPPALFVDLAERTLTEAGSAIIL
              *    *** **       *    *        *   ***      *    *
           -1 Gap                            n1=63+32-1=94 (Ex. a)
1AMK,      63 SAENA-IAKSGAFTGEVSMPILKDIGVHWVIIGHSERRTYYGETDEIVAQKVSEACKQGF
1AW2A,     63 GAQNTDLNNSGAFTGDMSPAMLKEFGATHIIIGHSERREYHAESDEFVAKKFAFLKENGL
               * *     ******  *   **  *    ******** *  * ** ** *       *
                        32 residues          n2=63+32+95 (Ex. a)

1AMK,     122 MVIACIGETLQQREANQTAKVVLSQTSAIAAKLTKDAWNQVVLAYEPVWAIGTGKVATPE
1AW2A,    123 TPVLCIGESDAQNEAGETMAVCARQLDAVINTQGVEALEGAIIAYEPIWAIGTGKAATAE
                  ****   * **  *  *   *  *        *      **** ******* ** *
                       -1 Gap           n1=182+26=208 (Ex. b)
1AMK,     182 QAQEVHLLLRKWVSENIGTDVAAKLRILYGGSVNAANAATLYAKPDINGFLVGGASL
1AW2A,    183 DAQRIHAQIRAHIAEK-SEAVAKNVVIQYGGSVKPENAAAYFAQPDIDGALVGGAAL
               **  *   *    *     **    * *****   ***   * *** * ***** *
                       26 residues      n2 =183+26-1=208 (Ex. b)
```

Figure 7. An illustration of the displacement problem between two highly-similar proteins (1AMK & 1AW2A). The gap length is subtracted from the start position of the upper boundary (n1 of Ex. a) and the lower boundary (n2 of Ex. b), since contact maps have no representation of gaps.

IV. RESULTS AND DISCUSSION

A. The Big Picture

To see the big picture of the problem, an all-against-all pair-wise analysis is performed on the benchmark Skolnick dataset, yielding several hundreds of pairwise alignment instances. The entire results of sequence and contact map similarity of each pairwise instance are presented as a 2D scatter plot to study the correlation between them, as shown in Figure 8. This figure draws a clear distinction between the correlation between sequence similarity and their contact map similarity in the diagonal area (short-range contacts), and the correlation between sequence similarity and their contact map similarity in the off-diagonal areas (long-range contacts).

Firstly, for long-range contacts, no matter how high the sequence similarity is the majority of the corresponding contact map similarity is very low (less than 20%). Thus, even high sequence similarity cannot help to suggest corresponding similarity for the long-range contacts. Secondly, for the short-range contacts, the plot reveals two different trends: 1) when sequence similarity is low (less than 60%), contact map similarity is indiscriminately dispersed between a very low similarity level (35%) and a very high one (90%), making it hard to reliably associate low sequence similarity to short-range contact map similarity. 2) When sequence similarity is high (greater than 60%), contact map similarity is apparently clustered in the upper-right corner of the plot (around 80%), suggesting a high correlation between local sequence similarity and short-range contact map similarity.

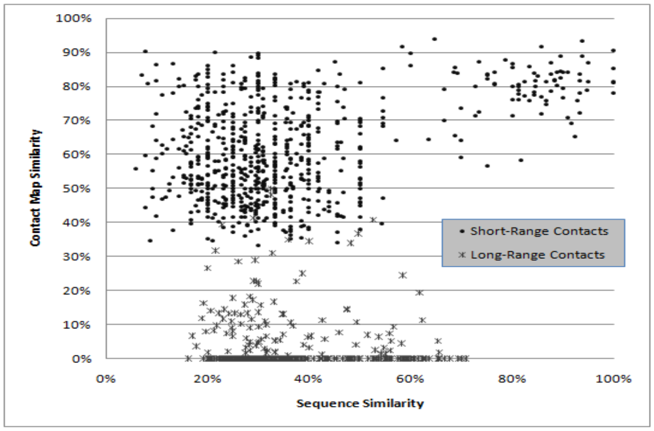


Figure 8. A 2D scatter plot showing the correlation between sequence similarities and their corresponding contact map similarities in the diagonal area (short-range contacts) and the off-diagonal areas (long-range contacts).

B. *Harrell-Davis Quantiles*

In an effort to improve performance in locating similar patterns in the diagonal regions of contact map pairs, evolutionary information (represented in SCOP family information) is proposed to be incorporated with the sequence information. As described in [36], the 40 protein domains of the Skolnick dataset are classified into five SCOP families. Based on SCOP family information, the results are distributed into four different groups: 1) the first group includes the results of pairs of protein subsequences that are most similar and of the same SCOP family. 2) The second group includes the results of pairs of protein subsequences that are most similar and of a different SCOP family. 3) The third group includes the results of pairs of protein subsequences that are least similar and of the same SCOP family. 4) The last group includes the results of pairs of protein subsequences that are least similar and of a different SCOP family.

Quantile-based analysis is performed to compare the four different groups. The q^{th} quantile of a dataset is defined as the value where the *q*-fraction of the data is below q and the (1- *q*) fraction of the data is above q. Some *q*-quantiles have special names: the 2-quantile (0.5 *q*) is called the median (or the 50th percentile), the 4-quantiles (0.25 *q*) are called quartiles, the 10-quantiles (0.1 *q*) are called deciles, and the 100-quantiles are called percentiles. The 0.01 quantile = the 1st percentile = the bottom 1% of the dataset, and the 0.99 quantile = the 99th percentile = the top 1% of the dataset.

Using the online R statistics software in [46], the Harrell-Davis method for 100-quantile estimation is computed for this study. The Harrell-Davis method [48] is based on using a weighted linear combination of order statistics to estimate quantiles. The standard error associated with each estimated value of a quantile is also computed and plotted as error bars, as shown in Figure 10. Error bars are commonly used on graphs to indicate the uncertainty, or the confidence interval in a reported measurement. Figure 10(a) clearly shows that the results of contact map similarity of the same family are much better (higher) than those of a different family as in Figure 10(b). This supports the previous hypothesis that incorporating evolutionary information with sequence information improves the performance of locating remarkably better (highly-similar) diagonal contact map region. Comparing Figure 10(a) and Figure 10(c) reveals that low sequence information considerably deteriorates the method performance, even for the results of the same SCOP family. Whereas, comparing Figure 10(c) and Figure 10(d) demonstrates that with low sequence information, the performance is almost the same (poor), no matter if the protein pairs are of the same or of a different SCOP family.

C. *Bagplots*

A bagplot, initially proposed by Rousseeuwet et al. [49], is a bivariate generalization of the well-known boxplot [50]. In the bivariate case, the "box" of the boxplot changes to a convex polygon forming the "bag" of the bagplot. As shown in Figure [9], the bag includes 50% of all data points. The fence is the external boundary that separates points within the fence from points outside the fence (outliers), and is simply computed by increasing the bag by a given factor. Data points between the bag and fence are marked by a light-colored loop. The loop is defined as the convex hull containing all points inside the fence. The hull center is the center of gravity of the bag. It is either one center point (the median of the data) or a region of more than one center points, usually highlighted with a different color. Therefore, the classical boxplot can be considered as a special case of the bagplot, particularly when all points happen to be on a straight line. The bagplot provides a visualization of several characteristics of the data: its location (the median), spread (the size of the bag), correlation (the orientation of the bag), and skewness (the shape of the bag) [49].

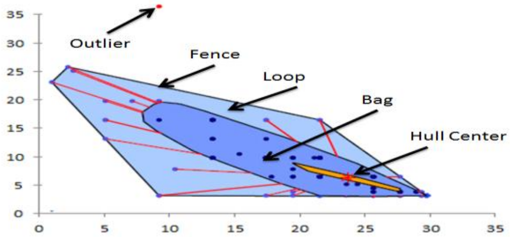


Figure 9. Basic elements of a generic bagplot.

In this statistical analysis, we study the effect of the global sequence similarity on the method performance. Thus, the factor that varies in this analysis is the global similarity information, while other factors will be fixed at their best settings obtained from Figure 10(a). In particular, 1) for the local similarity information, the subsequence pairs of the most local similarity will be used. 2) For the region of similarity, short-range contacts in the diagonal area will be considered. 3) For the evolutionary information, protein pairs will be of the same protein SCOP family. According to the global similarity information of the four categories of the Skolnick dataset (shown in Table I), the pair-wise results are further grouped into four clusters. Namely, 1) Low *vs*. Low, 2) Med *vs*. Med, 3) High *vs*. High (Short), and 4) High *vs*. High (Long). Using the online R statistics software in [47], the bagplots are computed for each cluster, in an effort to perform an in-depth correlation study of the experimental results between short-range contacts and most similar local subsequences at different ranges of global similarity. Although the available samples at the best settings are found to be considerably few, the global sequence information does appear to affect the method performance, as shown in Figure 11. For example, in Figure 11(a), even at the best settings, the center of gravity of the bag is fairly low (around ~62% for contact map similarity) in the case of low global similarity (15-30%). As for the rest of plots, the center of gravity is higher and remains almost the same (around 80% for contact map similarity), when global sequence similarity is medium and high. Samples of the retrieved contact map regions with highly-similar substructural patterns using the proposed pattern matching approach are shown in Figure 12.

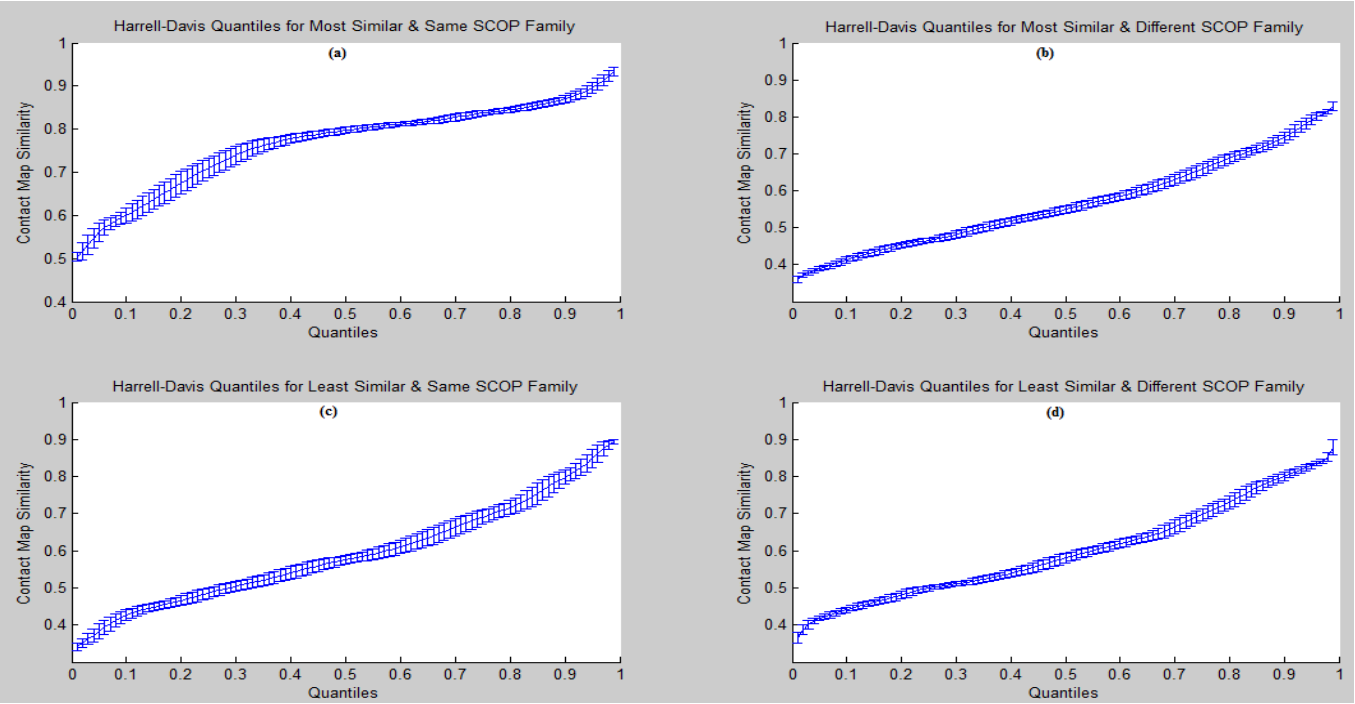


Figure 10. Harrell-Davis quantiles for different categories of the results, along with the error bars of the associated standard error for each reported quantile. (a) Shows the first category of the results of pairs of protein subsequences that are most similar and of the same protein class. (b) Shows category 2 of pairs of protein subsequences that are most similar and of the different protein class. (c) Shows category 3 for pairs of protein subsequences that are least similar and of the same protein class. (d) Shows the last category of pairs of protein subsequences that are least similar and of the different protein class.

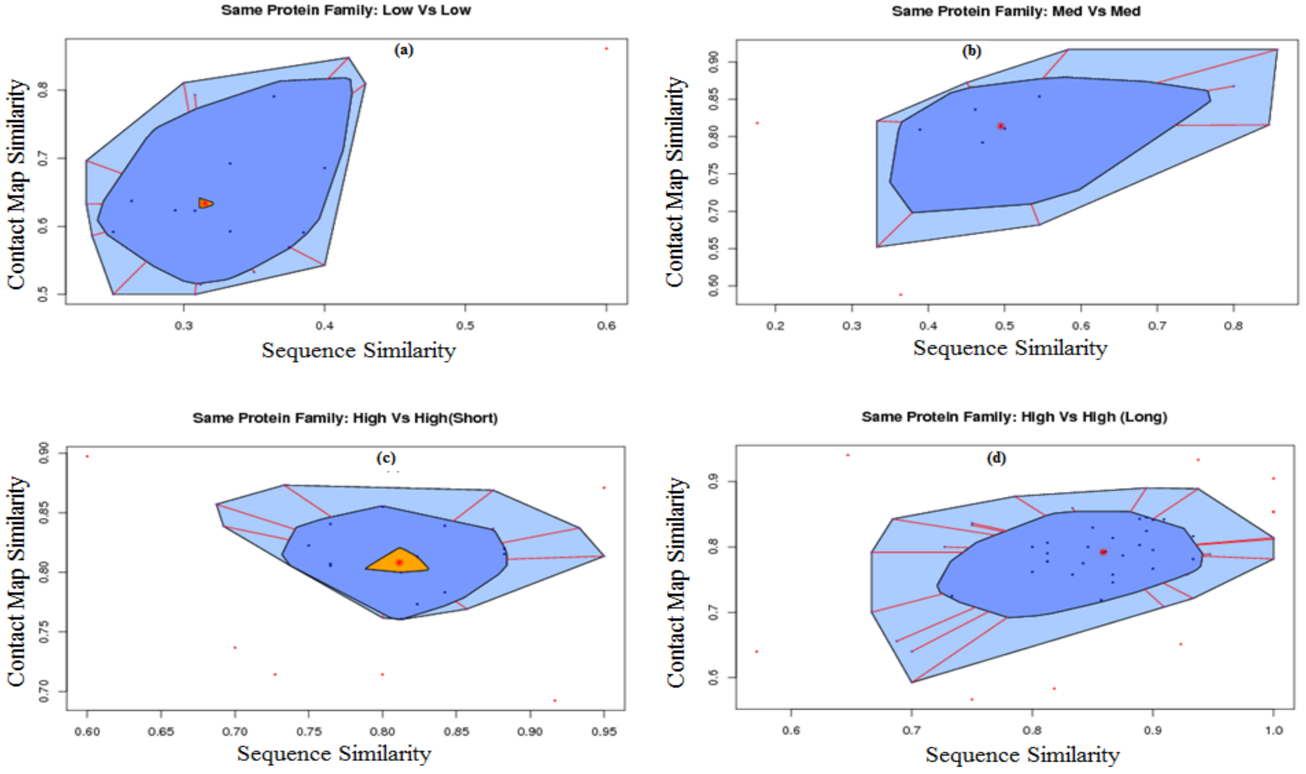


Figure 11. Bagplots for different clusters of the pair-wise results of most similar local subsequences and short-range contacts. (a) Shows the results of first cluster of pairs of protein sequences that are of low global sequence similarity (15-30%). (b) Shows the results of pairs of protein sequences that are of medium global sequence similarity (7 – 70%). (c) Shows the results of pairs of protein sequences that are of high global sequence similarity (35 – 90%) and short length (99 residues). (d) Shows the results of pairs of sequences that are of high global sequence similarity (30-90%) and long length (250 residues).

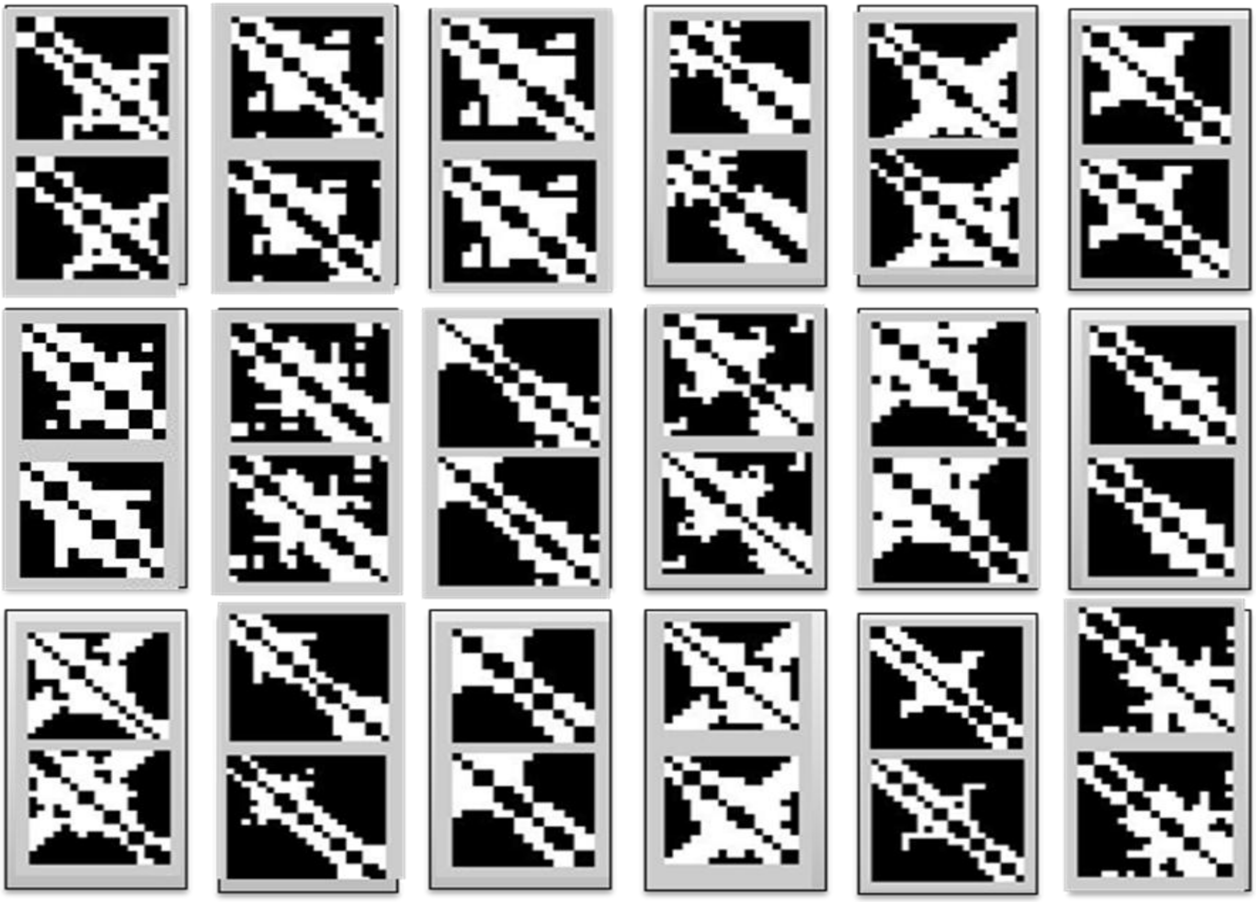


Figure 12. Samples of retrieved contact map regions with similar substructural patterns of 18 contact map pairs with similaty ⩾ 75%.

V. Conclusion And future work

This paper presented the case-based reasoning framework for protein structure prediction from predicted contact maps with the focus on locating similar substructural patterns between the query and template contact maps, as a necessary step for the *Adaptation* phase of the framework. In this paper, a structural pattern matching approach that incorporates both protein sequence and evolutionary information is proposed, with the ultimate goal of identifying the building blocks of proteins in a computational bottom-up approach to protein structure prediction from contact maps. A standard benchmark dataset of carefully-selected 40 large protein domains (Skolnick dataset) is adopted for this study as the experimental dataset.

To the best of our knowledge, this is the first-of-its-kind study to utilize sequence and evolutionary information in locating similar contact map patterns, with no comparable state-of-the-art results. The paper provides an extensive analysis for the three different factors believed to affect the performance of short-range pattern matching in the diagonal area, in particular, 1) local sequence information, 2) evolutionary information, and 3) global sequence information. Firstly, for local sequence information, high sequence similarity (above 60%) has demonstrated (using a scatter-plot analysis) to be a good indicator of a corresponding high diagonal contact map similarity (around 70-90%). This correlation, however, does not appear to be suitable when contacts are long-range (i.e., in the off-diagonal areas of contact maps), or when local sequence similarity is low (less than 60%). Secondly, for evolutionary information, the results proved (using a quantile-based analysis) to be considerably higher when protein pairs have a clear evolutionary relationship, i.e. when they are of the same SCOP family. Lastly, for global sequence information, the results are observed (using a bagplot analysis) to be superior when the global sequence similarity is not low (more than 30%).

Possible future work to improve pattern matching in the diagonal area would be to perform a dynamic expandable multi-regional analysis of contact maps to reduce any possibility of region displacement. That is, we may consider looking further in the neighborhood of the corresponding regions of similar local subsequences. As for the off-diagonal areas, alternative approaches could be employed instead of sequence and evolutionary information that both did not appear helpful in these areas due to the fuzzy nature of long-range contacts at the off-diagonal areas of contact maps [26]. We are currently looking into exploring *Swarm Intelligence* techniques [51] as a promising way to tackle the problem in the off-diagonal areas of contact maps, where the most uncertain, yet important, long-range contacts exist.

References

[1] H. R. Ahmed and J. I. Glasgow, "Incorporating Protein Sequence and Evolutionary Information in a Structural Pattern Matching Approach for Contact Maps", The 3rd International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO'11), Venice, Italy, 2011.

[2] F. Collins, M. Morgan, and A. Patrinos, "The human genome project: lessons from large-scale biology," *Science*, vol. 300, 2003, pp. 286–290.

[3] R. D. Schaeffer and V. Daggett, "Protein folds and protein folding," *Protein Engineering, Design and Selection*, Vol. 24, no. 1-2, 2010, pp. 11-19.

[4] A. C. Anderson, "The process of structure-based drug design," *Chemistry and Biology*, vol. 10, 2003, pp. 787–797.

[5] J. Drenth, "Principles of protein X-ray crystallography," *Springer-Verlag*, New York, 1999, ISBN 0-387-98587-5.

[6] M. Schneider, X. R. Fu, and A. E. Keating, "X-ray versus NMR structures as templates for computational protein design," *Proteins*, vol. 77, no. 1, 2009, pp. 97–110.

[7] A. Kolinski (Ed.), "Multiscale approaches to protein modeling," 1st Edition, Chapter 10, *Springer*, 2011, ISBN 978-1-4419-6888-3.

[8] P. R. Daga, R. Y. Patel, and R. J. Doerksen, "Template-based protein modeling: recent methodological advances," *Current Topics in Medicinal Chemistry*, vol. 10, no. 1, 2010 , pp. 84-94.

[9] C. Yang et al., "Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation," *Bioinformatics*, 2011, doi:10.1093/bioinformatics/btr00.

[10] F. Birzele and S. Kramer, "A new representation for protein secondary structure prediction based on frequent patterns," *Bioinformatics*, vol. 22, 2006, pp. 2628–2634.

[11] G. Pollastri and P. Baldi, "Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners," *Bioinformatics*, vol. 18, 2002, pp. 62–70.

[12] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio, "Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations," *Proteins*, vol. 5, 2001, pp. 157–62.

[13] Y. Zhao and G. Karypis, "Prediction of contact maps using support vector machines," *Proceedings of 3rd IEEE International Symposium on BioInformatics and BioEngineering (BIBE)*, Bethesda, MD, 2003, pp. 26–36.

[14] M. Zaki, J. Shan and C. Bystroff, "Mining residue contacts in proteins using local structure predictions," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 33, no. 5, 2003, pp. 789-801.

[15] D. Thomas, G. Casari and C. Sander, "The prediction of protein contacts from multiple sequence alignments," *Protein Eng*. vol. 9, 1996, pp. 941–948.

[16] M. Singer, G. Vriend and R. Bywater, "Prediction of protein residue contacts with a PDB-derived likelihood matrix," *Protein Eng.*, vol. 15, 2002, pp. 721–725.

[17] Y. Shao and C. Bystroff, "Predicting interresidue contacts using templates and pathways," *Proteins*, vol. 53, 2003, pp. 497–502.

[18] J. Kolodner, "An introduction to case-based reasoning," *Artificial Intelligence Review*, vol. 6, 1992, pp. 3-34.

[19] J. Glasgow, T. Kuo, and J. Davies, "Protein structure from contact maps: a case-based reasoning approach," *Information Systems Frontiers*, Special Issue on Knowledge Discovery in High-Throughput Biological Domains, Springer, vol. 8, no. 1, 2006, pp. 29-36.

[20] I. Walsh, A. Vullo, and G. Pollastri, "XXStout: improving the prediction of long range residue contacts," *ISMB 2006*, Fortaleza, Brazil.

[21] M. Vendruscolo, E. Kussell and E. Domany, "Recovery of protein structure from contact maps," *Folding and Design*, vol. 2, no. 5, 1997, pp. 295-306.

[22] P. Fariselli and R. Casadio, "A neural network based predictor of residue contacts in proteins," *Protein Eng.* vol. 12, 1999, pp.15–21.

[23] L. Mirny and E. Domany, "Protein fold recognition and dynamics in the space of contact maps," *Proteins,* vol. 26, 1996, pp. 391–410.

[24] M. Berrera, H. Molinari, and F. Fogolari, "Amino acid empirical contact energy definitions for fold recognition in the space of contact maps," *BMC Bioinformatics,* vol. 4, no. 8, 2003.

[25] M. Vassura et al., "Reconstruction of 3D structures from protein contact maps," *Proceedings of 3rd International Symposium on Bioinformatics Research and Applications*, Berlin, Springer, vol. 4463, 2007, pp. 578–589.

[26] X. Yuan and C. Bystroff, "Protein contact map prediction," *in Computational Methods for Protein Structure Prediction and Modeling,* Springer, 2007, pp. 255-277, doi:10.1007/978-0-387-68372-0_8.

[27] O. Gotoh, "An improved algorithm for matching biological sequences," *J. Mol. Biol.,* vol. 162, 1982, pp. 705-708.

[28] S. F. Altschul and B. W. Erickson, "Optimal sequence alignment using affine gap costs," *Bull. Math. Biol*., vol. 48, 1986, pp. 603-616.

[29] M. Dayhoff, "A model of evolutionary change in proteins, "*Atlas of Protein Sequence and Structure*, vol. 5, no. 3, 1978, pp. 345 – 352.

[30] E. Krissinel, "On the relationship between sequence and structure similarities in proteomics," *Bioinformatics*, vol. 23, 2007, pp. 717–723.

[31] Dictionary of secondary structure of proteins: available at http://swift.cmbi.ru.nl/gv/dssp/, 11.06.2012.

[32] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, vol. 247, 1995, pp. 536–540.

[33] Gunnar W. Klau, "Comparing structural information in the life sciences: from RNA to metabolic networks," *Symposium on Bioinformatics and Biomathematics*, 2007.

[34] G. Lancia, R. Carr, B. Walenz, and S. Istrail, "101 Optimal PDB structure alignments: A branch-and-cut algorithm for the maximum contact map overlap problem," *Proceedings of Annual International Conference on Computational Biology (RECOMB)*, 2001, pp. 193-202.

[35] W. Xie and N. V. Sahinidis, "A branch-and-reduce algorithm for the contact map overlap problem," *Proceedings of RECOMB of Lecture Notes in Bioinformatics*, Springer, vol. 3909, 2006, pp. 516-529.

[36] P. Lena, P. Fariselli, L. Margara, M. Vassura, and R. Casadio, "Fast overlapping of protein contact maps by alignment of eigenvectors," *Bioinformatics*, vol. 26, no. 18, 2010, pp. 2250-2258. doi: 10.1093

[37] H. Xiaoquin and W. Miller, "A time-efficient, linear-space local similarity algorithm," *Advances in Applied Mathematics*, vol. 12, 1991, pp. 337-357.

[38] SIM: Alignment Tool for Protein Sequences, available at http://ca.expasy.org/tools/sim-prot.html, 11.06.2012.

[39] V. Arjunan, S. Nanda, S. Deris, and M. Illias, "Literature survey of protein secondary structure prediction," *Journal Teknologi*, vol. 34, 2001, pp. 63-72.

[40] Y. Xu, D. Xu, and J. Liang (Eds.), "Computational methods for protein structure and modeling," *Springer*, Berlin, 2007, ISBN: 978-1-4419-2206-9

[41] Henikoff and Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the national academy of sciences*, USA, vol. 89, 1992, pp. 10915-10919.

[42] S. F. Altschul and B. W. Erickson, "Optimal sequence alignment using affine gap costs," *Bull. Math. Biol*., vol. 48, 1986, pp. 603-616.

[43] K Teknomo, Similarity Measurement, available at: http:\\people.revoledu.com\kardi\ tutorial\Similarity, 11.06.2012.

[44] L. Lee, "Measures of distributional similarity," *Proceedings of the 37th annual meeting of ACL*, 1999, pp. 25–32.

[45] H. R. Ahmed and J. I. Glasgow, "Multi-regional analysis of contact maps towards locating common substructural patterns of proteins," *J Communications of SIWN*, vol 6, 2009, pp.90-98.

[46] P. Wessa, "Harrell-Davis quantile estimator", *in Free Statistics Software*, Office for Research Development and Education, 2007, URL: http://www.wessa.net/rwasp_harrell_davies.wasp/, 11.06.2012.

[47] P. Wessa, "Bagplot," *in Free Statistics Software*, Office for Research Development and Education, 2009, URL: http://www.wessa.net/rwasp_bagplot.wasp/, 11.06.2012.

[48] F. E. Harrell and C. E. Davis, "A new distribution-free quantile estimator," *Biometrika*, vol. 69, 1982, pp. 635-640.

[49] P. J. Rousseeuw, I. Ruts, and J. W. Tukey, "The bagplot: A bivariate boxplot," *The American Statistician*, vol. 53, 1999, pp. 382–387.

[50] D. F. Williamson, R. A. Parker, and J. S. Kendrick, "The box plot: a simple visual method to interpret data," *Ann Intern Med*, vol. 110, 1989, pp. 916-921.

[51] S. Das, A. Abraham and A. Konar, "Swarm Intelligence Algorithms in Bioinformatics," *Studies in Computational Intelligence*. vol. 94, 2008, pp. 113–147.