Incorporating Protein Sequence and Evolutionary Information in a Structural Pattern Matching Approach for Contact Maps

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*— Protein structure prediction from the primary sequence remains a major challenging problem in bioinformatics. The main issue here is that it is computationally complex to reliably predict the full three-dimensional structure of a protein from its one-dimensional sequence. A two-dimensional contact map has, therefore, been used as an intermediate step in this problem. A contact map is a simpler, yet representative, alternative for the three-dimensional protein structure. In this paper, we propose a pattern matching approach to locate similar substructural patterns between protein contact map pairs using protein sequence information. These substructural patterns are of particular interest to our research, because they could ultimately be used as building blocks for a bottom-up approach to protein structure prediction from contact maps. We further demonstrate how to improve the performance of identifying these patterns by incorporating both protein sequence and evolutionary information. The results are benchmarked using a large standard protein dataset. We performed statistical analyses (e.g., Harrell-Davis Quantiles and Bagplots) that show sequence information is 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; protien contact maps; structural pattern matching; evolutionary information; harrell-davis quantiles.

I. INTRODUCTION

Since the human genome sequence was revealed in April 2003, the need to predict protein structures from protein sequences has dramatically increased [1]. Proteins are macromolecules with a wide range of biological functions that are vital for any living cell. They transport oxygen, ions, and hormones; they protect the body from foreign invaders; and they catalyze 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 [2], 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 [3].

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 [4] and NMR spectroscopy [5]) to less-expensive computational protein modeling methods for protein structure prediction (e.g., ab-initio protein modeling [6], comparative protein modeling [7], and side-chain geometry prediction [8]). While 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 [9]. 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 [26].

The paper is organized as follows: Section II provides the reader with the background material required to understand the concepts used in this 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 this study and shows the performance of the proposed method using statistical analyses, including a quantile-based analysis as well as a correlation analysis. The final section highlights the contributions and summarizes the main results of the study. It also presents a set of potential directions for future research.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-137-3

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 [10].

A. Distance and Contact Maps

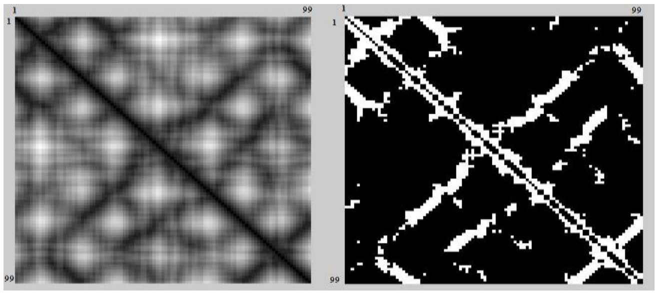
A distance map, *D*, for a protein of *n* amino acids is a two-dimensional *n* x *n* matrix that represents the distance between each pair of alpha-carbon atoms of the protein, as shown in Figure 1(a). The darker the region is, the closer the distance of its corresponding atom pairs is. The distance information can be 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, 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 1(b). According to extensive experimental results presented in [11], 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.

An element of the i^{th} and j^{th} residues of a contact map, C*(i,j)*, can be defined as follows:

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

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



(a) (b)

Figure 1. (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 (1 nm) on its distance map. (local contacts < 3.8 Å are ignored – refer to Section III-C for details.)

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 [12].

C. The Classification of Protein Domains

The Structural Classification of Proteins (SCOP) database was designed by G. Murzin et al. [15] 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 [13]. This suggests that sequence similarity could generally indicate structure similarity. Furthermore, a pair of proteins with similar structure has similar contact maps [14]. Therefore, as shown in Figure 2, 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).

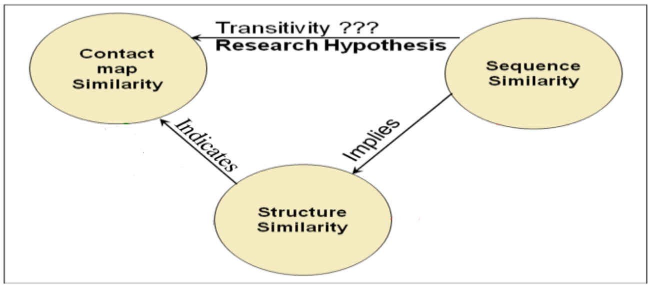


Figure 2. Protein similarity relationships at different levels of detail.

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 [16]. The dataset has been used in several recent studies related to structural comparison of proteins [16][17][18].

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 [19] 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 [20].

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 [21]). 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 [22]. 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, BLOSUM62 [23] is used as the similarity metric to score sequence alignment. As for gaps, the open and extended gap penalties are set to 10 and 1 respectively. This is because a large penalty for opening a gap and a much smaller one for extending it have generally proven to be effective [24]. An open gap penalty is a penalty for the first residue in a gap, and an extended gap penalty is a penalty for every additional residue in it. To analyze pairs of sequences, the best 100 local subsequence 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).

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). Based on the fact that the minimum distance between any pair of different residues cannot be less than 3.8 Å [22], every local contact of each residue and itself that is less than this threshold is ignored.

Jaccard's Coefficient (J) [25] is used as a similarity metric to score contact map regions. J is suitable for measuring contact map similarity, because it does not consider counting zero elements in the matrix (no contacts) of both contact maps, removing the effect of the "double absence" that has neither meaningful contribution to the similarity, nor the dissimilarity, of contact maps.

$$J = \frac{C_{11}}{S - C_{00}} \tag{1}$$

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).

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 3. 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 4, 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 3 and Figure 5.

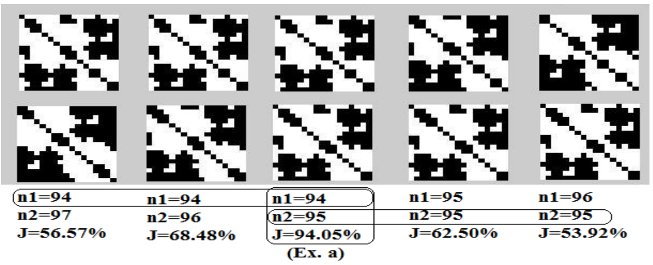
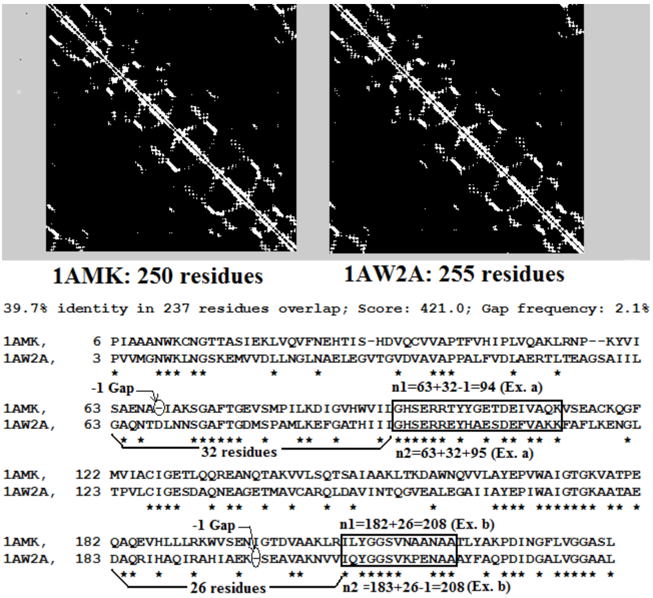


Figure 3. One example of the calculated region boundaries (n1 = 94 & n2 = 95) shows that the selected boundaries have the maximum Jaccard's coefficient (J = 94%) as opposed of 68% and 56% if the lower boundary is shifted by only one residue at a time, or 62% and 53% if the upper boundary is shifted by one residue at a time, instead.



```
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 4. 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.

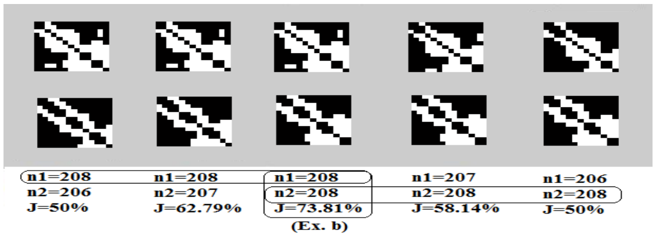


Figure 5. Another example of the calculated region boundaries of (Ex. b) also shows that the selected boundaries have the maximum Jaccard's coefficient (J = 73%) as opposed of 62% and 50% if the lower boundary is shifted by only one residue at a time, or 58% and 50% if the upper boundary is shifted by one residue at a time, instead.

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 6. 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.

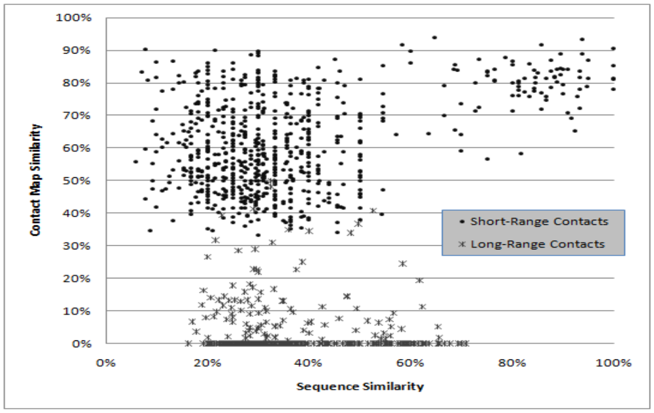


Figure 6. A 2D scatter plot showing the correlation between sequence similarities and their contact map similarities in the diagonal area (short-range contacts) and the off-diagonal areas (long-range contacts).

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 [18], 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 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 (or the 0.5 quantile) is called the median (or the 50th percentile), the 4-quantiles are called quartiles, the 10-quantiles are called deciles, and the 100-quantiles are called percentiles. For example, 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 [27], the Harrell-Davis method for 100-quantile estimation is computed for this study. The Harrell-Davis method [29] 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 7. Error bars are commonly used on graphs to indicate the uncertainty, or the confidence interval in a reported measurement. Figure 7(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 7(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 7(a) and Figure 7(c) reveals that low sequence information considerably deteriorates the method performance, even for the results of the same SCOP family. Whereas, comparing Figure 7(c) and Figure 7(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. [30], is a bivariate generalization of the well known boxplot [31]. In the bivariate case, the “box” of the boxplot changes to a convex polygon forming the “bag” of the bagplot. 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 centre 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) [30].

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 7(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 [28], 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 8. For example, in Figure 8(a), even at the best settings, the centre 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.

V. CONCLUSION AND FUTURE WORK

The paper proposes a pattern matching approach that incorporates both protein sequence and evolutionary information, with the goal of locating similar substructural patterns between contact map pairs. These patterns could ultimately be used as building blocks for a computational bottom-up approach to protein structure prediction from contact maps [9]. 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

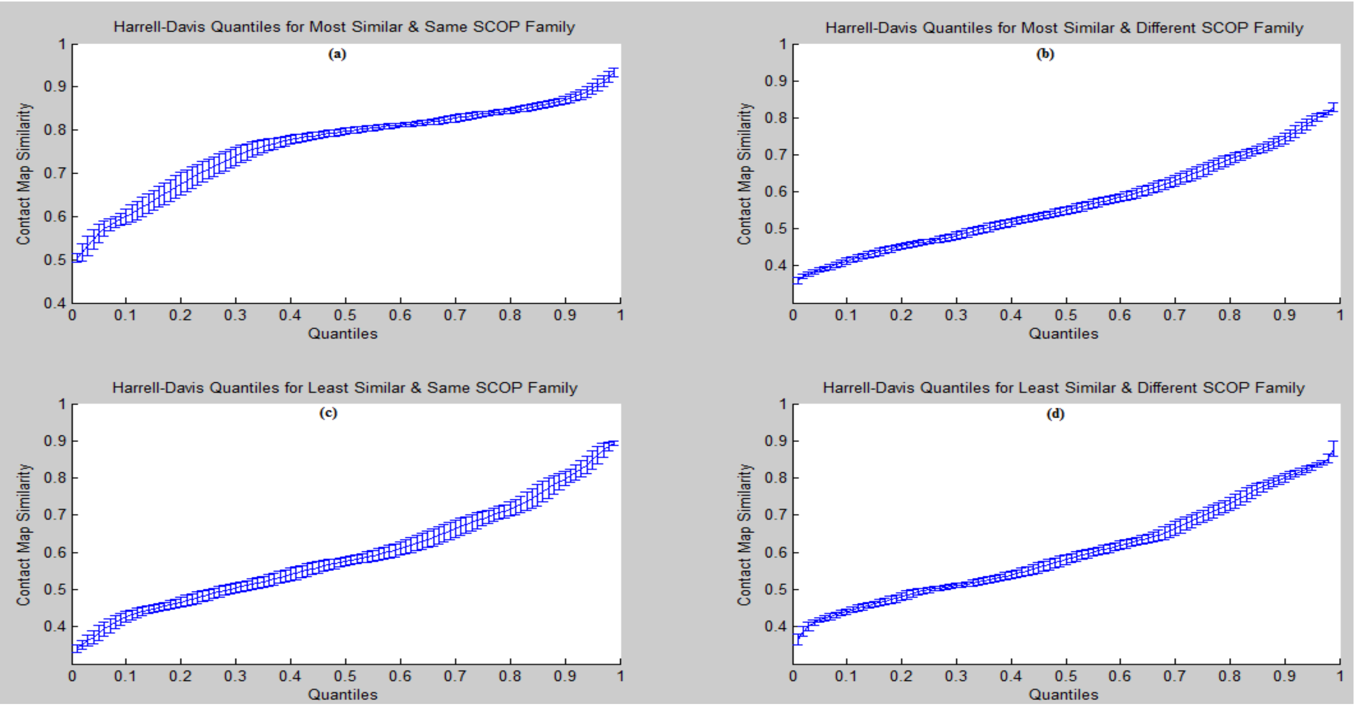


Figure 7. Harrell-Davis quantiles for different categories of the results, along with the error bars of the associated standard error for each reported quantile. (a) Shows the first category of the results of pairs of protein subsequences that are most similar and of the same protein class. (b) Shows 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.

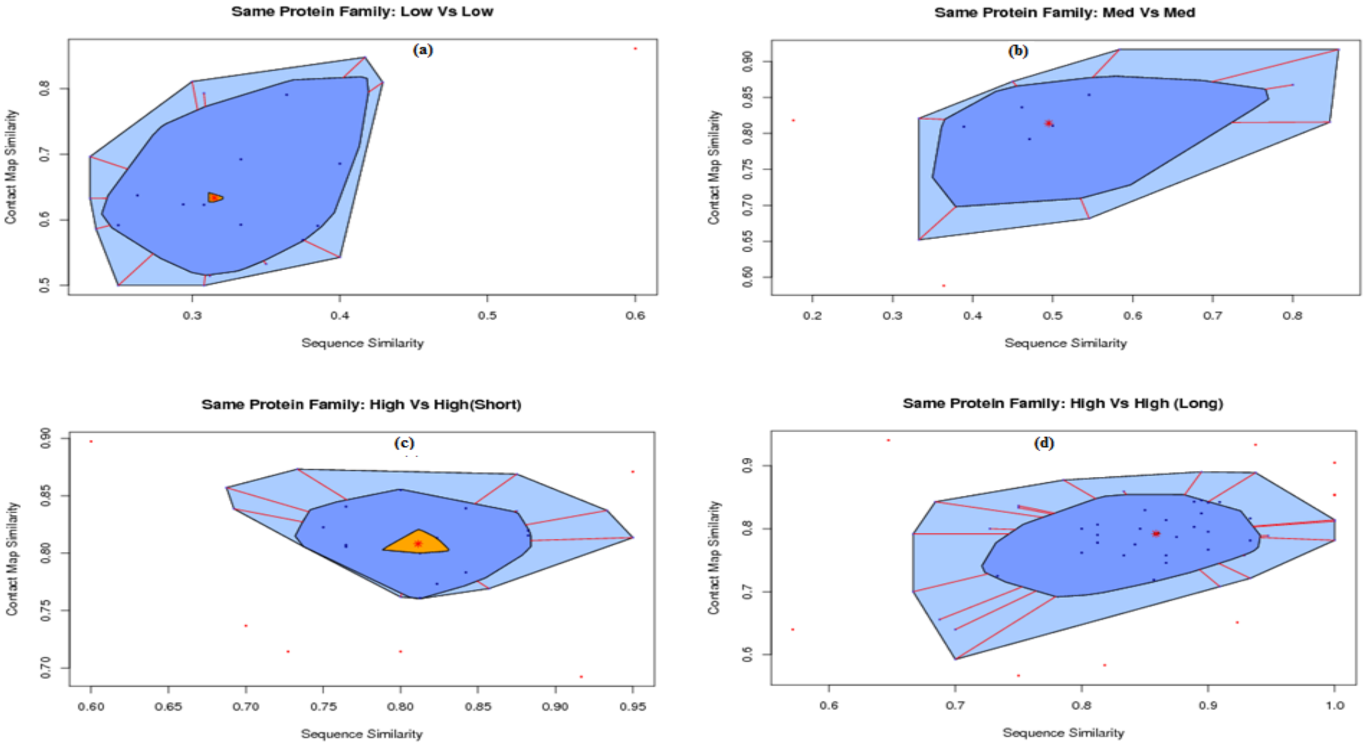


Figure 8. Bagplots for different clusters of the pair-wise results of most similar local subsequences and short-range contacts. (a) Shows the results of first cluster of pairs of protein sequences that are of low global sequence similarity (15-30%). (b) Shows the results of pairs of protein sequences 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).

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. We are currently looking into exploring *Swarm Intelligence* techniques [32] 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] F. Collins, M. Morgan, and A. Patrinos, “The human genome project: lessons from large-scale biology,” *Science*, vol. 300, 2003, pp. 286–290.

[2] 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.

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

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] Dictionary of secondary structure of proteins: available at http://swift.cmbi.ru.nl/gv/dssp/, 14.03.2011.

[15] 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.

[16] 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.

[17] 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.

[18] 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

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

[20] SIM: Alignment Tool for Protein Sequences, available at http://ca.expasy.org/tools/sim-prot.html, 14.03.2011.

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

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

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

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

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

[26] 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.

[27] 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/, 14.03.2011.

[28] P. Wessa, “Bagplot,” *in Free Statistics Software*, Office for Research Development and Education, 2009, URL: http://www.wessa.net/rwasp_bagplot.wasp/, 14.03.2011.

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

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

[31] 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.

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