Mining Dense Patterns from Off Diagonal Protein Contact Maps

M. Om Swaroopa
Department of Computer Science and Engineering
V.R.Siddhartha Engineering College
Vijayawada, Andhra Pradesh, India

K. Suvarna Vani
Department of Computer Science and Engineering
V.R.Siddhartha Engineering College
Vijayawada, Andhra Pradesh, India

ABSTRACT
The three dimensional structure of proteins is useful to carry out the biophysical and biochemical functions in a cell. Protein contact maps are 2D representations of contacts among the amino acid residues in the folded protein structure. Proteins are biochemical compounds consisting of one or more polypeptides, facilitating a biological function. Many researchers make note of the way secondary structures are clearly visible in the contact maps where helices are seen as thick bands and the sheets as orthogonal to the diagonal. In this paper, we explore several machine learning algorithms to data driven construction of classifiers for assigning protein off diagonal contact maps. A simple and computationally inexpensive algorithm based on triangle subdivision method is implemented to extract twenty features from off diagonal contact maps. This method successfully characterizes the off-diagonal interactions in the contact map for predicting specific folds. NaiveBayes, J48 and REPTree classification results with Recall 76.38%, 91.66% and 80.32% are obtained respectively.

General Terms
Protein Contact Map.

Keywords
Protein Contact Maps, Classification, Protein Data Bank, SCOP, J48, REPTree and Naive Bayes.

1. INTRODUCTION
Proteins are an important class of biological macromolecules present in all organ-isms. These are synthesized in the cell as linear chain of amino acids, which then fold to different secondary structures (alpha-helices and beta-sheets) through short and long-range interactions which gives rise to the final three-dimensional shapes useful to carry out the biophysical and biochemical functions. It is known that folding mechanisms are largely determined by a protein's topology rather than it's inter atomic interactions. The SCOP [1] (Structural Classification of Proteins) classification identifies four major structural classes of protein structures, viz. all-alpha, all-beta, alpha/beta and alpha + beta. The Protein Data Bank (PDB) [21] is an archive of experimentally determined three-dimensional structures of biological macromolecules that serves a global community of researchers, educators, and students. The data contained in the archive include atomic coordinates, crystallographic structure factors and NMR experimental data. Aside from coordinates, each deposition also includes the names of molecules, structure information, appropriate ligand and biological assembly information, structure solution, and bibliographic citations. Three-dimensional structures of proteins can be represented well by its two-dimensional distance map and its contact approximation. They are more easily predicted by machine learning methods. The advantage is that contact maps are invariant to rotations and translations. It provides useful information about protein's structure. For example, secondary structure can easily be recognized from it. Alpha-helices appear as dense patterns along the main diagonal since they involve contacts between one amino acid and its four successors, while Beta sheets are dense patterns parallel or anti-parallel to the main diagonal, etc [6]. Over the years, a variety of different approaches have been developed for contact map prediction, including statistical methods using correlated mutations [7], machine learning [6] [8] [9] [10] and threading template based voting scheme. Bioinformatics is an emerging field, undergoing rapid and exciting growth. Data mining techniques will play an increasingly important role in the analysis and discovery of sequence, structure and functional patterns. One of the grand challenges of bioinformatics still remains, namely the protein folding and protein fold prediction problem.

2. RELATED WORK
Contact maps have been extensively used as a simplified representation of protein structures. They capture most important features of a protein's fold, being preferred by a number of researchers for the description and study of protein structures. As per literature [17], contact maps convey strong information about 3D structure and are more compact than distance matrix. Contact maps are suited for prediction based on known properties learned from training data. They are used to predict long range inter-residue contacts and properties of contact maps are attractive for protein structure comparisons. Barah and Sinha [2] are the first to indicate that contact map analysis can be used for protein fold prediction. They demonstrated how the conserved contact patterns within proteins are visually similar. They also emphasized the hypothesis that proteins belonging to the same fold have similar contact maps. This indicates that a closer study of contact maps help in deriving features that pertain to fold information. The information can then be tested on empirically predicted contact maps to identify specific folds [8] [11]. Fraser and Glasgow [12] conducted a study to identify specific regions within contact maps. The authors have chosen contacts corresponding to alpha-alpha interactions in 171 proteins and demonstrate that these exhibit high similarity with the help of Jacquard and cosine metrics. Zaki [4] et al carried out mining of protein of protein contact maps. They discovered dense patterns using sliding window technique and used hashing for storing the results. Amer [18] et al contact maps are constructed by analyzing the dense regions of the map. They analyzed the shapes of clusters using geometric methods. Shi and Zhang [3] extracted secondary

structure features from the distance maps of Protein Contact Network (PCN) to carry out structural class prediction. Recently, SuvarnaVani.K [5] extracted rules based on the diagonal information and the off-diagonal interactions in the contact map for EF-Handlike and Cytochrome-c folds. The objective of this paper is to demonstrate that analyzing patterns in contact maps will contribute to insight positioning of proteins within a specific structural class/fold and folding information. It may sound farcical that contact maps which are in fact constructed from the protein structure/fold prediction. There exist good predictive algorithms that build contact maps from protein sequences using the amino acid features [13]. It will be a good idea to study contact maps closely and derive features pertaining to fold information. If these methods give satisfactory results, the methods can be tested on empirically predicted contact maps to identify dense regions. This would give a new route to fold discovery from the protein sequence via contact maps. Work in this direction has got initiated recently [16] [3] Shamim. In this paper, we carry out a systematic investigation to extract conserved patterns from the contact maps to identify secondary structural elements. We implemented triangle subdivision method (TSM), which captures the locations of the dense clusters. Here, three types of classifiers are used on the dataset. Also, the proteins with chain length up to 1000 can be used with this work.

3. MATERIALS AND METHODS

3.1 Dataset

We consider 96 proteins from All Alpha protein family, which constitutes 48 DNA/RNA binding 3 helical bundle fold and 48 Globin-like fold. The protein 3D structure PDB files are downloaded from Protein Data Bank [21] using the search criteria like no. of chains, chain length and X-Ray resolution. The proteins include those provided by Ding and Dubchak [13] and additionally more proteins are taken from the consensus database of Daggett group called Dynameomics [19]. The reason for addition of proteins to dataset is that, unless the data set size is large enough to effectively sample the distribution. The choice of the folds is made based on the abundance of structures available in the site after removing sequences with more than 90% of identity. This database is designed such that protein domains are classified as the protein fold if they agree in at least two of three classification systems.

3.2 Generation of Protein Contact Map

Contact map is defined as pair-wise, inter residue, two dimensional, symmetric, Boolean matrix of protein 3D structure. We constructed contact map by considering structural data available at Protein Data Bank (PDB). Fig: 1 shows how the PDB file is converted to Off -diagonal matrix via 2D contact map. Only the C atom of each amino acid is chosen and distance is calculated between any two C_α atoms. Distance map is a symmetric square matrix, in which the entry (i, j) represents the distance between the amino acids i and j along the protein primary sequence chain from the N to C terminals. The distance between two residues, has various definitions in the literature. A threshold distance Rc ($7A^o$) is maintained between any C_α -C_α atoms. Thus, a protein structure having N residues, can be represented by a two-dimensional matrix (contact map) of order N× N, whose elements are A (i, j), where i, j are the amino acid residues in the protein sequence. A (i, j) = 1 if the two residues i and j are within the threshold distance, otherwise A (i, j) = 0

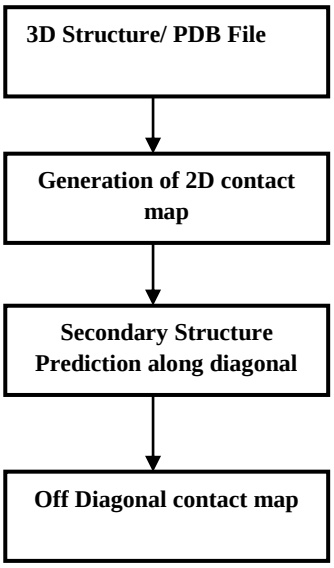

Figure 1: Flow diagram

3.3 Evaluation Measures for Performance Prediction

A confusion matrix contains information about the actual and predicted classifications done by a classification algorithm. Performance of algorithm is evaluated using the data available in the confusion matrix. Table:1 shows the structure of confusion matrix.

Table 1: Confusion Matrix

	Predicted Yes	**Predicted No**
Actual Yes	TP	FN
Actual No	FP	TN

True Positive Rate (TP) is the proportion of positive cases that were correctly predicted. False positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive. The true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly. False negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative. The predictions are evaluated by using the fivefold cross-validation and carried out to show the robustness of the results. Recall (True Positive Rate) gives the proportion of positives out of the total positions predicted as positive and is calculated as

$$\text{Recall} = TP / ((TP + FN)) \qquad (1)$$

3.4 Feature Extraction from off- Diagonal Contact Maps

The challenge is to distinguish the differences in the patterns and automate the procedures such that these can be run on huge data sets. A band of optimal width W is masked along the diagonal in the contact map to focus the study on the Off -diagonal region. We implemented a novel feature extraction scheme, namely, the TSM, which helps in identifying clusters. The contact network is a symmetric square matrix. Without loss of generality, consider the lower tri-angular matrix for this study. Divide the triangle into four equal triangles and assign a label L for left, M for middle, R for right, and T for top triangles. The interactions are labeled using the sub-triangles and it can be seen that majority of the off-diagonal

interactions for 1MGT belong to top, right and middle sub-triangles as shown in Figure: 2. In the data set of all-alpha proteins, DNA/RNA-binding 3 helical bundle fold of 1MGT exhibits Off-diagonal clusters parallel to the diagonal and hence top, right and middle sub-triangles of the lower half of the contact map are occupied.

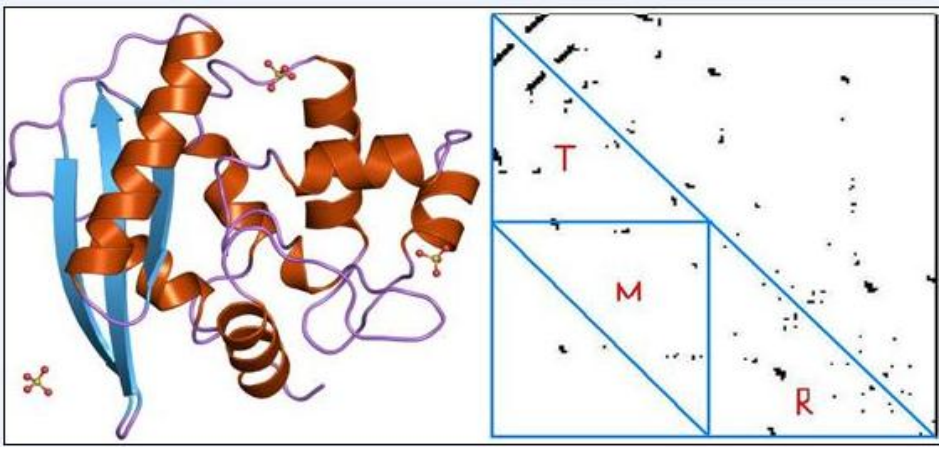

Figure 2: 1MGT, All , DNA/RNA – binding 3 helical bundle, chain length 174

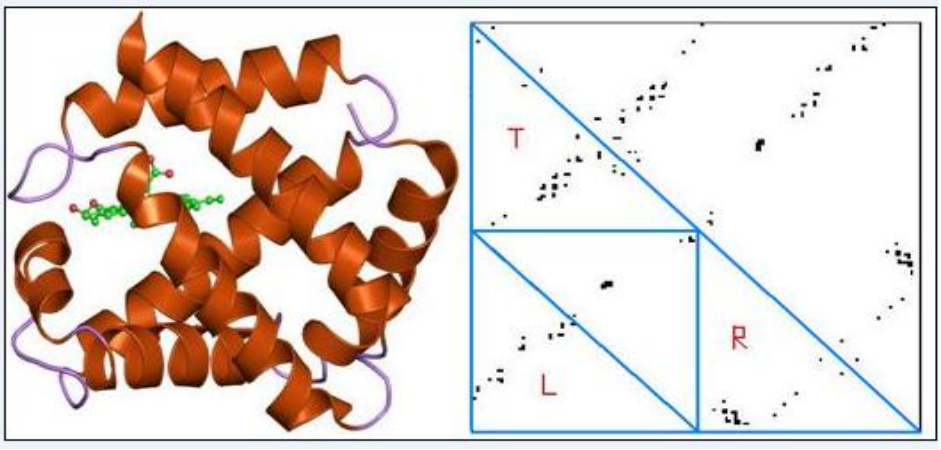

Figure 3: 2HBG, All , Globin-like, chain length 147

These results are tabulated in the Table: 2. In contrast, it can be clearly seen that for 2HBG of Globin-like fold belong to top, right and left sub-triangles of the contact map as shown in Figure 3. These results are tabulated in the Table: 3. hence, these triangle clusters may certainly prove to be useful features. The results in the tables 2 and 3 clearly proves that for DNA/RNA-binding 3 helical bundle fold the dense patterns are concentrated in T,R,M regions in the Off - diagonal contact map and for Globin-like fold in T,R,L regions. These patterns are none other than helices and sheets. Here, we extracted the dense pattern regions. So, the sub patterns are Helix-Helix and Helix-Sheet. An automated procedure needs to be developed that extracts these patterns and further utilize them for fold prediction. The protein contact map can be represented using density of the contacts in the sub triangles. This method is independent of the length of the protein and can be used to build a feature vector. In order to discriminate effectively, we need to iterate this procedure to one more level and subdivide each triangle into four sub-triangles and annotate the clusters within L with labels LL, LM, LR, and LT. Similarly, the triangles within R, M, and T are labeled with RL, RM, RR, RT; ML, MM, MR, MT; and TL, TM, TR, TT, respectively. Each protein contact map is represented as a feature vector of length of 20.

Table 2: The density of interactions in the off -diagonal region (T, R and M sub triangles) in proteins of, DNA/RNA - binding 3 helical bundle fold

Pid	Top	Middle	Left	Right
1FNN	165	30	23	131
1GV2	21	9	0	14
1GVJ	51	19	5	45
1GXQ	30	18	15	23
1IN4	173	19	5	145
1KGS	119	1	0	114
1MGT	82	19	7	46
10N2	44	5	1	16
1T0F	145	32	7	76
1P2F	128	9	0	74

Table 3: The density of interactions in the off -diagonal region (T, R and L sub triangles) in proteins of Globin-like fold

Pid	Top	Middle	Left	Right
1A6M	24	4	17	18
1DLY	27	3	8	17
1HDS	29	3	17	24
1I3D	31	1	11	29
1MYT	30	2	19	23
1S69	29	2	8	26
1SPG	37	3	18	31
1V4X	29	7	16	26
2HBG	36	13	21	24
2D5X	28	4	16	27

4. CLASSIFICATION RESULTS

Machine learning methods, such as support vector machine and neural networks, are powerful classifiers which are being used for protein structure prediction and fold prediction problems with features based on amino acid sequence [15]. In this paper, we illustrated classifiers, namely decision tree with the principal objective of gaining understanding of the results generated for each class [19] [14]. A Bayes classifier NaiveBayes and decision tree with J48, REPTree learning algorithm that is available in the open source software system Waikato Environment for Knowledge Analysis (WEKA) is used for the classification task [20]. We consider the dataset containing 96 proteins in all-alpha class constituting 48 DNA/RNA-binding 3 helical bundles like fold proteins and 48 Globin-like fold proteins. Fivefold cross-validation is carried out in all the experiments to remove any bias that a portion of the data set may impose on the classifier. That is, the total data set is divided into five equal parts: four parts are considered for training to build the model and the remaining one part is taken to test the model. This experiment is repeated five times by considering different parts as test set. Tables 4, 5, 6 show the results in terms of the performance of the classifier with parameters like TP, TN, FP and FN. The class label of 1 is given to Globin-like fold and 4 to DNA-binding 3 helical bundle fold. Clearly from the tables 4, 5, 6 the J48 Decision tree has highest number of correctly classified instances i.e. lowest number of false positives committed by the classifier. Table: 7 show the comparison of results with the previous methods [22, 23, 24,25]. The tabulated results show the accuracy of patterns of a fold extracted using J48 classification algorithm. Results are obtained for binary classification with an average percentage of correctly classified instances value of 95.8% forJ48, 85.8% for Naive Bayes and 82.2% for RepTree classifier. The graph in Figure: 4 shows performance measure, Actual True Positive rate (Recall). J48 has highest True Positive Rate followed by Reptree.

Table 4: Five Fold Cross-validation results for J48 classification of Globin-like fold and DNA/RNA - binding 3 helical bundle

S.No	TP	FP	FN	TN	Correctly classified Instances	In Correctly classified Instances
1	11	0	0	13	100	0
2	11	0	2	11	91.6667	8.3333
3	11	0	1	12	95.8333	4.1667
4	11	0	2	11	91.6667	8.3333
5	11	0	0	13	100	0
Average	11	0	1	12	**95.83334**	4.16666

Table 5: Five Fold Cross-validation results for Naive Bayes classification of Globin-like fold and DNA/RNA- binding 3 helical bundle fold

S.No	TP	FP	FN	TN	Correctly classified Instances	In Correctly classified Instances
1	11	0	4	9	83.3333	16.6667
2	11	0	2	11	91.6667	8.3333
3	11	0	4	9	83.3333	16.6667
4	11	0	4	9	83.3333	16.6667
5	11	0	3	10	87.5	12.5
Average	11	0	3.4	9.6	85.83332	14.16668

Table 6: Five Fold Cross-validation results for RepTree classification of Globin-like fold and DNA/RNA- binding 3 helical bundle fold

S.No	TP	FP	FN	TN	Correctly classified Instances	In Correctly classified Instances
1	10	1	0	13	95.8333	4.1667
2	10	1	4	9	79.1667	20.8333
3	9	2	1	12	87.5	12.5
4	10	1	3	10	83.3333	16.6667
5	10	1	4	9	79.1667	20.8333
Average	9.8	1.2	2.4	106	82.2916	15

Table 7: Results Comparison Table

S.No	Method	Dimension of Feature	Fold (%)
1	Ref[13]	20	49.4
2	Ref[13]	125	56.5
3	Ref[22]	125	58.18
4	Ref[23]	125	61.04
5	Ref[24]	1007	65.5
6	Ref[25]	1007	69.6
7	Ref[3]	3/9	74.55
8	Our method	20	95.83

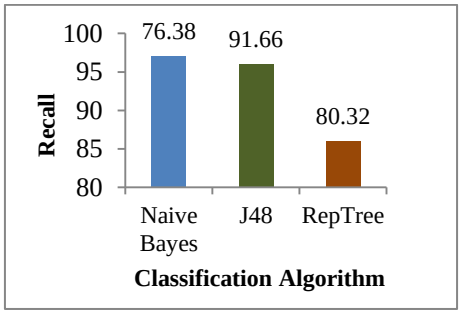

Figure 4: Classification results of NaiveBayes, J48 and Reptree

5. CONCLUSION

This paper validates the hypothesis that contact maps contain useful information that can be utilized to classify the dense regions. The comparison results show the strength of the algorithm used to classify the dense regions in off diagonal contact maps. The classification results biologically imply that EF-Hand and DNA/RNA-binding 3 helical bundle folds exhibit the similar top three dense regions i.e. (TMR and TRM). Also, Cytochrome-C and Globin-like folds exhibit the similar top three dense regions i.e. (TLR and TRL). Further work is carried on All-Alpha class, which includes extracting the frequent substructures from protein contact maps.

The occurrences of helix and sheet in the contact map are predicted using secondary structure prediction algorithm. These predicted secondary structure positions are read and mapped into graphical representation.

6. REFERENCES

[1] Lo Conte L, et al. 2000. SCOP:A structural Classification of Proteins database. Nucleic Acids Res.;28:257-259

[2] Barah P, Sinha S. 2008: Analysis of protein folds using protein contact networks, Pramana, 71 (2):369-78.

[3] Shi J-Y, Zhang Y-N. Fast SCOP classification of structural class and fold using secondary structure mining in distance matrix, PRIB2009, LNBI 2009, 5780:344-353.

[4] Hu J, Shen X, Shao Y, Bystro C, Zaki MJ. 2002: Mining protein contact maps. In: Zaki MJ, Wang JTL, Toivonen HTT, Eds. Second BIOKDD Workshop on Data Mining in Bioinformatics. Edmonton, Alberta, Canada; 310.

[5] S. D. Bhavani and K.Suvarnavani, Somdatta Sinha. 2011. Mining of protein contact maps for protein fold prediction. WIREs Data Mining and Knowledge Discovery, John Wiley & Sons, Volume 1, Pages 362-368, July-August.

[6] N. Gupta, N. Mangal and S. Biswas, 2005: Evolution and similarity evaluation of protein structures in contact map space, Proteins, vol59(2), pp. 196-204.

[7] U. Gobel, C. Sander, R. Schneider, A. Valencia, , 1994: Correlated mutations and residue contacts in proteins, Proteins, vol. 18(4). Pp.309-317.

[8] Y. Zhao and G. Karypis, , 2003:Prediction of Contact Maps Using Support Vector Machines, in proc of third IEEE Symposium on Bioinformatics and Bioengineering, pp. 22-23.

[9] A. Vullo, I. Walsh and G. Pollastri, , 2006: A two-stage approach for improved pre-diction of residue contact maps, BMC Bioinformatics, vol. 7:180.

[10] P. Fariselli, O. Olmea, A. Valencia and R. Casadio, , 2001: Prediction of contact maps with neural networks and correlated mutations, Protein Engineering, vol 14(11), pp. 835843.

[11] Vendruscolo M, Subramanian B, Kanter I, Domany E,Lebowitz J. 1999: Statistical properties of contact maps. Phys Rev E, 59:977984.

[12] Fraser R, Glasgow J, 2007: A demonstration of clustering in protein contact maps for alpha helix paris, ICANNGA 2007, LNCS, 4431: 758-766.

[13] Ding C H Q, Dubchak I., 2001: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 17:349-358.

[14] Alpaydin E. 2005. Introduction to Machine Learning. MIT Press, Prentice-Hall of India, Eastern Economy Edition series.

[15] Shamim MTA, Anwaruddin M, Nagarajaram HA. 2007: Support vector machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs, Bioinformatics, 23:24 3320-3327.

[16] M.J.Zaki, V.Nadimpally, D.Bardhan,and C.Bystroff., 2005: Predicting protein folding pathways. In Data Mining in Bioinformatics, pages 127-141. Springer-Verlag London Ltd.

[17] L.Bartoli, E.Capriotti, P.Fariselli, P.Martelli, and R.Casadio.2007: The pros and cons of predicting protein contact maps. Methods Mol Biol, 413:199-217, September.

[18] M.A.K.Amer F. Al-Badarneh and M.A.Al-Hami. 2008: Improving protein 3D structure prediction accuracy using dense regions areas of secondary structures in the contact map. American Journal of Biochemistry and Biotechnology, 4: 375-384, December.

[19] Dynameomics http://www.dynameomics.org/

[20] Weka http://www.cs.waikato.ac.nz/ml/weka/

[21] Protein Data Bank http://www.rcsb.org/pdb/home/home.do

[22] Chinnasamy, A., Sung, W.K., Mittal, 2005 A.: Protein Structure and Fold Prediction Using Tree-Augmented Naïve Bayesian Classifier. Journal of Bioinformatics and computational Biology 3, 803-820

[23] Shi, J-Y., Zhang, S.-W., Pan, Q., Liang, Y. 2006. Protein Fold Recognition with Support vector machines Fusion Network. Progress in Biochemistry and Biophysics 33, 155-162.

[24] Huang, C.D., Lin, c.-T., Pal, N.R. 2003: Hierarchical Learning Architecture with Automatic Feature Selection for Multiclass Protein Fold classification. IEEE Transactions on NanoBioscience 2,221-232.

[25] Lin, K.L., Lin, C-Y., Huang, C.D., Chang, H.-M., Yang, C.,-Y., Lin, C.-T., Tang, C.Y.,Hsu, D.F. 2007.:Feature Selection and Combination Criteria for Improving Accuracy in Protein Structure Prediction. IEEE Transactions on NanoBioscience 6, 186-196