Automated Identification of Protein Structural Features

Chandrasekhar Mamidipally[1], Santosh B. Noronha[1], Sumantra Dutta Roy[2]

[1] Dept. of Chemical Engg., IIT Bombay, Powai, Mumbai - 400 076, INDIA.
chandra_m_sekhar@hotmail.com, noronha@che.iitb.ac.in
[2] Dept. of Electrical Engg., IIT Delhi, Hauz Khas, New Delhi - 110 016, INDIA.
sumantra@cse.iitd.ac.in

Abstract. This paper proposes an iterative clustering procedure for finding protein motifs which do not necessarily correspond to secondary structures, or to features along the protein backbone. We show the applicability of our method to two important applications namely, protein structure matching, and automatically identifying active sites and other biologically significant sub-structures in proteins.

Key words: Protein Motifs, Robust Features, Protein Structure Comparison, Clustering, Active sites, Biologically Significant Substructures.

1 Introduction

Proteins are composed of different proportions of amino acids (residues) assembled using peptide bonds into unique 3-D structures. Existing protein structure comparison methods can be subdivided into three major categories: methods relying on subsequent overlapping polypeptide fragments: DALI [1] and CE [2]; methods based on Secondary Structure Elements (SSEs) such as VAST [3], TOP [4] and GRATH [5];and methods that search for generic 3-D patterns that are conserved: Clique detection approaches [6, 7], selective groups of amino acids based on physio-chemical properties like Russell [8] and Vishweshwara [9]. Russell through his analysis infers that functional residues within a diameter of 12.0Å are likely to interact. Details of other classification methods can be found in Eidhammer et al. [10] and others [11, 12].

In this work, we propose a iterative nearest-neighbour clustering method of generating meaningful cluster motifs (which may not necessarily include secondary structures or preserve sequential order along backbone). We observe correlation of our motifs with biologically significant protein fragment sizes, and suggest two broad implications: automatic detection of strongly conserved motifs including biologically significant sub-structures (active sites), and matching of protein structures using motif alignment. The subsequent sections of this paper are organized as follows: we describe our method for generating cluster motifs in Sec. 2, and analyse the efficacy of our method. Sec. 3 presents two important applications of our clustering method for finding motifs.

2 Clustering in Protein Feature Identification

We apply the Nearest-Neighbour clustering method for identifying groups of residues (nearest neighbours) as features, for the purpose of protein structure comparison in 3-D space. In this context, residues (C$_\alpha$ atoms) when represented as points in 3-D space can be thought to be interacting if they fall within a diameter of d_{thresh} (Å). Importantly, when *a priori* information of class labels is unavailable, unsupervised learning (clustering) methods are required to partition the collection.

We represent a protein as a collection of its C$_\alpha$ atoms, where each C$_\alpha$ atom is representative of the amino acid residue it lies in. We use a distance d_{thresh} to identify neighbouring C$_\alpha$ atoms. Our method uses three phases. (Fig. 1(a) outlines our method.) In the first phase, Nearest Neighbour clustering is used to create a joint participation matrix $V(q,r)$ where q and r represent random indices of C$_\alpha$ atoms. The joint participation of a pair of points a_q and a_r is expressed in terms of a binary matrix $V(q,r)_{N\times N}$. If a_q and a_r fall in same cluster then the value 1 is assigned to $V(q,r)$ (otherwise 0). A consensus of joint participation of pairs over M iterations $\bigcup_{q,r}^{N} s(q,r)$: (Fig. 1(a), second phase):

```
(* --- First Phase: Basic NN --- *)
1. SELECT a random $C_\alpha$ atom, assign it to a new cluster
2. WHILE there are unexamined $C_\alpha$ atoms left REPEAT
    A. SELECT a new random $C_\alpha$
    B. FIND its distance $d_{\min}$ to centroid of nearest cluster
    C. IF $d_{\min} \leq d_{thresh}$ THEN
            ASSIGN it to the corresponding cluster
            RECOMPUTE the cluster centroid
       ELSE assign it to a new cluster
(* --- Second Phase: --- *)
REPEAT steps 1, 2 $M$ times (* Sec 2 *)
COMPUTE $V(q,r)$ each time
(* --- Third Phase: Best Clusters --- *)
OUTPUT consensus clusters (* Sec 2 *)
```

(a)

Threshold (Å)	S(K) Mean	GK Mean
4.0	0.066	0.985
4.5	0.121	0.980
5.0	0.177	0.977
5.5	0.246	0.974
6.0	0.317	0.971
6.5	0.413	0.966
7.0	0.559	0.960
7.5	0.727	**0.952**
8.0	**0.878**	0.944
8.5	1.204	0.932

(b)

Fig. 1. (a) Robust, repeatable and fast clustering: Sec. 2, and (b) Validity indices computed for different thresholds on 236 randomly selected proteins of varying sizes.

$$s(q,r) = \frac{1}{M}\sum_{m=1}^{M}\sum_{\forall q,r}^{N} V(q,r) \tag{1}$$

M is an important parameter governing the time complexity of the entire procedure: our experiments with a very large number of PDB files show that it is reasonable to take M as a small constant. The final (third) phase involves a ranked search (in decreasing order) to identify high scoring $s(q,r)$ pairs. A final partition $P = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ is constructed by reading points in the order of decreasing score using the same clustering algorithm, where K is the optimal number of clusters obtained in the consensus procedure. We refer to the centroids of these consensus clusters as 'consensus centroids' earlier work by Ghosh et al. on a artificial data set [13]. We observe that an empirically determined value of 100 iterations works well for almost all proteins in the PDB database. A threshold value of 6.0 Å has been used for individual attempts of clustering and further issues are considered in the next section.

We show the efficacy of our clustering method using standard cluster validation indices such as the Scattering Density Index and the Goodman-Kruskal Index [14–16]. Fig. 1(b) shows the variation in the above cluster validation indices with the clustering threshold. A low value of $S(K)$ indicates a low average variance i.e., an indication that the clusters are compact. This is desirable in any clustering procedure. Fig. 1(b) indicates that this behaviour occurs for clustering thresholds of up to ∼8.0 Å. The ratio then exceeds 1.0 indicating average variation across clusters is more than the variations as a whole (i.e., the variation of the residues from the protein centroid). A large value of GK indicates good clustering and the value will always be such that $GK \in [-1, 1]$. A value of -1 is indicative of all atoms being wrongly assigned; and a value of 1 is indicative of correct clustering. In Fig. 1(b), we observe that we get a large value of the Goodman-Kruskal Index for small clustering thresholds (<7.5Å, with smaller thresholds resulting in correctly classified points. We note from our analysis of clustering indices that suitable features can be obtained within a clustering threshold range of 6.0 to 7.5 Å. This threshold range ensures that the maximum separation between atoms in a cluster would be of the range 12.0 to 15.0 Å. *An interesting observation from our cluster threshold analysis is that for a threshold of 6 Å, the average number of C_α atoms per cluster is 5 for α and $\alpha + \beta$ and 4 for β class proteins.* Our feature size therefore compares well with methods like DALI [1] (which uses hexapeptide backbone based fragments), and the empirical value of 12.0 Å used to identify interacting residues in the detection of active sites [8]. We make a case for choosing a clustering threshold of about 6 Å, a value which turns out to be biologically significant [8], in addition to being statistically favourable as determined above using cluster validity indices.

3 Cluster Significance: Two Important Applications

This section examines the application of our fast and robust cluster-finding strategy for two important tasks - identifying biologically significant sub-structures in proteins, and comparing two protein structures.

3.1 Automatic Detection of Biologically Significant Substructures

We observe that biologically significant clusters superimpose well due to the retained sub-structure, with observed substitutions typically involving amino acids with similar properties. Fig. 2(a) shows results of experiments with clusters in the Phospholipase family (alpha class structures), with a clustering threshold of 6 Å. Each row in the table represents sample clusters in the protein family, with a cluster represented by single-letter codes corresponding to its constituent amino acid residues. Clusters comprising other class structures found are not

Amino Acid Labels	Secondary structure	Proteins
CFV$\underline{\text{H}}$D	$\alpha\alpha\alpha\alpha\alpha$	1JIA, 1PSJ, 1A2A, 1VIP, 1VPI
CFV$\underline{\text{H}}$K		1PPA
CEC$\underline{\text{D}}$K	$\alpha\alpha\alpha\alpha\alpha$	1J1A, 1CL5
CEC$\underline{\text{D}}$R		1A2A, 1PSJ, 1VIP, 1VPI
DATDRC	c$\alpha\alpha\alpha\alpha\alpha$	1JIA, 1PPA, 1PSJ, 1VIP, 1VPI
DGTDRC		1A2A
1KKMTGK	$\alpha\alpha\alpha\alpha\alpha$cc	1JIA
1KEETGK		1A2A
1LEETGK		1CL5
1LQETGK		1PPA
1LQKTGK		1VPI
IVKMTGK		1VIP

(a)

1JIA	1VIP	SW_{score}	RMSD (Å)
LLQFRK	LFQFAE	15	0.20
PDILC	PPSQC	11	0.06
CECDK	CECDR	31	0.17
IKKMTGK	IVKMTGK	28	0.16
AICFRD	ATCFRD	29	0.22
CYEKV	CYEKV	30	0.14
PVVSYA	PLSSYS	18	0.56
TYS-VCG	SYS-VCG	31	0.31
DATDRC	DATDRC	35	0.08
WKNGTI	FQNGGI	16	0.30
CFVHD	CFVHD	33	0.06
NLKTY	NLNTY	22	0.09

(b)

Fig. 2. (a) Sample clusters of common substructure 'matches' occurring in some proteins of the Phospholipase family. Biologically significant motifs remain conserved, with substitutions typically involving amino acids with similar properties. The underlines letters H (His) and D (Asp) indicate residues in the active site. (b) Comparison of sub-structures in 1JIA and 1VIP using sequence alignment and the Kabsch superimposition method. Sequence alignment scores are generated using the Smith-Waterman matching algorithm, with the BLOSUM62 scoring matrix - details in Sec. 3.1.

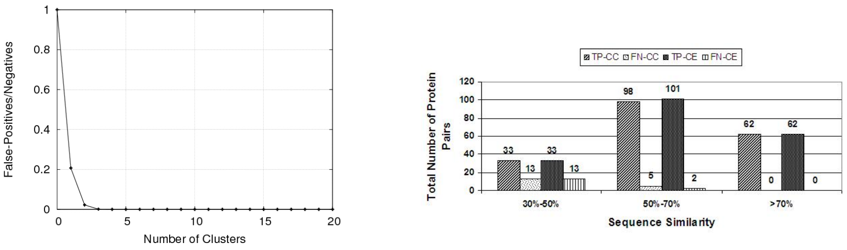
shown due to space constraints. Phospholipase A2 is a lipolytic enzyme that is involved in the biosynthesis of prostaglandin and other mediators of inflammation. The underlined amino acids in the table are part of the active site, and are implicated in the catalytic mechanism. The neighbourhood of these amino acids is also preserved with allowed substitutions of amino acids with similar properties. For example, the CECDK motif in 1J1A has lysine (K) substituted by arginine (R) in 1A2A; these are both positively charged polar amino acids. Superposition of CECDK motif in 1J1A and CECDR motif in 1A2A yields an rmsd of 0.15 Å indicating a good fit. In comparison, the overall sequence similarity between these two proteins is 59%. It is also evident from Fig. 2(a), that the secondary structure motifs for related clusters are conserved, and that the clusters in this case are likely parts of secondary structural elements (SSEs).

The quantitative extent to which the motifs are similar can be identified as follows. As an example, we compare clusters from 1JIA and 1VIP: these phospholipases share 63.6% sequence identity. We wish to compute an optimal cluster correspondence while accounting for possible amino acid substitutions (algorithm in section 3.2. Fig. 2(b) shows quantitative results for the cluster correspondence in proteins 1JIA and 1VIP. We analyze clusters having a minimum of 4 amino acid residues. Nonsequential residues in a cluster are separated by a '-' symbol. Matched residues are then superposed using the Kabsch method and corresponding rmsd values are shown.

3.2 Protein Structure Matching

In this section, we show the utility of clusters as features for fast and efficient protein structure comparison, using dynamic programming and the Kabsch rotation matrix. The method involves three steps: Representation of clusters, matching, and refinement of matches. We search for inter-cluster chemical similarity across a pair of proteins using the Smith-Waterman (SW) local alignment approach. Only those clusters constituted of at least 4 residues are chosen for alignment. A suitable threshold for the SW score ($SW_{score} >= 10$), derived empirically, is used to detect similarity. Simultaneously, the matched residues are superimposed to within a threshold rmsd value ($R < 0.8$). The resulting pair of matching subsequences are retained as if there are at least 4 residue equivalences.

Towards evaluating the efficacy of our structure comparison approach we have considered 213 similar protein pairs and 868 dissimilar pairs covering five major families: Hemoglobin, Immunoglobulin, Lysozyme, Phosphilipase and Plastocyanins. Similar proteins are observed to have similar substructures even when the extent of sequence similarity is low (data not shown due to space constraints). The minimum number of consensus clusters required to correctly classify protein pairs as similar or dissimilar is seen to be 3 from Figure 3(a). This therefore corresponds approximately to at least a 12 residue match given a minimum cluster size of 4 residues. It should be reiterated that our comparison procedure discounts clusters with less than 4 residues, and also cluster pairs with less than 4 residue equivalences. *It is therefore inappropriate to compare rmsd values from a full protein alignment with the value derived from the structure comparison approach described above.* In Fig. 3(b) we show a comparison of the structure comparison capabilities of our method versus that of CE. We require that homologous pairs have at least three matching cluster pairs in our method. For CE, similarity is claimed given a Z-score > 5 and an rmsd < 4.0 Å[2].



(a) (b)

Fig. 3. (a) Cut-off threshold for similarity across pair of proteins at 6Å across 1081 pairs of protein structures. (b) Number of homologous protein pairs classified as true positives and false negatives. TP-CC, FN-CC represent true positives and false negatives predicted by our method. Similarly TP-CE, FN-CE represents predictions of CE

.

We further analyzed some difficult protein pairs: 1SCE and 1PUC possess high sequence identity but were treated as unique structures by DALI as observed by Shindyalov et al. [2]. On the same proteins we obtained 9 aligned cluster pairs (45 residues) as compared to CE (93/1.26Å). Similarly a comparison of 2SEC and 1EGP yielded 11 residue matches while CE detected 34 (1.2Å). Alignment of the sequentially dissimilar protein pair 1A2Y and 1A4J revealed a potentially strong motif WVRQPPGK–EWL (36-43,46-48) and WYLQKPGQ–KLL (40-47,50-52) respectively. 1HDA and 1PBX, proteins that share 48% sequence identity resulted in a 72 residue alignment as compared to 141 residues aligned by CE. In Immunoglobulins two homologous pairs 1BBD - 1DFB and 1BBD - 4FAB gave 29 and 25 residue matches and were correctly classified as similar; while CE aligned them with rmsd 4.27Å and 4.42Å respectively.

References

1. Holm, L., Sander, C.: Protein Structure Comparison by Alignment of Distance Matrices. Journal of Molecular Biology **233** (1993) 123 – 138
2. Shindyalov, I.N., Bourne, P.E.: Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path. Protein Engineering **11** (1998) 739 – 747
3. Gibrat, J.F., Madej, T., Spouge, J.L., Bryant, S.H.: The VAST Protein Structure Comparison Method. Biophysics Journal **72** (1997) MP298
4. Lu, G.: TOP: a new method for protein structure comparisons and similarity searches. Journal of Applied Crystallography **33** (2000) 176 – 183
5. Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J., Orengo, C.: Recognizing the fold of a protein structure. Bioinformatics **19** (2003) 1748 – 1759
6. Grindley, H.M., Artymiuk, P.J., Rice, D.W., Willett, P.: Identification of Tertiary Structure Resemblance in Proteins Using a Maximal Common Subgraph Isomorphism Algorithm. Journal of Molecular Biology **229** (1993) 707 – 721
7. Koch, I., Lengauer, T., Wanke, E.: An algorithm for finding maximal common subtopologies in a set of protein structures. Journal of Computational Biology **3** (1996) 289 – 306
8. Russell, R.B.: Detection of Protein Three-Dimensional Side-chain Patterns: New Examples of Convergent Evolution. Journal of Molecular Biology **279** (1998) 1211 – 1227
9. Kannan, N., Vishveshwara, S.: Identification of side-chain clusters in protein structures by a graph spectral method. Journal of Molecular Biology **292** (1999) 441 – 464
10. Eidhammer, I., Jonassen, I., Taylor, W.R.: Structure Comparison and Structure Pattern. Journal of Computational Biology **7** (2000) 685 – 716
11. Kolodny, R., Petrey, D., Honig, B.: Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Current Opinion Structural Biology **16** (2006) 393–398
12. Dundas, J., Binkowski, T., Dasgupta, B., Liang, J.: Topology independent protein structural alignment. BMC Bioinformatics **8** (2007) 388
13. Strehl, A., Ghosh, J.: Cluster ensembles – a knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research **3** (2002) 583 – 617
14. Goodman, L., Kruskal, W.: Measures of associations for cross-validations. Journal of American Statistical Association **49** (1954) 732 – 764
15. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: part II. SIGMOD Record **31** (2002) 19 – 27
16. Bolshakova, N., Azuaje, F.: Cluster Validation Techniques for Genome Expression Data. Signal Processing **83** (2003) 825 – 833