Analysis and Visualization of Long-Range Contacts and Networks in Homologous Families of Proteins

K.M. Saravanan and S. Selvaraj*

Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli 620 024, Tamil Nadu, India

Abstract: Proteins are grouped into various families according to their evolutionary origin. Analyzing such types of families based on their inter residue interactions is crucial because algorithms that search for pair wise homologies can miss important relations and produce false hits. Several statistical models have been created to aid in the classification but so far had only partial success. In this work, we have analyzed the variation of long-range contacts in different bin intervals as well as characterized the long-range order in a set of 37 families of homologous proteins belonging to different structural classes. The results reveal the family specific long-range contacts as well as variation of long-range order in different structural classes. The pair-wise residue preference to form long-range contacts reveals the dominance of hydrophobic residues irrespective of the structural class. We also provide visual examples of long-range contact network pattern in the different structural classes.

Keywords: Inter-residue interaction, homologous family, long-range order, contact network.

BACKGROUND

Proteins evolved from a common ancestor are said to be homologues and to constitute a "family" with potentially similar structures, functions, and interactions. Analysis of a set of similarly folded proteins with distinct amino acid sequences, such as homologues, can help in identifying residues and regions of polypeptide chains that are likely to be important in the formation and stability of the fold. The problem of identifying "real" protein families based on amino acid sequence conservation has been the subject of extensive debate, because algorithms that search for pair wise homologies can miss important relations and produce false hits. Automatic classification of proteins into homologous super families, by looking at their three dimensional structure has been a long goal for scientists studying proteins. Several statistical models have been created to aid in the classification but so far had only partial success. Correct functional and evolutionary classification of new structures is difficult for distantly related proteins and error-prone using simple statistical scores based on sequence or structure similarity.

There are databases, which contains homologous families of proteins that have been classified by their structural classes and folds. A fully automated database of protein sequences patterns derived from the analysis of the conserved residues that are predicted to be functional in structurally-aligned homologous families is the HOMSTRAD database [1] and PALI [2] is a database that consists of 1922 protein families containing over 13,500 protein domains. The SCOP (Structural Classification of Proteins) [3] database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known, including all entries in the Protein Data Bank (PDB) maintained by Research Collaboratory for Structural Bioinformatics (RCSB) consortium [4]. The structural classifications began to emerge, after release of SCOP with DALI [5] and CATH databases [6]. It is available as a set of tightly linked hypertext documents which make the large database comprehensible and accessible.

There has been considerable progress made over the past years in linking experimental and theoretical approaches to protein folding. Recent experiments have set benchmarks for testing new models and progress has been made in developing theoretical models for interpreting and predicting experimental results [7].

Evolutionarily close residues appear to have similar contact propensities and leads one to postulate that the extent of similarity between the contact propensities [3]. Proteins are classified to reflect both structural and evolutionary relatedness. Many levels exist in the hierarchy, but the principal levels are Family (Clear evolutionarily relationship), Super family (Probable common evolutionary origin) and Fold (Major structural similarity). Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Homologous proteins of residue identity from 100% to 20% have regions which retain the same general fold and regions where the folds differ. Further the extent of the structural changes is directly related to the extent of sequence changes [8].

The degree of success to be expected in predicting the structure of proteins from their sequence using the known structure of homologous proteins depends on the extent of

*Address correspondence to this author at the Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli 620 024, Tamil Nadu, Índia; E-mail: sel_emi@yahoo.co.uk

1874-1991/09

the residue identity. However the active sites of distantly related proteins can have very similar geometries. This is due to coupling of structural changes during evolution. Thus the structure of the active site in a protein may provide a good model for those in related protein even if all the residue identities are low [9]. For most of the protein families, the changes in structural similarity are linearly dependent on changes in sequence similarity. Families with lower structure/sequence correlations must have other sources of apparent structural variation that are not accounted by sequence change. Most of the evolutionary structural change in a protein family is linearly related to changes in sequence similarity, when plotted in terms of statistical significance or as RMSD versus percent identity [10].

Inter-residue interactions play an essential role in driving protein folding, and analysis of these interactions increases our understanding of protein folding and stability and facilitates the development of tools for protein structure prediction. In this work, we systematically characterized the common inter-residue interactions at a certain sequence separation cutoffs using homologous family of proteins. A marginal part of the failures of secondary structure predictions may be attributed to the influence of long-range interactions [11]. Thus, inter-residue interactions have been one of the main focuses to understand the mechanisms of protein folding and stability [12]. Contact exploration in proteins could be of great interest from different perspectives, e.g. to develop potentials [13], to identify amino acid side-chain clusters playing structural and functional roles [14]. For instance, different distributions of contacts in proteins reflect their different environments, the extra cellular and the intracellular ones [15]. Interestingly, inter-residue interactions can be characterized by contact order (CO) [16] and long-range order (LRO) parameters that have a strong correlation with the folding rate of small proteins [17].

The statistical potentials often correctly rank-order the relative strengths of inter residue interactions, but they do not reflect the true underlying energies because of systematic errors arising from the neglect of excluded volume in proteins [18]. The residue-residue contact preference potentials are demonstrated to successfully discriminate correct sequences in inverse folding experiments. A comparative analysis has been done for the structures of distantly related proteins that reveal with effect of residue-residue preference that have occurred during evolution.

Network type of representation is one of the most successful type of representation to understand the three dimensional structure of proteins. The native state protein structures can be modelled, using a graph-theoretical approach, as coarse-grained networks of amino acid residues as 'nodes' and the inter-residue interactions/contacts as 'links'. Using the network representation of protein structures and their 2D contact maps, it is easy to identify the conserved contact patterns [19]. The residue segments with hydrophobic clusters have high thermal stability [20]. Further, these clusters are formed and stabilized through long-range interactions. Specifically, a network of long-range contacts connects adjacent β-strands of the $(\alpha/\beta)^8$ barrel domain and the hydrophobic clusters.

In the present work, we have analyzed the inter-residue interactions in 37 families of homologous proteins of considerable sequence divergence and the variation of long-range order. The results reveal family specific long-range contact patterns in different bin intervals and structural class dependent variation of the long-range order. We also provide visual representation of long-range contact networks in proteins belonging to different structural classes.

MATERIALS AND METHODS

Dataset

The proteins of 37 homologous super families with varying percentage identity have been taken from the DALI [5] database forms the source for our present study. The proteins under each superfamily are selected by considering the residue identity and number of residues in the protein. The residue identity of each protein in a homologous family ranges from 100% to 25% and the number of residues for all the proteins in a family is nearly similar. Since there are lesser number of superfamilies with varying percentage identity in all-α, all-β and α+ β structural class, we considered minimum number of homologous superfamilies in these classes compared to the α/β structural class. The coordinates of the proteins were obtained from the PDB [4]. We obtained the information about the structural classes of all proteins from SCOP [3] database.

Computation of Distribution of Long-Range Contacts in Various Intervals

Each residue in the protein molecule is represented by its Cα atom. The centre is fixed at the Cα atom of the first (N-terminal) residue and the distance between this atom and the rest of the Cα atoms in the protein molecule is computed. The composition of surrounding residues associated with this residues were calculated with a sphere of radius 8Å, which has been shown to be the required volume of the medium within which a residue in a protein molecule is known to exert a detectable influence [21-23]. The procedure was repeated each time by moving the centre to the successive carbon atom along the polypeptide chain to compute the composition of surrounding residues, for all residues in a given protein. From the computation of surrounding residues within the sphere of 8Å radius, the contribution due to short range (Cα atom ±2 residues along the sequence), medium range, (Cα atom ±3 or ±4 residues along the sequence) and long-range (Cα atom >±4 residues) interactions were computed [24].

The long-range contacts (Cα atom >±4 residues) are further classified into intervals with a step of 10 (4-10; 11-20; 21-30; 31-40; 41-50; 51-60; >60). The number of long-range contacts in each interval for all homologous family of proteins belonging to different structural classes is computed. Also, the percentage of long-range contacts for all the proteins in each interval was calculated. The cross correlation of long-range contact distribution for 37 homologous super families was measured in the form of a 37X37 matrix. Further cluster analysis is performed for the cross correlated average long-range distribution matrix and the distance tree is drawn.

Computation of Long-Range Order

A novel parameter long-range order (LRO) based on the number of long-range contacts in the three dimensional structures of proteins was developed by Gromiha and Selvaraj, [17] to predict the folding rates of two-state proteins.

Long-range order is defined as,

$$LRO = \sum n_{ij} / N \quad (1)$$

$$n_{ij} = 1, \text{ if } |i-j| > 12$$

$$= 0, \text{ Otherwise}$$

Where i and j are two residues for which the C_α- C_α distance is ≤ 8 Å and N is the total number of residues in a protein.

Computation of Pairwise Residue-Residue Preferences to form Long-Range Contacts

The residues that form long-range contacts as detailed above were analyzed for pair-wise residue preferences for proteins belonging to all the four structural classes. The values were normalized using the frequencies of the 20 amino acid residues occurring under each structural class.

Plotting of Inter-Residue Interaction as Network Graphs

The contacts for each residue in a protein at residue separation >±4 are computed with a sphere of radius 8Å. A program is coded in Graph Modeling Language [25] to plot each amino acid residue as a node and their contacts as edges. Cytoscape system biology software [26] is used to view the GML file and to perform various operations such as, to color the nodes according to their physical properties, to change the node size, label size and to zoom in and out the network etc., The maximum numbers of contacts rising from a group of nodes in a circle may be crucial to understand domain structure in proteins. Also we represent the homologous family of proteins as a single circular graph, which shows us the crucial patterns, which forms structural homologues in a family of four different structural classes of proteins.

RESULTS AND DISCUSSION

Distribution of Long-Range Contacts at Various Intervals in Homologous Family of Proteins

The list of 37 homologous super families and the distribution of long-range contacts at various intervals in all-α,

Table 1. Percentage of Long-Range Contacts at Different Residue Intervals in all-α Proteins

S. No	Family	4 - 10	11 - 20	21 - 30	31 - 40	41 - 50	51 - 60	> 60
1	Globin	21.22	4.72	6.16	17.68	8.06	16.31	25.87
2	Annexin	25.60	11.12	2.10	31.76	10.82	0.00	18.59
3	Phospholipase	28.10	17.41	6.35	2.54	3.44	5.83	36.32
4	DPS Protein	13.51	14.91	9.45	8.05	7.98	20.64	25.47
5	Cytothesin	26.95	10.83	8.14	25.86	13.85	2.27	12.09
	Average	23.08	11.80	6.44	17.18	8.83	9.01	23.67
	Standard Deviation	5.95	4.82	2.78	12.10	3.86	9.02	9.06

Table 2. Percentage of Long-Range Contacts at Different Residue Intervals in all- β Proteins

S. No	Family	4 - 10	11 - 20	21 - 30	31 - 40	41 - 50	51 - 60	> 60
1	Lectin	15.16	17.05	8.53	6.61	1.38	2.28	49.00
2	Ferritin	14.79	27.53	5.17	3.96	10.33	14.01	24.20
3	Trypsin	15.02	17.77	12.43	8.48	8.07	6.40	31.83
4	Cupredoxin	11.59	18.26	35.65	14.84	3.95	2.11	13.61
5	Amylase	17.45	14.92	16.40	15.52	8.28	9.11	18.31
6	HLA Histo Compatibility Antigen	20.38	26.37	11.95	6.95	9.33	6.85	18.17
7	Leukotriene	13.82	14.72	30.02	6.86	4.20	2.89	27.49
8	Tumour necrosis factor	11.25	12.57	10.92	5.87	10.48	8.89	40.01
9	Herpes simplex virus II Protease	14.58	17.57	8.13	2.99	1.73	6.96	48.04
	Average	14.89	18.53	15.46	8.01	6.42	6.61	30.07
	Standard Deviation	2.79	5.11	10.43	4.38	3.62	3.86	13.11

all-β, α+β and α/β classes of proteins is shown in Table **1**, Table **2**, Table **3** and Table **4**, respectively. It reveals the opposite trends between the folding of all-α and all-β proteins. The all-α class proteins have more long-range contacts in the 4-10 interval whereas all-β class of proteins have more number of long-range contacts in 11-20 interval. This is due to hydrogen bonds between helices and strands in these classes of proteins. In the case of α/β class of proteins, they have maximum number of long-range contacts in 21-30 interval whereas in α+β class of proteins the maximum number of long-range contacts is in 4-10 interval. The helical and strand segments are present as separate domains in α+ β class of proteins and the proteins in this class behave like either all-α or all-β type. In α/β class, the helices and strands occur alternatively which may lead them to behave like this. The similar trend was also observed by Gromiha and Selvaraj [12].

The average distribution of long-range contacts of four different structural classes between residue intervals such as 4-10, 11-20, 21-30, 31-40, 41-50, 51-60 and above 60 is shown in Fig. (**1**). The overall trend in the long-range contact distribution is similar beyond the 41-50 intervals in all structural classes of proteins. In the interval 11-20, the average long-range contact distribution for all-α and α/β

Table 3. Percentage of Long-Range Contacts at Different Residue Intervals in α+β Proteins

S. No.	Family	4-10	11-20	21-30	31 - 40	41 - 50	51 - 60	> 60
1	Stromelycin	25.98	10.83	5.91	27.71	8.70	1.40	19.47
2	Vigilin	28.21	18.41	9.56	12.82	17.48	12.59	0.93
3	Cysteine proteinases	14.88	11.88	13.42	17.19	9.74	0.71	32.14
4	ADP RIBO	13.00	13.79	8.10	10.09	7.87	5.27	41.93
5	Lysozyme Like	31.95	25.46	7.71	9.00	3.42	0.00	21.53
	Average	22.80	16.07	8.94	15.36	9.44	3.99	23.20
	Standard Deviation	8.39	6.00	2.82	7.59	5.10	5.22	15.36

Table 4. Percentage of Long-Range Contacts at Different Residue Intervals in α/β Proteins

S. No.	Family	4 - 10	11 - 20	21 - 30	31 - 40	41 - 50	51 - 60	> 60
1	Enolase	19.12	11.15	15.67	6.63	6.57	6.59	34.28
2	Keto Acyl Reductase	15.54	10.06	21.23	5.45	15.91	12.37	19.45
3	Enoyl CoA Reductase	15.93	11.29	14.73	19.34	12.55	12.58	13.58
4	Ribulose Phospho Epimerase	13.25	12.49	40.05	20.47	0.71	0.49	12.54
5	Lactate Dehydrogenase	17.26	13.15	23.92	12.57	3.31	1.64	28.14
6	ADP Ribosylation Factor	17.53	10.28	4.58	30.33	13.80	3.68	19.80
7	ARPG	15.35	9.96	21.16	4.98	17.74	11.77	19.05
8	Triose Phosphate Isomerase Glycosomal	13.77	13.70	22.28	24.25	12.94	0.62	12.45
9	Ovalbumin	14.63	13.03	7.87	4.14	5.48	5.51	49.34
10	Dephospho kinase	21.03	4.65	2.49	5.78	5.61	12.70	47.73
11	Hydrogenase small unit	23.95	10.45	7.44	24.24	11.91	5.49	16.51
12	Sepiapterin Reductase	16.10	10.85	16.42	6.59	12.67	12.00	25.37
13	Flavodoxin	14.78	8.09	14.85	34.73	12.69	10.53	4.32
14	WBPP	13.30	6.23	14.69	8.89	12.69	8.02	36.17
15	C Terminal Binding Protein	11.75	8.87	48.35	2.59	5.20	7.53	15.71
16	Nudix Homolog	22.43	15.29	11.75	9.70	7.29	5.10	28.45
17	Flouroscent Protein	16.16	21.36	18.26	4.33	3.48	2.86	33.55
18	Lumazine Synthetase	16.17	4.09	5.36	34.77	5.81	8.24	25.57
	Average	16.56	10.83	17.28	14.43	9.24	7.10	24.56
	Standard Deviation	3.25	3.99	11.74	11.12	4.92	4.23	12.21

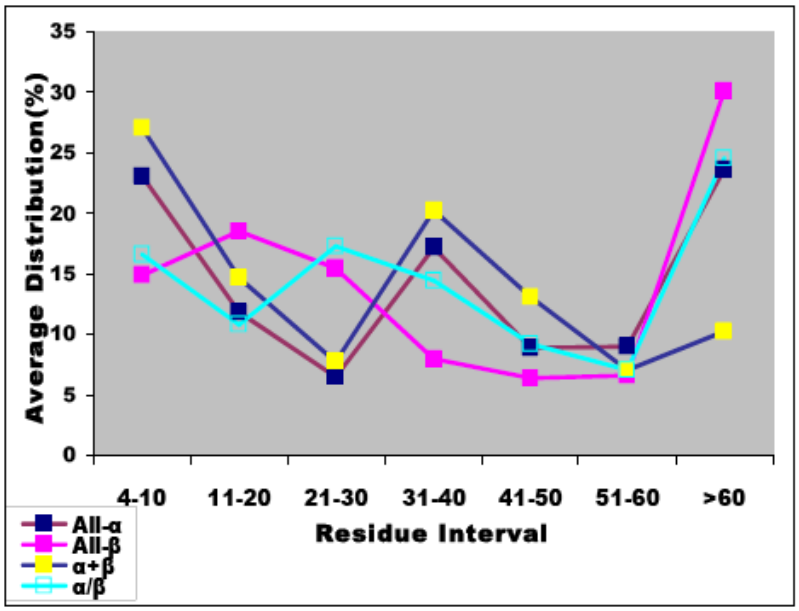

Fig. (1). Distribution of long-range contacts in different bin intervals.

proteins is consistent and it is 11.8 and 12.6 respectively. In α+β and α/β classes of proteins, the average long-range contact distribution trend is similar at 41-50 intervals, which is 7.9 and 7.6, respectively. The average long-range contact distribution for all-β and α/β proteins is similar and it is 6.1 and 6.2 respectively.

The average long-range contact distribution trend is opposite in all-α and all-β class of proteins. The average long-range contact distribution trend for all-α class of proteins goes down for two consecutive intervals and comes up at 31- 40 interval and again comes down and stays same for 51-60 interval and goes to maximum at above 60 intervals. The trend for all-β proteins is the average distribution goes down for two consecutive intervals and comes up at 31-40 interval and again goes down for couple of consecutive intervals and ends with maximum.

In α+β class of proteins, the long-range contact distribution is different only in 51-60 interval. The contact distribution is even in all other intervals. The trend between α+β and α/β are found similar except in the third interval (31-40). For α+β, the average long-range contact distribution at the third interval is 12.9 whereas in case of α/β, it is 20.56.

The cross correlation coefficient of average long-range contact distribution is shown in Table **5**. The comparison of long range contact distribution shows us that there is similar kind of distribution across different structural classes. The annexin family belonging to all-α structural class, stromelycin family belonging to α+β structural class and hydro-

Table 5. Cross Correlation Coefficient of Percentage of Long-Range Contacts at Different Residue Intervals for Homologous Family of Proteins

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
1	1.00	0.57	0.28	0.22	0.57	-0.05	0.27	-0.29	-0.45	0.16	-0.26	-0.42	-0.53	-0.09	0.60	0.41	0.05	0.00	0.17	0.11	-0.33	0.61	-0.23	-0.13	0.49	-0.37	-0.16	0.01	0.80	0.66	-0.11	0.45	0.03	-0.41	0.23	-0.39	0.67
2	0.57	1.00	0.28	-0.44	0.97	0.31	-0.18	0.14	-0.11	0.49	0.10	-0.22	-0.42	-0.01	0.98	0.47	0.74	0.58	0.48	0.11	-0.56	0.76	0.04	0.18	0.96	-0.54	0.51	0.14	0.30	0.98	-0.43	0.73	-0.03	-0.46	0.47	-0.07	0.84
3	0.28	0.28	1.00	0.31	0.38	0.83	0.58	0.77	0.03	0.55	0.81	0.23	0.60	0.85	0.38	0.86	0.25	0.75	0.93	0.77	0.12	-0.09	-0.04	0.37	-0.01	0.08	-0.12	0.96	0.69	0.36	0.46	-0.36	0.07	-0.04	0.92	0.65	-0.12
4	0.22	-0.44	0.31	1.00	-0.51	0.11	0.57	0.03	-0.38	-0.23	0.18	-0.25	0.20	0.46	-0.41	0.11	-0.75	-0.14	0.06	-0.02	0.10	-0.49	-0.45	-0.36	-0.54	-0.18	-0.85	0.26	0.47	-0.41	0.10	-0.55	-0.52	-0.19	-0.04	0.03	-0.34
5	0.57	0.97	0.38	-0.51	1.00	0.35	-0.22	0.21	-0.04	0.55	0.15	-0.09	-0.29	0.02	0.98	0.58	0.79	0.60	0.54	0.29	-0.34	0.74	0.11	0.30	0.88	-0.32	0.55	0.24	0.37	0.99	-0.20	0.64	0.21	-0.32	0.59	0.02	0.74
6	0.05	0.31	0.83	0.11	0.35	1.00	0.60	0.96	0.42	0.78	0.92	0.48	0.55	0.91	0.40	0.55	0.51	0.91	0.94	0.70	-0.06	-0.01	0.31	0.62	0.08	-0.11	0.25	0.90	0.22	0.32	0.20	-0.17	-0.14	0.12	0.88	0.88	-0.04
7	-0.27	-0.18	0.58	0.57	-0.22	0.60	1.00	0.64	-0.20	-0.02	0.81	-0.09	0.67	0.79	-0.18	0.46	-0.25	0.54	0.56	0.12	-0.18	-0.69	-0.42	-0.17	-0.32	-0.18	-0.48	0.68	0.17	-0.22	0.03	-0.67	-0.58	-0.31	0.36	0.49	-0.49
8	-0.29	0.14	0.77	0.03	0.21	0.96	0.64	1.00	0.49	0.69	0.96	0.58	0.72	0.90	0.23	0.51	0.46	0.87	0.88	0.70	0.12	-0.20	0.33	0.63	-0.07	0.09	0.25	0.90	0.07	0.14	0.32	-0.34	-0.04	0.24	0.83	0.95	-0.26
9	-0.45	-0.11	0.03	-0.38	-0.04	0.42	-0.20	0.49	1.00	0.70	0.25	0.95	0.23	0.26	0.00	-0.34	0.53	0.23	0.18	0.54	0.40	0.19	0.96	0.90	-0.15	0.33	0.70	0.20	-0.48	-0.10	0.31	0.10	0.35	0.85	0.29	0.70	-0.09
10	0.16	0.49	0.55	-0.23	0.55	0.78	-0.02	0.69	0.70	1.00	0.50	0.68	0.12	0.54	0.61	0.26	0.78	0.69	0.70	0.75	0.02	0.52	0.74	0.90	0.32	-0.06	0.65	0.57	0.13	0.53	0.18	0.32	0.20	0.40	0.78	0.71	0.35
11	-0.26	0.10	0.81	0.18	0.15	0.92	0.81	0.96	0.25	0.50	1.00	0.36	0.77	0.93	0.16	0.61	0.29	0.86	0.88	0.59	0.04	-0.36	0.06	0.40	-0.12	0.36	0.02	0.93	0.17	0.10	0.27	-0.47	-0.18	0.03	0.78	0.85	-0.34
12	-0.42	-0.22	0.23	-0.25	-0.09	0.48	-0.09	0.58	0.95	0.68	0.36	1.00	0.45	0.39	-0.09	-0.16	0.43	0.25	0.29	0.72	0.61	0.04	0.89	0.91	-0.33	0.54	0.53	0.39	-0.28	-0.17	0.58	-0.14	0.49	0.92	0.43	0.79	-0.28
13	-0.53	-0.42	0.60	0.20	-0.29	0.55	0.67	0.72	0.23	0.12	0.77	0.45	1.00	0.70	-0.36	0.43	-0.08	0.41	0.51	0.56	0.57	-0.73	0.01	0.26	-0.60	0.58	-0.22	0.75	0.07	-0.38	0.68	-0.85	0.17	0.33	0.49	0.74	-0.83
14	-0.09	-0.01	0.85	0.46	0.02	0.91	0.79	0.90	0.26	0.54	0.93	0.39	0.70	1.00	0.08	0.53	0.11	0.73	0.84	0.65	0.05	-0.33	0.10	0.41	-0.26	-0.02	-0.14	0.93	0.33	0.01	0.33	-0.50	-0.24	0.11	0.76	0.84	-0.33
15	0.60	0.98	0.38	-0.41	0.98	0.40	-0.18	0.23	0.00	0.61	0.16	-0.09	-0.36	0.08	1.00	0.51	0.78	0.62	0.56	0.27	-0.46	0.79	0.15	0.32	0.90	-0.46	0.54	0.24	0.36	0.99	-0.30	0.70	0.06	-0.34	0.58	0.05	0.82
16	0.41	0.47	0.86	0.11	0.58	0.55	0.46	0.51	-0.34	0.26	0.61	-0.16	0.43	0.53	0.51	1.00	0.24	0.66	0.80	0.51	0.00	-0.03	-0.36	0.04	0.24	0.02	-0.17	0.76	0.75	0.55	0.30	-0.23	0.12	-0.38	0.78	0.28	0.01
17	0.05	0.74	0.25	-0.75	0.79	0.51	-0.25	0.46	0.53	0.78	0.29	0.43	-0.08	0.11	0.78	0.24	1.00	0.67	0.52	0.43	-0.11	0.65	0.60	0.70	0.69	-0.11	0.91	0.27	-0.13	0.73	-0.09	0.59	0.31	0.16	0.59	0.41	0.53
18	0.00	0.58	0.75	-0.14	0.60	0.91	0.54	0.87	0.23	0.69	0.86	0.25	0.41	0.73	0.62	0.66	0.67	1.00	0.94	0.52	-0.33	0.54	-0.05	0.25	0.74	-0.32	0.31	0.50	0.58	0.94	-0.08	0.40	0.09	-0.41	0.75	0.16	0.58
19	0.60	0.90	0.64	-0.25	0.94	0.52	0.07	0.38	-0.15	0.55	0.38	-0.13	-0.07	0.29	0.93	0.79	0.52	0.94	1.00	0.41	-0.33	0.54	-0.05	0.25	0.74	-0.32	0.31	0.50	0.58	0.94	-0.08	0.40	0.09	-0.41	0.75	0.16	0.58
20	0.11	0.11	0.77	-0.02	0.29	0.70	0.12	0.70	0.54	0.75	0.59	0.72	0.56	0.65	0.27	0.51	0.43	0.52	0.72	1.00	0.56	0.10	0.52	0.79	-0.15	0.49	0.25	0.79	0.42	0.23	0.76	-0.23	0.55	0.56	0.86	0.77	-0.17
21	-0.33	-0.56	0.12	-0.10	-0.34	-0.06	-0.18	0.12	0.40	0.02	0.04	0.61	0.57	0.05	-0.46	0.00	-0.11	-0.25	-0.06	0.56	1.00	-0.32	0.34	0.37	-0.65	0.99	-0.02	0.20	0.01	-0.43	0.93	-0.53	0.81	0.77	0.14	0.33	-0.65
22	0.61	0.76	-0.09	-0.49	0.74	-0.01	-0.69	-0.20	0.19	0.52	-0.36	0.04	-0.73	-0.33	0.79	-0.03	0.65	0.11	0.05	0.10	-0.32	1.00	0.43	0.37	0.78	-0.36	0.65	-0.24	0.13	0.78	-0.31	0.93	0.23	-0.02	0.17	-0.21	0.92
23	-0.23	0.04	-0.04	-0.45	0.11	0.31	-0.42	0.33	0.96	0.74	0.06	0.89	0.01	0.10	0.15	-0.36	0.60	0.15	0.11	0.52	0.34	0.43	1.00	0.91	0.01	0.27	0.77	0.08	-0.39	0.06	0.26	0.32	0.43	0.82	0.27	0.54	0.13
24	-0.13	0.18	0.37	-0.36	0.30	0.62	-0.17	0.63	0.90	0.90	0.40	0.91	0.26	0.41	0.32	0.04	0.70	0.47	0.49	0.79	0.37	0.37	0.91	1.00	0.05	0.30	0.71	0.46	-0.10	0.24	0.42	0.17	0.46	0.73	0.63	0.77	0.08
25	0.49	0.96	-0.01	-0.54	0.88	0.08	-0.32	-0.07	-0.15	0.32	-0.12	-0.33	-0.60	-0.26	0.90	0.24	0.69	0.40	0.22	-0.15	-0.65	0.78	0.01	0.05	1.00	-0.61	0.56	-0.14	0.08	0.90	-0.61	0.85	-0.09	-0.51	0.20	-0.27	0.90
26	-0.37	-0.54	0.08	-0.18	-0.32	-0.11	-0.18	0.09	0.33	-0.06	0.02	0.54	0.58	-0.02	-0.46	0.02	-0.11	-0.25	-0.09	0.49	0.99	-0.36	0.27	0.30	-0.61	1.00	-0.02	0.16	-0.03	-0.42	0.90	-0.53	0.81	0.70	0.10	0.28	-0.67
27	-0.16	0.51	-0.12	-0.85	0.55	0.25	-0.48	0.25	0.70	0.65	0.02	0.53	-0.22	-0.14	0.54	-0.17	0.91	0.37	0.17	0.25	-0.02	0.65	0.77	0.71	0.56	-0.02	1.00	-0.05	-0.46	0.48	-0.14	0.65	0.35	0.37	0.26	0.29	0.48
28	0.01	0.14	0.96	0.26	0.24	0.90	0.68	0.90	0.20	0.57	0.93	0.39	0.75	0.93	0.24	0.76	0.27	0.80	0.93	0.79	0.20	-0.24	0.08	0.46	-0.14	0.16	-0.05	1.00	0.47	0.20	0.49	-0.47	0.04	0.10	0.91	0.82	-0.29
29	0.80	0.30	0.69	0.47	0.37	0.22	0.17	0.07	-0.48	0.13	0.17	-0.28	0.07	0.33	0.36	0.75	-0.13	0.18	0.46	0.42	0.01	0.13	-0.39	-0.10	0.08	-0.03	-0.46	0.47	1.00	0.42	0.33	-0.13	0.12	-0.31	0.51	-0.05	0.15
30	0.66	0.98	0.36	-0.41	0.99	0.32	-0.22	0.14	-0.10	0.53	0.10	-0.17	-0.38	0.01	0.99	0.55	0.73	0.56	0.51	0.23	-0.43	0.78	0.06	0.24	0.90	-0.42	0.48	0.20	0.42	1.00	-0.27	0.69	0.11	-0.39	0.54	-0.05	0.81
31	-0.11	-0.43	0.46	0.10	-0.20	0.20	0.03	0.32	0.31	0.18	0.27	0.58	0.68	0.33	-0.30	0.30	-0.09	-0.03	0.24	0.76	0.93	-0.31	0.26	0.42	-0.61	0.90	-0.14	0.49	0.33	-0.27	1.00	-0.60	0.73	0.66	0.42	0.47	-0.61
32	0.45	0.73	-0.36	-0.55	0.64	-0.17	-0.67	-0.34	0.10	0.32	-0.47	-0.14	-0.85	-0.50	0.70	-0.23	0.59	0.02	-0.14	-0.23	-0.53	0.93	0.32	0.17	0.85	-0.53	0.65	-0.47	-0.13	0.69	-0.60	1.00	0.01	-0.18	-0.09	-0.38	0.95
33	0.03	-0.03	0.07	-0.52	0.21	-0.14	-0.58	-0.04	0.35	0.20	-0.18	0.49	0.17	-0.24	0.06	0.12	0.31	-0.15	-0.02	0.55	0.81	0.23	0.43	0.46	-0.09	0.81	0.35	0.04	0.12	0.11	0.73	0.01	1.00	0.60	0.23	0.11	-0.12
34	-0.41	-0.46	-0.04	-0.19	-0.32	0.12	-0.31	0.24	0.85	0.40	0.03	0.92	0.33	0.11	-0.34	-0.38	0.16	-0.15	-0.07	0.56	0.77	-0.02	0.82	0.73	-0.51	0.70	0.37	0.10	-0.31	-0.39	0.66	-0.18	0.60	1.00	0.12	0.53	-0.38
35	0.23	0.47	0.92	-0.04	0.59	0.88	0.36	0.83	0.29	0.78	0.78	0.43	0.49	0.76	0.58	0.78	0.59	0.85	0.96	0.86	0.14	0.17	0.27	0.63	0.20	0.10	0.26	0.91	0.51	0.54	0.42	-0.09	0.23	0.12	1.00	0.74	0.06
36	-0.39	-0.07	0.65	0.03	0.02	0.88	0.49	0.95	0.70	0.71	0.85	0.79	0.74	0.84	0.05	0.28	0.41	0.69	0.73	0.77	0.33	-0.21	0.54	0.77	-0.27	0.28	0.29	0.82	-0.05	-0.05	0.47	-0.38	0.11	0.53	0.74	1.00	-0.37
37	0.67	0.84	-0.12	-0.34	0.74	-0.04	-0.49	-0.26	-0.09	0.35	-0.34	-0.28	-0.83	-0.33	0.82	0.01	0.53	0.15	0.04	-0.17	-0.65	0.92	0.13	0.08	0.90	-0.67	0.48	-0.29	0.15	0.81	-0.61	0.95	-0.12	-0.38	0.06	-0.37	1.00

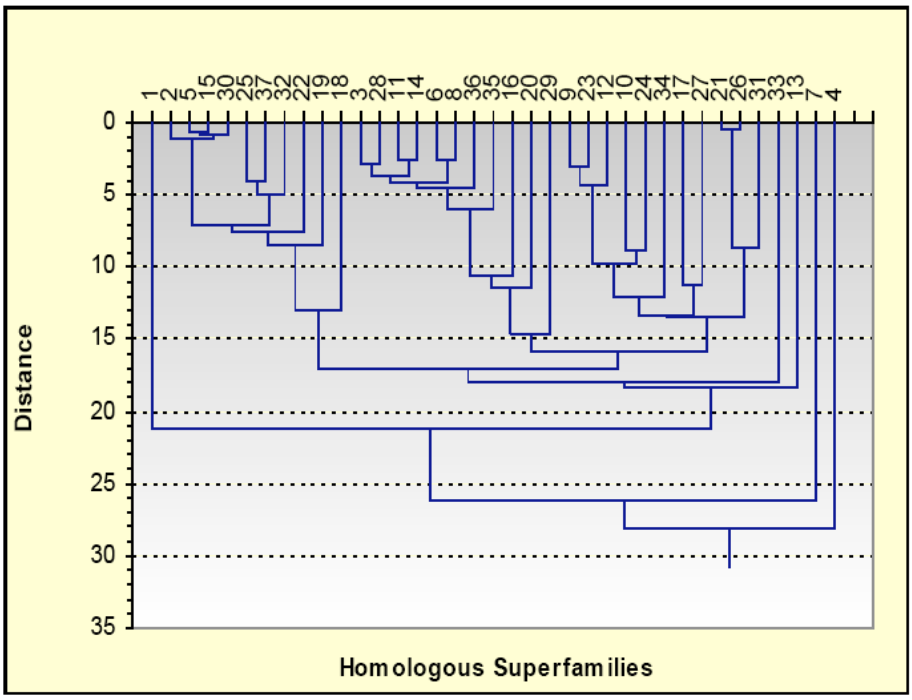

Fig. (2). Cross correlation distance tree of long-range contact distribution in different intervals.

genase small unit belonging to α/β structural class has similar kind of long range contact distribution. Also the cytothesin family of all-α class and Lectin family of all- β class has similar kind of long-range distribution with other structural classes.

The distance tree with the cluster of families with similar kind of long-range contact distribution is shown in Fig. (**2**). The figure shows the cluster of families 5, 15 and 30 has similar kind of long-range distribution. Interestingly, the family 5 belongs to all-α class, 15 belongs to α+β structural class and 30 belong to α/β class.

Sequence Identity and Long-Range Order in Different Structural Classes of Proteins

ALL-α CLASS

The PDB ID (PDB), Percentage Identity (%), Number of residues (N), Total number of long-range contacts (NLONG) and long-range order (LRO) for homologous family of proteins belonging to four different structural classes such as all α, all β, α+β and α/β classes respectively is shown in Table **6**, Table **7**, Table **8** and Table **9** respectively. Table **6** reveals that all alpha proteins in general possess lower values of LRO across different families. This reveals that the inter-

Table 6. Sequence Identity and Long-Range Order of all-α Protein Families

S. No.	PDB	%	N	NLONG	LRO
1. GLOBIN					
1	1HBR A	100	141	238	1.688
2	1NS9A	60	141	244	1.730
3	1HDAA	59	141	238	1.688
4	1FHJA	57	141	214	1.518
5	1A4FA	56	141	236	1.674
6	1JEBA	56	141	206	1.161
7	1QPWA	56	141	244	1.730
8	1HDSA	50	141	236	1.674
9	1LA6A	47	142	246	1.732
10	1CG8A	42	141	268	1.901

(Table 6) Contd…..

S. No.	PDB	%	N	NLONG	LRO
2. ANNEXIN					
1	1DK5A	100	316	594	1.880
2	1N00A	71	309	574	1.805
3	1AXN	32	312	618	1.913
4	1ANN	30	308	574	1.822
5	1DM5A	30	309	570	1.810
6	1ALA	32	310	542	1.715
7	1ANW	31	308	546	1.712
8	1W45A	30	307	592	1.850
3. PHOSPHOLIPASE					
1	1DPYA	100	117	248	2.120
2	1FE5A	91	117	254	2.153
3	1G2XA	84	117	268	2.271
4	1U4JA	81	117	268	2.271
5	1PO8A	72	117	252	2.136
6	1M8TA	56	116	248	2.084
7	2NOT	52	116	262	2.202
8	1MH2B	48	117	232	1.950
4. DPS PROTEIN					
1	1N1Q	100	149	264	1.772
2	1JIG	57	145	284	1.945
3	1JI5	55	142	238	1.676
4	1JI4	33	143	266	1.847
5	2BK6	41	143	236	1.650
6	1UMN	34	143	280	1.958
7	1O9R	28	148	292	1.973
8	1VEI	25	147	226	1.537
9	1DPS	27	145	266	1.834
5. CYTOTHESIN					
1	1BC9	100	200	322	1.610
2	1R8ME	77	189	288	1.477
3	1XT0B	36	182	326	1.630
4	1RE0B	34	186	428	2.642
5	1KU1A	33	187	346	1.640

residue contacts of all alpha proteins are formed mainly from local interactions. Even when the sequence identity gets down to minimum value of 25% in the DPS Protein homologous family its LRO value stands same, indicating the conservation of long-range interactions.

ALL-β CLASS

In all-β class of proteins 9 homologous super families namely Lectin, Ferritin, Trypsin, Cupredoxin, Amylase, HLA Histo compatibility antigen, Leukotriene, tumour necrosis factor and Herpes simplex trans type II have been considered. The LRO values with their sequence identity are given in Table **7**. The average LRO value for this structural class is 3.903. This is the highest average LRO value compared to other three structural classes. This indicates that the complexities of three dimensional structures for all-β proteins are higher than the other structural classes. When we compare the average LRO values for all-β class with α/β class of proteins, both of them possess similar range of values. Also, the conservation of long-range contacts is the

reason for possessing similar three dimensional structures in homologous super families even when their sequence identity goes very low.

α+β CLASS

In this class of proteins (Table **8**), there are five homologous super families namely Stromelycin, Vigilin, Cysteine

Table 7. Sequence Identity and Long-Range Order of All-β Protein Families

S. No.	PDB	%	N	NLONG	LRO
1. LECTIN					
1	1FAYA	100	236	1046	4.432
2	1WBFA	63	231	1082	4.565
3	1V00	58	233	1172	4.565
4	1G8WA	49	229	1082	4.644
5	1FNZA	48	229	1138	4.863
6	1N47	47	227	1056	4.532
7	1IOA	39	219	1090	4.781
2. FERRITIN					
1	1BVKA	72	106	452	4.185
2	1FGVL	69	106	434	4.056
3	1B0W	67	106	438	4.056
4	1WTLA	63	106	472	4.370
5	1C08A	57	106	450	4.206
3. TRYPSIN					
1	1FXYA	100	228	988	4.333
2	1TRNA	69	220	994	4.438
3	1H4WA	62	220	1038	4.494
4	1V6DA	55	219	978	4.386
5	1J15A	52	219	1014	4.547
6	1TGSZ	52	219	984	4.373
7	1A0JA	49	219	1002	4.493
8	1UTJA	45	218	996	4.486
4.CUPREDOXIN					
1	1PMY	100	123	472	3.837
2	1BQK	52	123	490	3.952
3	1ADWA	46	123	504	4.098
4	4PAZ	43	123	458	3.724
5. AMYLASE					
1	1HVXA	100	483	1848	3.826
2	1W9XA	67	480	1818	3.780
3	1WPCA	66	480	1894	3.938
4	1OB0A	65	478	2060	4.071
5	1E43A	65	480	1960	4.058
6	1UD2A	58	480	1886	3.929
6. HLA HISTO COMPATIBILITY ANTIGEN					
1	1AO7	100	209	744	2.715
2	1J8H	68	206	468	2.600
3	1OGA	73	204	774	2.804

(Table 7) Contd…..

S. No.	PDB	%	N	NLONG	LRO
6. HLA HISTO COMPATIBILITY ANTIGEN					
4	1SBB	69	204	984	4.134
5	1MI5	63	204	760	2.744
6	1NFD	53	205	728	3.586
7. LEUKOTRIENE					
1	1V3T	100	333	1212	3.640
2	1VJ1	38	315	1188	3.568
3	1GUF	20	315	1324	3.637
8. TUMOR NECROSIS FACTOR					
1	1TNF	100	152	600	3.947
2	2TNF	80	148	618	4.176
3	1TNR	35	138	598	4.153
4	1D4V	22	139	650	4.114
5	1IQA	21	139	675	4.405
9. HERPES SIMPLEX VIRUS TYPE II PROTEASE					
1	1AT3A	100	217	706	3.253
2	1VZV	53	204	612	2.900
3	1O6EB	30	208	708	3.147
4	1IEDA	28	200	690	3.224
5	2WPOA	28	204	754	3.278

Proteinases, ADP Ribo and Lysozyme like Proteins. In stromelycin family the parameter long-range order remains fixed, even the sequence identity decreases up to 50%. The average LRO value of this structural class of proteins is 3.478. These average LRO values of this structural class are high compared to all-α class and lower than all-β class of proteins. The similar kind of average standard deviation values is seen for both α/β and α+β structural class of proteins.

α/β CLASS

In α/β class of proteins, the number of long-range contacts and LRO values with their sequence identity are given in Table **9**. There are 18 homologous families in this

Table 8. Sequence Identity and Long-Range Order of α+β Protein Families

S. No.	PDB	%	N	NLONG	LRO
1. STROMELYCIN					
1	1Q3A	100	160	414	2.654
2	1HOV	56	160	388	2.425
3	1GKD	55	156	420	2.642
4	1RM8	50	157	450	2.663
5	1BUV	50	157	476	2.736
2. VIGILIN					
1	1VIH	100	71	150	2.113
2	1KHM	22	70	118	1.686
3	1J4W	21	65	136	2.092

(Table 8) Contd…..

S. No.	PDB	%	N	NLONG	LRO
3. CYSTEINE PROTEINASES					
1	2PAD	100	212	1050	4.953
2	1PPN	98	212	1076	5.075
3	1PPO	66	212	1154	5.44
4	1MEG	65	216	1086	5.027
4. ADP RIBONUCLEASE					
1	1PAX	100	361	1338	3.706
2	1WOK	86	350	1326	3.788
3	1GSO	45	349	1502	4.303
4	2PA9	36	344	1320	3.837
5. LYSOZYME LIKE					
1	3LYZ	100	129	472	3.658
2	1HEO	99	129	470	3.643
3	2IHL	95	129	468	3.627
4	1LMQ	61	127	456	3.590

structural class. The average LRO value for this structural class is low compared to all-β class and higher than all-α structural class. The similarity in LRO values of all-β and α+β may be due to the presence of hydrophobic residues in β strands, further the helices in the core region of α/β domain have one face toward the hydrophobic β-structure. The

Table 9. Sequence Identity and Long-Range Order of α/β Protein Families

S. No.	PDB	%	N	NLONG	LRO
1. ENOLASE					
1	1P48A	100	436	1608	3.688
2	1TE6	61	433	1660	3.825
3	1OEPA	59	414	1478	3.502
4	1E9ID	50	421	1644	3.823
5	1IYX A	49	414	1582	3.671
2. KETO ACYL REDUCTASE					
1	1W4ZB	100	261	912	3.494
2	1IY8A	35	246	922	3.574
3	1EDOA	37	242	838	3.434
4	1Q7CA	37	241	818	3.366
5	1GEG	33	248	930	3.647
6	1GEE	31	244	874	3.349
3. ENOYL COA REDUCTASE					
1	1DUB	100	260	840	3.231
2	1UIY	35	248	698	2.815
3	1HZD	33	256	832	3.128
4	1NZY	28	250	774	2.877

(Table 9) Contd…..

S. No.	PDB	%	N	NLONG	LRO
3. ENOYL COA REDUCTASE					
5	1EF9	28	215	720	2.759
6	1DCI	27	255	806	2.931
7	1XX4	21	213	742	2.921
4. RIBULOSE PHOSPAHTE 3 EPIMERASE					
1	1RPX	100	230	792	3.443
2	1TQJ	67	216	770	3.548
3	1H1Z	42	211	756	3.452
4	1TQX	35	210	768	3.475
5. LACTATE DEHYDROGENASE					
1	1Y6J	100	289	958	3.315
2	1LDN	44	289	1028	3.253
3	9LDT	40	287	972	2.937
4	1EZ4	38	289	988	3.218
5	1LLD	37	287	964	3.080
6	1HYG	36	286	1052	3.361
6. ADP RIBOSYLATION FACTOR					
1	1E0S	100	173	468	2.705
2	1MR3	65	172	596	3.153
3	1HUR	65	172	532	2.956
4	1MOZ	52	169	516	2.899
5	1FZQ	45	169	514	2.920
6	1KSG	40	152	520	2.889
7. ARPG					
1	1XG5	100	257	888	3.455
2	1CYD	31	236	802	3.314
3	1EDO	26	237	838	3.434
4	1IPF	24	243	884	3.413
5	1AE1	23	243	822	3.355
6	1XKQ	22	242	900	3.309
8. TRIOSE PHOSPHATE ISOMERASE GLYCOSOMAL					
1	1SUX	100	250	834	3.336
2	1TPF	73	250	770	3.080
3	1HTI	52	246	812	3.274
4	1TIM	49	246	790	3.198
5	1YPI	46	246	788	3.190
6	1HG3	22	206	828	3.696
9. OVALBUMIN					
1	1OVA	100	385	1344	3.491
2	1XQJ	31	367	1368	3.598

(Table 9) Contd…..

S. No.	PDB	%	N	NLONG	LRO
9. OVALBUMIN					
3	1HP7	29	362	1210	3.218
4	1SEK	24	360	1364	3.628
10. DEPHOSPHOKINASE					
1	1JJV	100	194	422	2.175
2	1VIY	48	190	418	2.010
3	1UF9	26	179	362	1.926
11. HYDROGENASE SMALL UNIT					
1	1E3D	100	262	826	3.153
2	1UBR	71	262	854	3.199
3	1FRV	65	260	828	3.160
4	1FRF	62	259	832	3.188
5	1CC1	42	256	848	3.084
12. SEPIAPTERIN REDUCTASE					
1	1OAA	100	259	830	3.205
2	1CYD	24	228	802	3.314
3	2AHI	21	225	716	3.598
4	1HXH	21	228	842	3.351
5	1BDB	21	220	918	3.438
13. FLAVODOXIN					
1	1OBV	100	169	638	3.775
2	1D04	69	168	602	3.562
3	1AG9	45	168	608	3.474
4	1B1C	21	146	546	3.289
14. WBPP					
1	1SB8	100	341	1256	3.683
2	1R6D	28	309	1132	3.663
3	1KEW	28	312	1064	3.410
4	1ORR	27	311	1104	3.550
5	1GY8	25	317	942	2.972
15. C TERMINAL BINDING PROTEIN					
1	1HKU	100	330	1160	3.505
2	1YGY	32	313	1130	3.610
3	2NAD	26	319	956	2.997
4	1PSD	26	311	1050	3.376
5	1J49	25	311	1066	3.428
16. NUDIX HOMOLOG					
1	1K2EA	100	152	410	2.697
2	1SU2A	23	119	394	3.311
3	1NQZ	22	109	302	2.771
4	1JKN	20	119	364	3.059

(Table 9) Contd…..

S. No.	PDB	%	N	NLONG	LRO
17. FLOUROSCENT PROTEIN					
1	1MOV	100	218	938	4.303
2	1GGXA	65	216	932	4.315
3	1G7K	63	217	864	4.075
4	1UIS	47	217	948	4.232
5	1XQM	46	217	946	4.167
6	1XA9	43	215	962	4.295
7	1KP5A	23	214	1112	4.595
18. LUMAZINE SYNTHETASE					
1	1C41A	100	165	546	3.309
2	1KYVE	56	151	486	3.284
3	1EJBA	50	159	546	3.25
4	1NQV	39	148	556	3.61
5	1RVVA	37	148	532	3.455
6	1C2YA	33	148	570	3.677

structure of α/β structural class is very complex due to the position of opposite side of the helix is usually at the protein surface. The amino acid sequence should contain a set of hydrophobic residues placed to reflect the periodicity of the helix. There are two types of β-sheets regularly seen in α/β proteins, planar and cylindrical. Planar β-sheets contain two types of strands. Those on the interior of the sheet are well shielded from solvent and show hydrophobic residues at the interior of the sequential positions in order to provide two hydrophobic faces [27].

The overall average of LRO values and their standard deviation for all the four structural classes are given in Table **10**. The average LRO values for all α, all β, α+β and α/β are

Table 10. Average and Standard Deviation of LRO for all Individual Families

All-α Proteins		
S. No.	**Family**	**Average LRO**
1	Globin	1.650
2	Annexin	1.813
3	Phospholipase	2.148
4	DPS Protein	1.616
5	Cytothesin	1.800
	Average	1.805
	Standard Deviation	0.211
All-β Proteins		
S. No.	**Family**	**Average LRO**
1	Lectin	4.626
2	Ferritin	4.041
3	Trypsin	4.444
4	Cupredoxin	3.903
5	Amylase	3.934
6	HLA Histo Compatibility Antigen	3.097

All-β Proteins		
S. No.	**Family**	**Average LRO**
7	Leukotriene	3.761
8	Tumour necrosis factor	4.159
9	Herpes simplex virus type II Protease	3.160
	Average	3.903
	Standard Deviation	0.515
α+β Proteins		
S. No.	**Family**	**Average LRO**
1	Stromelycin	2.624
2	Vigilin	2.107
3	Cysteine Proteinases	5.123
4	ADP Ribo	3.908
5	Lysozyme Like	3.629
	Average	3.478
	Standard Deviation	1.175
α/β Proteins		
S. No.	**Family**	**Average LRO**
1	Enolase	3.701
2	Keto Acyl Reductase	3.477
3	Enoyl CoA Reductase	2.952
4	Ribulose Phospho Epimerase	3.442
5	Lactate Dehydrogenase	3.194
6	ADP Ribosylation Factor	2.920
7	ARPG	3.380
8	Triose Phosphate Iso Glycosomal	3.291
9	Ovalbumin	3.398
10	Dephospho kinase	2.251
11	Hydrogenase small unit	3.157
12	Sepiapterin Reductase	3.381
13	Flavodoxin	3.525
14	WBPP	3.456
15	C Terminal Binding Protein	3.383
16	Nudix Homolog	2.821
17	Flouroscent Protein	4.283
18	Lumazine Synthetase	3.431
	Average	3.302
	Standard Deviation	0.414

1.805, 3.903, 3.478 and 2.286, respectively. The average long-range order for all β and α+β classes of proteins are similar while those of all-α and α+β differs, it is 1.8 for all-α and 2.28 for α/β.

Influence of Residue-Residue Distance of Conserved Residues in Homologous Family of Proteins

The homologous super family cupredoxin is considered and their conserved residue contacts and their residue-

Table 11. Conserved Residue-Residue Contact Distance for Cupredoxin Super Family

1ADW					1BQK					1PMY					4PAZ				
				Distance					Distance					Distance					Distance
16	MET	7	MET	7.62	16	MET	7	MET	7.62	16	MET	7	MET	7.67	16	MET	7	MET	7.71
16	MET	8	LEU	5.83	16	MET	8	LEU	5.89	16	MET	8	LEU	5.69	16	MET	8	LEU	5.83
16	MET	9	ASN	4.55	16	MET	9	ASN	4.57	16	MET	9	ASN	4.49	16	MET	9	ASN	4.41
16	MET	10	LYS	5.92	16	MET	10	LYS	5.55	16	MET	10	SER	5.81	16	MET	10	LYS	5.54
16	MET	11	GLY	7.55	16	MET	11	GLY	7.67	16	MET	11	GLY	7.49	16	MET	11	GLY	7.4
16	MET	81	HIS	7.06	16	MET	81	HIS	6.76	16	MET	81	HIS	6.75	16	MET	81	HIS	6.65
16	MET	84	MET	6.95	16	MET	84	MET	6.87	16	MET	84	MET	6.75	16	MET	84	MET	6.75
16	MET	85	GLY	6.97	16	MET	85	GLY	6.93	16	MET	85	GLY	7.04	16	MET	85	GLY	6.89
16	MET	86	MET	5.22	16	MET	86	MET	5.15	16	MET	86	MET	5.27	16	MET	86	MET	5.28

residue distances are shown in Table **11** and Fig. (**3**). The difference between the reference residue and the contacting residue is known as residue separation. The contacting residues and residue separation value of four proteins 1ADW, 1BQK, 1PMY and 4PAZ are similar. Table **10** shows that the distance between reference residue and contacting residue remains same for all the four proteins. The contacts shown in the Table **11** is computed for the residues coming within 8Å sphere and with residue separation greater than 4. The fully conserved residue methionine which is at 16th residue of the protein 1ADW has contacts with 7Met, 8Leu, 9Asn, 10Lys, 11Gly, 81His, 84Met, 85Gly and 86Met. The similar residue-residue contacts are present in the other three proteins 1BQK, 1PMY and 4PAZ of the cupredoxin superfamily. The distances between the conserved residues are also found to be similar. This clearly indicates that the similar kind of long-range contacts and the conserved residue-residue distance is the reason for forming structural homologues with high sequence divergence.

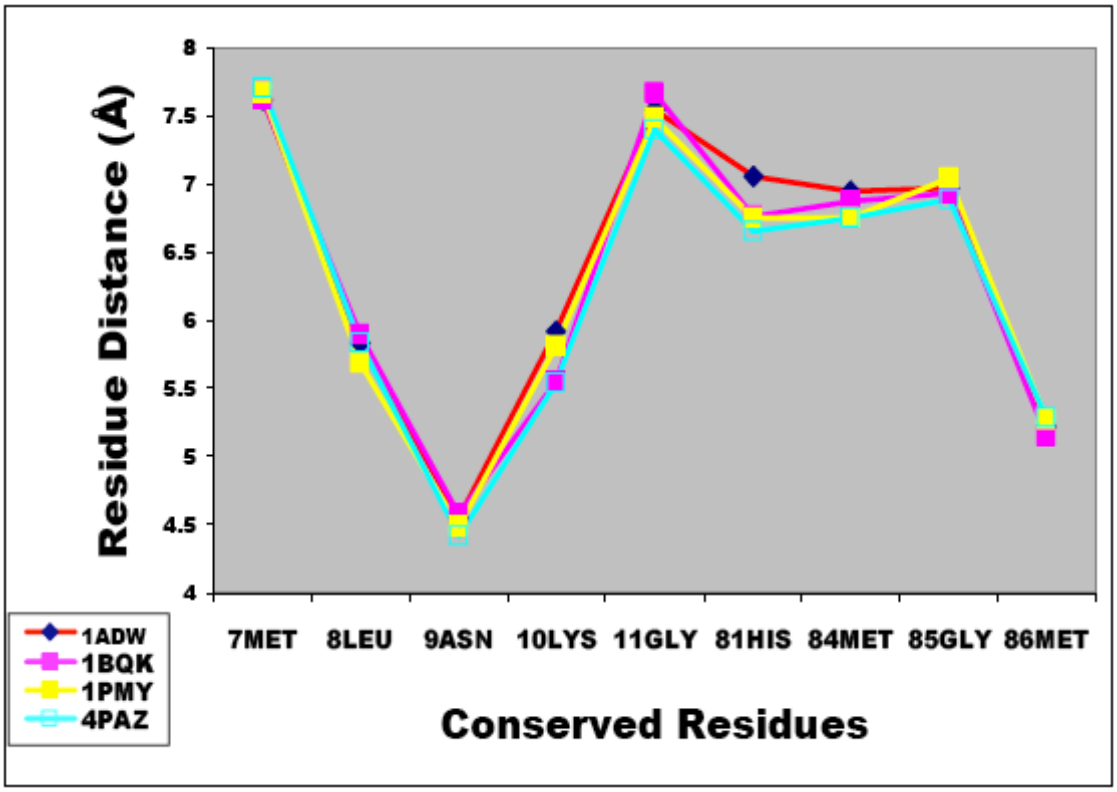

Fig. (3). Influence of residue-residue distance in conserved residues.

Influence of Residue-Residue Contact Preferences in Homologous Family of Proteins

The pairwise residue-residue preferences to form long-range contacts in all the four structural classes of homologous families of proteins are given in Table **12**. It is seen that hydrophobic residues dominate to form pair-wise long range contacts irrespective of the structural class. Specifically the residues Ala, Val and Leu contribute more

Table 12. Pairwise Residue-Residue Preferences in Homologous Family of Proteins in Four Structural Classes

ALL-α Proteins																				
	ALA	**ARG**	**ASN**	**ASP**	**CYS**	**GLN**	**GLU**	**GLY**	**HIS**	**ILE**	**LEU**	**LYS**	**MET**	**PHE**	**PRO**	**SER**	**THR**	**TRP**	**TYR**	**VAL**
ALA	90	52	31	55	22	34	82	36	27	52	122	49	25	33	17	47	48	9	28	70
ARG	52	10	14	29	9	15	31	18	8	27	51	24	8	10	9	13	32	4	17	26
ASN	31	14	34	20	10	17	13	17	9	25	48	26	8	21	15	16	17	4	9	29
ASP	55	29	20	24	14	14	50	22	14	24	45	34	10	18	10	23	23	1	22	26
CYS	22	9	10	14	24	15	10	13	5	7	21	5	3	7	6	16	25	0	19	8
GLN	34	15	17	14	15	10	20	16	4	21	38	23	5	14	3	17	18	5	8	14
GLU	82	31	13	50	10	20	46	31	32	29	84	41	7	29	15	29	32	3	23	40
GLY	36	18	17	22	13	16	31	32	8	17	34	33	5	17	13	20	17	3	19	22
HIS	27	8	9	14	5	4	32	8	38	9	30	21	5	19	8	11	20	5	17	24
ILE	52	27	25	24	7	21	29	17	9	22	72	32	11	19	7	20	18	5	13	25
LEU	122	51	48	45	21	38	84	34	30	72	124	58	18	48	23	49	38	7	30	100
LYS	49	24	26	34	5	23	41	33	21	32	58	24	14	20	16	26	21	6	23	23
MET	25	8	8	10	3	5	7	5	5	11	18	14	2	11	2	10	8	1	5	9
PHE	33	10	21	18	7	14	29	17	19	19	48	20	11	8	10	10	16	4	11	26
PRO	17	9	15	10	6	3	15	13	8	7	23	16	2	10	0	13	13	1	12	16
SER	47	13	16	23	16	17	29	20	11	20	49	26	10	10	13	16	24	2	18	33
THR	48	32	17	23	25	18	32	17	20	18	38	21	8	16	13	24	24	5	24	19
TRP	9	4	4	1	0	5	3	3	5	5	7	6	1	4	1	2	5	0	2	9
TYR	28	17	9	22	19	8	23	19	17	13	30	23	5	11	12	18	24	2	18	20
VAL	70	26	29	26	8	14	40	22	24	25	100	23	9	26	16	33	19	9	20	54

ALL-β Proteins																				
	ALA	**ARG**	**ASN**	**ASP**	**CYS**	**GLN**	**GLU**	**GLY**	**HIS**	**ILE**	**LEU**	**LYS**	**MET**	**PHE**	**PRO**	**SER**	**THR**	**TRP**	**TYR**	**VAL**
ALA	118	59	33	66	21	55	60	91	37	58	117	50	19	59	56	61	99	42	61	104
ARG	59	40	22	44	7	30	51	59	11	29	73	23	11	33	27	39	53	21	31	51
ASN	33	22	26	28	6	17	37	52	11	35	31	31	10	28	30	36	40	12	27	46
ASP	66	44	28	58	11	27	28	87	28	36	63	43	14	50	44	57	66	30	51	62
CYS	21	7	6	11	16	17	8	29	5	9	25	5	8	8	9	16	19	7	15	24
GLN	55	30	17	27	17	26	33	51	18	19	40	24	12	21	32	30	53	20	35	33
GLU	60	51	37	28	8	33	38	57	17	33	60	39	5	27	38	31	46	25	25	60
GLY	91	59	52	87	29	51	57	152	19	54	73	58	13	55	62	86	118	26	73	90
HIS	37	11	11	28	5	18	17	19	8	21	32	16	10	18	10	16	33	11	17	34
ILE	58	29	35	36	9	19	33	54	21	30	63	29	10	36	30	52	52	17	36	66
LEU	117	73	31	63	25	40	60	73	32	63	110	51	18	49	51	66	92	30	69	112
LYS	50	23	31	43	5	24	39	58	16	29	51	24	6	31	30	46	54	31	35	51
MET	19	11	10	14	8	12	5	13	10	10	18	6	6	14	6	6	25	8	15	18
PHE	59	33	28	50	8	21	27	55	18	36	49	31	14	52	32	51	50	18	29	74
PRO	56	27	30	44	9	32	38	62	10	30	51	30	6	32	30	35	47	17	29	61
SER	61	39	36	57	16	30	31	86	16	52	66	46	6	51	35	72	78	25	25	72
THR	99	53	40	66	19	53	46	118	33	52	92	54	25	50	47	78	88	29	59	87
TRP	42	21	12	30	7	20	25	26	11	17	30	31	8	18	17	25	29	12	34	34
TYR	61	31	27	51	15	35	25	73	17	36	69	35	15	29	29	25	59	34	52	53
VAL	104	51	47	62	24	33	60	90	34	67	112	51	18	75	62	72	88	35	53	126

(Table 12) Contd…..

α+β Proteins																				
	ALA	**ARG**	**ASN**	**ASP**	**CYS**	**GLN**	**GLU**	**GLY**	**HIS**	**ILE**	**LEU**	**LYS**	**MET**	**PHE**	**PRO**	**SER**	**THR**	**TRP**	**TYR**	**VAL**
ALA	48	16	31	33	15	31	35	36	27	46	68	38	13	27	23	47	26	18	30	61
ARG	16	14	21	17	15	15	21	37	6	35	30	20	4	11	9	26	20	12	27	19
ASN	31	21	24	28	17	19	16	51	8	39	40	27	5	9	11	39	27	15	37	24
ASP	33	17	28	28	7	19	21	40	8	31	30	34	12	13	25	32	22	4	18	31
CYS	15	15	17	7	16	5	4	23	2	8	14	7	0	3	3	14	4	10	6	11
GLN	31	15	19	19	5	18	14	32	2	20	38	22	2	7	13	29	9	4	24	36
GLU	35	21	16	21	4	14	18	28	6	27	45	25	6	11	12	26	14	6	24	34
GLY	36	37	51	40	23	32	28	72	18	48	62	38	12	14	33	50	39	20	50	43
HIS	27	6	8	8	2	2	6	18	8	13	22	10	6	12	6	17	10	2	14	11
ILE	46	35	39	31	8	20	27	48	13	74	71	47	10	24	17	39	23	4	45	49
LEU	68	30	40	30	14	38	45	62	22	71	102	53	16	24	25	54	31	15	50	56
LYS	38	20	27	34	7	22	25	38	10	47	53	28	15	18	19	31	19	7	26	43
MET	13	4	5	12	0	2	6	12	6	10	16	15	2	6	5	7	4	4	6	8
PHE	27	11	9	13	3	7	11	14	12	24	24	18	6	6	6	19	9	4	15	18
PRO	23	9	11	25	3	13	12	33	6	17	25	19	5	6	22	20	20	5	26	17
SER	47	26	39	32	14	29	26	50	17	39	54	31	7	19	20	28	28	9	42	41
THR	26	20	27	22	4	9	14	39	10	23	31	19	4	9	20	28	18	6	20	24
TRP	18	12	15	4	10	4	6	20	2	4	15	7	4	4	5	9	6	6	6	13
TYR	30	27	37	18	6	24	24	50	14	45	50	26	6	15	26	42	20	6	26	32
VAL	61	19	24	31	11	36	34	43	11	49	56	43	8	18	17	41	24	13	32	54

α/β Proteins																				
	ALA	**ARG**	**ASN**	**ASP**	**CYS**	**GLN**	**GLU**	**GLY**	**HIS**	**ILE**	**LEU**	**LYS**	**MET**	**PHE**	**PRO**	**SER**	**THR**	**TRP**	**TYR**	**VAL**
ALA	596	201	203	189	91	167	281	455	68	341	481	245	82	166	153	247	222	45	120	498
ARG	201	86	62	112	27	71	163	163	40	127	205	74	46	72	61	122	99	24	58	155
ASN	203	62	92	93	43	66	107	153	42	150	160	112	39	83	66	89	113	23	53	192
ASP	189	112	93	106	41	76	109	175	56	148	197	134	35	93	75	137	104	32	55	192
CYS	91	27	43	41	30	25	35	87	15	76	86	29	16	42	36	38	48	8	30	82
GLN	167	71	66	76	25	50	98	128	20	88	165	90	42	59	70	103	90	11	36	122
GLU	281	163	107	109	35	98	170	191	50	155	248	147	56	84	78	149	142	29	65	230
GLY	455	163	153	175	87	128	191	348	79	242	355	197	61	154	161	235	227	41	125	344
HIS	68	40	42	56	15	20	50	79	24	62	85	44	20	31	24	51	52	12	35	94
ILE	341	127	150	148	76	88	155	242	62	244	322	136	77	109	111	159	184	32	95	366
LEU	481	205	160	197	86	165	248	355	85	322	494	222	86	145	144	235	236	60	134	485
LYS	245	74	112	134	29	90	147	197	44	136	222	106	46	84	82	156	118	32	80	204
MET	82	46	39	35	16	42	56	61	20	77	86	46	22	40	43	52	47	12	35	80
PHE	166	72	83	93	42	59	84	154	31	109	145	84	40	66	58	103	79	21	58	148
PRO	153	61	66	75	36	70	78	161	24	111	144	82	43	58	58	104	94	20	60	164
SER	247	122	89	137	38	103	149	235	51	159	235	156	52	103	104	158	147	38	75	231
THR	222	99	113	104	48	90	142	227	52	184	236	118	47	79	94	147	112	25	88	225
TRP	45	24	23	32	8	11	29	41	12	32	60	32	12	21	20	38	25	4	19	38
TYR	120	58	53	55	30	36	65	125	35	95	134	80	35	58	60	75	88	19	42	126
VAL	498	155	192	192	82	122	230	344	94	366	485	204	80	148	164	231	225	38	126	508

towards the formation of long-range contacts and result in the formation of the hydrophobic core. The contribution of Ile is less compared to other hydrophobic residues. The all-α and α+β class of proteins prefer mostly Leu-Leu contacts as the top most residue pair. The all-β class of proteins prefer Gly-Gly contacts and α+β class of proteins prefer Ala-Ala

contacts as the top most residue pair. Overall the residue preference is similar to our earlier work [12].

Long-Range Contact Networks

Using the network representation of protein structures it is possible to identify the contact patterns i.e. groups of contacts which are present as a common feature for the homologous family of proteins. Fig. (**4a**) shows the contact network of Globin family where yellow nodes represents the hydrophobic residues. There are three groups of nodes where maximum numbers of edges are raising. These areas in particular contain hydrophobic residues as major content. The enhanced view of globin contact network is shown in

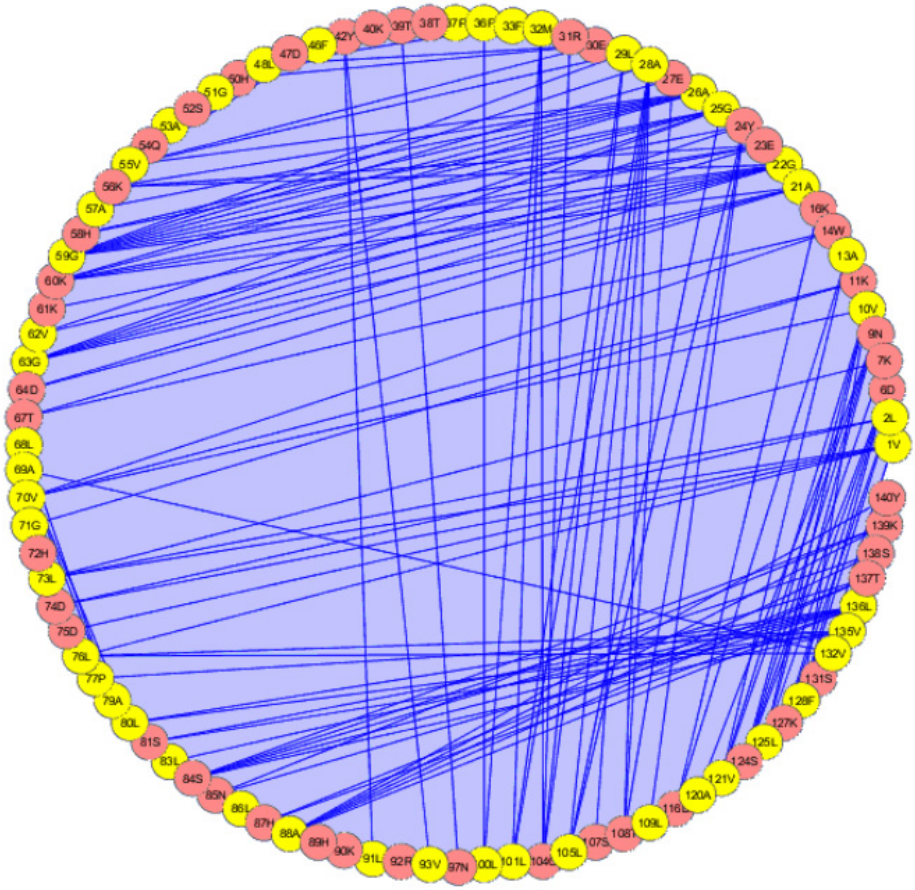

Fig. (4a). All-α structural class - globin family network diagram.

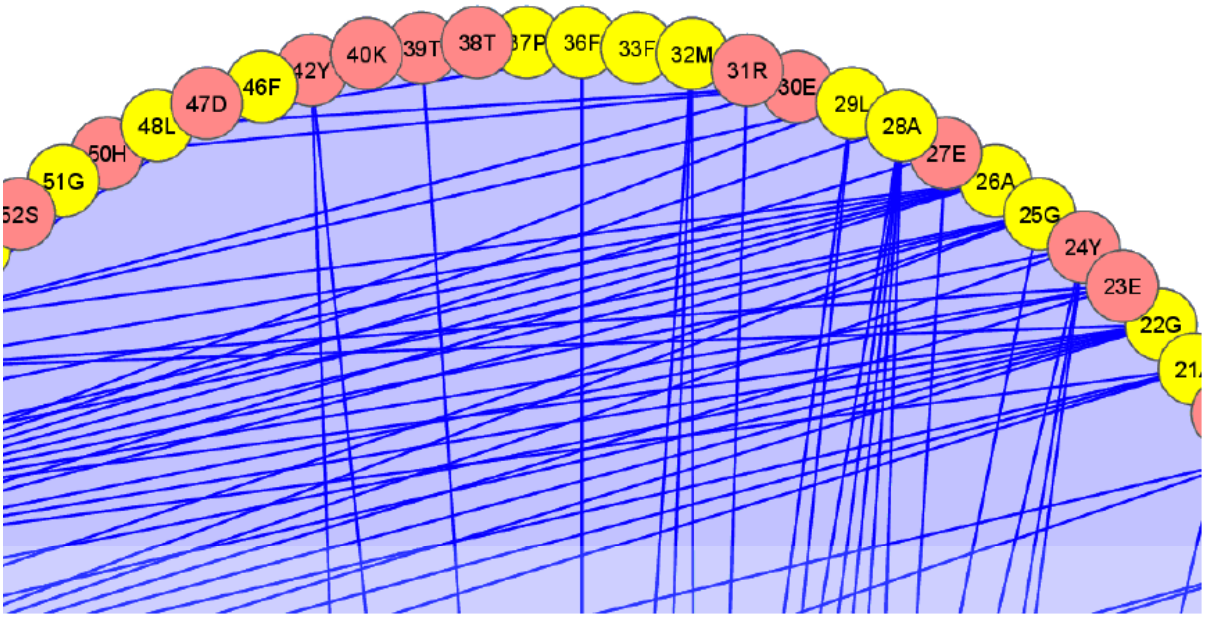

Fig. (4b). Enhanced view of Globin family's long-range contact networks.

Fig. (**4b**). Since all alpha proteins tertiary structure is less complex, the network looks very clear. In all-β structural class of proteins, the contact network of Lectin family is plotted. Fig. (**5a**) represents the network diagram of Lectin family, which contains five groups of nodes where the maximum numbers of edges are present. So these groups of nodes are important for determining the folded three dimensional structure. Since all-β proteins have highest average LRO value the network seems to be complex when compared to other three structural classes. Fig. (**5b**) shows the enhanced view of the Lectin family contact network.

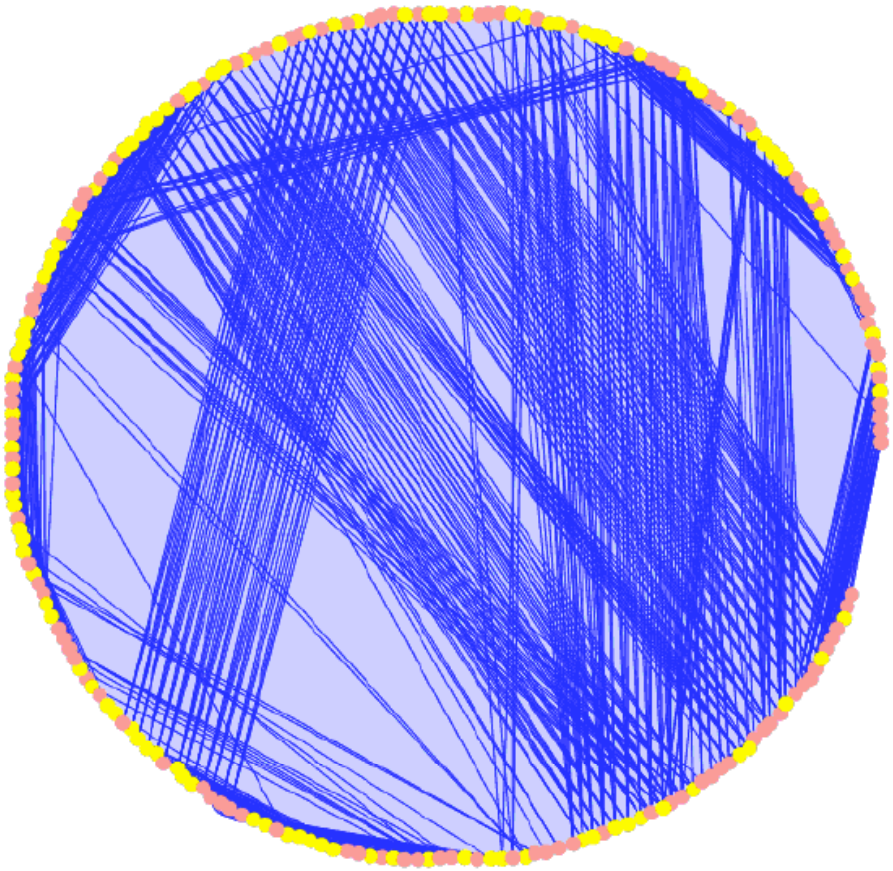

Fig. (5a). All-β structural class - lectin family network diagram.

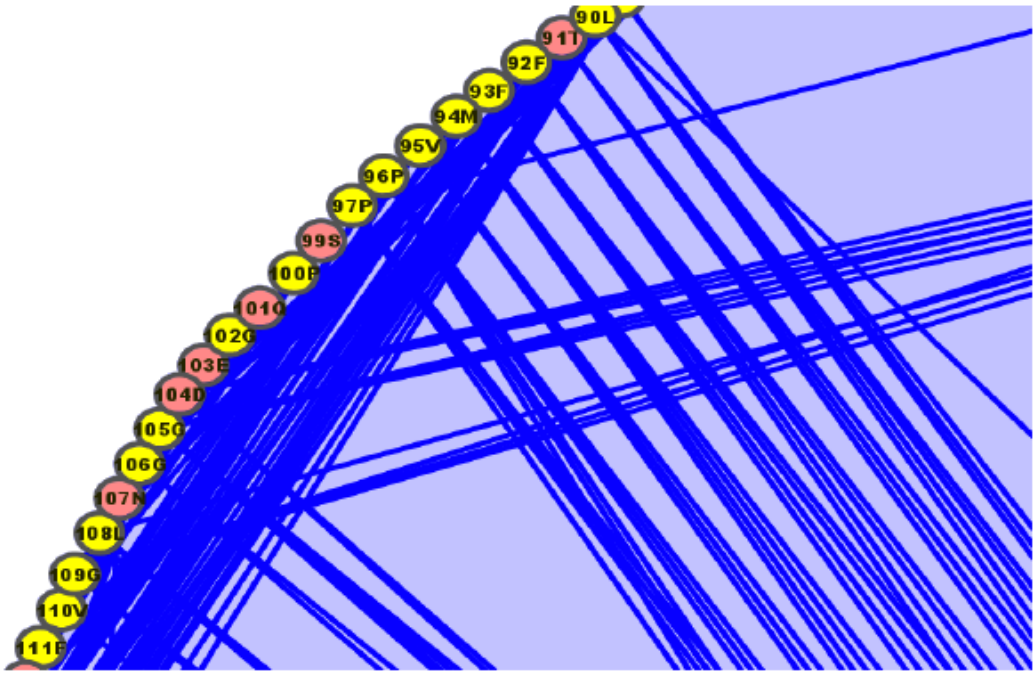

Fig. (5b). Enhanced view of Lectin family's long-range contact networks.

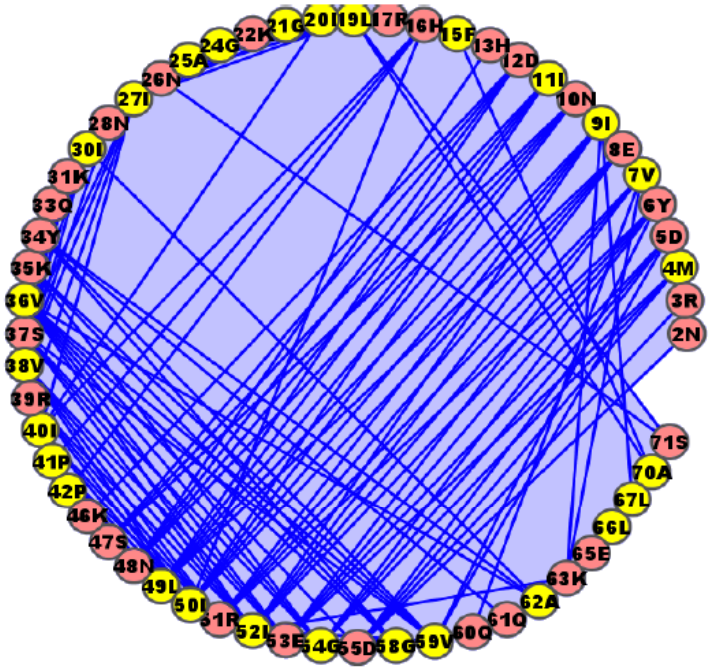

Fig. (6). α+β Structural class - vigilin family network diagram.

Fig. (**6**) represents the contact network of α+β structural class of proteins, the family considered here is Vigilin. In this family of proteins the network looks very clear and the most number of edges rises from one particular group of nodes. But in case of α/β class of proteins, the fluorescent family of proteins has the maximum number of groups of nodes and each group contains long stretch of nodes (i.e. residues). This indicates the complex topology of this particular structural class of proteins. The network contact diagram is shown in Fig. (**7a**) and the enhanced view of the network is shown in Fig. (**7b**). The difference between the contact topology between four structural classes is understood by considering the long-range network graphs.

CONCLUSION

The comparison of the three-dimensional structures of proteins is a complex algorithmic problem. In the present

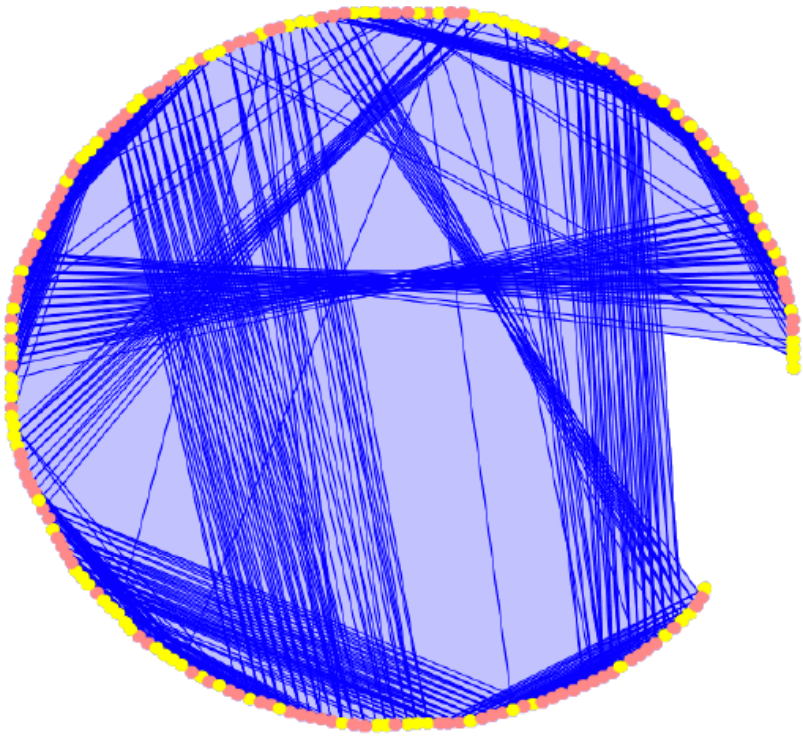

Fig. (7a). α/β Structural class – fluorescent family network diagram.

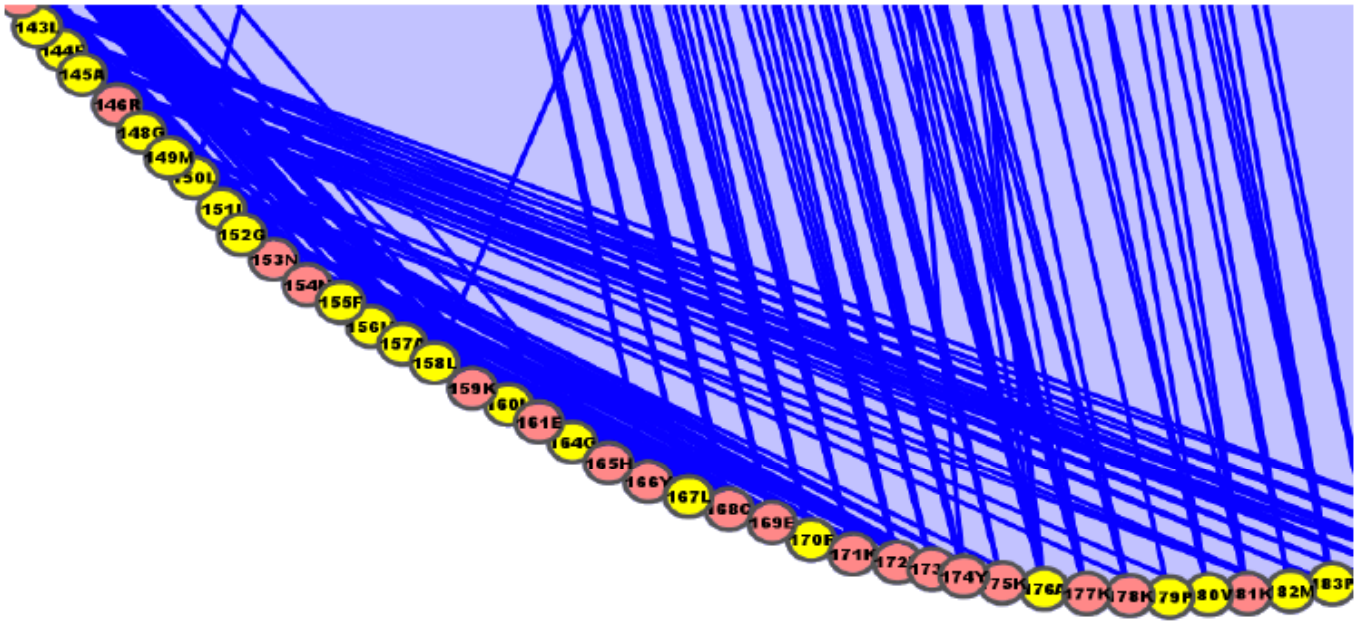

Fig. (7b). Enhanced view of fluorescent family long-range contact networks.

work we have made an analysis based on inter-residue interaction to study the homologous family of proteins. A comparative analysis has been done for the structures of related proteins that reveal the variation of long-range contacts that has occurred during evolution. The long-range contact distribution in homologous family of proteins of different structural classes is the main reason to possess structure homologues even when the sequence divergence is very high. The parameter LRO for distantly related proteins shows the importance of long-range contacts to have similar geometries in homologous family of protein. The present analysis also indicates the variation of long-range order across different structural classes and to provide a qualitative measure of the topological complexity of different protein structures. It also reveals that the conserved residue contacts and their residue-residue distance seem to be similar for all proteins in a homologous super family of proteins. Pair-wise residue preference to form long-range contacts indicates the dominance of hydrophobic residues. The network type representation of inter-residue interactions provides a visual representation of the complex three-dimensional structures of proteins and the nodes with maximum number of edges may be very crucial for the folding, stability and function.

ACKNOWLEDGEMENT

We gratefully acknowledge DST, New Delhi, India for financial assistance through Research Project Grant. SR/SO/ BB-09/2006.

REFERENCES

[1] Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998; 7: 2469-71.

[2] Sujatha S, Balaji S, Srinivasan N. PALI-a database of alignments and phylogeny of homologous protein structures. Bioinformatics 2001; 17: 375-6.

[3] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995; 247: 536-40.

[4] Westbrook J, Feng Z, Chen L, Yang H, Berman HM. The Protein Data Bank and Structural genomics. Nucleic Acids Res 2003; 31: 1489-99.

[5] Holm L, Kaariainen S, Rosenstrom P, Schenkel A. Searching protein structure databases with DaliLite v.3. Bioinformatics 2008; 24: 2780-81.

[6] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH - A hierarchical classification of protein domain structures. Structure 1997; 5: 93-108.

[7] Alm E, Baker D. Matching theory and experiment in protein folding. Curr Opin Struct Biol 1999; 9: 189-96.

[8] Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986; 5: 823-6.

[9] Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. A database of protein structure families with common folding motifs. Protein Sci 1993; 1: 1691-98.

[10] Wood TC, Pearson WR. Evolution of protein sequences and structures. J Mol Biol 1999; 291: 977-95.

[11] Faure1 G, Bornot1 A, Alexandre G. Protein contacts, inter-residue interactions and side-chain, Modeling. Biochimie 2008; 90: 626-39.

[12] Gromiha MM, Selvaraj S. Importance of long-range interactions in protein folding. Biophys Chem 1999; 77: 49-68.

[13] Zhang L, Skolnick J. How do potentials derived from structural databases relate to "true" potentials? Protein Sci 1998; 7: 112-22.

[14] Heringa J, Argos P. Side-chain clusters in protein structures and their role in protein folding. J Mol Biol 1991; 220: 151-71.

[15] Gugolya Z, Dosztanyi Z, Simon I. Inter-residue interactions in protein classes. Proteins 1997; 27: 360-66.

[16] Paci E, Lindorff LK, Dobson CM, Karplus M, Vendruscolo M. Transition state contact orders correlate with protein folding rates. J Mol Biol 2005; 352: 495-500.

[17] Gromiha MM, Selvaraj S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins. J Mol Biol 2001; 310: 27-32.

[18] Thomas PD, Dill KA. Statistical potentials extracted from protein structures: How accurate are they? J Mol Biol 1996; 257: 457-69.

[19] Barah P, Sinha S. Analysis of protein folds using protein contact networks. J Phys 2008; 71: 369-78.

[20] Selvaraj S, Gromiha MM. Role of hydrophobic clusters and long-range contact networks in the folding of $(\alpha/\beta)^8$ barrel proteins. Biophys J 2003; 84: 1919-25.

[21] Manavalan P, Ponnusamy PK. A study of the preferred environment of amino acid residues in globular proteins. Arch Biochem Biophys J 1977; 184: 476-87.

[22] Manavalan P, Ponnusamy PK. Hydrophobic character of amino acid residues in globular proteins. Nature 1978; 275: 673-4.
[23] Ponnusamy PK. Hydrophobic characteristics of folded proteins. Prog Biophys Mol Biol 1993; 59: 57-103.
[24] Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol 2004; 86: 235-77.
[25] Herman I, Melancon G and Marshall MS. IEEE Trans Vis Comp Graph 2000; 6: 24-43.
[26] Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: A software environment for integrated models of bimolecular interaction networks. Genome Res 2003; 13: 2498-504.
[27] Cohen FE, Abarbanel RM, Kuntz ID, Fletterick RJ. Secondary structural assignment for alpha/beta proteins by a combinatorial approach. Biochemistry 1983; 22: 4894-904.

Received: December 26, 2008 Revised: February 24, 2009 Accepted: March 02, 2009