# OPTIMAL PATHS IN RESIDUE NETWORKS IDENTIFY COMMUNICATION PATHWAYS IN PROTEINS

by

Murat Mülayim

Submitted to Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
July 2009

.

# OPTIMAL PATHS IN RESIDUE NETWORKS IDENTIFY COMMUNICATION PATHWAYS IN PROTEINS

Murat Mülayim

MAT,MSc Thesis,2009

Thesis supervisor: Prof. Dr. Canan Atılgan

Keywords: Residue Networks, communication on Residue networks, random walks, biased random walks

## Abstract

Navigation of information flows in networks is studied. As real-life systems, residue networks constructed from the coordinates deposited in the protein data bank are targeted. The cost of the navigation between neighbors are measured by residue-residue interaction potentials. By constructing all paths between initial/target nodes according to selected criteria, structurally and/or functionally important residues in the network are implicated. In particular, strong paths that minimize the weights along all possible pathways are found to differentiate between the functional nodes in protein families with high overall structural similarity, but low sequence similarity scores. To determine factors that drive the usage of strong paths in the network, a biased random walk scheme is deviced where the probability of edge selection is based on a balance between the knowledge of the location of the destination and the energy of interaction with the immediate neighbors. Since long range communication between two distantly placed functional regions in the protein calls for the gradient of information flow, strong paths emerge by satisfying the competition of local and global knowledge while navigating along the structure.

PROTENLERDEKİ İLETİŞİM YOLLARINI BELİRLEYEN RESIDÜ AĞ YAPILARINDAKİ OPTİMAL YOLLAR

Murat Mülayim

MAT,Master Tezi,2009

Tez Danışmanı: Prof. Dr. Canan Atılgan

Anahtar Kelimeler: Residü ağ yapıları, Residü ağ yapılarındaki iletişim, rastgele yürüyüşler, eğimli rastgele yürüyüşler

## Özet

Ağ yapılarındaki bilgi akışları incelendi. Gerçek yaşamdan alınmış, residü ağ yapıları protein veri bankasında depolanmış kordinatlardan yapılandırıldı. Komşular arasındaki yönlenme maliyeti residü-residü etkileşim maliyetleri ile hesaplandı. Seçilmiş kritelere göre yapılandırılan ilk/hedef düğümleri arasında yapılandırılan bütün yollar aracılığı ile, ağ yapısındaki yapısal ve/veya işlevsel önemli residüler sezdirildi. Bilhassa, yüksek yapısal benzerliğe ve düşük dizi benzerliğine sahip protein ailelerinde bütün olası yollardaki yükleri minimize eden güçlü yolların farklılaştığı bulundu. Güçlü yolları kullanmaya sürükleyen faktörleri belirleyebilmek için, kenar seçme olasılığı varış yeri bilgisi ve ilk komşuların energi ilişiklendirilmesi arasındaki dengeye bağlı olarak olasılıklandırılan eğimli rastgele yürüyüşler yapılandı. Uzak yerleştirilmiş iki fonksiyonel grubun birbiri ile olan uzun mesafeli iletişimi bilgi akışının eğimli olmasını gerektirdiğinden, güçlü yollar bölgesel ve global bilgi arasındaki çekişmenin sonucu olarak ortaya çıktı.

## Aknowledgements

I would like to thank my thesis advisor Prof. Canan Atılgan for her support, guidance and patience.

I also thank to Prof. Ali Rana Atılgan for their guidance, useful discussions and support.

And, of course I am indebt to all my friends at MIDST group, Deniz Turgut, Gökhan Kaçar, Ayşe Özlem Sezerman, Osman Burak Okan, Ibrahim Inanç and last but not least Anastassia Zakhariouta for their interest and support.

Finally, I devote this thesis to my family for their everlasting support.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Literature Search

Understanding and predicting structure and function relationships in proteins is an area of intense scientific research. Most proteins perform their function by binding other molecules (i.e. ions, nucleic acids) or other proteins. They regularly experience perturbations in their crowded environment, yet they function efficiently, accurately and rapidly [3]. They also have conformational flexibility which in return signifies the concerted action of residues within the structure [4]. These attributes of proteins make them effective information transmitters in the environment of the cell [5]. Research using different methodologies reveals the shroud surrounding these highly specific organic molecules. Recent research has made progress in expressing the protein structures with a network representation; this provides a simplified model of biological systems [6]. Both residue-residue interaction and protein-protein interaction networks are investigated to contribute to our understanding of protein structure/function relationship [7]. Below, we describe the background of these approaches.

Interactions, delay, and feedback are the three key characteristics of complex fluids. Using these features, entities at different time and length scales communicate with great accuracy, efficiency, and speed [5]. Proteins being Self-assembling molecules are complex fluids with robust and adaptable architectures that incorporate nanoscopic and mesoscopic length scales decisive on their emergent properties over different timescales. Their internal motions which are crucial on their folding, stability, and function, are exquisite examples of these [8–10].

Proteins are tolerant to mutations with their liquid-like free volume distributions [11]; however, the average packing density in a protein is comparable to that inside crystalline solids [12]. It has been shown that the interiors of proteins are more like randomly packed spheres near their percolation threshold and that larger proteins are packed more loosely than smaller proteins [13]. At physiological temperatures, the conformational flexibility is essential for biological activity that requires a concerted action of residues located at different regions of the protein [3]. This cooperation requires an infrastructure that permits a plethora of fast communication protocols. Highly transitive local packing arrangements, giving rise to regular packing geometries [14] cannot provide such short distances between highly separated residues for fast information sharing. On average, random packing of hard spheres similar to soft condensed matter is obtained for a set of representative proteins [15]. This architecture is capable of organizing short average path lengths between any two nodes in a structure, but it cannot warrant a high clustering similar to regular packing.

Proteins regularly experience perturbations in their environment-e.g., in the cell where other small and large molecules are densely and heterogeneously distributed-or in the test tube with only water around, displaying ceaseless fluctuations around their folded structure. Since proteins function efficiently, accurately, and rapidly in the crowded environment of the cell, they are expected to be effective information transmitters by design. Whether the protein is functional or not depends on the size and location of these fluctuations, making use of the concerted action of residues positioned at different regions of the protein [5]. It is, therefore, of utmost interest to investigate how proteins respond to changes in the environment under physiological or extreme conditions.

The response of any structure to perturbations depends on its general architecture. For proteins, local, regular packing geometries [14] cannot provide short distances between highly separated residues for fast information transmission. In fact, it has been shown that random packing of hard spheres similar to soft condensed matter is observed in a set of representative proteins [15]. Consistent with the concurrent requirement of order and randomness in the protein structure, it has been shown that proteins are organized within the small-world network (SWN) topology. A network is referred to as 'small-world'if the average shortest path between any two vertices scales logarithmically with the total number of vertices, provided that a high local clustering is observed [5]. The former property of short paths is responsible for the name 'small world'. Neither regular configurations nor random orientations seem to exhibit these two intrinsic properties. However, such properties are common in many real-world complex networks [16], and there are examples from a diverse pool of applications such as the world wide web [17], the internet [18], math coauthorship [19], power grid [20], and residue networks [4]. In recent years, proteins have been treated as networks of interacting amino acid pairs to determine their network structure and to identify the adaptive mechanisms in response to perturbations [4,21]. In fact, similar network treatments of proteins predict collective domain motions, hot spots, and conserved sites [22]. For these networks, we employ the term 'residue networks' [4] to distinguish them from 'protein networks'. The latter are used to describe systems of interacting proteins [23].

With their ordered secondary structural units made up of $\alpha$ -helices and $\beta$ -sheets on the one hand, and their seemingly unstructured loops on the other, it was predicted that proteins may have the SWN organization [4]. Later, a statistical analysis showed that proteins may in fact be treated within the small-world network topology, balancing efficiency and robustness. The local and global properties of these networks with their spatial location in the three-dimensional structure of the protein were determined [4]. The same local organization of core residues appears irrespective of the protein size. Moreover, a remarkable correlation was found to exist between residue fluctuations and shortest path lengths. Recent developments of elastic network models for studying large amplitude motions in proteins have been successful in predicting functional mechanisms [21,24]. In particular, the cohesive domain-like behavior of proteins is well understood by these models. In the residue networks treatment, a similar network construction based on the average structure is used with a different perspective. Instead of a statistical mechanical approach whereby the system energy is described by the additive local interactions of harmonic springs, a graph theoretical viewpoint is taken by considering pathways of interconnections. Thus, the two approaches, both originating from packing characteristics, lead to different information.

In the past few years, the network treatment of residues in proteins has been further adopted to study their various features such as conserved long-range interactions [25], functional residues [26], protein-protein association, and detection of structural elements [27,28] . In all these treatments, which have been successful in describing many important properties of proteins and provide insight as to how they function, the identities of individual amino acids are omitted in the calculations. In other words, specificity is taken into account in an indirect manner by assuming that the locations of the different amino acid types along the contour of the polymeric chain have been operational in determining the particular average three-dimensional structure. In this viewpoint, the interactions between different pairs, triplets, etc. of amino acids are assumed to be smeared out, and the observed behavior once the protein is folded is driven by the overall structure. In fact, it has been noted that the residue non-specific interactions (i.e., those depending on the relative placement of residue pairs, irrespective of their identity) contribute more to the overall stability of proteins by a factor of about five, compared to distinct residue-residue interactions [27]. The question remains, however, as to the extent to which such a coarsened description of the folded protein may be used to determine other crucial properties, especially those pertaining to dynamics. Recently, the paths between residue pairs have been elaborated upon, which are termed 'information pathways,'to understand how they relate to dynamic phenomena in proteins [5].

In particular, it is of interest to understand allosteric interactions mediated through the changes in the dynamic fluctuations around the average structure, both in the presence and absence of conformational changes, the latter having recently been shown to exist in proteins through a series of NMR experiments [29]. To this end, weights have been attributed to the links between residue pairs using knowledge-based potentials [1, 30] and the relationship between dynamic phenomena occurring in proteins and the optimal path lengths obtained from these weighted networks have been discussed. It has been shown that it is possible to extract minimal subgraphs from the fully connected networks of residues, where a few designed interactions overlaying the backbone are sufficient to display communication path lengths similar to that of the full residue network [5]. A demonstration of the application of these ideas using a non-redundant data set of interacting proteins have been made and residue pairs on the interface of the receptor/ligand that frequently appear along information pathways have been extracted in the same study.

Most theoretical and computational biophysical methods available today will give information on equilibrium states. The non-equilibrium dynamical information is usually inferred from the study of different equilibrium states and interpolation [4]. The idea of following pathways on networks is an attractive one for studying not-far-from-equilibrium phenomena such as the attainment of new equilibrium states upon binding. However, one first needs to validate the limitations of coarse graining. In particular, the extent to which quantum mechanical effects can be neglected or incorporated into the models must be assessed; e.g., in CO binding to myoglobin [31] the relaxation pathway in the protein is of utmost interest [32]. Consequently, this unifying network perspective lets us explore protein dynamics such that, apart from distinguishing structurally important residues in folding, binding, and stability, it will be possible to locate the routes through which a perturbation is communicated in a protein, and estimate the time scales on which a response is generated. As such, it will complement newly developing experimental techniques such as femtosecond spectroscopy. The spatiotemporal nature of the hypothesized process calls for deeper investigation on particular proteins. The global rules deduced for proteins are also expected to have applications in bioinformatics problems such as identifying interaction surfaces in protein docking and distinguishing misfolded states.

By taking a network perspective of analyzing proteins, it was shown that residue specificity plays an important role in protein functioning. Inhomogeneity is introduced into the residue networks by assigning each edge a weight that is determined by amino acid pair potentials. Two methodologies are utilized to calculate the optimal path lengths (APLs) between pairs in these weighted networks: to minimize i), the maximum weight in the **strong** APL, and ii), the total weight in the **weak** APL. A statistical analysis on nearly 600 non-homologous proteins has led to define key quantities for discriminating the underlying structure that make the protein robust in the environment where it is functional. In particular, a quantity has been uniquely defined for finding a critical threshold value to determine the key interactions in the protein, if it is to survive extreme events and to continue carrying out its function. Those results also support the finding that optimized protein sequences can tolerate relatively large random errors in pair potentials obtained using a variety of methodologies [33].

It was proposed that in events involving small perturbations, the total energy to traverse that path will be important and information will flow through the optimal paths with weak disorder, similar to that in the homogeneous network. On the other hand, when large perturbations are involved, such events require surpassing the largest energy barriers along the paths. In this approach, the same pair potentials are used as thermodynamic measures in the former case and as kinetic measures in the latter. If a pair of residues has high contact energy, it may be assumed that the energy that must be used to separate them will be commensurate with its value to a first approximation. Due to other effects such as the size and the shape of the residues, slight modifications may be included. The strong paths, therefore, were predicted to set a limit on the protein whereby the robust structure resists large amounts of external perturbations and preserves its protein-like communication pathways. Furthermore, using this approach, we have been able to define key contacts that form bridges between interacting proteins. Note that nearly half the surface area of the total protein, and therefore an overwhelming number of residue pairs, is involved in protein-protein interactions.

## 1.2 Aim

In this study, we systematically construct paths between pairs in residue networks to understand how they relate to dynamic phenomena in proteins furthering previous studies carried out in our group [4, 5]. We carefully select data sets that have the characteristics of high structural alignment, but low sequence alignment to emphasize the impact of residue interaction to information transfer between highly separated residues.

We also search factors that drive the usage of strong paths in the network. A biased random walk scheme is deviced, where the probability of edge selection is based on a balance between the knowledge of the location of the destination and the energy of interaction with the immediate neighbors.

The thesis is organized as follows: In Chapter 2 we briefly describe how residue networks are constructed from their Protein Data Bank coordinates [34]. Therein, we also define various path length measures used in this work and describe how these paths are calculated. In Chapter 3, we present results from the TIM Barrel superfamily of proteins and Calcium Binding family of protein pairs. These are selected to represent communication paths in single chains and interacting pairs of chains, respectively. In Chapter 4 we use biased random walk algorithms on weighted networks using different weight assignments to local neighbors as well as a global knowledge of the destination node. We discuss the conditions under which the system can be maneuvered from weak-like to strong-like paths. In Chapter 5, concluding remarks are presented on the usage of SAPL versus WAPL in information communication in residue networks. Implications for protein function are discussed and future work is suggested.

## 2 Background on Residue Networks

Proteins in this study are treated as networks by taking every residue in the protein structure as a single node and interaction among them as edges [5]. These networks are based on protein structure data obtained from the Protein Data Bank [34]. Each residue in the structure is represented as a node centered on $C_\beta$ atoms spatial position. In the case of Glycine, $C_\alpha$ represents the coordinates of the node. These nodes are then considered as connected if they are positioned within the first coordination shell of each other, which is 6.7 Å [5]. This procedure enables us to generate the $N$x$N$ adjacency matrix for each protein, where $N$ is the number of residues. The elements of these matrices have the values zero or one depending if there exist a contact between residues or not. This can be mathematically expressed as

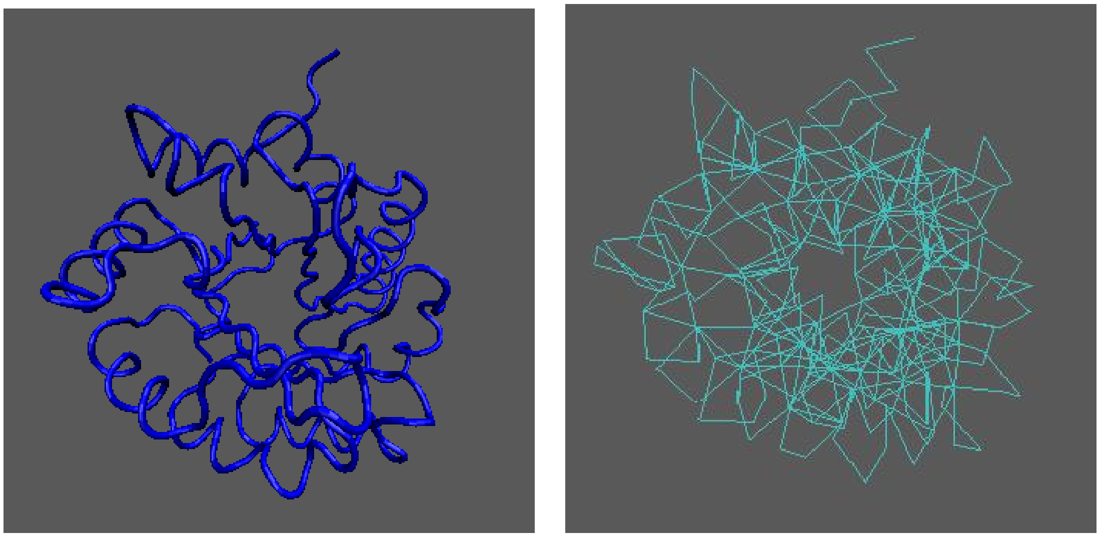
$$A_{ij} = \begin{cases} H(r_c - r_{ij}) & i \neq j \\ 0 & i = j \end{cases} \tag{1}$$

where $r_{ij}$ is the distance between the $i^{th}$ and $j^{th}$ nodes, $H(x)$ is the Heavyside step function given by $H(x) = 1$ for $x > 0$ and $H(x) = 0$ for $x \leq 0$, and $r_c$ is the given cut-off distance, which is an upper limit for the separation between two residues in contact. Since proteins are linear polymers of 20 different amino acids, chain connectivity is also conserved and mathematically expressed as:

$$A_{ij} = 1 \qquad if \quad i = j \pm 1 \tag{2}$$

An example network representation of the protein D-Ribulose-5-Phosphate 3-Epimerase from Solanum Tuberosum Chloroplasts (PDB code:1RPX) shown in Figure 2.1 for

$r_c = 6.7$ Å.



(a) Protein structure of 1RPX (Tube representation)

(b) Derived network structure of 1RPX

Figure 2.1: Protein and derived network structure of 1RPX

The networks are classified by local and global parameters, all of which can be derived from the adjacency matrix. The most general descriptor of the network structure is the average connectivity of a node. The connectivity $k_i$ is the number of neighboring residues of residue $i$ which is given as;

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{3}$$

The average shortest path through the network is another widely used network descriptor. Dijkstra algorithm is used to compute the shortest paths, i.e. the number of minimum steps between a pair of residues [35]. The shortest path lengths of a homogeneous network, where the edges have no weight, is termed as *Homogeneous Average Path Length (HAPL)* in this study.

Constructed residue-residue networks can also be represented as weighted networks once we assign weights to the edges. We use residue-residue interaction potentials of Miyazawa - Jernigan (MJ) and Thomas - Dill (TD) as attributes of edges; data given in

Appendix A [1,2]. These two potentials have been extensively tested in threading algorithms, protein stability and designability studies, folding and binding energetics and amino acid clasification [30, 36–38]. In these weighted networks, we use two definations to calculate shortest paths: Weak average path lengths (WAPL), and strong average path lengths (SAPL). In the former, the optimal path connecting residues $i$ and $j$ is the length of the path that minimizes the sum of the weights along the path. Dijkstra algorithm for the weighted graphs is used to compute the WAPL [35]. Minimization of the sum of the weights along the path requires weights to be positive, thus a positive value is added to residue-residue potentials. This value is set as three for Thomas - Dill and as eight for Miyazawa - Jernigan interaction potentials . For calculating SAPL, we sort the attributes of edges in descending order and systematically remove the connection beginning with the highest weight until a *bottleneck* value is reached, whose removal results in loss the of connection between nodes $i$ and $j$. Below we represent both HAPL, WAPL and SAPL and bottleneck edge on a toy model constructed by a 3x3 grid structure. On Figure 2.2 each node is considered as a residue; thus, to each edge attributes (TD potetials) are assigned. In this toy model both HAPL, WAPL and SAPL are shown with a different color scheme connecting the start and target nodes. We use blue, yellow and red lines to represent paths for HAPL, WAPL and SAPL respectively. Note that the edge $MET8 \Leftrightarrow HIS13$ is the bottleneck edge and whose removal result in loss of connection as shown on Figure 2.3.

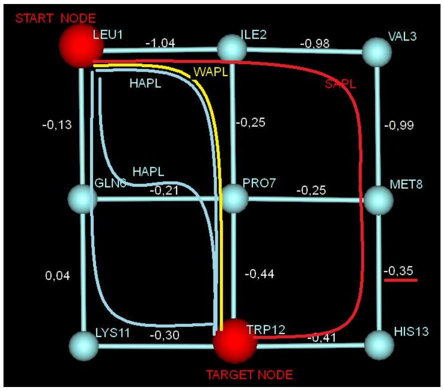


Figure 2.2: HAPL, WAPL AND SAPL on a 3x3 grid

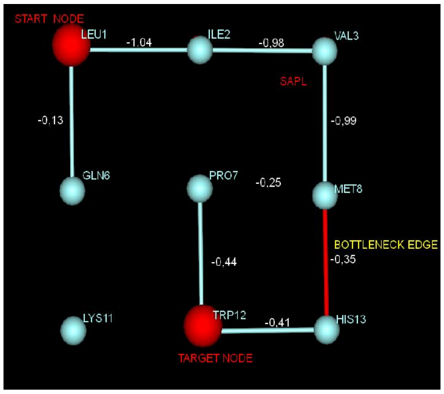


Figure 2.3: Reduced network structure on a 3x3 grid structure

The characteristic path length of the network is the average

$$L^{\dagger} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} L_{ij}^{\dagger} \tag{4}$$

where the dagger symbol, †, represents the homogeneous, weak or strong paths. Note that $L^{\dagger}$ is the measure of the global properties under the imposed constraints.

In this work, we systematically compare the HAPL, WAPL and SAPL in selected single proteins and pairs of interacting protein to derive relationships between their structure and function.

# 3 Protein Structures of Residue Networks and Protein Functionality

## 3.1 Paths in Single Chains

We first conduct a systematic study of the paths on TIM barrels. TIM barrel family is chosen owing to several reasons. The members of this protein family exhibit high structural similarity, whereas they lack sequence similarity. They have the most common tertiary fold observed in high resolution protein crystal structures; approximately 10 % of all known enzymes have this domain [39]. 584 structural hits were observed among the 55546 protein structures in PDB [34]. The members of this large family of proteins catalyze very different reactions, including five of the six primary classes of enzymes [40]. As the evolutionary history of TIM barrels is still being unrevealed, the fact that such a variety of sequences acquire the same fold puts them under scrunity [40–42].

TIM barrels acquire a canonical $(\beta/\alpha)_8$-barrel fold consisting of inner eight parallel $\beta$-strands wrapped by an outer wheel comprising eight $\alpha$-helices. They vary in size from 200 to 400 amino acids. TIM barrels have phosphate binding sites formed by loop 7, loop 8 and a small helix (helix-$\acute{8}$) [40]. They also bear other active sites of metal-binding located on $\beta$-sheet5 $\alpha$-helix5, a catalytic site on $\beta$-sheet5 $\alpha$-helix5 and $\beta$-sheet1 $\alpha$-helix1, as shown in Figure 3.1 [40]. Residues located on these sites will be used as starting and destination nodes for constructing paths, in this study.

We select two different superfamilies of this fold, namely Ribulose-phosphate-binding TIM barrels and TIM barrel glycosyl hydrolases according to SCOP classification to further investigate information pathways within the structure. We first seek structurally highly aligned proteins within and between the superfamilies. We use MultiProt to align pairs of proteins [43].The multiple alignments are achieved by simultaneous structural superposition of input molecules in all possible ways under the condition that at least

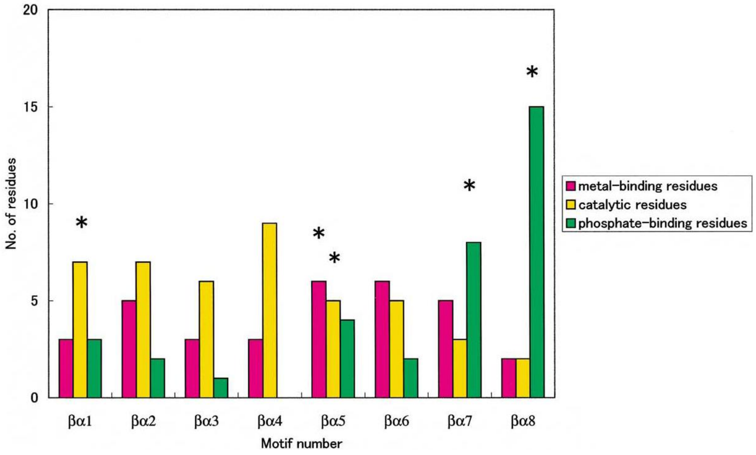


Figure 3.1: Active site residues at the eight $\beta/\alpha$ motifs.

short contiguous fragments (three amino acids or more) of the backbone chains should be structurally similar. The method computes the best scoring structural alignments, which can be either according to a sequence order, like in sequence alignment, or be sequence-order independent in order to seek geometric patterns which do not follow the sequence order [43].

Results indicate that structural alignment of Ribulose-phosphate-binding TIM barrels vary between 54% and 90% identity with an average value of 69%. Onthe other hand intersuperfamily aligment of proteins yield an average value of 29%. We then compute HAPL and SAPL for proteins whose alignment within and between superfamilies show good agrement. We use two parameters to select proteins for further examination. First root mean square deviation (RMSD) values are compared, second the numbers of residues aligned is taken into acount by defining a match ratio, the latter is the ratio of aligned residues to the total number of residues of the smaller protein. The protein pairs selected are 1PII and 1RPX of Ribulose-phosphate-binding TIM barrels and 1CWY, 1BAG and 1CEO of TIM barrel glycosyl hydrolases. In Table 3.1 we show RMSD values and match ratios of selected proteins belonging to different superfamilies. Further data related are given in Appendix B.

| RMSD<br>Match Ratio | 1PII | 1RPX | 1CWY | 1BAG | 1CEO |
|---|---|---|---|---|---|
| 1PII | | 1.71 | 1.76 | 1.84 | 1.78 |
| 1RPX | 0.7 | | 1.79 | 1.79 | 1.79 |
| 1CWY | 0.54 | 0.54 | | 2.05 | 1.87 |
| 1BAG | 0.22 | 0.54 | 0.48 | | 1.90 |
| 1CEO | 0.33 | 0.60 | 0.40 | 0.41 | |

Table 3.1: RMSD (Å, upper diagonal) and match ratio (lower diagonal) of proteins 1PII, 1RPX, 1CWY, 1BAG, 1CEO

We select residues from the phosphate binding site of each protein as starting node, residue no. 236, 207 and 451 for 1pii, 1rpx and 1cwy, respectively. We compute paths to every secondary structure of the considered proteins to verify differences in terms of node selection and secondary structure usage. If structural data are not available, those from the structurally aligned counterpart is chosen. The process flowchart is given Appendix B.

We find that distinguishing features are captured by SAPL and not WAPL. We therefore present results from the former only. The *bottleneck edges*, i.e. those having highest weight in paths constructed by SAPL, are listed for 1RPX in Appendix B. They have a weighted average Thomas Dill (TD) potential of $-0.31k_BT$. We select top 50 bottleneck edges in order of percent usage, which represent % 46 of total bottleneck edges, and characterize their proximity to the surface of the protein structure. We show percent usage of bottleneck edges on Figure 3.2. Also, the top three bottleneck edges, whose total usage adds up to %5.8, shown on as balls Figure 3.3. 1PII and 1RPX,the two members of Ribulose-phosphate-binding TIM barrels, differ in terms of residue usage in these paths, whereas 1PII and 1CWY have common nodes in terms of spatial positions of nodes used. Having such a difference between members of the same superfamily or similarity between different superfamily members signify the effect of bottleneck edges to direct information pathways.

Our first observation is the excessive use of $\beta$-sheet secondary structures within SAPL computed. This kind of path behavior is the result of two important features of this fold.

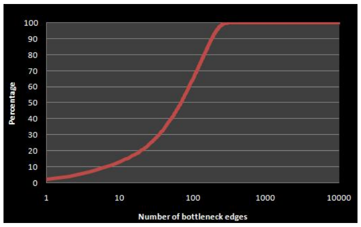
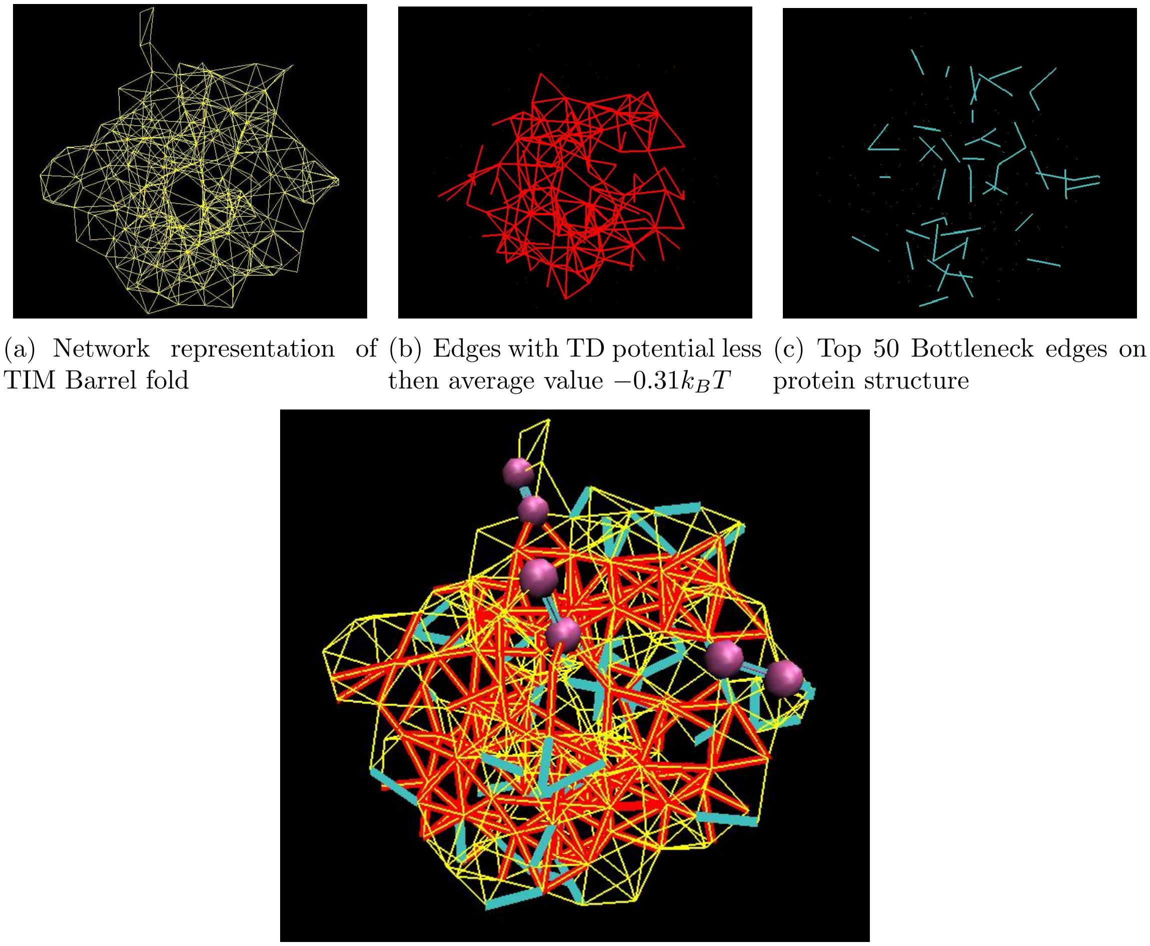


Figure 3.2: Percent usage of bottleneck edges for 1RPX

Although the protein has a donut like structure, interactions between residues, residing on $\beta$-sheets grant passages between distantly located residues. Adding to this, the number of interactions between adjacent $\alpha$-helices are limited owing to spatial positions of these secondary structures. The other important factor is the residue distribution within the tertiary structure of the fold. The core region, mostly formed of hydrophobic residues, valine, leucine, and isoleucine, comprise about 40% of the total residues, favoring paths in SAPL once the bottleneck edge is reached [44]. Thus, the spatial positon of the bottleneck edge and its residue-residue interaction potential determine constructed paths, hence are the residues used for information passage. A sample path from phosphate binding site of each protein to a distantly located loop on structurally aligned protein structure is given in Figure 3.4 (a). Usage of each node on calculated paths are normalized and visualized by nodes having different diameters owing to the existance of more than one path to connect these residue pairs. 1RPX, 1PII, 1CWY are represented by different color scheme green, brown and blue respectively.

For each protein, path characteristics differ in terms of nodes visited and the bottleneck edges used. 1RPX and 1PII, being members of Ribulose-phosphate-binding TIM barrels, are expected to have similar pathways. Although spatial position of bottleneck edges for these proteins are aligned, which are [Gly-207 ↔ Val-210] for 1RPX and [Gly-236 ↔ Ala-238] for 1PII, the pathways connecting start and destination residues differ. Considering these three pathways, similarity arises only in the usage of core residues for communication. Another constructed path is also given in Figure 3.4(b) where start and destination nodes are phosphate binding site and catalytic residue located and



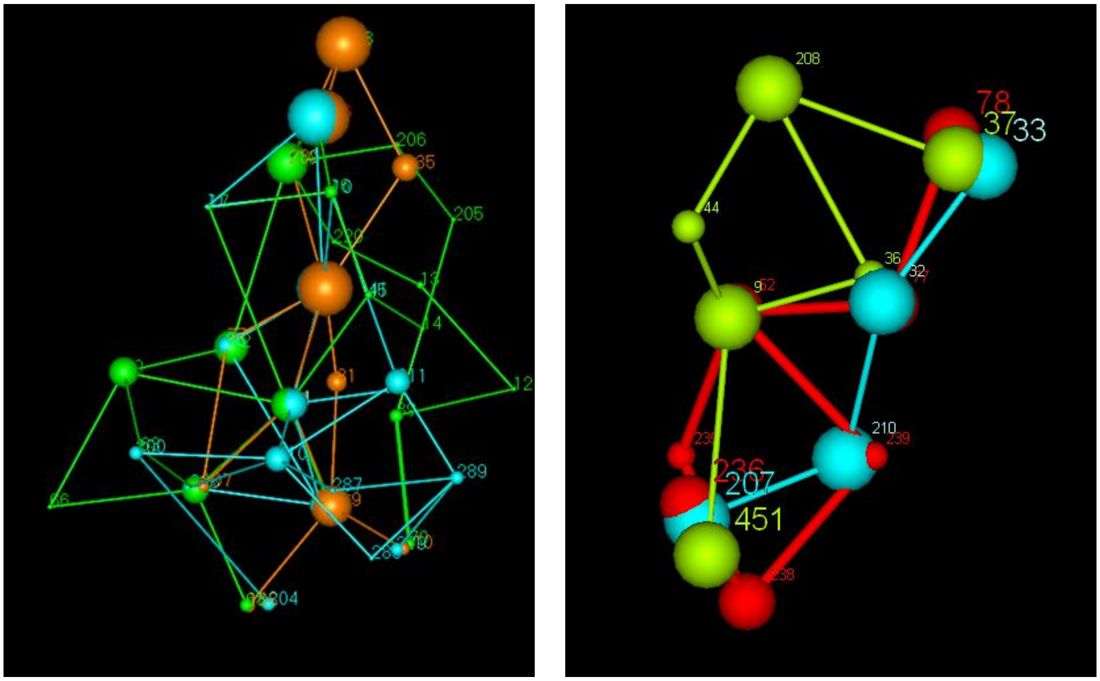
(a) Network representation of TIM Barrel fold

(b) Edges with TD potential less then average value $-0.31k_BT$

(c) Top 50 Bottleneck edges on protein structure

(d) Superimposed network structures of Figures (a),(b),(c)). Top three bottleneck edges represented with purple nodes.

Figure 3.3: TIM Barrel fold and bottleneck edges represented on the TIM Barrel fold for the protein 1RPX



(a) Phosphate binding site to loop formed by $\alpha_2$ and $\beta_3$ secondary structure

(b) Phosphate binding site to loop formed by $\alpha_1$ and $\beta_2$ secondary structure

Figure 3.4: Paths starting from phosphate binding site of each protein to distant nodes

aligned on $\alpha_1$ and $\beta_2$ secondary structure for each protein. The pathways constructed for 1RPX and 1PII , shown in blue and red respectively, have the same bottleneck edge as in paths to $\alpha_2 and \beta_3$ secondary structure and also uses more nodes which are also aligned, yet the path constructed for 1CWY differs in terms of the aligned node usage.

## 3.2 Paths In Complex Structures

We have further studied a data set of nine proteins clustered according to the similarities of the global structure of the chains [45]. Hence, interfaces derived from these proteins also have similar structures [46]. These nine calcium binding proteins belong to the superfamily EF-hand, has EF hand like fold, forming all alpha proteins according to SCOP definition.

In this data set, we applied both homogeneous average path length (HAPL) and strong average path length (SAPL) methodologies to investigate paths which are favored for information transfer within the two chains of the proteins [5]. All paths starting from chain A and ending at chain B were computed for every residue of each chain and statistical data were gathered. We display the top six residue pairs that appear in the HAPL and SAPL, in the Appendix C, where residue pairs that are structurally aligned are marked with **X** and nearly aligned pairs with **I**. The amount of the match between these protein pairs vary, but in general interface edges that appear in SAPL match more often. For reference, we also list the sequence dissimilarities, which yield the proportion of amino acids that are different in both sequences, of these nine proteins in the whole structure and along the interface in Table 3.2, as calculated by the structural alignment of STRAP program [47]. In this Table, the upper right triangle contains data for the whole protein, whereas the data for the the aligment of only the interface residues are displayed in the lower triangle.

Of these nine proteins, we select three of them for detailed analysis. These three proteins are 1KSO, 1B4C and 1BT6. The overall sequence dissimilarities differ from that of the interface, especially for 1KSO-1BT6 and 1KSO-1B4C given in bold numbers in Table 3.2. In other words, these three proteins are more alike considering interface alignment. Note that the number of aligned residues for the pairs 1KSO-1B4C, 1KSO-1BT6 and 1BT6-1B4C are 22, 28 and 25 respectively.

| Protein<br>Interface | 1E8A | 1MR8 | 1BT6 | 1YUT | 1PSR | 1B4C | 1A03 | 1KSO | 1NSH |
|---|---|---|---|---|---|---|---|---|---|
| 1E8A | | 0.60 | 0.69 | 0.67 | 0.74 | 0.69 | 0.70 | 0.95 | 0.94 |
| 1MR8 | 0.56 | | 0.74 | 0.78 | 0.76 | 0.69 | 0.77 | 0.95 | 0.96 |
| 1BT6 | 0.68 | 0.75 | | 0.74 | 0.81 | **0.76** | 0.74 | **0.92** | 0.97 |
| 1YUT | 0.69 | 0.78 | 0.75 | | 0.80 | 0.80 | 0.75 | 0.96 | 0.91 |
| 1PSR | 0.73 | 0.71 | 0.74 | 0.75 | | 0.81 | 0.79 | 0.94 | 0.94 |
| 1B4C | 0.66 | 0.63 | **0.71** | 0.84 | 0.73 | | 0.63 | **0.93** | 0.95 |
| 1A03 | 0.71 | 0.74 | 0.77 | 0.79 | 0.81 | 0.68 | | 0.94 | 0.95 |
| 1KSO | 0.69 | 0.78 | **0.77** | 0.69 | 0.85 | **0.70** | 0.70 | | 0.75 |
| 1NSH | 0.71 | 0.66 | 0.68 | 0.71 | 0.66 | 0.63 | 0.75 | 0.69 | |

Table 3.2: Dissimilarity scores of whole (upper right triangle) and interface (lower right triangle) structures of nine calcium binding family proteins

In the residue networks, edges whose connecting nodes reside on separate chains of the dimers are termed as *interface edges.* They are considered to play a significant role for information transfer between the two chains. Bottlenecks of SAPL are separately labeled to determine how paths, and in particular interface edge usages differ with respect to the methodology used. Frequencies of interface edges used in both methodologies for selected proteins are listed in Table 3.3, in Table C.1, the same data for the whole data set is given. The results for HAPL and WAPL display the same top pairs and hence are not listed separately.

For each protein, the most frequently used interface edges are common in both HAPL and SAPL. However, observed frequencies differ significantly for some of the interface edges whereas others remain relatively unchanged. Moreover, as in the case of 1KSO the interface edge [A77 ↔ B77] which has 10.5% usage in HAPL totally vanishes in SAPL. This kind of behavior of interface edges can be attributed to the effect of global structure and residue-residue interaction potential on protein-protein interaction. We may thus have two types of interface edges: Those that are structurally strategically positioned appear with high usage both in HAPL and SAPL. Others that are kinetically important in information communication between chains appear with high usage in SAPL.

| | Strong APL | | Homogeneous APL | |
|---|---|---|---|---|
| | Interface Edge | **% Usage** | Interface Edge | **% Usage** |
| 1B4C | A70-B82 | 14.6 | A70-B82 | 13.2 |
| | A82-B70 | 11.7 | A3-B39 | 8.7 |
| | A78-B74 | 10.6 | A78-A71 | 8.1 |
| | A3 -B39 | 10.0 | A39-B3 | 8.1 |
| | A39-B3 | 9.9 | A82-B70 | 6.8 |
| | A11-B87 | 6.2 | A78-B74 | 6.1 |
| 1BT6 | A76-B72 | 12.5 | A4 -B38 | 8.0 |
| | A4 -B38 | 11.0 | A38-B4 | 8.0 |
| | A38-B4 | 10.7 | A4 -A37 | 7.8 |
| | A72-B76 | 8.5 | A80-B68 | 5.0 |
| | A80-B68 | 7.8 | A68-B80 | 3.8 |
| | A12-B82 | 7.3 | A76-B72 | 3.8 |
| 1KSO | A80-B72 | 9.7 | A77-B77 | 10.5 |
| | A72-B83 | 8.7 | A80-B72 | 7.0 |
| | A76-B76 | 6.7 | A72-A83 | 6.7 |
| | A73-B77 | 6.3 | A27-B93 | 5.3 |
| | A27-B93 | 5.6 | A76-B76 | 5.1 |
| | A76-B79 | 4.6 | A76-B79 | 4.1 |

Table 3.3: Percent usage of top five interface edges in SAPL and HAPL

We also used two different structure comparison algorithms to locate and compare interface edges and bottlenecks within the protein complex. MultiProt, a sequence order independent structural comparison algorithm, and STRAP (ClustalW 3D), a structural comparison algorithm which takes into account both sequence identity and protein structure [43]. Both result in the same structural alignment owing to the nature of cluster between selected proteins [43]. This approach enables us to identify edges which are located at the same spatial position. In Figure 3.5, most frequently used interface edges that are positioned at common sites are shown for HAPL and SAPL, respectively, above and below the two arrow headed lines.

Considering each of these proteins has higher similarity along their interface then the global structure, each protein pair can be further analyzed. For pair 1B4C and 1BT6, interface edges [A3 ↔ B39 and A4 ↔ B38] and [A39 ↔ B3 and A38 ↔ B4] are not

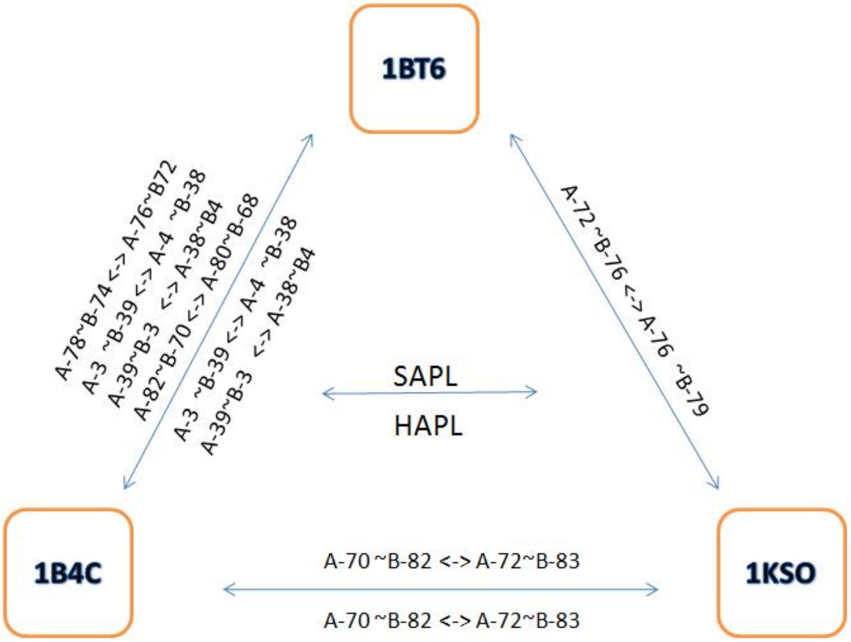


Figure 3.5: Aligned interface edges for both SAPL and HAPL

only structurally aligned, they also have approximately the same percentage usage for both methodologies, on the other hand [A78 ↔ B74 and A76 ↔ B72] and [A82 ↔ B70 and A80 ↔ B68] only appear in SAPL. For pair 1BT6 and 1KSO, interface edges [A72 ↔ B76 and A75 ↔ B79] are only seen with SAPL where no substitution for these edges appears with HAPL. For pair 1B4C and 1KSO, interface edges [A70 ↔ B82 and A72 ↔ B83] are located with both methodologies where usage percentage with SAPL is slightly reduced by 1.4 % and 2.1 %, respectively. Finally, for the protein pair 1BT6 and 1KSO, only one common interface edge appears in SAPL and none in HAPL. The procedure followed is given as a flowchart in Appendix C (Figure B.1.

The phenomenon can better be visualized if paths using these interfaces are shown on the three dimensional structures. Since our algorithm outputs a vast amount of pathways (e.g. for 1KSO the number of all paths in SAPL exceeds 97700) only some of the paths bearing the above characteristics are shown in Figure 3.6. Starting nodes are chosen from the cluster of conserved residues of each protein and structural alignment of these residues are also taken into consideration. Ending nodes are chosen from amongst residues that are either on the surface or close to the surface of the opposite chain. Paths starting from residue Leu62 of Chain A of 1KSO connect to the residues B46-B47-B56-B88 and B92 through the interface edge A72 ↔ B83. Structurally aligned counterparts of these residues are Leu60 of Chain A of 1B4C and residues B45-B50 and B54, whereas residues B55-B58-B62 and B81 also use the same interface to connect starting residue. In Figure 3.6, paths between Leu62 to B46 and B56 for 1KSO (pink and red) and Leu60 to B45 and B54 (blue and light blue) and also the nodes which are not structurally aligned, but using same interface edge to connect to the destination node; of each residue network are shown. Note that this interface edge is the common one in the two proteins and appear with a large frequency in the statistics of all SAPL and HAPL paths.

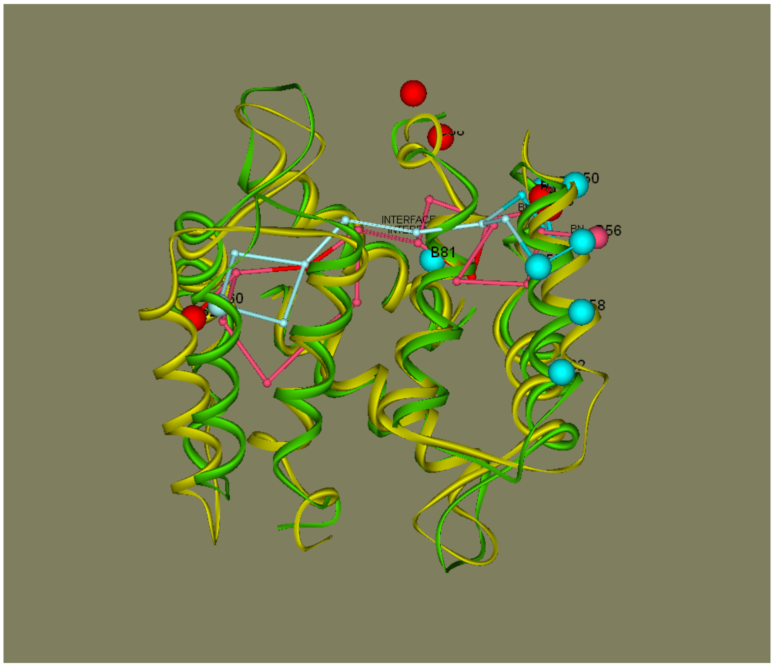
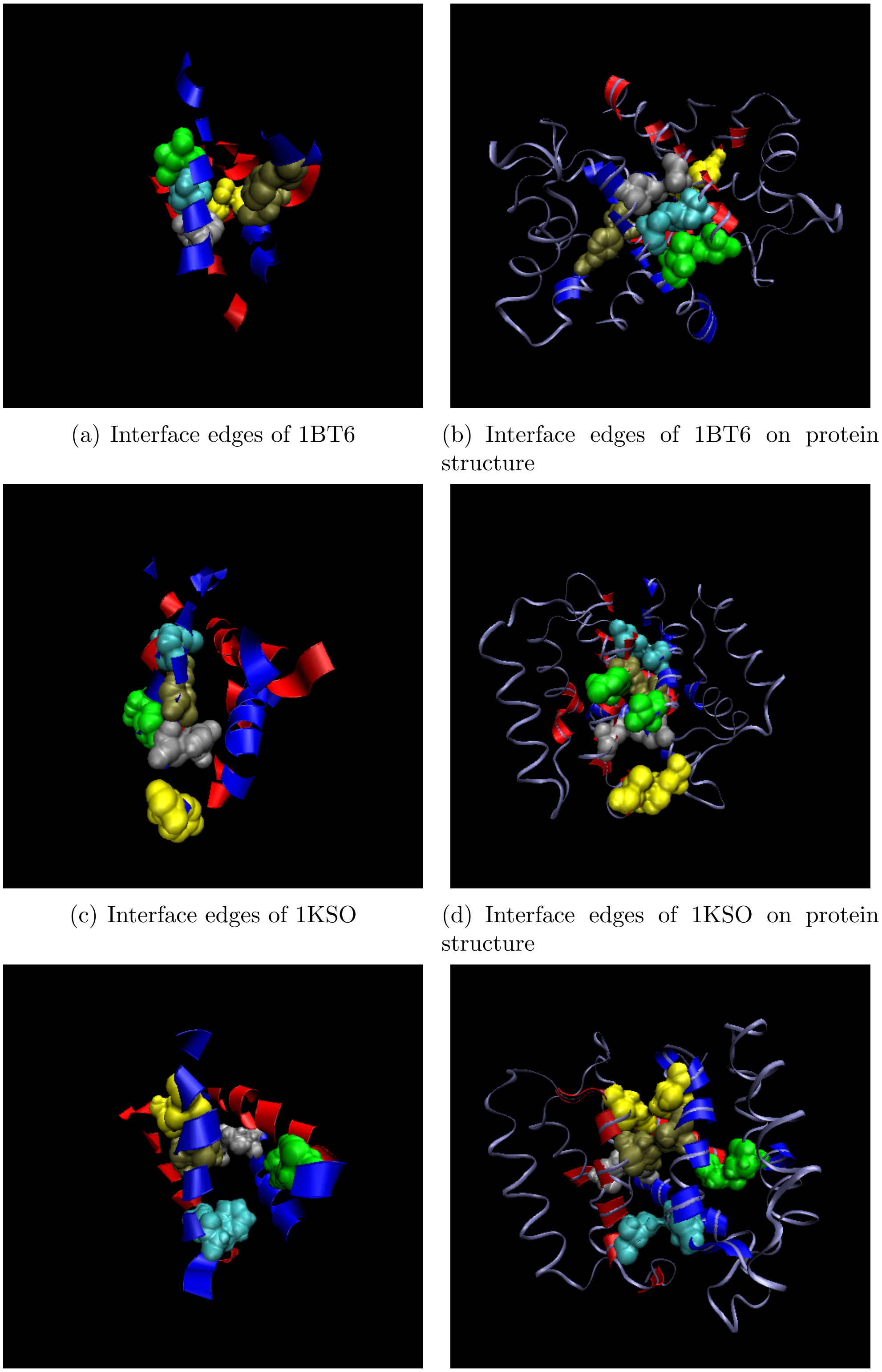


Figure 3.6: SAPL using the same interface edge and structurally aligned residues of proteins 1KSO and 1B4C

The same approach for homogeneous paths result in different characteristics. Paths starting from Leu62 of Chain A of 1KSO and ending at residue B46 resulted in almost the same kind of path behavior with one of the six different paths connecting these residues, differing only slightly at the interface region whereas paths starting from Leu62 of Chain A of 1KSO and ending at residue B56 differs significantly from SAPL methodology. Of the 27 different paths connecting these residues, only four of them uses the same interface, but are shorter in terms of the number of steps.

In Figure 3.6 having listed some example paths between structurally aligned regions of these selected protein, we next study the overall placement of the important interface edges on the three selected proteins. We present the top five frequently used interface edges and their spatial positions on these proteins along with the interface residues. The whole structure is represented in ice blue, whereas the interface residues are marked with the red and blue ribbon structures on chain A and B respectively. The important interface edges that appear in the top five of the SAPLs are shown in the different colored atomic clusters (ice-blue, yellow, tan, silver, and green in the order of decreasing percent usage). We find that these residue pairs emerge on both faces of the interface for 1BT6 and 1B4C, whereas they are clustered along one side in 1KSO. In Figure 3.2, we display the interface and the frequently used pairs only, in a view where the interface is rotated by $90^o$ along the z-axis.

These three proteins, 1KS0, 1BT6 and 1B4C have 52, 36 and 40 interface residues respectively and structural alignment of these interface residues are given in Table 3.3. As it is shown in Figure 3.7, each three of the top five interface edges are located between the fourth $\alpha$-helix structure of each chain. The remaining two interface edges have the same characteristic for 1BT6 and 1B4C; they connect each chain to each other by the interaction between their first and second $\alpha$-helix structures. In the case of 1KSO the loop structure between the first and the second $\alpha$-helices of chain A and the fourth $\alpha$-helix of chain B connect the two chains with high percentage usage in constructed paths. Even though 1KSO bears more interface residues, hence interface edges, polar and charged residues residing on the surface structure act as bottlenecks. This kind of replacement of bottleneck edges close to the surface, away from the interface result in



(a) Interface edges of 1BT6

(b) Interface edges of 1BT6 on protein structure

(c) Interface edges of 1KSO

(d) Interface edges of 1KSO on protein structure

(e) Interface edges of 1B4C

(f) Interface edges of 1B4C on protein structure

Figure 3.7: Top five interface edges connecting different chains of proteins 1B4C-1KSO-1BT6

following nearly shortest paths within the core and less hydrophobic regions of the protein once they are reached. Disappearance of the highest used interface edge of HAPL, [A-77 $\leftrightarrow$ B-77], for 1KSO happens because paths previously using this edge in HAPL have bottlenecks close to the surface and have residue-residue interaction potentials lower than that of [A-77 $\leftrightarrow$ B-77]. Also the percent usage of interface edge [A-5 $\leftrightarrow$ B-41] of 1KSO, which is neither aligned nor has significant usage in terms of appearance , but reside at the same secondary structures as [A-3 $\leftrightarrow$ B-39] of 1B4C and [A-4 $\leftrightarrow$ B-38] of 1BT6 have higher percent usage in SAPL. This kind of SAPL data signifies how bottlenecks orient or control the information paths within the protein structure.

Thus, a close examination of these protein pairs shows that for some proteins, edges that are structurally positioned along the interface are used for cross-talk between the two chains. The three residue pairs that appear between the fourth $\alpha$-helices are examples of such cases. These appear with high usage percentage in both HAPL and SAPL. However, there are other residue pairs that emerge in alternative locations. Such shifts in positioning is due to the lowering in the overall energy cost during cross-talk. The structure directs the communication along longer paths in exchange of lower barrier-crossing energies.

# 4 Biased Random Walks on Residue Networks

Residue network paths; that we have been examining; are based on protein structure data derived from PDB. All methodologies we utilize stem from adjacency matrices calculated for the proteins and residue interaction potentials (see Chapter 2). Thus, these approaches are all global approaches and can be derived once related data are available. In principle, most information traveling on a network has access to 'local'knowledge; i.e the identities of the direct neighbors. In addition, 'global'knowledge may also be available, such as the location of the final destination. The later scenario is particularly plausible if there is a gradient towards the destination node.

In this chapter we systematically investigate how local structure affects information sharing on residue networks and how this information determines paths between distantly located residues. We therefore begin by a random walk procedure where all neighbors are treated equally. On this model, we then superpose modified probability distributions on neighbors using local potentials (TD or MJ) and directionality of the target node. We analyze the relative contribution of each factor by monitoring how close the procedure mimics the calculated HAPL, WAPL and SAPL.

## 4.1 Random walks

Our first approach is to perform random walks on the protein structure with the following criteria:

1. Generate a random number between zero and one.
2. Assign the $k$ nearest neighbors in order of appearance in the adjacency matrix for the selected node.
3. Generate intervals of size $\frac{1}{k}$ for each nearest neighbor by normalizing with the total number of neighboring residues.

4. Select the appropriate node to move to, based on the generated random number.

5. Repeat the above procedure for the newly arrived node until the destination node is reached.

This kind of random walks with no self avoidance result in a distribution of paths each with number of steps varying between HAPL, which is very seldom, to thousands. This is evidently not suitable for effective information transfer mechanism within proteins. It is also known that the number of nodes crossed during a random walk is proportional to the number of neighbors [7]. We have verified that this limit is reached in our numerical results. In Figure 4.1 we show how this procedure runs for a selected node.

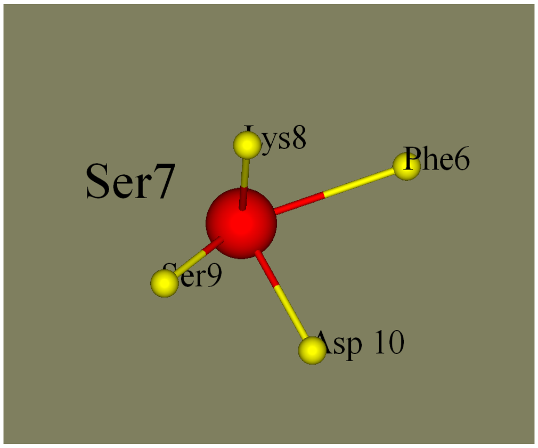


Figure 4.1: An example of random walk by simple projection next step selection. Each link has equal selection probability of 0.25

## 4.2 Simple Projections

A more complicated approach is to set a destination node as an anchor and orient each step according to the relative positioning of this node with the current location. To achieve this, we used the the cosine of the angle formed between two vectors $\vec{r_{ij}}$ and $\vec{r_{ik}}$. These vectors are reconstructed at every step. The criteria for simple projection also starts with random number generation, but probabilities of selecting the next steps are derived from Boltzmann distribution ;

$$P(j) = \frac{e^{-E_j/k_B T}}{\sum_{i=1,k} e^{-E_i/k_B T}} \tag{5}$$

where the summation runs over the neighbors of $n_j$, $k_B$ is the Boltzmann constant, $T$ is a characteristic temperature parameter that affects the efficiency, $E_j$ is the energy assigned to the link between $n_i$ and its neighbor $n_j$ and $P(j)$ is the probability assigned to that link. We consider $E_j$ as the cosine of angle between the vector $r_{ij}$ and the vector $r_{ik}$ and assign $k_B T$ as unity. Thus, the nodes closer in space to the destination node have higher probabilities owing to the exponential dependence of the Boltzmann distribution. In Figure 4.2, a path from the protein 1RPX with starting and destination residues ($n_i = 1$) and ($n_k = 22$), respectively, are shown. The red line represents the path between nodes 1 and 22, yellow lines represent the constructed direction vectors connecting destination nodes and current steps in path. The different colored nodes connected to each red node are the nearest neighbors.

This kind of approach to calculate paths by connecting distant nodes significantly reduces the step size. However, the global structure sometimes prohibits reaching the destination node, especially for closely located node pairs and for those nodes located at the surface of the proteins. In Figure 4.2 this phenomenon is also shown. The path between $Ser1 \leftrightarrow Phe22$ which is $Ser1 \leftrightarrow Arg2 \leftrightarrow Pro70 \leftrightarrow Leu71 \leftrightarrow Leu63 \leftrightarrow Val59 \leftrightarrow Ile55$ does not converge to destination node Phe22, the latter resides on an adjacent loop to the alpha helix secondary structure on which Val59 and Ile55 appear. This structural restriction forms a trap where the consecutive moves are stuck in a region and the destination node is never reached. In our calculations, we omit such trapped pathways which constitute 42 % of all paths.

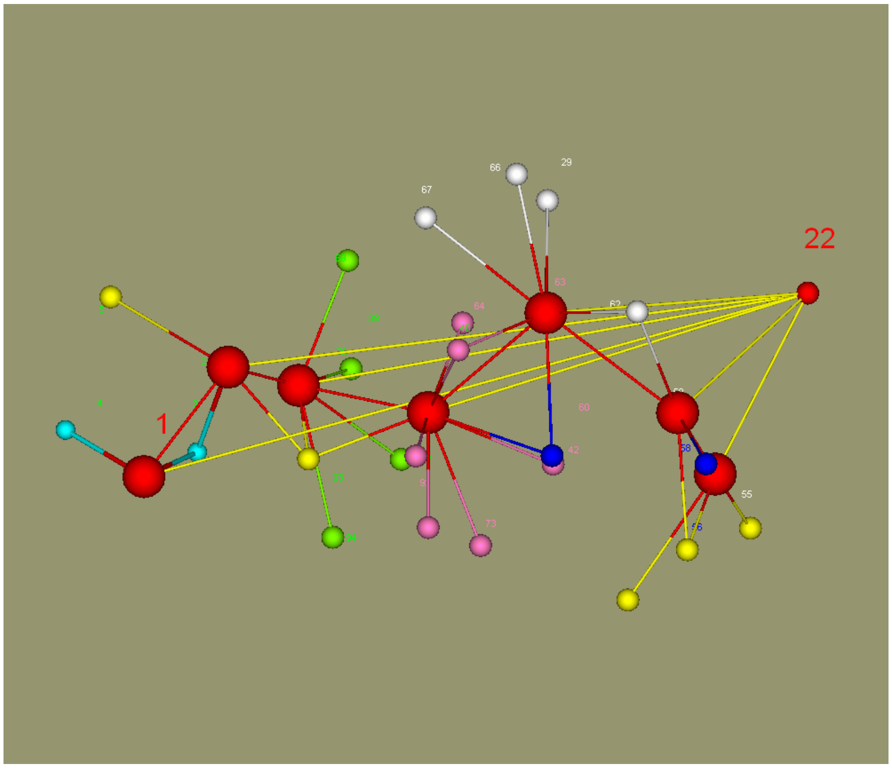


Figure 4.2: Path in 1RPX starting residue $n_i$=1 , ending residue $n_k$=55 with destination residue 22 generated by simple projection

## 4.3 Interaction potentials (TD-MJ)

We also studied residue-residue interaction potentials to achieve global destinations from local properties only. A similar procedure to simple projections in section 4.2 is followed, but here we use interaction potentials of TD or MJ as explained in section 2.1 as the energy assigned to each edge [1, 2]. Resulting paths of this approach are also shorter than average random walks in terms of the number of steps but, since no directionality is taken into account, path size is strictly determined by residue types at the spatial position. Traps within the paths are also observed in 64 . An important note about potentials is, there is no observable difference in the overall results obtained by TD or MJ thus, we present TD potentials in all the calculations, following our previous work.

| Node | Interaction Potential ($k_BT$) | Probability |
|---|---|---|
| Phe6 | 0.14 | 0.2208 |
| Lys8 | 0.18 | 0.2121 |
| Ser9 | -0.13 | 0.2892 |
| Asp10 | -0.09 | 0.2779 |

Table 4.1: An example of random walk next step selection by simple projection

## 4.4 Simple Projection + Interaction potentials

We lastly studied paths derived by both simple projection and interaction potentials. In this case the probability assigned to a link is formulated as;

$$E_i = A * (E_{Projection}) + B * (E_{Potential}) \tag{6}$$

where $A$ and $B$, varying between zero and five ($0 \leq A, B \leq 5$), are the constants to be optimized. In this formulation $A*(E_{Projection})_i$ represents the global knowledge whereas $B*(E_{Potential})_i$ represents the local knowledge and $A$ and $B$ are amplification factors to be determined to mimic HAPL, WAPL and SAPL. We used $(E_{Projection})_i$ as the cosine of the angle between the vector $r_{ij}$ and vector $r_{ik}$ and TD residue-residue interaction potentials as $(E_{Potential})_i$. as both described in Section 4.2 and 4.3. Note that we get similar results with MJ potentials.

We run the algorithm for a data set of 76 proteins having folds $\alpha$ , $\beta$, $\alpha/\beta$ and $\alpha + \beta$. The proteins used and the corresponding fold types with $C$ and $L$ values are listed in Appendix D. For each protein, $L$ matrices (equation 4) are calculated for different values of $A$ both in the range of 0.0 and 1.0 with increment 0.05. These $L$ matrices are then compared to those $L$ matrices calculated by HAPL, WAPL and SAPL.

In order to determine the extent of similarity between the actual and constructed $L$ matrices, we compare their eigenvalue structures. The eigenvalue distribution of a sample protein ($1RPX$) is given in Figure 4.3 for HAPL, WAPL and SAPL. We find that the eigenvalue distributions of HAPL and WAPL are very similar, each having a distantly located eigenvalue at 1173.2 and 1135.7 and the next two at 9.00 and 8.00 for HAPL and 11.6 and 11.3 for WAPL. For SAPL, the largest eigenvalue is more separate, appearing at 1355.1, and the following eigenvalues are at 59.2 and 31.0. Thus, the eigenvalue structure of SAPL is quite distinct from that of HAPL, whereas that of WAPL is very similar to the latter.

For each distribution, the three highest valued eigenvalues are selected and the overlap of their eigenvectors with those of the $L$ matrices, constructed for each $A$ and $B$ combination, are calculated. We set different similarity scores for each APL as 0.8 for HAPL, 0.7 for WAPL and 0.3 for SAPL and distribution of $A$ and $B$ combinations having dot product higher values then these previously set similarity scores are calculated. For cases where dot products does not exceed predefined set values, top three dot products and their $A$ and $B$ combination are taken into consideration. Results are presented in Figure 4.4.a, 4.4.b, 4.4.c for HAPL, WAPL and SAPL respectively. Distribution of $A$ and $B$ indicates that when $A = 3.75$ and $B = 4.75$ the proposed algorithm best fits SAPL when path lengths are taken into consideration. Values of $A$ and $B$ fitting HAPL and WAPL calculated to be $A = 5.0$ and $B = 1.0 \pm 0.5$ for HAPL and $A = 4.5$ and $B = 1.0 \pm 0.5$ for WAPL.

The difference between HAPL, WAPL and SAPL signifies the effect of residue-residue interaction potentials on determining path length, hence the paths connecting two distantly separated residues. Although WAPL uses residue-residue interaction potentials, its characteristics is almost same as HAPL in terms of the number of steps. Since both methodologies show good agreement with simple projection + interaction potentials approach for $A = 5.0$ and $B = 1.0$ values, these two methodologies show almost fully projection characteristics, thats is paths are forced to follow certain direction between two nodes.

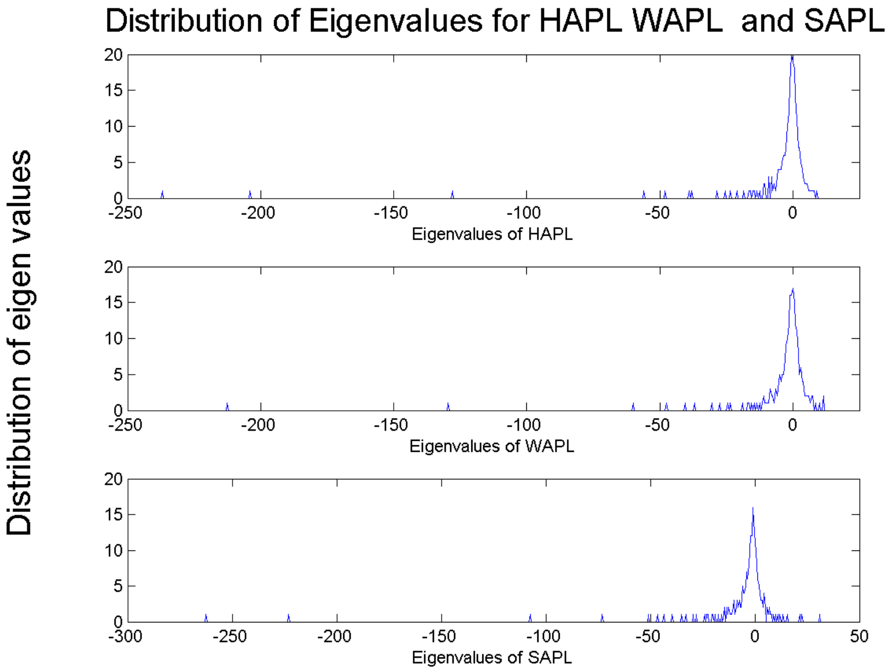
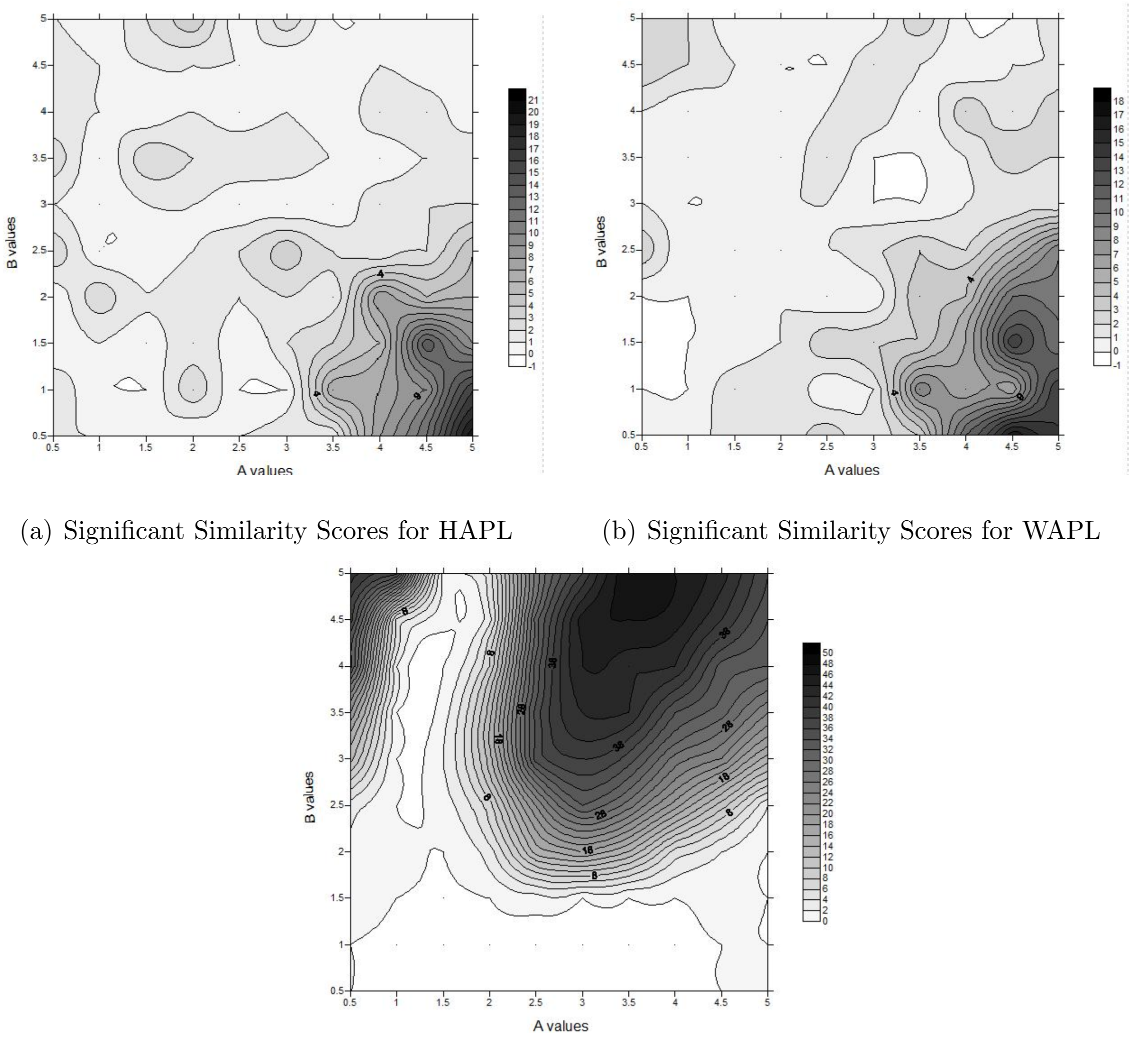


Figure 4.3: Eigenvalue distribution of L matrices of HAPL, WAPL and SAPL

On the other hand the observed $A$ and $B$ values for SAPL data fitting signifies that even directionality is taken into account by higher percentage there exists an effect of residue-residue interaction potentials on paths and this effect appears especially if residue-residue interaction potentials are enhanced with relatively high $B$ values.



(a) Significant Similarity Scores for HAPL

(b) Significant Similarity Scores for WAPL

(c) Significant Similarity Scores for SAPL

Figure 4.4: Frequency Distribution of Significant Similarity Scores for Various $A$ and $B$ values.

# 5 Conclusions and Future Work

In this study we have systematically investigated paths along residue networks to uncover structurally and functionally important residues. The paths that are constructed according to three different criteria: (i) HAPL connect a pair of nodes in the shortest number of edges; (ii)WAPL are the ones that minimize the total cost of navigating between a pair of nodes; (iii) SAPL minimize the most costly single link occurring on all paths connecting the pair of nodes. For (ii) and (iii) the cost of navigation between directly connected nodes is taken as the residue interaction potentials.

We find that weak paths converge to the homogeneous paths due to the special distribution of contact energies that have evolved along the protein structure. There is a hierarchical distribution of link weights in the protein whereby the mainly hydrophobic, low cost contacts are located in the core of the structure and high energy contacts progressively occur towards the surface.

By studying two families of proteins, TIM barrels for single chain systems and Ca-binding proteins for interacting pairs of chains, we find that key locations in the structure may be located by scrutinizing the strong paths. For single chains, bottleneck edges determine evolutionary important hot regions that are located on paths connecting the active site to distantly located secondary structural units. For interacting proteins, interface edges that are most frequently used in strong paths while navigating between chains are found to be affected by the identity of residues that are far from the interface.

To determine the factors that derive the usage of strong versus weak paths in the structure, we have used a biased random walk scheme where the probability of edge selection is based on local and/or global knowledge of structure. We find that a combined local contact/global destination approach may be optimized to generate the strong paths, while solely local knowledge is enough to mimic paths.

Allosteric communication in proteins necessitates gradients for information flows in the structure. It is then plausible to assume that strong paths appear due to the competition between local and global knowledge at a given node in the structure. Homogeneous or weak paths are not likely to point to evolutionary/functionally important residues along the structure, since they are only the product of the general network structure.

In our study with protein we continuously seek characteristic features of paths between two nodes. These features in the case of a single chain are, start and destined nodes, bottleneck edge or edges and paths passing through these features with minimum number of steps. We also seek interface edges for protein complexes. We used PDB codes of proteins to construct residue network which require full knowledge of the protein structure. We can apply same other algorithms to achieve same goals. This algorithm may include not only nearest neighbors but also second or even third nearest neighbors and their cost of navigation to archive same pathways as SAPL.

As future work, it is of interest to find alternative "smarter" approaches that combine local and global features of the network to efficiently locate distantly located destinations on the network. One candidate is the Battleship algorithm where the aim is to locate occupied locations by ships of different classes on a confined surface with the minimum number of tries. These use the global features of the system. The maze algorithms, on the other hand, are based on traversing graphs with obstacles, and operate on local information. In the protein case the problem turns into finding the same features with less computational time or even with less knowledge about global structure. Maze and battleship algorithms combined will not only try to find these structural features, but will also target finding paths connecting these features.

# References

[1] P.D. Thomas and K.A. Dill. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci.*, 93:11628–11633, 1996.

[2] S. Miyazawa and R.L. Jernigan. Residue-residue potentials with a favorable contact pair term and an favorable high packing density term, for simulation and threading. *J.Mol.Biol*, 256:623–644, 1996.

[3] C. Baysal and A.R. Atilgan. Relaxation kinetics and glassiness of proteins: the case of bovine pancreatic trypsin inhibitor. *Biophysical Journal*, 83:699–705, 2002.

[4] A.R. Atilgan, P. Akan, and C. Baysal. Small-world communication of residues and significance for protein dynamics. *Biophysical Journal*, 86:85–91, 2004.

[5] A.R. Atilgan, D. Turgut, and C. Atilgan. Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication. *Biophysical Journal*, 92:3052–3062, 2007.

[6] K.V. Brinda and S. Vishveshwara. A network representation of protein structures: Implications for protein stability. *Biophysical Journal*, 89:4159–4170, 2005.

[7] T. Can, O. Camoglu, and A.K. Singh. Analysis of protein-protein interaction networks using random walks. *Proceedings BIOKDD*, 2005.

[8] W.M. Gelbart and A. Ben-Shaul. The new science of complex fluids. *J. Phys. Chem.*, 100:13169–13189, 1996.

[9] R. Piazza. Interactions and phase transitions in protein solutions. *Curr. Opin. Colloid Interface Sci.*, 5:38–43, 2000.

[10] G. M. Whitesides and R. F. Ismagilov. Complexity in chemistry. *Science*, 284:89–92, 1999.

[11] W. A. Baase, N.C. Gassner, X.-J.Zhang, R.Kuroki, L.H.Weaver, D.E. Tronrud, and B.W. Matthews. *How much sequence variation can the functions of biological molecules tolerate? In Simplicity and Complexity in Proteins and Nucleic Acid.* Dahlem University Press, Berlin, Germany, 1999.

[12] A. M. Tsai, D. A. Neumann, and L. N. Bell. Molecular dynamics of solid-state lysozyme as effected by glycerol and water: a neutron scattering study. *Biophys. J.*, 79:2728–2732, 2000.

[13] J. Liang and K. A. Dill. Are proteins well packed? *Biophysical Journal*, 81:751–766, 2001.

[14] G. Raghunathan and R. Jernigan. Ideal architecture of residue packing and its observation in protein structures. *Protein Sci*, 6:2072–2083, 1997.

[15] A. Soyer, J. Chomilier, J.-P. Mornon, R. Jullien, and J.-F. Sadoc. Voronoi tessellation reveals the condensed matter character of folded proteins. *Phys. Rev. Lett.*, 85:3532–3535, 2000.

[16] M. E. J. Newman. Models of the small world. *J. Stat. Phys.*, 101:819–841, 2000.

[17] L. A. Adamic and B. A. Huberman. Growth dynamics of the world-wide web. *Nature*, 401:131–131, 1999.

[18] A. Vazquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the internet. *Phys. Rev. E.*, 65:066130–066130, 2002.

[19] A.L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, , and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A.*, 311:590–614, 2002.

[20] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[21] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.

[22] C. Baysal and A. R. Atilgan. Coordination topology and stability for the native and binding conformers of chymotrypsin inhibitor 2. *Proteins.*, 45:62–70, 2001.

[23] H. Jeong, S. P. Mason, A.L. Barabasi, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature.*, 411:41–42, 2001.

[24] I. Bahar, A.R. Atilgan, M.C. Demirel, and B. Erman. Vibrational dynamics of folded proteins: significance of slow and fast modes in relation to function and stability. *Phys. Rev. Lett.*, 80:2733–2736, 1998a.

[25] V.A. Higman and L.H. Greene. Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology. *Physica A.*, 368:595–606, 2006.

[26] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovsky. Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, 344:1135–1146, 2004.

[27] A. del Sol, H. Fujihashi, D. Amoros, and R. Nussinov. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Prot. Sci.*, 15:2120–2128, 2006.

[28] T. J. Taylor and I. I. Vaisman. Graph theoretic properties of networks formed by the delaunay tessellation of protein structures. *Phys. Rev. E.*, 73:041925, 2006.

[29] N. Popovych, S. Sun, R.H. Ebright, and C. G. Kolodimos. Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.*, 13:831–838, 2006.

[30] S. Miyazawa and R.L.Jernigan. An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins*, 36:357–369, 1999.

[31] J. M. Kriegl, K. Nienhaus, P. Deng, J. Fuchs, and G. U. Nienhaus. Ligand dynamics in a protein internal cavity. *Proc. Natl. Acad. Sci.*, 100:7069–7074, 2006.

[32] A. Ansari, J. Berendzen, S. F. Bowne, H. Frauenfelder, I. E. T. Iben, T. B. Sauke, E. Shyamsunder, and R. D. Young. Protein states and protein quakes. *Proc. Natl. Acad. Sci.*, 82:5000–5004, 2006.

[33] M. R. Betancourt and D. Thirumalai. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.*, 1999.

[34] H.M. Berman, J. Westbrook, Z.Feng, G. Gilliland, T. N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[35] E.W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[36] H. Cao, Y. Ihm, C.-Z.Wang, J.R.Morris, M.Su, and D. Dobbs. Three-dimensional threading approach to protein structure recognation. *Polymer*, 45:687–697, 2004.

[37] H. Li, C. Tang, and N.S.Wingreen. Designability of protein structures: Lattice-model study using the miyazawa-jernigan matrix. *Journal of Molecular Biology*, 256:623–644, 1996.

[38] J.G. Esteve and F. Falceto. A general clustering approach with applications to the miyazawa-jernigan potentials for amino acids. *Proteins*, 55:999–1004, 2004.

[39] G. K.Farber. An a/b-barrel full of evolutionary trouble. *Current Opinion in Structural Biology*, 3:104–412, 1993.

[40] N.Nagano, C.A. Orengo, and J.M. Thornton. One fold with many functions: The evolutionary relationship between tim barrel families based on their sequence, structure and functions. *Journal of Molecular Biology*, 321:741–765, 2002.

[41] A.E. Todd, C.A. Orengo, and J.M. Thornton. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, 307:1113–1143, 2001.

[42] M.C.Vega, E. Lorentzen, A. Linden, and M. Wilmanns. Evolutionary markers in the $(\beta/\alpha)_8$ barrel fold. *Current Opinion in Chemical Biology*, 7:694–701, 2003.

[43] M. Shatsky, , R. Nussinov, and H.J. Wolfson. A method for simultaneous alignment of multiple protein structures. *Proteins: Structure, Function, and Bioinformatics.*, 56(1):143–56, 2004.

[44] R.K. Wierenga. The tim-barrel fold: a versatile framework for efficient enzymes. *FEBS Letters*, 492:193–198, 2001.

[45] H. Wolfson, C.J. Tsai, O. Keskin, and R. Nussinov. A new, structurally nonredundant, diverse data set of proteinprotein interfaces and its implications. *Protein Sci.*, 13(4):1043–1055, 2004.

[46] U. Ogmen, O. Keskin, S. Aytuna, R.Nussinov, and A.Gursoy. Prism: Protein interactions by structural matching. *Nucleic Acids Research*, 33:331–336, 2005.

[47] C.h Gille and C. Frommel. Strap: editor for structural alignments of proteins. *Bioinformatics*, 17:377–378, 2001.

# A The residue interaction potentials used in this work

| | Cys | Met | Phe | Ile | Leu | Val | Trp | Tyr | Ala | Gly | Thr | Ser | Gln | Asn | Glu | Asp | His | Arg | Lys | Pro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cys | -1.79 | -1.23 | -0.98 | -0.48 | -0.69 | -0.94 | -0.3 | -0.96 | -0.3 | -0.42 | -0.38 | -0.2 | -0.49 | -0.32 | 0.04 | 0.55 | -0.82 | -0.4 | 0 | 0.07 |
| Met | -1.23 | 0.36 | -1.03 | -0.41 | -0.31 | -0.94 | -0.07 | -1.1 | 0.05 | 0 | 0.06 | -0.47 | -0.54 | 0.31 | 0.02 | 1.07 | -0.35 | -0.43 | 0.55 | -0.25 |
| Phe | -0.98 | -1.03 | -0.61 | -0.66 | -1.02 | -0.78 | -0.89 | -0.82 | -0.05 | 0.21 | -0.19 | 0.14 | 0.1 | -0.02 | 0.19 | 0.2 | -0.75 | -0.22 | -0.17 | -0.43 |
| Ile | -0.4 | -0.41 | -0.66 | -0.71 | -1.04 | -0.98 | -0.89 | -0.87 | -0.64 | 0.4 | -0.29 | -0.13 | -0.39 | 0.39 | -0.2 | 0.04 | -0.52 | -0.08 | -0.26 | 0.25 |
| Leu | -0.69 | -0.31 | -1.02 | -1.04 | -1.14 | -1.03 | -0.97 | -0.6 | -0.57 | -0.08 | -0.39 | -0.07 | -0.13 | -0.1 | -0.05 | 0.5 | -0.36 | -0.1 | 0.1 | 0.09 |
| Val | -0.94 | -0.94 | -0.78 | -0.98 | -1.03 | -1.15 | -0.6 | -0.7 | -0.6 | -0.2 | 0.06 | -0.31 | -0.09 | -0.24 | -0.02 | 0.25 | -0.35 | -0.48 | -0.08 | -0.08 |
| Trp | -0.3 | -0.07 | -0.89 | -0.89 | -0.97 | -0.6 | 0.02 | -0.99 | -0.08 | -0.14 | 0.07 | -0.2 | 0.4 | -0.68 | 0.32 | 0.24 | -0.41 | -0.78 | -0.3 | -0.44 |
| Tyr | -0.9 | -1.1 | -0.82 | -0.87 | -0.6 | -0.7 | -0.99 | 0.35 | -0.37 | -0.32 | -0.23 | 0.25 | -0.39 | -0.74 | 0.22 | 0.11 | -0.67 | 0.21 | -0.2 | -0.45 |
| Ala | -0.3 | 0.05 | -0.05 | -0.64 | -0.57 | -0.6 | -0.08 | -0.37 | -0.08 | -0.09 | -0.22 | -0.01 | -0.11 | -0.14 | 0.03 | 0.1 | -0.15 | 0.07 | 0 | 0.41 |
| Gly | -0.42 | 0 | 0.21 | 0.4 | -0.08 | -0.2 | -0.14 | -0.32 | -0.09 | 0.04 | 0.13 | -0.04 | 0.12 | -0.18 | 0.4 | -0.06 | 0 | -0.15 | 0.1 | 0.4 |
| Thr | -0.38 | 0.06 | -0.19 | -0.29 | -0.39 | 0.06 | 0.07 | -0.23 | -0.22 | 0.13 | 0.26 | 0.05 | -0.17 | -0.27 | 0.15 | -0.03 | -0.27 | -0.17 | 0.09 | 0.36 |
| Ser | -0.2 | -0.47 | 0.14 | -0.13 | -0.07 | -0.31 | -0.2 | 0.25 | -0.01 | -0.04 | 0.05 | -0.13 | 0.4 | 0.37 | 0.3 | -0.09 | -0.59 | 0.61 | 0.18 | 0.44 |
| Gln | -0.4 | -0.54 | 0.1 | -0.39 | -0.13 | -0.09 | 0.4 | -0.39 | -0.11 | 0.12 | -0.17 | 0.4 | -0.08 | -0.05 | 0.62 | 0.46 | 0.05 | 0.62 | 0.04 | -0.21 |
| Asn | -0.32 | 0.31 | -0.02 | 0.39 | -0.1 | -0.24 | -0.68 | -0.74 | -0.14 | -0.18 | -0.27 | 0.37 | -0.05 | -0.86 | -0.25 | -0.12 | 0.06 | 0.04 | 0.18 | 0.11 |
| Glu | 0.04 | 0.02 | 0.19 | -0.2 | -0.05 | -0.02 | 0.32 | 0.22 | 0.03 | 0.4 | 0.15 | 0.3 | 0.62 | -0.25 | 0.21 | 0.68 | -0.53 | -0.26 | -0.09 | 0.33 |
| Asp | 0.5 | 1.07 | 0.2 | 0.04 | 0.5 | 0.25 | 0.24 | 0.11 | 0.1 | -0.06 | -0.03 | -0.09 | 0.46 | -0.12 | 0.68 | 0.6 | -0.06 | -0.15 | -0.09 | 0.84 |
| Arg | -0.4 | -0.43 | -0.22 | -0.08 | -0.1 | -0.48 | -0.78 | 0.21 | 0.07 | -0.15 | -0.17 | 0.61 | 0.62 | 0.04 | -0.26 | -0.15 | -0.01 | 0.23 | 0.3 | -0.02 |
| Lys | 0 | 0.55 | -0.17 | -0.26 | 0.1 | -0.08 | -0.3 | -0.2 | 0 | 0.1 | 0.09 | 0.18 | 0.04 | 0.18 | -0.09 | -0.09 | 0.14 | 0.3 | 1.45 | 0.51 |
| Pro | 0.07 | -0.25 | -0.43 | 0.25 | 0.09 | -0.08 | -0.44 | -0.45 | 0.41 | 0.4 | 0.36 | 0.44 | -0.21 | 0.11 | 0.33 | 0.84 | -0.22 | -0.02 | 0.51 | 0.28 |

Table A.1: TD residue-residue interaction potential [1]

| | Cys | Met | Phe | Ile | Leu | Val | Trp | Tyr | Ala | Gly | Thr | Ser | Gln | Asn | Glu | Asp | His | Arg | Lys | Pro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cys | -5.44 | -4.99 | -5.8 | -5.5 | -5.83 | -4.96 | -4.95 | -4.16 | -3.57 | -3.16 | -3.11 | -2.86 | -2.85 | -2.59 | -2.27 | -2.41 | -3.6 | -2.57 | -1.95 | -3.07 |
| Met | -4.99 | -5.46 | -6.56 | -6.02 | -6.41 | -5.32 | -5.55 | -4.91 | -3.94 | -3.39 | -3.51 | -3.03 | -3.3 | -2.95 | -2.89 | -2.57 | -3.98 | -3.12 | -2.48 | -3.45 |
| Ph | -5.8 | -6.56 | -7.26 | -6.84 | -7.28 | -6.29 | -6.16 | -5.66 | -4.81 | -4.13 | -4.28 | -4.02 | -4.1 | -3.75 | -3.56 | -3.48 | -4.77 | -3.98 | -3.36 | -4.25 |
| Ile | -5.5 | -6.02 | -6.84 | -6.54 | -7.04 | -6.05 | -5.78 | -5.25 | -4.58 | -3.78 | -4.03 | -3.52 | -3.67 | -3.24 | -3.27 | -3.17 | -4.14 | -3.63 | -3.01 | -3.76 |
| Leu | -5.83 | -6.41 | -7.28 | -7.04 | -7.37 | -6.48 | -6.14 | -5.67 | -4.91 | -4.16 | -4.34 | -3.92 | -4.04 | -3.74 | -3.59 | -3.4 | -4.54 | -4.03 | -3.37 | -4.2 |
| Val | -4.96 | -5.32 | -6.29 | -6.05 | -6.48 | -5.52 | -5.18 | -4.62 | -4.04 | -3.38 | -3.46 | -3.05 | -3.07 | -2.83 | -2.67 | -2.48 | -3.58 | -3.07 | -2.49 | -3.32 |
| Trp | -4.95 | -5.55 | -6.16 | -5.78 | -6.14 | -5.18 | -5.06 | -4.66 | -3.82 | -3.42 | -3.22 | -2.99 | -3.11 | -3.07 | -2.99 | -2.84 | -3.98 | -3.41 | -2.69 | -3.73 |
| Tyr | -4.16 | -4.91 | -5.66 | -5.25 | -5.67 | -4.62 | -4.66 | -4.17 | -3.36 | -3.01 | -3.01 | -2.78 | -2.97 | -2.76 | -2.79 | -2.76 | -3.52 | -3.16 | -2.6 | -3.19 |
| Ala | -3.57 | -3.94 | -4.81 | -4.58 | -4.91 | -4.04 | -3.82 | -3.36 | -2.72 | -2.31 | -2.32 | -2.01 | -1.89 | -1.84 | -1.51 | -1.7 | -2.41 | -1.83 | -1.31 | -2.03 |
| Gly | -3.16 | -3.39 | -4.13 | -3.78 | -4.16 | -3.38 | -3.42 | -3.01 | -2.31 | -2.24 | -2.08 | -1.82 | -1.66 | -1.74 | -1.22 | -1.59 | -2.15 | -1.72 | -1.15 | -1.87 |
| Thr | -3.11 | -3.51 | -4.28 | -4.03 | -4.34 | -3.46 | -3.22 | -3.01 | -2.32 | -2.08 | -2.12 | -1.96 | -1.9 | -1.88 | -1.74 | -1.8 | -2.42 | -1.9 | -1.31 | -1.9 |
| Ser | -2.86 | -3.03 | -4.02 | -3.52 | -3.92 | -3.05 | -2.99 | -2.78 | -2.01 | -1.82 | -1.96 | -1.67 | -1.49 | -1.58 | -1.48 | -1.63 | -2.11 | -1.62 | -1.05 | -1.57 |
| Gln | -2.85 | -3.3 | -4.1 | -3.67 | -4.04 | -3.07 | -3.11 | -2.97 | -1.89 | -1.66 | -1.9 | -1.49 | -1.54 | -1.71 | -1.42 | -1.46 | -1.98 | -1.8 | -1.29 | -1.73 |
| Asn | -2.59 | -2.95 | -3.75 | -3.24 | -3.74 | -2.83 | -3.07 | -2.76 | -1.84 | -1.74 | -1.88 | -1.58 | -1.71 | -1.68 | -1.51 | -1.68 | -2.08 | -1.64 | -1.21 | -1.53 |
| Glu | -2.27 | -2.89 | -3.56 | -3.27 | -3.59 | -2.67 | -2.99 | -2.79 | -1.51 | -1.22 | -1.74 | -1.48 | -1.42 | -1.51 | -0.91 | -1.02 | -2.15 | -2.27 | -1.8 | -1.26 |
| Asp | -2.41 | -2.57 | -3.48 | -3.17 | -3.4 | -2.48 | -2.84 | -2.76 | -1.7 | -1.59 | -1.8 | -1.63 | -1.46 | -1.68 | -1.02 | -1.21 | -2.32 | -2.29 | -1.68 | -1.33 |
| His | -3.6 | -3.98 | -4.77 | -4.14 | -4.54 | -3.58 | -3.98 | -3.52 | -2.41 | -2.15 | -2.42 | -2.11 | -1.98 | -2.08 | -2.15 | -2.32 | -3.05 | -2.16 | -1.35 | -2.25 |
| Arg | -2.57 | -3.12 | -3.98 | -3.63 | -4.03 | -3.07 | -3.41 | -3.16 | -1.83 | -1.72 | -1.9 | -1.62 | -1.8 | -1.64 | -2.27 | -2.29 | -2.16 | -1.55 | -0.59 | -1.7 |
| Lys | -1.95 | -2.48 | -3.36 | -3.01 | -3.37 | -2.49 | -2.69 | -2.6 | -1.31 | -1.15 | -1.31 | -1.05 | -1.29 | -1.21 | -1.8 | -1.68 | -1.35 | -0.59 | -0.12 | -0.97 |
| Pro | -3.07 | -3.45 | -4.25 | -3.76 | -4.2 | -3.32 | -3.73 | -3.19 | -2.03 | -1.87 | -1.9 | -1.57 | -1.73 | 1.53 | -1.26 | -1.33 | -2.25 | -1.7 | -0.97 | -1.75 |

Table A.2: MJ residue-residue interaction potential [2]

# B Structural alignment between selected TIM Barrel proteins and other members of the TIM barrel family.

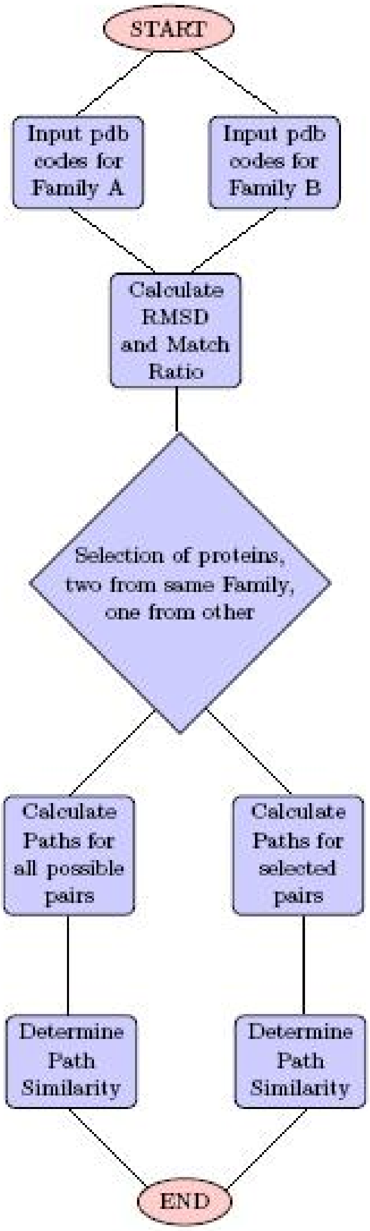

START
Input pdb codes for Family A
Input pdb codes for Family B
Calculate RMSD and Match Ratio
Selection of proteins, two from same Family, one from other
Calculate Paths for all possible pairs
Calculate Paths for selected pairs
Determine Path Similarity
Determine Path Similarity
END

Figure B.1: Flowdiagram of procedure followed for TIM barrels

| Super Family | PDB Code | Aligment Size | RMSD | $\#ofatoms$ | Match Ratio |
|---|---|---|---|---|---|
| Ribulose phosphate binding TIM barrels | **1PII** | ——— | ——— | 452 | |
| | 1igs | 222 | 1.22 | 247 | 0.90 |
| | 1nsj | 179 | 1.26 | 205 | 0.87 |
| | 1dv7 | 149 | 1.68 | 212 | 0.70 |
| | 1dqx | 144 | 1.92 | 267 | 0.54 |
| | 1dbt | 142 | 1.95 | 237 | 0.60 |
| | 1ubs | 147 | 1.72 | 257 | 0.57 |
| | 1rpx | 161 | 1.71 | 230 | 0.70 |
| TIM barrel glycosyl hydrolases | 1cyg | 132 | 1.96 | 257 | 0.51 |
| | 1ciu | 133 | 1.97 | 452 | 0.29 |
| | 1cwy | 92 | 1.76 | 452 | 0.20 |
| | 1vjs | 131 | 2.00 | 452 | 0.29 |
| | 1aqm | 119 | 1.92 | 452 | 0.26 |
| | 1bag | 95 | 1.84 | 425 | 0.22 |
| | 1dhkA | 111 | 2.05 | 452 | 0.25 |
| | 1smd | 130 | 1.93 | 452 | 0.29 |
| | 1jae | 120 | 1.80 | 452 | 0.27 |
| | 2aaa | 113 | 1.93 | 452 | 0.25 |
| | 7taa | 105 | 1.89 | 452 | 0.23 |
| | 1ava | 128 | 1.95 | 452 | 0.28 |
| | 1uok | 119 | 1.95 | 452 | 0.26 |
| | 2amg | 127 | 1.91 | 415 | 0.31 |
| | 1bf2 | 114 | 2.13 | 452 | 0.25 |
| | 1sma | 131 | 1.95 | 452 | 0.29 |
| | 1bvz | 140 | 1.95 | 452 | 0.31 |
| | 1byb | 141 | 1.97 | 452 | 0.31 |
| | 1b1y | 142 | 1.91 | 452 | 0.31 |
| | 1b9z | 111 | 1.98 | 452 | 0.25 |
| | 1qba | 140 | 1.88 | 452 | 0.31 |
| | 1cbg | 111 | 2.00 | 452 | 0.25 |
| | 1bgg | 140 | 1.97 | 452 | 0.31 |
| | 1gow | 109 | 2.09 | 452 | 0.24 |
| | 1qvb | 116 | 1.95 | 452 | 0.26 |
| | 1pbg | 52 | 2.17 | 452 | 0.12 |
| | 2myr | 106 | 2.10 | 452 | 0.23 |
| | 1ceo | 111 | 1.78 | 332 | 0.33 |
| | 1edg | 105 | 2.02 | 380 | 0.28 |
| | 1eceA | 118 | 2.01 | 358 | 0.33 |
| | 7a3hA | 116 | 2.04 | 300 | 0.39 |
| | 1egzA | 125 | 2.10 | 291 | 0.43 |
| | 1cz1A | 96 | 2.02 | 394 | 0.24 |
| | 1ex1A | 92 | 1.89 | 452 | 0.20 |
| | 1bqcA | 119 | 2.06 | 302 | 0.39 |
| | 1xyzA | 121 | 2.05 | 320 | 0.38 |
| | 1clxA | 123 | 1.92 | 345 | 0.36 |
| | 1bg4 | 119 | 1.98 | 302 | 0.39 |
| | 1gok | 130 | 2.08 | 301 | 0.43 |
| | 1exp | 127 | 2.04 | 312 | 0.41 |
| | 1bglA | 120 | 2.02 | 452 | 0.27 |
| | 1bhgA | 126 | 2.10 | 452 | 0.28 |
| | 1ghsA | 105 | 2.08 | 306 | 0.34 |
| | 1aq0A | 110 | 2.09 | 306 | 0.36 |

Table B.1: Structural Alignment of 1PII inter and intra superfamilies

| Super Family | PDB Code | Aligment Size | RMSD | $\#ofatoms$ | Match Ratio |
|---|---|---|---|---|---|
| Ribulose phosphate binding TIM barrels | **1RPX** | ——— | ——— | 230 | |
| | 1igs | 163 | 1.58 | 230 | 0.71 |
| | 1nsj | 148 | 1.87 | 205 | 0.72 |
| | 1dv7 | 149 | 1.79 | 212 | 0.70 |
| | 1dqx | 155 | 1.78 | 230 | 0.67 |
| | 1dbt | 153 | 1.95 | 230 | 0.67 |
| | 1ubs | 142 | 1.92 | 230 | 0.62 |
| | 1pii | 161 | 1.71 | 230 | 0.70 |
| TIM barrel glycosyl hydrolases | 1cyg | 114 | 2.10 | 230 | 0.50 |
| | 1ciu | 108 | 2.18 | 230 | 0.47 |
| | 1cwy | 124 | 1.79 | 230 | 0.54 |
| | 1vjs | 109 | 1.91 | 230 | 0.47 |
| | 1aqm | 112 | 1.96 | 230 | 0.49 |
| | 1bag | 112 | 1.75 | 230 | 0.49 |
| | 1dhkA | 133 | 2.10 | 230 | 0.58 |
| | 1smd | 110 | 1.99 | 230 | 0.48 |
| | 1jae | 109 | 2.06 | 230 | 0.47 |
| | 2aaa | 126 | 2.06 | 230 | 0.55 |
| | 7taa | 123 | 2.06 | 230 | 0.53 |
| | 1ava | 130 | 2.08 | 230 | 0.57 |
| | 1uok | 103 | 1.90 | 230 | 0.45 |
| | 2amg | 108 | 2.13 | 230 | 0.47 |
| | 1bf2 | 115 | 2.03 | 230 | 0.50 |
| | 1sma | 142 | 2.07 | 230 | 0.62 |
| | 1bvz | 143 | 2.17 | 230 | 0.62 |
| | 1byb | 117 | 2.03 | 230 | 0.51 |
| | 1b1y | 112 | 1.99 | 230 | 0.49 |
| | 1b9z | 108 | 1.84 | 230 | 0.47 |
| | 1qba | 130 | 2.01 | 230 | 0.57 |
| | 1cbg | 117 | 1.98 | 230 | 0.51 |
| | 1bgg | 116 | 1.96 | 230 | 0.50 |
| | 1gow | 114 | 1.97 | 230 | 0.50 |
| | 1qvb | 124 | 2.02 | 230 | 0.54 |
| | 1pbg | 119 | 1.95 | 230 | 0.52 |
| | 2myr | 118 | 2.02 | 230 | 0.51 |
| | 1ceo | 138 | 1.79 | 230 | 0.60 |
| | 1edg | 108 | 1.86 | 230 | 0.47 |
| | 1eceA | 107 | 1.93 | 230 | 0.47 |
| | 7a3hA | 107 | 1.99 | 230 | 0.47 |
| | 1egzA | 123 | 2.10 | 230 | 0.53 |
| | 1cz1A | 111 | 1.85 | 230 | 0.48 |
| | 1ex1A | 72 | 2.04 | 230 | 0.31 |
| | 1bqcA | 114 | 2.01 | 230 | 0.50 |
| | 1xyzA | 105 | 1.99 | 230 | 0.46 |
| | 1clxA | 116 | 2.03 | 230 | 0.50 |
| | 1bg4 | 104 | 1.96 | 230 | 0.45 |
| | 1gok | 100 | 2.20 | 230 | 0.43 |
| | 1exp | 110 | 1.98 | 230 | 0.48 |
| | 1bglA | 155 | 2.11 | 230 | 0.67 |
| | 1bhgA | 123 | 2.08 | 230 | 0.53 |
| | 1ghsA | 133 | 2.17 | 230 | 0.58 |
| | 1aq0A | 131 | 2.15 | 230 | 0.57 |

Table B.2: Structural Alignment of 1RPX inter and intra superfamilies

# C Data for the interacting Calcium binding proteins

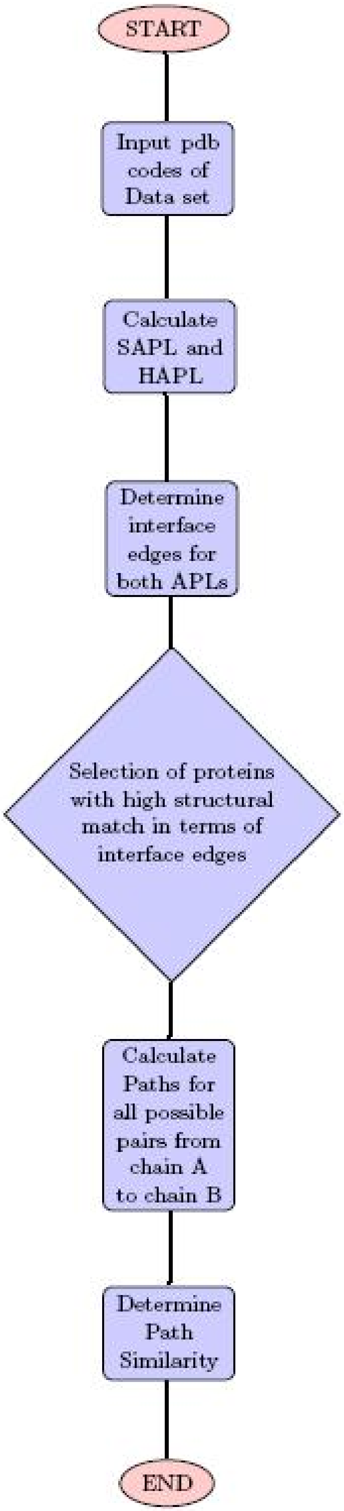

START
Input pdb codes of Data set
Calculate SAPL and HAPL
Determine interface edges for both APLs
Selection of proteins with high structural match in terms of interface edges
Calculate Paths for all possible pairs from chain A to chain B
Determine Path Similarity
END

Figure C.1: Flowdiagram of procedure followed for Ca-Binding Protein Data set

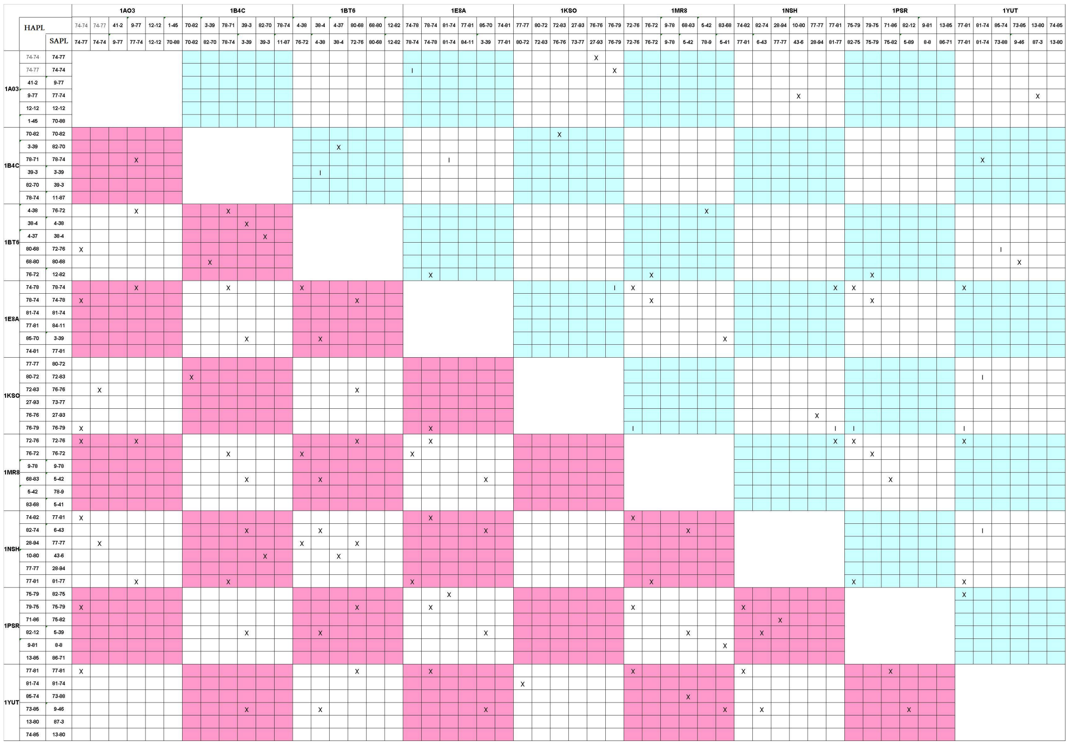


Figure C.2: Structurally aligned interface edges for HAPL (upper diagonal) and SAPL (lower diagonal). Outer axis represents edges calculated by HAPL, inner axis represents edges calculated by SAPL

| | Strong APL | | Homogeneous APL | |
|---|---|---|---|---|
| | Interface Edge | **% Usage** | Interface Edge | **% Usage** |
| 1B4C | A70-B82 | 14.6 | A70-B82 | 13.2 |
| | A82-B70 | 11.7 | A3-B39 | 8.7 |
| | A78-B74 | 10.6 | A78-A71 | 8.1 |
| | A3 -B39 | 10.0 | A39-B3 | 8.1 |
| | A39-B3 | 9.9 | A82-B70 | 6.8 |
| | A11-B87 | 6.2 | A78-B74 | 6.1 |
| 1BT6 | A76-B72 | 12.5 | A4 -B38 | 8.0 |
| | A4 -B38 | 11.0 | A38-B4 | 8.0 |
| | A38-B4 | 10.7 | A4 -A37 | 7.8 |
| | A72-B76 | 8.5 | A80-B68 | 5.0 |
| | A80-B68 | 7.8 | A68-B80 | 3.8 |
| | A12-B82 | 7.3 | A76-B72 | 3.8 |
| 1KSO | A80-B72 | 9.7 | A77-B77 | 10.5 |
| | A72-B83 | 8.7 | A80-B72 | 7.0 |
| | A76-B76 | 6.7 | A72-A83 | 6.7 |
| | A73-B77 | 6.3 | A27-B93 | 5.3 |
| | A27-B93 | 5.6 | A76-B76 | 5.1 |
| | A76-B79 | 4.6 | A76-B79 | 4.1 |
| 1A03 | A76-B72 | 12.1 | A4-B38 | 8.0 |
| | A4 -B38 | 11.0 | A38-B4 | 8.0 |
| | A38-B4 | 10.7 | A4-A37 | 7.8 |
| | A72-B76 | 8.4 | A80-B68 | 5.0 |
| | A80-B68 | 7.7 | A68-B80 | 3.8 |
| | A12-B82 | 7.3 | A76-B72 | 3.8 |
| 1E8A | A78-B74 | 13.4 | A74-B78 | 11.7 |
| | A74-B78 | 9.2 | A78-B74 | 8.4 |
| | A81-B74 | 7.7 | A81-A74 | 6.4 |
| | A84-B11 | 7.1 | A77-B81 | 5.7 |
| | A3 -B39 | 6.6 | A85-B70 | 5.7 |
| | A77-B81 | 6.5 | A74-B81 | 5.2 |
| 1YUT | A77-B81 | 12.7 | A77-B81 | 11.9 |
| | A81-B74 | 6.5 | A81-B74 | 7.8 |
| | A73-B88 | 5.8 | A85-A74 | 6.3 |
| | A9 -B16 | 5.2 | A73-B85 | 4.1 |
| | A9 -B46 | 4.9 | A13-B80 | 3.8 |
| | A87-B3 | 4.6 | A74-B85 | 3.8 |
| 1MR8 | A72-B76 | 12.5 | A72-B76 | 12.1 |
| | A76-B72 | 9.9 | A76-B72 | 10.9 |
| | A9 -B78 | 6.6 | A9 -A78 | 5.8 |
| | A5 -B42 | 6.6 | A68-B83 | 4.9 |
| | A78-B9 | 6.2 | A5 -B42 | 4.9 |
| | A5 -B41 | 8.3 | A83-B68 | 4.6 |
| 1NSH | A77-B81 | 8.3 | A74-B82 | 6.9 |
| | A6 -B73 | 7.2 | A82-B74 | 6.1 |
| | A77-B77 | 6.6 | A28-A94 | 5.4 |
| | A43-B6 | 5.9 | A10-B80 | 4.4 |
| | A28-B94 | 5.7 | A77-B77 | 4.2 |
| | A81-B77 | 5.4 | A77-B81 | 3.8 |
| 1PSR | A82-B75 | 9.4 | A75-B79 | 15.5 |
| | A75-B79 | 9.0 | A79-B75 | 13.9 |
| | A75-B82 | 8.7 | A71-A86 | 5.5 |
| | A5 -B39 | 6.3 | A82-B12 | 5.1 |
| | A8 -B8 | 5.9 | A9 -B81 | 4.6 |
| | A86-B71 | 5.6 | A13-B85 | 4.4 |

Table C.1: Percent usage of top six interface edges in SAPL and HAPL

# D Data set of proteins with different fold types and their network parameters

| *Pr* | *Length* | *Type* | *C* | *L* | *Pr* | *Length* | *Type* | *C* | *L* |
|---|---|---|---|---|---|---|---|---|---|
| *1aep* | 153 | $\alpha$ | 0.41 | 4.24 | *1amp* | 291 | $\alpha/\beta$ | 0.4 | 4.8 |
| *1ash* | 146 | $\alpha$ | 0.41 | 4.3 | *1chd* | 198 | $\alpha/\beta$ | 0.41 | 4.19 |
| *1bcf* | 158 | $\alpha$ | 0.42 | 4.46 | *1cnv* | 283 | $\alpha/\beta$ | 0.4 | 5.13 |
| *1bip* | 122 | $\alpha$ | 0.43 | 4.43 | *1cse* | 63 | $\alpha/\beta$ | 0.41 | 3.17 |
| *1bmt* | 246 | $\alpha$ | 0.43 | 5.32 | *1ctt* | 294 | $\alpha/\beta$ | 0.43 | 5.25 |
| *1bp2* | 123 | $\alpha$ | 0.43 | 4.11 | *1cus* | 197 | $\alpha/\beta$ | 0.43 | 4.26 |
| *1ccr* | 111 | $\alpha$ | 0.41 | 3.99 | *1cyd* | 242 | $\alpha/\beta$ | 0.41 | 4.78 |
| *1cmb* | 104 | $\alpha$ | 0.4 | 4.38 | *1dea* | 266 | $\alpha/\beta$ | 0.4 | 5.02 |
| *1dsb* | 188 | $\alpha$ | 0.42 | 4.89 | *1dhr* | 236 | $\alpha/\beta$ | 0.39 | 4.81 |
| *1etc* | 106 | $\alpha$ | 0.44 | 3.99 | *1dih* | 272 | $\alpha/\beta$ | 0.42 | 6.04 |
| *1fc2* | 206 | $\alpha$ | 0.42 | 6.13 | *1din* | 233 | $\alpha/\beta$ | 0.41 | 4.63 |
| *1hrc* | 104 | $\alpha$ | 0.43 | 3.77 | *1dpb* | 243 | $\alpha/\beta$ | 0.41 | 5.94 |
| *1hrz* | 73 | $\alpha$ | 0.4 | 4.52 | *1dyr* | 205 | $\alpha/\beta$ | 0.4 | 4.91 |
| *1hul* | 108 | $\alpha$ | 0.42 | 5.35 | *1ecp* | 237 | $\alpha/\beta$ | 0.41 | 4.68 |
| *1irl* | 133 | $\alpha$ | 0.37 | 4.4 | *1ede* | 310 | $\alpha/\beta$ | 0.37 | 5.16 |
| *1lfb* | 77 | $\alpha$ | 0.39 | 3.58 | *1eny* | 268 | $\alpha/\beta$ | 0.42 | 5.03 |
| *1lis* | 131 | $\alpha$ | 0.38 | 4.52 | *1eri* | 261 | $\alpha/\beta$ | 0.41 | 5.64 |
| *1lki* | 172 | $\alpha$ | 0.37 | 4.54 | *1erw* | 105 | $\alpha/\beta$ | 0.4 | 3.52 |
| *1lpe* | 144 | $\alpha$ | 0.41 | 4.61 | *1esc* | 302 | $\alpha/\beta$ | 0.42 | 5.11 |
| *1mse* | 105 | $\alpha$ | 0.41 | 4.39 | *1hjr* | 158 | $\alpha/\beta$ | 0.4 | 4.33 |
| *2sas* | 185 | $\alpha$ | 0.41 | 4.93 | *1hsl* | 238 | $\alpha/\beta$ | 0.39 | 5.04 |
| *1abr* | 267 | $\beta$ | 0.4 | 5.83 | *1lau* | 228 | $\alpha/\beta$ | 0.37 | 4.89 |
| *1arb* | 263 | $\beta$ | 0.43 | 4.65 | *1nar* | 289 | $\alpha/\beta$ | 0.39 | 5.19 |
| *1bpl* | 179 | $\beta$ | 0.41 | 5.71 | *3chy* | 128 | $\alpha/\beta$ | 0.41 | 3.83 |
| *1bw4* | 125 | $\beta$ | 0.43 | 4.05 | *3dfr* | 162 | $\alpha/\beta$ | 0.4 | 4.5 |
| *1cau* | 184 | $\beta$ | 0.4 | 5.5 | *5p21* | 166 | $\alpha/\beta$ | 0.42 | 4.17 |
| *1cfb* | 205 | $\beta$ | 0.42 | 6.54 | *153l* | 185 | $\alpha+\beta$ | 0.42 | 4.29 |
| *1cid* | 177 | $\beta$ | 0.4 | 5.13 | *1aps* | 98 | $\alpha+\beta$ | 0.4 | 3.56 |
| *1ctm* | 250 | $\beta$ | 0.42 | 6.3 | *1atl* | 200 | $\alpha+\beta$ | 0.41 | 4.53 |
| *1cyx* | 158 | $\beta$ | 0.39 | 4.24 | *1bri* | 107 | $\alpha+\beta$ | 0.43 | 3.93 |
| *1dlh* | 180 | $\beta$ | 0.43 | 5.34 | *1cew* | 108 | $\alpha+\beta$ | 0.43 | 4.07 |
| *1dup* | 136 | $\beta$ | 0.44 | 4.75 | *1chk* | 238 | $\alpha+\beta$ | 0.41 | 5.37 |
| *1dyn* | 113 | $\beta$ | 0.44 | 3.86 | *1cks* | 78 | $\alpha+\beta$ | 0.34 | 8.13 |
| *1exg* | 110 | $\beta$ | 0.48 | 3.8 | *1cns* | 243 | $\alpha+\beta$ | 0.42 | 5.04 |
| *1fnf* | 368 | $\beta$ | 0.41 | 10.37 | *1com* | 118 | $\alpha+\beta$ | 0.41 | 3.88 |
| *1gpr* | 158 | $\beta$ | 0.42 | 4.1 | *1cyu* | 98 | $\alpha+\beta$ | 0.46 | 3.86 |
| *1hbq* | 176 | $\beta$ | 0.4 | 4.74 | *1doi* | 128 | $\alpha+\beta$ | 0.41 | 3.84 |
| *1hce* | 118 | $\beta$ | 0.44 | 3.88 | *1esl* | 157 | $\alpha+\beta$ | 0.42 | 4.95 |
| *1hng* | 175 | $\beta$ | 0.41 | 6.08 | *1fim* | 102 | $\alpha+\beta$ | 0.42 | 3.78 |
| *1hvk* | 99 | $\beta$ | 0.37 | 3.83 | *1huc* | 205 | $\alpha+\beta$ | 0.44 | 5.39 |
| *1knb* | 186 | $\beta$ | 0.42 | 4.6 | *1ikl* | 69 | $\alpha+\beta$ | 0.44 | 3.68 |
| *1len* | 181 | $\beta$ | 0.43 | 5.4 | *1mol* | 94 | $\alpha+\beta$ | 0.44 | 3.75 |
| *1lxa* | 262 | $\beta$ | 0.44 | 5.66 | *1msc* | 129 | $\alpha+\beta$ | 0.44 | 6.17 |
| *2ncm* | 99 | $\beta$ | 0.44 | 3.84 | *1mut* | 129 | $\alpha+\beta$ | 0.42 | 4.27 |
| *2prd* | 174 | $\beta$ | 0.43 | 4.38 | *2aak* | 150 | $\alpha+\beta$ | 0.39 | 4.46 |
| *2stv* | 184 | $\beta$ | 0.42 | 4.86 | *2phy* | 125 | $\alpha+\beta$ | 0.41 | 3.81 |
| *4fgf* | 124 | $\beta$ | 0.44 | 3.75 | *7rsa* | 124 | $\alpha+\beta$ | 0.4 | 4.12 |
|  |  |  |  |  | *9pap* | 212 | $\alpha+\beta$ | 0.42 | 4.54 |
|  |  |  |  |  | *9rnt* | 104 | $\alpha+\beta$ | 0.42 | 3.68 |

Table D.1: Data set of 76 proteins having folds $\alpha$ , $\beta$, $\alpha/\beta$ and $\alpha+\beta$, their clustering coefficients (C) and shortest paths lengths (L) computed at a cut of distance of 6.7 Å.

| Rank | Bottleneck Edge | # of occurance | TD potential | % Usage | Rank | Bottleneck Edge | # of occurance | TD potential | % Usage |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 194 ↔ 228 | 666 | 0.21 | 2.17 | 26 | 67 ↔ 68 | 205 | -0.03 | 0.67 |
| 2 | 30 ↔ 33 | 594 | -0.09 | 1.94 | 27 | 81 ↔ 110 | 205 | -0.03 | 0.67 |
| 3 | 10 ↔ 12 | 542 | 0.04 | 1.77 | 28 | 157 ↔ 187 | 203 | -0.04 | 0.66 |
| 4 | 27 ↔ 30 | 397 | -0.09 | 1.29 | 29 | 116 ↔ 117 | 202 | -0.07 | 0.66 |
| 5 | 16 ↔ 19 | 363 | -0.13 | 1.18 | 30 | 50 ↔ 51 | 201 | -0.08 | 0.66 |
| 6 | 75 ↔ 97 | 354 | -1.03 | 1.15 | 31 | 84 ↔ 85 | 201 | -0.08 | 0.66 |
| 7 | 40 ↔ 41 | 306 | -0.52 | 1 | 32 | 124 ↔ 147 | 201 | -0.94 | 0.66 |
| 8 | 165 ↔ 168 | 286 | -0.09 | 0.94 | 33 | 14 ↔ 38 | 198 | -0.09 | 0.65 |
| 9 | 73 ↔ 75 | 261.5 | -1.03 | 0.85 | 34 | 14 ↔ 72 | 198 | -0.09 | 0.65 |
| 10 | 106 ↔ 107 | 255 | -0.36 | 0.83 | 35 | 16 ↔ 43 | 198 | -0.09 | 0.65 |
| 11 | 146 ↔ 184 | 239 | -0.98 | 0.78 | 36 | 24 ↔ 27 | 198 | -0.09 | 0.65 |
| 12 | 81 ↔ 82 | 226 | 0.46 | 0.74 | 37 | 118 ↔ 119 | 198 | -0.09 | 0.65 |
| 13 | 213 ↔ 214 | 225 | 0.41 | 0.73 | 38 | 120 ↔ 142 | 198 | -0.09 | 0.65 |
| 14 | 101 ↔ 102 | 224 | 0.4 | 0.73 | 39 | 148 ↔ 185 | 198 | -0.09 | 0.65 |
| 15 | 190 ↔ 223 | 223 | 0.25 | 0.73 | 40 | 149 ↔ 156 | 198 | -0.09 | 0.65 |
| 16 | 22 ↔ 23 | 219 | 0.14 | 0.71 | 41 | 164 ↔ 168 | 198 | -0.09 | 0.65 |
| 17 | 154 ↔ 156 | 217 | 0.12 | 0.71 | 42 | 212 ↔ 213 | 198 | -0.09 | 0.65 |
| 18 | 155 ↔ 156 | 217 | 0.12 | 0.71 | 43 | 218 ↔ 221 | 198 | -0.09 | 0.65 |
| 19 | 139 ↔ 140 | 216 | 0.1 | 0.7 | 44 | 219 ↔ 222 | 198 | -0.09 | 0.65 |
| 20 | 189 ↔ 191 | 216 | 0.1 | 0.7 | 45 | 29 ↔ 32 | 189 | -0.98 | 0.62 |
| 21 | 130 ↔ 131 | 213 | 0.09 | 0.69 | 46 | 19 ↔ 208 | 181 | -0.13 | 0.59 |
| 22 | 225 ↔ 226 | 212 | 0.05 | 0.69 | 47 | 25 ↔ 28 | 181 | -0.13 | 0.59 |
| 23 | 64 ↔ 65 | 208 | -0.02 | 0.68 | 48 | 102 ↔ 103 | 181 | -0.13 | 0.59 |
| 24 | 78 ↔ 79 | 208 | -0.02 | 0.68 | 49 | 159 ↔ 161 | 181 | -0.13 | 0.59 |
| 25 | 150 ↔ 153 | 208 | -0.02 | 0.68 | 50 | 20 ↔ 21 | 175 | -0.14 | 0.57 |

Table D.2: Bottleneck data for protein 1RPX (top 50 of 617 given)