On Interaction Patterns in Proteins

by
Gizem Özbaykal

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

Sabancı University
Spring 2014-2015

On Interaction Patterns in Proteins

APPROVED BY

Prof. Dr. Ali Rana ATILGAN .....
(Thesis Supervisor)

Assoc. Prof. Dr. Semih Onur Sezer .....

Assoc. Prof. Dr. Gizem Dinler-Doğanay .....

DATE OF APPROVAL: .....

Acknowledgments

I am grateful to my adviser Professor Ali Rana Atılgan for his guidance and support throughout many years. I want to thank Professor Canan Atılgan for endless fruitful discussions that widened my perspective. It has been an invaluable experience to learn the research discipline through their eyes. I owe my deepest gratitude to my family who has always encouraged me.

This work was supported by the Scientific and Technological Research Council of Turkey (grant numbers 110T624 and 113Z408).

On Interaction Patterns in Proteins

Gizem Özbaykal

MAT, M.Sc. Thesis, 2015

Thesis Supervisors: Professor Ali Rana ATILGAN

Keywords: complex systems, random graphs, clustering, atomic clusters, single-point mutations

Abstract

Proteins act like molecular machines that perform various functions in cellular activities. The physical laws determine the rules of atomic arrangements, however the organization of amino acids in proteins inherit evolutionary information. Understanding the three-dimensional structures of proteins are crucial for the exploration of the strong relationship between structure and functionality. This provides motivation to inspect how the network structure affects communication in global scale. In this thesis, we study the interaction patterns in proteins to explore what kind of local mechanisms and global properties they inherit. Using the spatial information of amino acids, simplified models of complex molecular systems are built. We generate synthetic structures that resemble proteins in terms of network properties such as degree distribution and clustering characteristics. The differences between synthetic structures and proteins are traced to distinguish proteins from non-protein structures. Such a differentiation points out patterns that are peculiar to proteins and reveal the randomness within the proteins. We introduce the Mutation-Minimization (MuMi) method which mimics single point alanine mutation scan to investigate how proteins respond to naturally occurring random perturbations. Our approach enables us to unravel motifs that are common in protein structures and point out amino acids that have significant functional roles in biological activities.

Protein Etkileşim Örüntüleri Üzerine

Gizem Özbaykal

MAT, Yüksek Lisans Tezi, 2015

Tez Danışmanı: Profesör Ali Rana ATILGAN

Anahtar Kelimeler: Karmaşık sistemler, rastgele ağlar, kümelenme, atomik kümeler, tek nokta mutasyonları

Özet

Proteinler, hücresel faaliyetlerin gerçekleşmesinde çeşitli roller oynayan moleküler makineler gibi hareket ederler. Fizik kanunları atomik düzenlenmeler üzerinde etkilidir. Ancak proteinler, amino asit örgütlenmeleri üzerinden evrimsel bilgiyi taşırlar. Proteinlerin üç boyutlu yapısını anlamak, onların şekilleri ve fonksiyonları arasındaki güçlü bağı keşfetmek için son derece önemlidir. Bu, aynı zamanda ağ yapılarının küresel ölçüde iletişimi nasıl etkilediğini incelemek için gerekli motivasyonu sağlar. Bu tezde proteinlerin etkileşim örüntüleri, bölgesel yapılanmaları ve küresel özellikleri anlamak için çalışılmıştır. Amino asitlerin sağladığı uzaysal bilgi sayesinde karmaşık protein sistemleri basitleştirerek modellenebilir. Bu doğrultuda, proteinleri temsil edecek yapay ağlar oluşturulur. Yapay ağların proteinleri en iyi şekilde temsil etmeleri için proteinlerin ağsal özellikleri onlara atfedilir; örneğin komşuluk dağılımı ve kümelenme karakteristiği gibi. Bu aşamadan sonra oluşturulan yapay ağlar ile proteinler arasındaki farkların izi sürülerek proteinlere has özellikler araştırılır. Söz konusu başkalaşımlar proteinlerin rastgeleliğe ne kadar yakın olduklarını da gözler önüne sermekte yardımcı olurlar. Ek olarak ilk kez bu tezde tanıtılan Mutasyon-Minimizasyon (MuMi) metodu, tek nokta alanin mutasyonlarının benzetimlenmesiyle, proteinlerin rastgele oluşan doğal karışıklıklara tepkisini inceleme imkanı sunar. Yaklaşımımız, proteinlere özgü örüntüleri keşfetmeyi ve biyolojik faaliyetlerde hususi görev

alan amino asitleri teşhis etmeyi mümkün kılmaktadır.

Table of Contents

List of Figures

List of Tables

1

Introduction

Biological, social, economical and many real life systems systems develop under the changing conditions of the surrounding environment and their components evolve accordingly. The major difficulty is to decide over many possible definitions of the system components and their interactions. Thus the challenge becomes, how a model, both simple and effective, can be constructed to define the rules in the system, predict the limitations on how individuals behave and produce the observed emergent properties. Networks are good representatives of many real life systems such as World Wide Web, scientific collaborations, cellular activities, ecosystems of interacting species, communication in social media, financial markets, linguistics, power grids, neural communication in brain and many others [5]. The interaction patterns in these systems play a pivotal role in the definition and characterization of the system. As it might be apparent, these interaction patterns are not formed by pure chance neither by uncompromising specific rules. These interaction patterns are complex: the components interact in such a way so that their collective behavior is not a simple combination of their individual behaviors [6].

In this thesis, we study the nature of the interaction patterns in proteins. These patterns can reveal the characteristics peculiar to proteins and therefore can be utilized to differentiate proteins from other non-protein structures.

In Chapter 2 we present the definitions of some concepts in network science that are extensively used throughout the thesis. Measures that allow the exploration of local and global properties of networks are analyzed in detail. The distinction between simple

and complex networks are presented and properties of different classes of networks are investigated. The detailed classes are selected for being representatives of systems that inherit different levels of randomness.

In Chapter 3, atomic systems are described as network structures. Besides graphs constructed from empirical data, we also utilize computer generated, *synthetic*, graphs to form a basis for the comparison between different classes of networks. To make such a comparison, we place two extreme cases at the ends of a *randomness scale*. At one end, we have random graphs where interactions are formed by pure chance and at the other end, we have crystal lattice networks which are examples of complete order and regularity. To tune between the two ends, we have generated synthetic systems with different proportions of clustering. We investigate how random the protein structures are by making use of their interaction patterns with other systems.

Developing useful methods for finding sites that are significant for biological functioning of a protein by using only its known three dimensional structure is useful to understand the organization of amino acids. It is becoming clear that proteins act like machines and positions away from the functioning sites have evolved to orchestrate the interactions in these machines. In Chapter 4, we present a method to mimic experimental alanine mutation scan studies and to pinpoint residues that are significant for protein function. We analyze our method by detailing the two case studies and validate our findings with experimental studies. We conclude with Chapter 5 by briefly summing up the main findings of this thesis.

2

Complex Networks

2.1 Definitions and Preliminaries

A graph is a set of vertices and edges where vertices define the elements of the system and edges specify a connection pattern for the vertices. A graph is represented by an adjacency matrix (denoted as $\mathbf{A}$). A_{ij} is a nonzero element for vertex pair i and j if they are connected and zero otherwise. In this thesis, the terms graph and network are used interchangeably similar for vertices-nodes and edges-links. Also, none of the networks used in this study has self loops or multiple edges between vertices. A network is directed when a link between any node pair has a direction; all networks studied in this work are undirected. If all links are identical regardless of their direction, the network is termed homogeneous. The total number of nodes in a network, network size, is denoted by N. Networks that have links with different weights are termed as weighted.

2.1.1 Simple versus Complex Networks

Regular networks, such as lattices are examples of simple networks. Since there is no exact definition for a simple network, the following sections are devoted to possible explanations of what happens to a simple system when some complexity is introduced. Grids have simple connection patterns and are mostly based on spatial information. They are good representatives of crystal structures which inherit almost perfect order and reg-

ularity. However, many real life systems such as social or biological networks, do not have such ordered interaction patterns. To have an understanding of the irregular interaction patterns of these real life networks, lattice structures are not good enough [7].

For complex systems, the whole is not just the sum of its parts, but also the interactions between the parts. To understand the nature of complex systems, the interaction of parts should be evaluated. Networks are extremely powerful for representing the system as a whole and the interaction pattern of its parts. They are extremely useful tools in exploring global properties as well as local mechanisms.

2.1.2 Degree Distribution

Degree of a node i, denoted as k_i, represents the number of nearest neighbors it has and it can be referred as connectivity of a node. k_i is simply equal to the sum of links node i has (equation 2.1), sum of the elements of $\mathbf{A}$ column wise (or row wise, since $\mathbf{A}$ is symmetric for undirected homogeneous graphs).

$$k_i = \sum_{j}^{N} A_{ij} \tag{2.1}$$

Degree distribution specifies a probability distribution function, $P(k)$, for k_i values, implying the probability of finding a node that has exactly k_i many degrees. For empirical networks, networks that are generated from given data, the degree distribution usually has some deviation from the actual probability function used to describe it. Two types of degree distributions are extremely important for modeling and analysis of real life networks: (i) Poisson degree distribution and (ii) power-law degree distribution. For networks with Poisson distributed degrees, k_i values fall in a narrow interval compared to a power law network where the gap between the highly connected and the least connected nodes is very large. In the latter case, the term *hub* is introduced for nodes with very high connectivity.

Degree Sequence provides the number of neighbors for each node in the network. A given degree sequence is called graphic if a graph can be generated by using the sequence [8]. In this thesis we utilize graphic sequences with Poisson distribution. Major distinc-

tions between classes of networks can be made as discussed in Section 2.2, where specific characteristics of networks with Poisson distributed degrees are also given in detail.

2.1.3 Clustering

Clustering of nodes is a useful measure for inspecting the local structure in the network. Clustering coefficient is a measure for specifying the probability of finding a common neighbor of any connected node pair. Thus, C takes value between zero and one. If the pair of nodes have a common neighbor, the three form a *triangle*. As the number of common neighbors increases for a node pair, the number of triangles also increases. Thus this number is normalized by the maximum possible number of triangles that a node can make with all of its neighbors. The symbol C_i is used for clustering coefficient of a node (equation 2.2) and C is for the average clustering coefficient of the whole network (equation 2.3).

$$C_i = \frac{\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} A_{ij}A_{ik}A_{kj}}{\binom{k_i}{2}} \tag{2.2}$$

$$C = \frac{1}{N}\sum_{i}^{N} C_i \tag{2.3}$$

The more C approaches to one, the denser the network is. With low levels of clustering (for example 0.1) and a given N, there are many possible configurations for a generated network but with $C = 1$ and any N, there is only one configuration where all nodes are connected to each other, sharing the same degree. It is possible to encounter two networks with same degree distribution while having huge differences between their connection patterns. These differences can be detected by using a local measure like C and global measures such as the average shortest path length as described in the following.

2.1.4 Shortest Paths

The shortest path length, denoted L_{ij}, between two nodes is the number of connections that needs to be crossed to reach node j from i. In this thesis, the shortest path

lengths are computed by Johnson's Algorithm implemented in the Bioinformatics Toolbox of MATLAB [9]. The average shortest path length, L, of node i, L_i, is then the average over the minimum number of steps that the node may be reached from all other nodes of the network.

$$L_i = \frac{1}{N-1} \sum_{i}^{N} L_{ij} \tag{2.4}$$

All networks which are used in this study are connected graphs, implying that each node has at least one neighbor. This ensures the existence of a path between any node pair in the network, thus a finite numbers of path lengths. L is a measure for global characteristics of the network:

$$L = \frac{1}{N} \sum_{i}^{N} L_i \tag{2.5}$$

L values differ greatly between networks from different classes which share the same number of nodes and links. Therefore, it is crucial to analyze how connection patterns and local motifs such as triangles affect the global properties such as navigability for a deeper understanding of the system. In addition, the number of possible routes (with the same length as the shortest path) exist between node i and j is beneficial for comparing graphs with different connection patterns. One way to utilize the number of alternative routes is defined by the measure *betweenness centrality*, explained in the following section.

2.1.5 Centrality

There are different measures for centrality such as degree centrality, eigenvector centrality, closeness centrality and betweenness centrality [10]. How different centrality measures assign highest centrality to nodes can be briefly listed as:

- Degree centrality: to nodes with high degree
- Eigenvector centrality: to nodes with central neighboring nodes

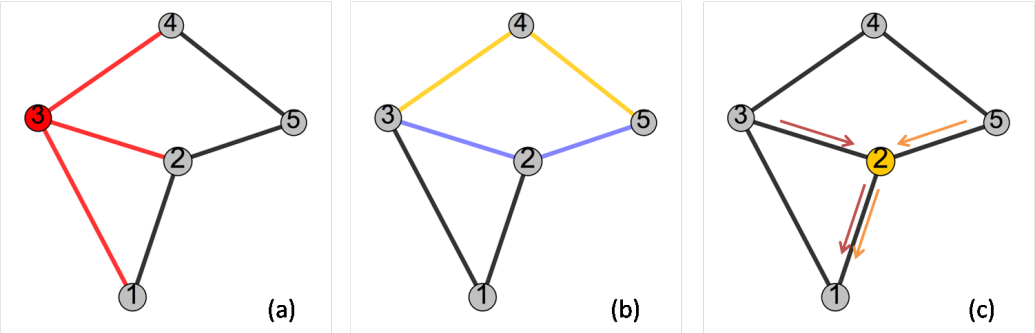


Figure 2.1: Example describing main network properties. A sample network for a chain of five nodes having non-bonded interactions between nodes 1 – 3 and 2 – 5 is displayed. (a) Node 3 has degree $k_3 = 3$ (red connections), (b) two sample shortest paths are displayed between nodes 3 and 5; average path length to node 2, L_2 is $= 5/4 = (L_{12} + L_{23} + L_{24} + L_{25})/4 = 5/4$ and (c) two sample paths from 3 to 1 and from 5 to 1 while crossing node 2 are shown, the betweenness centrality of node 2 is $BC_2 = 4/10 = 0.4$.

- Closeness centrality: to nodes that minimize distance to other nodes
- Betweenness centrality: to nodes that are traversed on more shortest paths

In this work, we use betweenness centrality (denoted as BC). It is computed for all nodes in a network using Dijkstra's algorithm [11]; the numbers are then normalized by $N(N-1)/2$.

The definitions of extensively used network measures are schematized in figure 2.1.

2.1.6 Neighborhood Overlap

The term *bridge* is used to define single links that connect two (or more) clusters (node groups) which otherwise would be disconnected. The triadic closure principle is defined as "If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future." [12]. However, it is expected and observed that probability of finding bridges is very low in many types of networks mainly due to the triadic closure principle [1]. Instead of single links there are a few links connecting groups of nodes, communities and these are named *local bridges*. Therefore the probability of these groups to become disconnected decreases

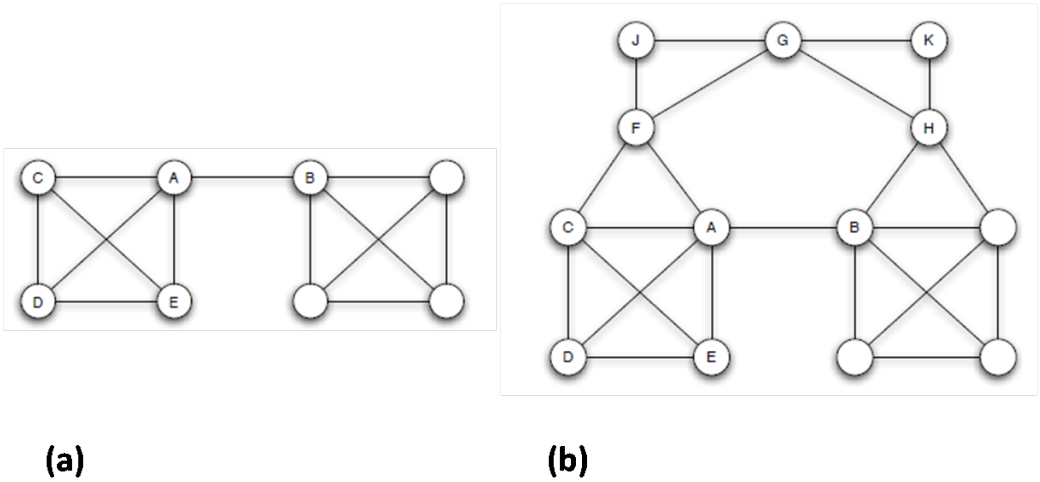


Figure 2.2: Two toy models to illustrate the notion of bridge and local bridge. (a) The link between node A and B is called a bridge because if $A-B$ link vanishes, there will be two separate graphs. (b) $A-B$ link is a called a local bridge.Although upon its removal there will be still one connected graph, the distance between A and B will increase to four from one: $A-F$ to $F-G$ to $G-H$ and $H-B$. Images from [1]

in case of random link failures in the network. The neighborhood overlap (denoted as NO) measure is introduced to detect local bridges. NO is defined through each link in the network by computing the ratio of:

$$NO = \frac{\text{number of nodes which are neighbors of both } i \text{ and } j}{\text{number of nodes which are neighbors of at least one of } i \text{ or } j} \tag{2.6}$$

When it is close to zero, the link is considered a *local bridge* and if it is equal to zero, a *bridge*. Figure 2.2 provides a visual for the definitions of bridge and local bridge.

2.1.7 Node Neighborhood Overlap

In this section, we report those subgraphs in residue networks which harbor evolutionary conserved residues. We propose a new measure with a slight modification on the conventional neighborhood overlap, NO. Rather than defining NO for edges (eq. 2.6), we introduce node neighborhood overlap, denoted NNO.

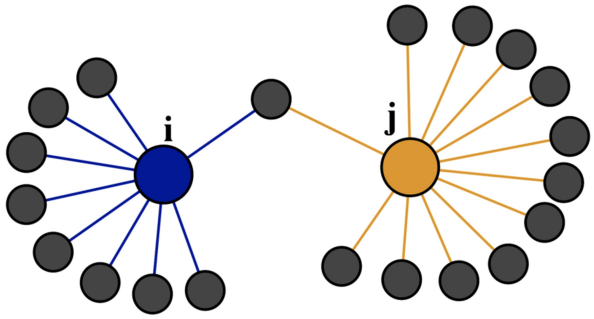


Figure 2.3: A toy graph for NNO measure is provided with numeric calculation NNO_{ij}. A sample NNO calculation for node pair $i - j$ where $n = 1$, $k_i = 9$, $k_j = 12$, results in $NNO_{ij} = 0.05$

$$NNO_{ij} = \frac{n}{k_i + k_j - n} \tag{2.7}$$

NNO is a pairwise measure which depends on the number of common neighbors of nodes i and j (denoted by n) and the degree of i and j (k_i and k_j) under the condition that i and j do not share a link. NNO measure can be computed for subgraphs with various configurations including different number of nodes. NNO_{ij} value is computed from equation 2.7. In other words, NNO gives a *weighted* value of how many different two step paths exist between nodes i and j that do not share a link. These results are collected in the $m \times m$ NNO sparse matrix, $\mathbf{N}$, where the indices of non-zero elements of $\mathbf{N}$ are identical with those of the squared adjacency matrix, $\mathbf{A^2}$. A descriptive scheme is provided in figure 2.3 and figure 2.4 visualizes a protein, its adjacency matrix and its NNO matrix.

2.1.8 Network Motifs

As introduced in [3], network motifs are defined as patterns that occur in the real network significantly more often than in the randomized networks. A motif can include many number of nodes and since it is a subgraph, a motif does not have to include all links between its nodes. Links in a motif can be directed or undirected and this affects the number of all possible configurations. In an undirected network the numbers of all

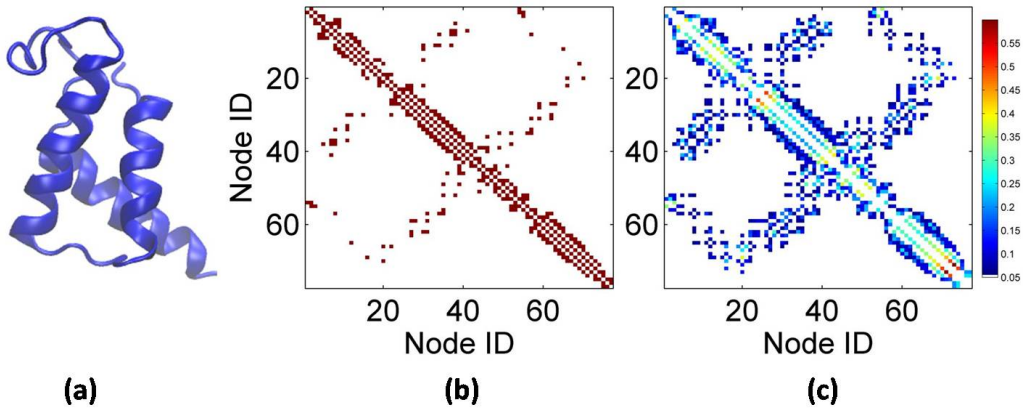


Figure 2.4: (a) The tertiary structure of 1LFB [2] is displayed. 1LFB is the homeodomain portion of transcription factor from rat liver nuclei. (b) Adjacency matrix, **A**, of the protein (c) **NNO** matrix of the protein.

possible configurations are as follows: (i) three-node-motifs: 2, (ii) four-node-motifs: 6, (iii) five-node-motifs: 21 and numbers increase for higher order motifs. If this was a directed network numbers would become: (i) three-node-motifs: 13, (ii) four-node-motifs: 199, (iii) five-node-motifs: > 9000. All possible configurations for four-node and five-node motifs in undirected graphs are shown in figure 2.5 Motifs are computed with the Network Motif Software, mfinder [3]. The user must provide an input adjacency matrix, specify whether the graph is undirected or directed and give the number of nodes in a motif to be searched for. Then (when default parameters in the software are used), the software generates 100 randomized networks by using link switching method. Link switching is made by randomly choosing 100-200 edges in the input network and changing their arrival/departure nodes. A schematic of the randomization and motif search process is provided in figure 2.6 Automatically repeating this procedure separately 100 times results in 100 different randomized networks. This provides a comparison between input and randomized input graphs instead of input and 100 completely random (and irrelevant) graphs. A sample run is provided below:

- First the program searches for all possible subgraphs with the given number of nodes, say 4, in the input graph.

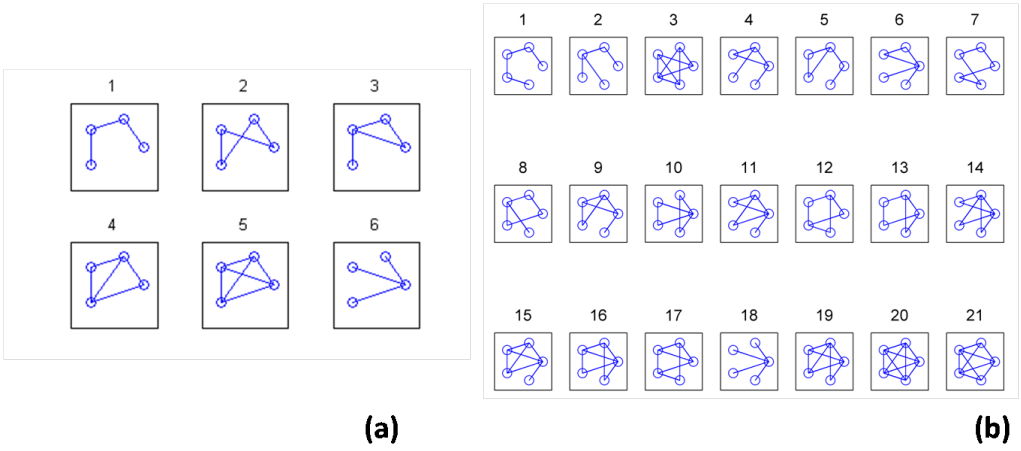


Figure 2.5: (a) Six possible configurations for four-node-motifs (b) 21 possible subgraphs for five-node-motifs.

- The result is a 1-by-6 vector since there are 6 different configurations in four-node motifs. This vector keeps the number of occurrences of each subgraph in the input network.
- Then same search is done in 100 randomized graphs resulting with 100-by-6 matrix.
- The mean and standard deviation (μ and σ) of the number of occurrences of each subgraph in the randomized graphs are calculated.
- All subgraphs that occur more than $\mu + 2\sigma$ times, are considered as significantly over-expressed, thus motifs, in the input network.

We utilize motif calculations by defining a motif distribution, $p(x)$ where x is the motif identity (ID). Motif distribution is a probability distribution which quantifies the probability of a subgraph becoming a motif in a class of networks. For instance, say there is a set containing 150 graphs of different sizes which share the same degree sequence and same average clustering coefficient. The software is fed one-by-one for 150 graphs and motif search is done for each. Then, for each graph, significantly over-expressed subgraphs (motifs) are recorded. If a subgraph, say four-node-motif with ID:2 in figure 2.5, is significantly over-expressed in 50 out of 150 graphs, then $p(2) = 50/150 = 0.3$. As

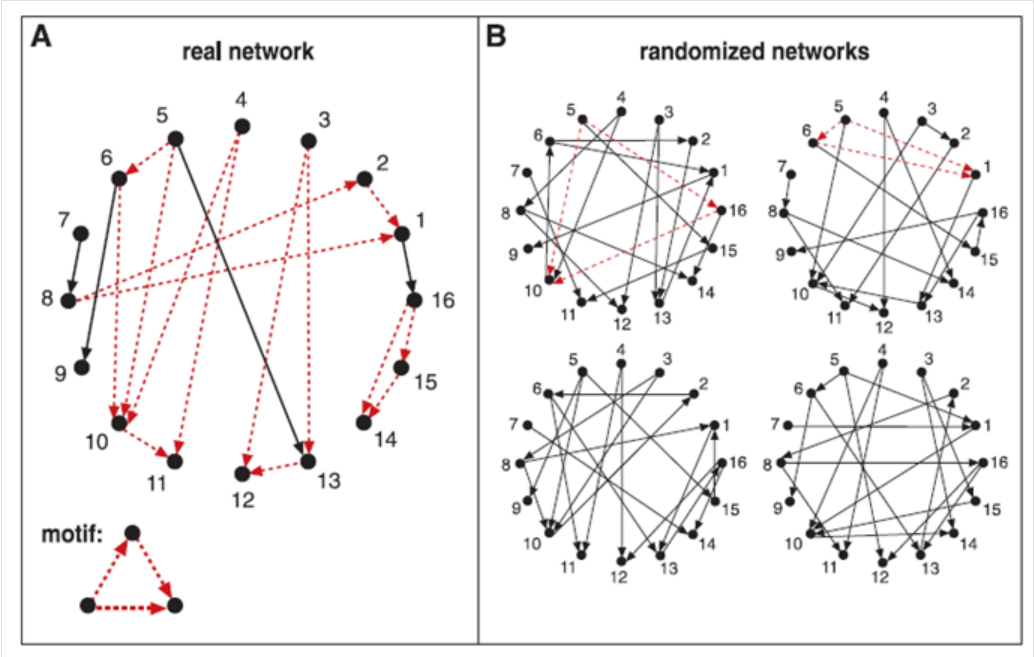


Figure 2.6: A representation of the motif search process (A) The input network is displayed with the subgraph being searched for (lower-left). On the network, the red dashed lines show links that contribute in the formation of the subgraph. (B) Four samples of randomized networks are given and again red dashed lines indicate that the subgraph is found. This subgraph is a motif for the input network displayed in (A) since it is found five times as much in the real network than in the randomized graphs. Figure is taken from [3].

a result, by using the motif distributions of different classes of networks, we are able to compare them with each other.

2.2 Classes of Networks

2.2.1 Random Networks

Random networks, also called ER graphs after Erdös and Rényi are central to the study of complex networks [13]. Not only can random graph be representative of some organizations in nature [5], it can also form the basis of comparison as a measure of complexity for many real life networks. A random graph can be generated by defining two parameters: (i) number of nodes, N, and (ii) the probability of two nodes having a link in between, p. The degree distribution of these graphs converge to a Poisson distribution with mean λ:

$$p_k = \frac{\lambda^k e^\lambda}{k!} \tag{2.8}$$

Random networks share short average path lengths, L, that is represented by the expression:

$$L = \frac{log(N)}{log(\lambda)} \text{ as } n \to \infty \tag{2.9}$$

ER model is a good representative of structures where objects are linked completely by chance. Therefore the probability of observing a link between two neighbors of a randomly selected node, $C \approx 0$ in ER graphs.

2.2.2 Small-World Networks

Small-World (SW) model, introduced by Watts and Strogatz [7], captures a property of real life networks which random networks cannot. The similarity between ER model and real life networks, where objects are not linked completely by chance, is that they both have short path lengths in between. However, the problem of clustering arises; ER model networks have almost zero clustering as opposed to heavy clustering in real life networks. On the other hand, regular graphs (which inherit perfect order and no randomness) can mimic the high clustering, but they cannot satisfy the low L property. Thus, at the two

extreme of a randomness scale, these two models are insufficient to provide high C and low L at the same time. SW model starts with a regular graph where nodes are arranged in a cyclic order and linked to two nearest neighbors. Each node has four neighbors and with a probability p, a randomly chosen link is rewired to a randomly chosen node. Starting from a regular graph where $p = 0$, as p gets close to 0.01, resulting rewired graphs have the properties of high clustering and short path lengths simultaneously. This result is remarkable because of two major reasons in the scope of this thesis: (i) by adjusting a single parameter, one can navigate between different levels of randomness and (ii) a model that generates graphs with the real life network properties is introduced.

2.2.3 Random Networks with Tunable C

ER graphs lack the necessary clustering to mimic complex networks such as transportation, internet or social networks [7, 14]. A possible solution for this problem can be adding/switching links in the graph that can increase the clustering. The task of increasing clustering in a random network is quite possible. One step further would be adjusting the clustering of the random graph so that it becomes the best representative of the properties of the real graph. Is it possible to have a graph with the given degree sequence and the given average clustering coefficient, C? The answer depends on the degree sequence and the value of C. If the parameter C is zero, there are many possible pure random graphs with the given degree sequence. If C is one, the graph must be fully clustered which means every neighbor of every node is connected to each other. This ends up in a single possible configuration, a fully connected graph, where each node has $N - 1$ neighbors (where N is the graph size). With the same degree sequence, the number of possible configurations decrease dramatically as C approaches 1.

The difficulty of sweeping C arises from traveling between two extremely different topologies: pure randomness and complete order. By keeping the degree sequence, a good model should travel between various randomness levels efficiently. There are many different methods/algorithms for network generation [15, 16, 17, 18, 19, 20, 21]. For the purpose of network generation, we use the algorithm *Clustering* (the details of the algorithm can be found in [21]).

3

Structural Patterns in Nature

In this section, we seek network properties that are specific to a protein structure to comprehend its physical nature better. Further, these properties can be used to distinguish a protein from another structure.

3.1 Networks from Atomistic Clusters

Protein Residue Networks

Proteins are the basic building blocks of the biological activities in organisms. With the protein-protein interactions, many cellular processes occur. We know by Central Dogma that proteins are the products of genes and are synthesized according to the information encoded in DNA. A similarity measure for proteins is homology; two genes or gene products (such as proteins) are called homologous if they are descendants from a common ancestral DNA sequence. We use use a set of 553 single chain proteins of various sizes with sequence homology less than 25% (see Appendix for a complete list and ref. [22] supplementary information). We have this limit to avoid over-learning some properties that might be specific to small groups.

We utilize the three-dimensional data provided in Protein Data Bank (PDB) [23] and construct protein residue networks. A residue network is constructed by considering each residue as a point located at its C_β atom (C_α in the case of glycine) and two residues are considered as interacting if the Euclidean distance between them is less than a cutoff

distance. The cutoff distance is taken as 6.7 Å following the first coordination shell of contacts in the radial distribution function, based on the findings in a previous study [24]. Protein amino acid networks are known to inherit small-world model characteristics, having highly clustered nodes with short path lengths in between [24]. As a result, an undirected $m \times m$ adjacency matrix $\mathbf{A}$, where m is the number of residues in the protein, is computed for each protein. The network approach has enabled the study of specific proteins and has helped reveal interesting features not directly evident from structure or sequence homology [25, 26]. For example, interaction conservation was utilized in phylogenetic analysis of remote homologs of the TIM barrel fold to reveal loop-based conserved interactions near the active site [27].

Residue networks have Poisson distributed degrees where $\lambda = 6$, $min(k_i) = 2$ and $max(k_i) = 15$. Thus the number of neighbors is distributed in a narrow range. The typical average clustering coefficient, C, is ≈ 0.35. We know many random or real life networks which have Poisson distributed degrees and $C \approx 0.3$. The essential point here is to realize the uniformity in the distribution of triangles. In a random network, observing a triangle in two randomly selected sites should be equal. However, this is not true for real networks, especially for those which inherit spatial information based on chemical interactions. The highly packed hydrophobic core is more clustered; two neighbors of a C_β atom are also neighbors with high probability. However, residues at the core region have also high connectivity. This causes the clustering coefficients of these residues to decrease, because the number of possible triangles at the neighborhood is a large number (see the denominator term in equation 2.2). For example typically a core residue has 10 neighbors, for which the number of possible triangles is $\frac{10\times9}{2} = 45$. Whereas a surface residue has about four neighbors, then the number of possible triangles becomes $\frac{4\times3}{2} = 6$. Thus, although less clustering is expected at the surface, we observe highly clustered nodes with low connectivity. The reverse happens in core residues; we observe nodes are less clustered compared to surface residues with increased connectivity. In the following subsections, we will see why this non-uniformity is essential.

Crystal Lattice Networks

Crystal lattices are examples of perfect order and regularity. We utilize Ag, CsCl, Zr and Al lattices which have the face-centered cubic (FCC), body-centered cubic (BCC), hexagonal-closed pack (HCP) and simple cubic (SC) structures respectively. By using the Accelrys Discovery Studio 3.1 program (Accelrys Inc., San Diego, CA) these lattice structures are repeated periodically until each forms a network of $\approx$ 400, 100, 400 and 100 atoms, respectively. Networks are constructed by considering atoms as nodes and a link is established if two atoms are first neighbors. Sample crystal structures with their adjacency matrices are illustrated in figure 3.1.

One can immediately recognize that crystal lattices stand at the complete opposite of random networks, thus the two constitute the opposite ends of a randomness scale. The graph set of crystal lattices consists of only four graphs (FCC, BCC, SC, HCP) because we do not expect to see any differences between two graphs of the same crystal lattice. We therefore take one sample for each unit cell.

Evolutionary Conservation

A protein has differences in its sequence between different life forms. For example, Heat Shock Protein 70 kDa is an important chaperon that functions in organisms with various complexity, from bacterium to human. A position in the amino acid sequence is called conserved if it is identical among many organisms. For the detection of a conserved position, there are many methods that perform multiple seuence alignments and statistical tests. We use the ConSurf scores [28] to quantify the evolutionary conservation information since it suggests a quite simple scaling system for the conservation. Scores are between one and 9, 9 implying highest conservation and one highest variability. If a pair of nodes i and j is under consideration, the evolutionary score is obtained from the sum of their individual scores, denoted by S_{ij}.

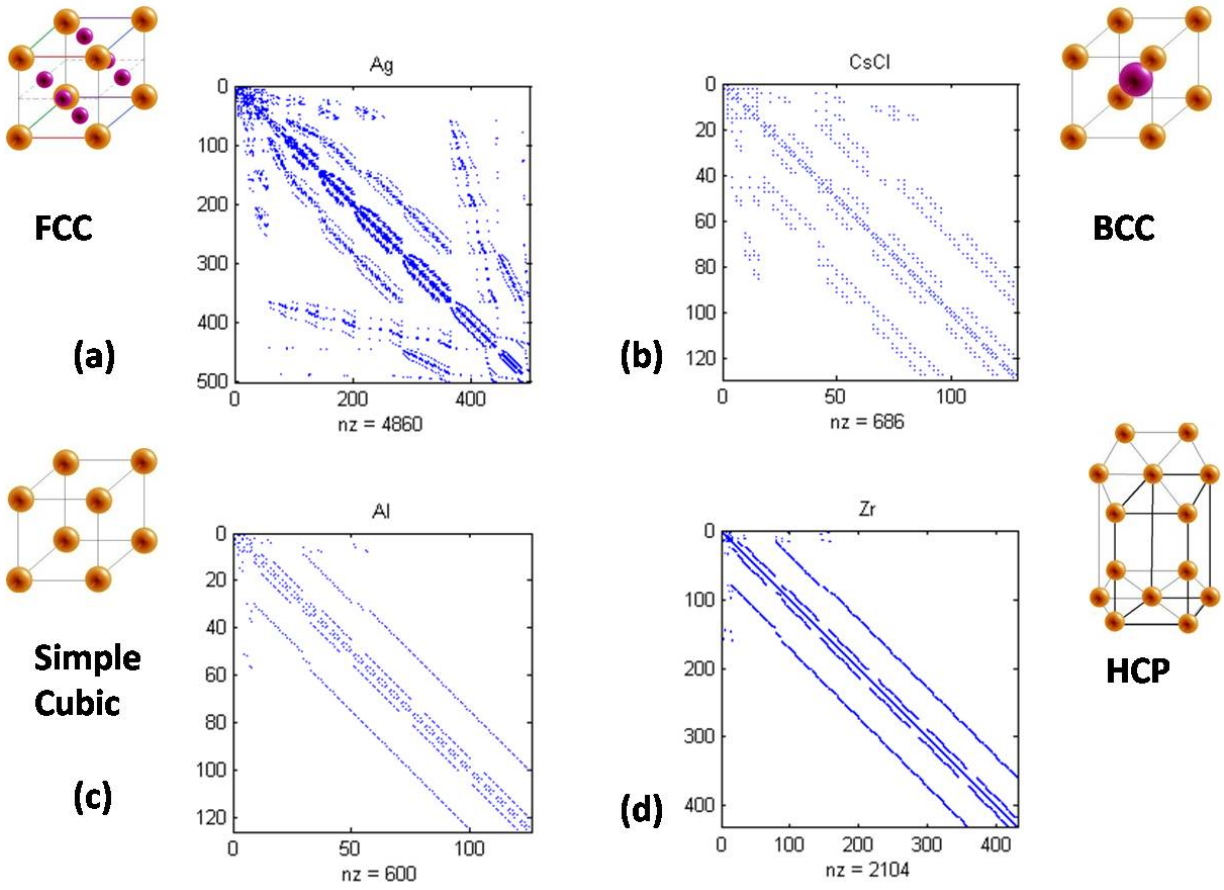


Figure 3.1: At the top left, the unit cells of three crystal structures are displayed along with their adjacency matrices: (a) for Ag (silver), a face-centered-cubic (b) for CsCl (caesium chloride), a body-centered-cubic (c) for Al (aluminum), a simple cubic (d) for Zr (zirconium), a hexagonal-close pack.

3.1.1 Subgraphs at Sites of High Evolutionary Conservation in Residue Networks

There is a relationship between evolutionary conservation and residue connectivity. Approximately 0.90% of the residues in our data set have $k_i < 11$ (marked by the horizontal dotted line in the cumulative degree distribution displayed in figure 3.2a) with conservation scores between one and 9. However, $\approx 0.80\%$ of the remaining residues which have $k_i > 11$ have conservation scores ≥ 7 (figure 3.2b). Therefore, a node with high connectivity ($k_i > 11$) is selected, this will be a evolutionary conserved residue with probability ≈ 0.80. This observation motivates us to develop a measure that can detect conserved sites, improving on the simple connectivity measure k.

We observe that pairs which have extreme low values of NNO exhibit high evolutionary conservation. Given that NNO_{ij} is between (0.035,0.045), the probability of observing a pair with $S_{ij} > 13$ (pairs with scores of 7, 8, 9) is 0.8 (figure 3.3) and probability decreases as NNO value increases. For pairs with low conservation, $S_{ij} < 7$ (pairs with scores of one, two and three), probability of occurrence stays very low in the (0.035,0.045) interval. These results are significant in two major aspects: (i) It is possible to recognize sequential conservation without using sequence data or specificity of amino acids, and (ii) highly conserved amino acids with high connectivity prefer to share low numbers of common neighbors. Nevertheless, we do not refer to NNO as a predictive measure as explained with an example. For instance, $n = 1$, $k_i = 12$ and $k_j = 14$ satisfies NNO_{ij} to be equal to 0.04. The prospective pairs that satisfy $NNO_{ij} = 0.04$ must have $n = 1$, which forces the denominator to be $k_i + k_j - n = 25$; $k_i + k_j = 24$. Since $max(k) = 15$, possible (k_i, k_j) pairings can be (9,15), (10,14), (11,13) and (12,12). As shown in figure 3.2, nodes with high connectivity are rare in proteins. As a result, the number of possible pairings that satisfy the NNO interval (0.035,0.045), is $\approx 3.5 \times 10^{-}3$ of the whole data. This observation motivates us to search for other patterns that may help us to further conceive the protein structure.

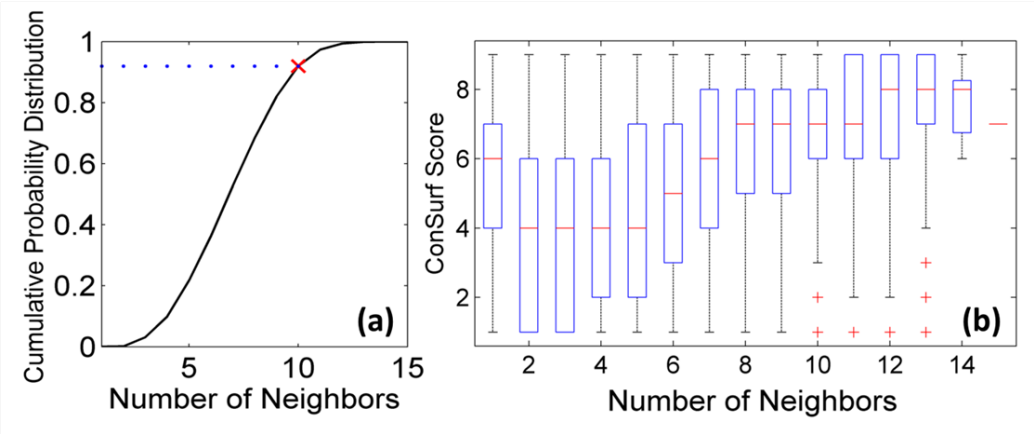


Figure 3.2: (a) Cumulative probability distribution of contact number of residues from our protein set. A Poisson distribution with mean 6 is obtained. (b) Boxplot of the relationship between residue connectivity and their conservation for the same protein set. Small red lines indicate the mean and red plus signs are outliers. ConSurf scores very between 1 (no conservation) and 9 (highest conservation).

3.2 Building Blocks of Proteins: Structural Patterns

We have the following information about a residue network from our set: (i) its degree distributions are Poisson, (ii) its clustering coefficient, C, is ≈ 0.35 and (iii) its average shortest path length, L, is ≈ 5.5 We now ask what differences exist between a residue network and a randomly generated network which has these three properties.

3.2.1 Proteins and Graphs with Tunable Clustering

We generate 11 computer generated *random* graphs with different C values for each protein in our set: the 11 and the protein always share the same network size. These graphs are computed using the algorithm described in Section 2.2.3 We have 11 different graphs for each protein because we wish to determine which amount of randomly introduced clustering best represent a residue network. Since the algorithm used to generate the graphs target these values, the actual C of generated networks may deviate. In these cases, the C observed for the synthetic networks are 0.05, 0.13, 0.2, 0.29, 0.35, 0.37, 0.40, 0.44, 0.48, 0.52 and 0.57.

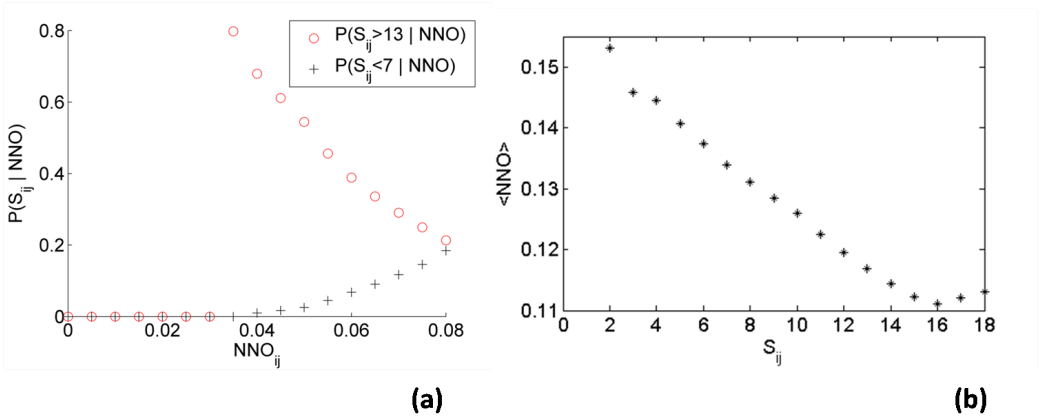


Figure 3.3: (a) NNO values are computed for each node pair in the subset of 553 proteins. With 0.8 probability, node pairs with NNO values (0.035,0.045) are found to have ConSurf scores 7, 8 or 9 (red curve, where $S_{ij} > 13$),while node pairs with scores one, two or three (black curve, where $S_{ij} < 7$) are observed with very low probability. As NNO approaches to 0.08, the probabilities for having high or low conservation gets closer and for values greater than 0.08 NNO they highly fluctuate (not displayed). This graph has $\approx 5.4 \times 10^5$ data points that constitute 20% of whole data. Our results are consistent for cutoff values between 7 ± 0.3 (data not shown). (b) The average NNO measures of node pairs $i - j$ in the dataset is shown with respest to their S_{ij} values. The graph clearly illustrates that highly conserved pairs tend to exhibit low NNO.

Out of the three properties $k_{Poisson}$, $C = 0.35$ and $L = 5.5$, the easiest feature to mimic is the degree sequence, $k_{Poisson}$. We saw earlier that ER random graphs already have Poisson distributed degrees. We calculated the parameter λ as six (as seen in figure 3.4a, black line) for residue networks; hence, the randomly generated Poisson sequences have $\lambda = 6$. Generating a graph with given a graphic degree sequence is an easy task [29].

Following the degree sequence, now we need to satisfy the second condition of generating random graphs with the given clustering coefficient. We mentioned earlier that ER random graphs have almost zero clustering ($C \approx \frac{k}{N}$). Thus, it will be very unlikely to generate an ER random graph with elevated C values such as 0.2 and above. As we increase the desired C, the resulting graph falls somewhere between ER graph and regular graph on a randomness scale. Finally, we keep the third feature, L, free and observe how it behaves with the given first two conditions.

In figure 3.4, black lines show the k, C and L distributions for all nodes in 553 proteins. Similarly, colored lines are computed for each set of graphs having the same input C value. Note that seven of the 11 generated synthetic networks are displayed in 3.4. Figure 3.4a illustrates the desired Poisson degree distributions are achieved for each input C. We also observe that the parameter $\lambda = 6$ is a quite good, though not perfect, fit for the black line representing proteins. Secondly, the middle graph shows some interesting features. Yellow, orange and red lines displays the clustering coefficient distributions of three graph sets with $C_i = 0.48$, $C_i = 0.52$ and $C_i = 0.57$ respectively. Maximum probability of occurrences in these three lines correspond to the mean clustering coefficient of each set. Similarly, graph sets with lower C (displayed in blue line for $C_i = 0.05$ and light blue line for $C_i = 0.20$) have peaks at ≈ 0.1. The cyan (for $C = 0.35$) and green (for $C = 0.40$) graphs, representing medium clustering, are peaked at $C \approx 0.3$ as expected. For all lines, probability of occurrence decreases as numbers are farther from the peak values. Proteins (black line) displays a different behavior. Since $C_{proteins} = 0.35$, if they were computer generated graphs with $C = 0.35$, their C distribution should have appeared in between the cyan and green lines. However, we see that C distribution of proteins has some different characteristics. For instance, $P(C_i \approx 0.30) = 0.25$ is higher in proteins than what we see in cyan and green lines $P(C_i \approx 0.30) = 0.17$. In addition, nodes with lower clustering

than 0.25 is less likely to occur than they do in the synthetic networks of similar C. It seems that protein graphs prefer to have nodes with clustering in the interval of (0.25,0.4) rather than nodes with very high or very low clustering (see the two extremes for proteins in figure 3.4b). Thirdly, in 3.4c), the average path length distributions of protein graphs and graphs with different C are displayed. We see significant differences between the black line and the cyan line. Since these the graph sets have the same C, one could expect that these two would display similar behavior. However, we see that black line is very similar to the yellow line; shortest paths in proteins are most alike to the ones in graphs with 0.48 clustering. Proteins have longer shortest path lengths than the computer generated graphs with same degree sequence and same clustering. We think the reason of this increase is the non uniformity in the distribution of clustered nodes. The best fitting line for the C distribution in residue networks is the one belonging to the $C = 0.40$ set with R-squared of 0.79 (highest among 14 sets) and RMSE (root mean square error) of 0.022 (lowest among 14 sets). Also $C = 0.37$ is almost equally well with $R^2 = 0.77$ and $RMSE = 0.024$. The best fit for the L distribution in proteins belong to $C = 0.46$ set with $R^2 = 0.95$ and $RMSE = 0.018$. The second best is $C = 0.48$ set with $R^2 = 0.934$ and $RMSE = 0.017$. We can argue that, some sites in the protein graphs appear in a randomly generated manner and some resemble regular graphs with dense clusters. Next, we examine what these *dense clusters* can look like.

3.2.2 Network Motifs Resolve How Random Protein Structures are

We next search for specific groupings that are peculiar to proteins. We have seen in figure 3.4 that proteins have shortest path length distribution similar to a random network with a higher C. As we mentioned earlier, almost zero clustering is found in pure random networks and as clustering increases, randomness decreases; thus a network becomes more and more like a regular graph. We briefly argue that a protein structure resembles a random graph in terms of its clustering and a regular graph in terms of shortest path lengths. On the other hand, we wish to achieve a computer generated graph (not based on

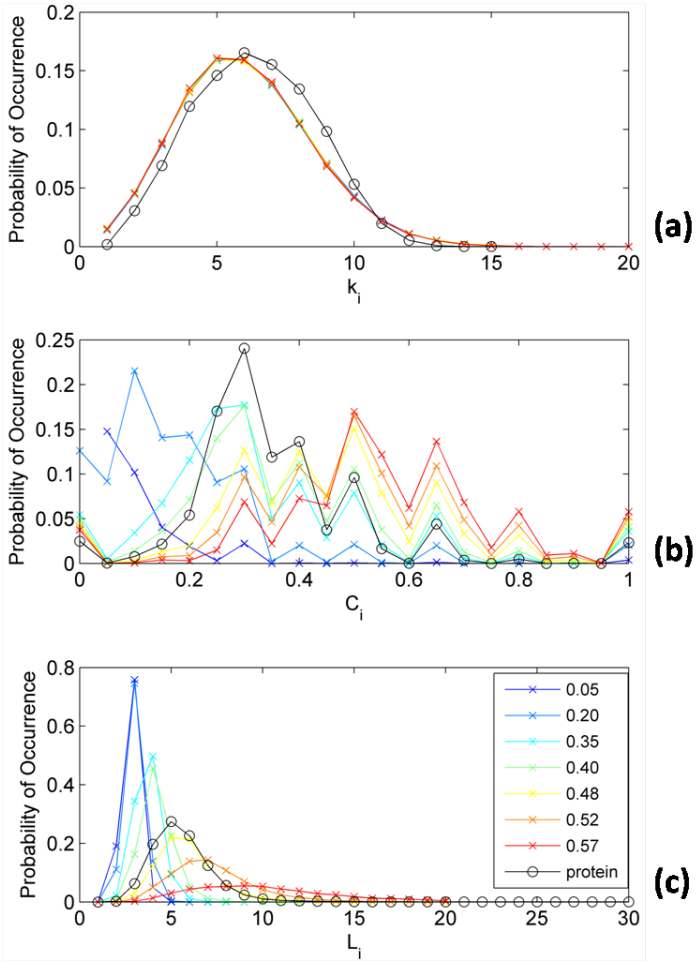


Figure 3.4: (a) The degree distribution of each group of networks is displayed with a different color. A degree distribution is calculated from a huge array which keeps the connectivity of each node in all of the networks in one group. Grouping is done according to the input C. These input C's are displayed in the legend of part (c) of the figure. There is one array for each C and one array for the residue networks; in total of 8 arrays; 8 lines. Since k_i values are integers, probability of occurrence, $P(k_i)$, is simply the number of occurrences of k_i divided by the total number of nodes. (b) Clustering coefficient, C_i, distributions of 8 network groups are displayed. Since C_i values are in the interval of (0,1), the $P(C_i)$ is calculated differently from $P(k_i)$. The interval (0,1) is divided into 21 sub-intervals of 0.05 length. Then the number of points that are in the sub-interval is counted and divided by the the total number of nodes. (c) Shortest path length,L_i distributions of 8 network groups are calculated as in the top graph. Out of 11 C values 7 are displayed to avoid crowd. Lines are added for a better view.

empirical data) which best represents the properties of a protein. We need to observe how close a network we can get by using the idea of tuning clustering in random graphs. In this section, we present our findings about utilization of network motifs and motif distributions (using the approach described in Section 2.1.8).

Motif Distributions

We compute the probability of over-expression of motifs shown in figure 2.5 for each graph set; 11 graph sets with different C values, the protein set and the crystal lattices set. We expect to observe significant differences between the expression patterns of motifs in proteins and other graph sets to identify what building blocks of proteins consist of.

The resulting motif distributions for 13 graph sets are displayed in figure 3.5 for four-node motifs and in figure 3.6 for selected five-node motifs. In top left graph (titled P) in figure 3.5, we see that the probabilities of over-expression of motifs one and six are zero. This means in none of the residue networks, motif1 is found to be significantly over-expressed. The opposite case happens for motifs three, four and five since they appear with probability one; thus in all of the residue networks they are found to be significantly over-expressed. For motif2, the probability value is ≈ 0.6 which means in 0.6×553 many residue networks, motif2 is found as a network motif (significantly over-expressed). We then compare motif distribution of proteins with other graph sets. For simplicity, we can focus on three sets where C values are: 0.05, 0.35 and 0.57 (Table 3.1 on page 27). These are chosen because $C = 0.35$ is closest to the C of proteins and $C = 0.05$ set and $C = 0.57$ are at the two ends of the clustering scale we study. We see that motifs one and six again appear with zero probability ($p = 0$) for the three graph sets. Probabilities for motifs three, four and five increase as C increases. However, for motif2, we realize that the p value for proteins ($p = 0.68$) is larger than all other graph sets. In addition, motif2 has $p = 0.5$ in the lattice set (labeled L). Here we notice that lattice graphs do not harbor many kinds of motifs but some specific patterns which agree well with the packing inside the unit cell. For instance, motif2 is the diamond motif; so it is definitely a building block for the simple cubic unit cell. As a result, we conclude that motif2 is an essential pattern found in proteins and not found in computer generated networks with various Cs. In addition,

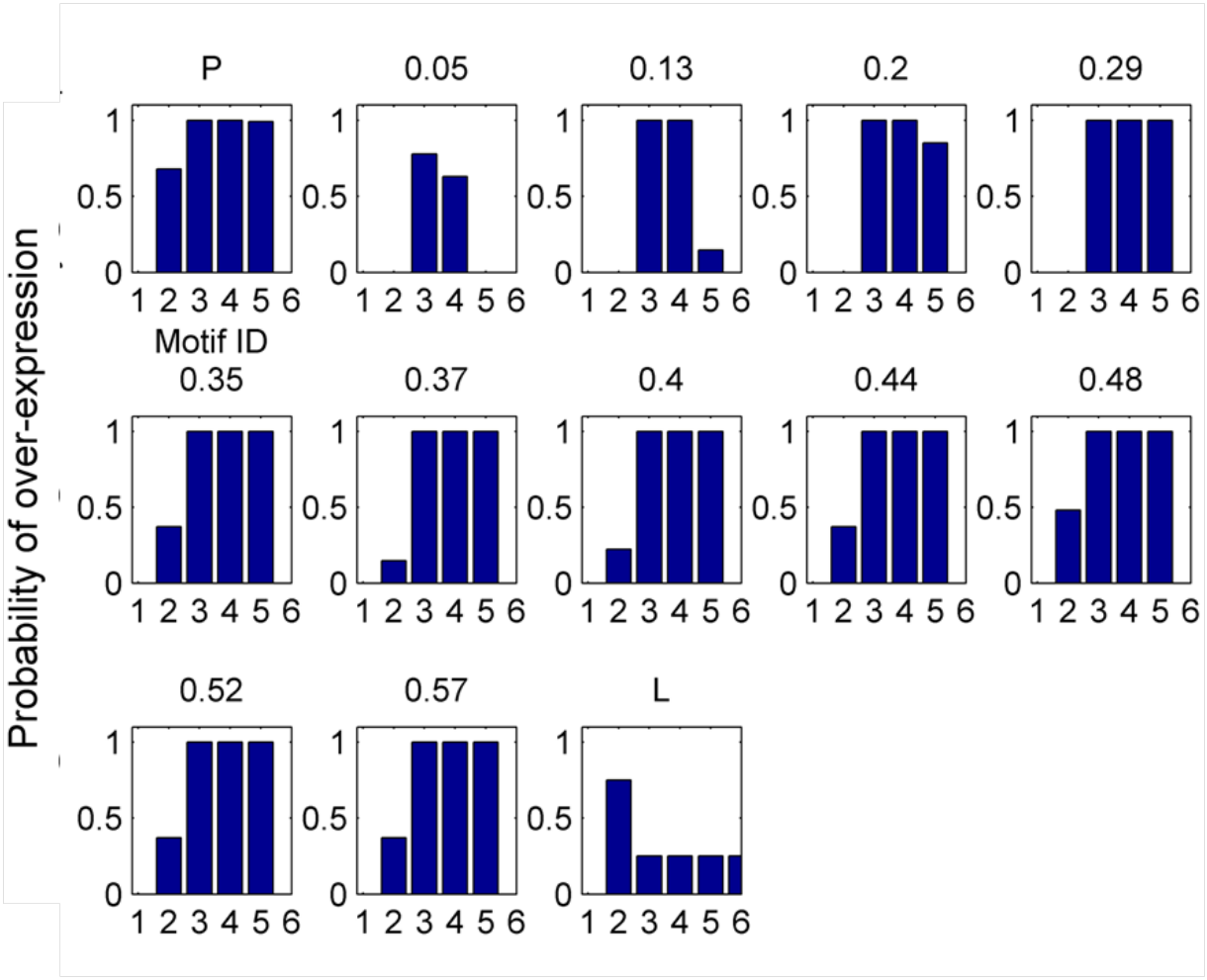


Figure 3.5: Probability of significant over-expression of the six 4-node motifs displayed in figure 2.5a. The title of each figure specifies the name of the graph set. For instance, P stands for the protein set, L for the lattice set, 0.44 for graphs that have $C = 0.44$.

	Motif1	Motif2	Motif3	Motif4	Motif5	Motif6
Proteins	0	0.68	1	1	0.99	0
C=0.05	0	0	0.78	0.63	0	0
C=0.35	0	0.30	1	1	1	0
C=0.57	0	0.37	1	1	1	0
Lattices	0	≈ 0.5	≈ 0.1	0	0	≈ 0.1

Table 3.1: The four-node-motif appearance behavior of five graph sets are grouped based on observation patterns. The probability values for motif2 display great deviation between different graph sets.

since motif2 is also common in lattice graphs, we can say that its appearance increases the regularity (thus decreases randomness) in residue networks. Table 3.1 summarizes the above observations. We can make the same reasoning as above for the five-node motifs by using the motif distributions. Since there are 21 different configurations for five-node motifs (instead of six in the case of four-node ones), interpretation gets more complicated. For this reason, we benefit from a motif specific comparison as displayed in figure 3.6. Once more, we observe that lattices have almost zero probability for each motif. So, motif structures do not agree well with the unit cell packing. Other graph sets display high and low probabilities for various motifs. The graphs of motifs 1, 2, 7, 15 and 18 are not displayed because they exhibit zero probability of over-expression in all graph sets; there would be empty graphs for these motifs. In Table 3.2 motifs are grouped based on the expression patterns. In the first column, listed five-node configurations are not found as motifs in any of the graph sets. For instance, the number of appearances of motif1 in the real network is not found to be significantly higher than the number of appearances in the *average randomized* network. The reason can be because of the simplicity of motif1; it is basically a five-node chain which has four links in total and zero clustering. Thus, the reason is not that motif1 is expressed very less in all of the networks but rather observing such a simple motif in a randomized network many times is highly probable. In the second, third and fourth columns, we see that those motifs appearing significantly more often than is expected in randomized networks. For $C = 0.05$, we do not

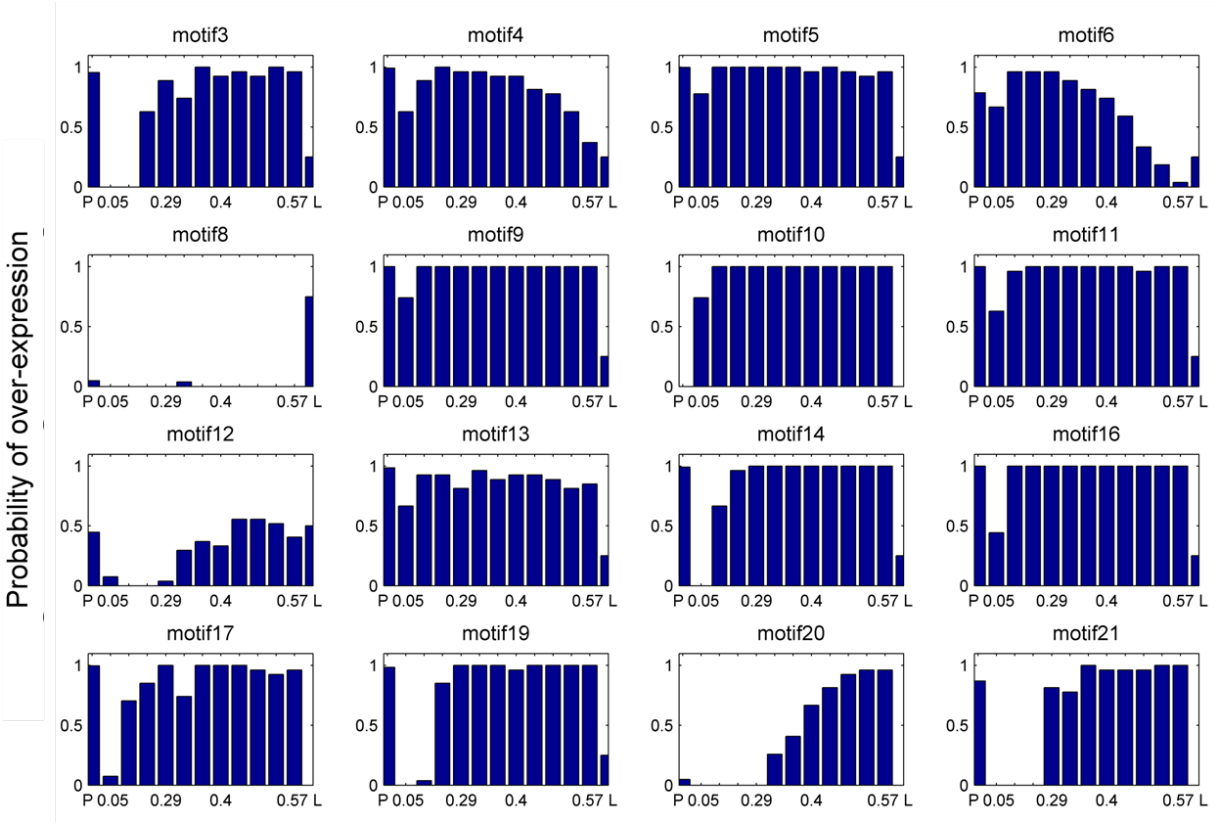


Figure 3.6: Probability of significant over-expression of the 21 5-node motifs displayed in figure 2.5b. The title of each figure specifies the motifID. For example, in the top-left graph titled motif3, we see the probability of motif3 to be significantly over-expressed among different graph sets. On the x-axis, the names of the graph sets are displayed: P stands for the protein set, L for the lattice set, 0.29 for graphs that have 0.29 C. To avoid confusion, some names in the x-axis are not displayed. A full labeling for x-axis will be: P, 0.05, 0.13, 0.2, 0.29, 0.35, 0.37, 0.40, 0.44, 0.48, 0.52, 0.57 and L.

	Motifs 1,2,7,8,15,18	Motifs 3,14,17,19,21	Motifs 5,9,11,13,16	Motifs 4,6	Motif 10	Motif 12	Motif 20
Proteins	0	≈ 1	≈ 1	≈ 1	0	0.50	0.05
C=0.05	0	0	≈ 0.5	≈ 0.5	0.75	0.07	0
C=0.35	0	≈ 1	≈ 1	≈ 1	1	0.33	0.25
C=0.57	0	≈ 1	≈ 1	≈ 0.5	1	0.40	0.97

Table 3.2: The five-node-motif appearance behavior of four graph sets are grouped based on observation patterns. Three separate columns for motifs 10, 12 and 20 are added since their appearance in proteins are much different than in graph sets with clustering coefficients of 0.05, 0.35 and 0.57.

expect to see much motif appearances since 0.05 is a very low value to harbor even lower order motifs (such as triangles). Thus, an essential property of protein structures arises. Motifs 10, 12 and 20 displays very different behavior than they do in other graph sets (see the configurations of motif10, motif12 and motif20 in figure 2.5). By observing the high probability values of motif10 in $C = 0.05$, $C = 0.35$ and $C = 0.57$, we would expect a similar value for proteins as well. However, we realize that motif10 never appears as a motif in residue networks although it frequently appears as one in computer generated graphs. This lack-of-appearance can be because this configuration may be unfavorable due to packing constraints. A similar case appears for motif20 (the complete graph where all nodes are linked to each other) where we do not see the over-expression pattern as we expected to. The two motifs share a common property: elevated clustering coefficient. The clustering coefficient of motif10 is 0.86 and for motif20, it is one. So it may be difficult for a protein to accommodate such dense packing. Nevertheless, the clustering coefficient of motif14 is also 0.80 but it appears as a motif with probability close to one. Another noteworthy point is that although the motif5 and motif8 share the same degree sequence, they exhibit very different behavior in terms of expression. The same occurs for motif12 and motif13. The major distinction arises from the differences in their clustering coefficients.

In contrast to motifs 10 and 20, we observe that the probability value for motif12 is

Table 3.3: The motif appearances in each crystal lattice are given in detail.

	4-node motifs ID's	5-node motif ID's
FCC	3, 4, 5	3, 4, 5, 6, 9, 11, 13, 14, 16, 19
HCP	2	8, 12
BCC	2, 6	8, 12
SC	2	8

higher than the most clustered graph set in the study. For graphs with $C = 0.57$, the probability of observing motif12 as a motif is 0.40 while it is 0.50 for proteins. We perceive this elevated probability value for motif12 as a preferred structural pattern in proteins.

Next, we question whether the distinction between the motif appearance behavior can be done using a measure instead of observation. We aim to find a quantitative way that results in a grouping based on motif-specific properties. For this reason we use three measures B1, B2 and B3 as described in [30], to express the complexity of each motif. B1, B2 and B3 are calculated according to equations 3.1, 3.2 and 3.3.

$$B1 = \sum_{i,j} \frac{\mathbf{A}_{ij}}{\mathbf{L}_{ij}} \tag{3.1}$$

$$B2 = \sum_{i} \frac{\mathbf{A}_{i}}{\mathbf{D}_{i}} \tag{3.2}$$

$$B3 = B2log_2B2 - \sum_{i} \frac{\mathbf{A}_{i}}{\mathbf{D}_{i}} log_2 \frac{\mathbf{A}_{i}}{\mathbf{D}_{i}} \tag{3.3}$$

The matrices $\mathbf{A}$ and $\mathbf{L}$ represent adjacency matrix and the shortest path length matrix as defined in Section 2.1. All three measures utilizes degree sequence (k_i) and the average reachability of nodes (L_i). Based on the results displayed in figure 3.8, we utilize only the B3 measure (see figure caption). As displayed in figure 3.4, we previously observed

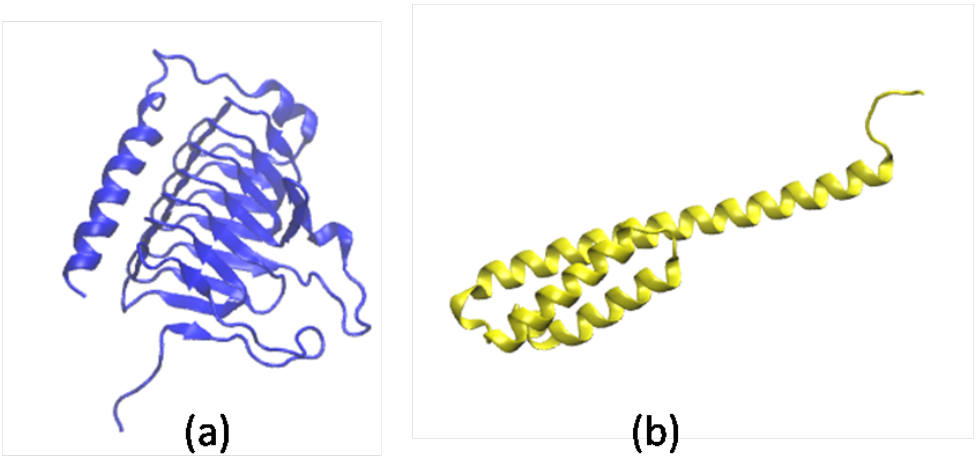
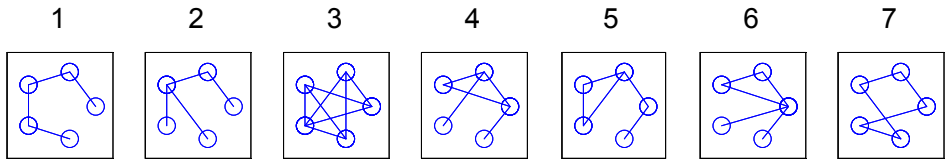
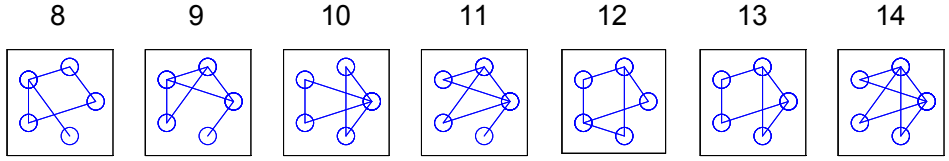


Figure 3.7: For motif appearances in secondary structures: (a) PDB Code: 1QRE for beta sheets and (b) PDB Code: 4B9Q (chain A and residues between 504 and 605) for alpha-helices are used. The appearance of four-node motifs is identical for both and found with ID's of 2, 3, 4 and 5. For five-node motifs in alpha helices: 5-10, 12, 17, 18, 21 and for five-node motifs in beta-sheets: 1, 4-7, 10-13, 16, 17, 19, 21.

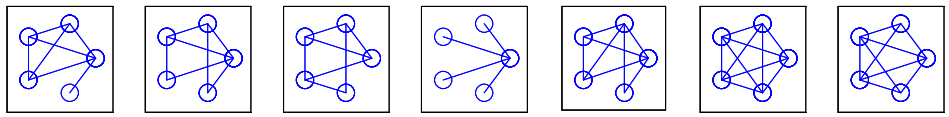
that L_i values in residue networks are large, and they follow a trend which is more likely to be observed in graphs with elevated clustering (such as 0.57). Thus, the complexity measure B3 can lead us to the reasons of this differentiation. Also, it would be helpful to build up a connection between the shortest path length and the clustering coefficient of the motifs. If some motifs are pointed out by B3 for some reason (such as extremely high or low complexity), and these are found to be expressed with a different pattern in proteins (such as motifs 10, 12 and 20 in table 3.2 on page 29), than we could have some reasoning. However, we are unable to extract those motifs that are essential for the proteins by using the complexity measures. According to B3 values, it seems impossible to perform a grouping which also suits for expression patterns of proteins. We aimed to explain the unclear points in these patterns, however it seems that the differences are not due to motif complexity. Perhaps, a complexity definition based on degree sequence and clustering coefficient rather than degree sequence and shortest path lengths can be more useful to differentiate motifs from each other. Such a discriminatory complexity measure can also perform better for motif classification.



	1	2	3	4	5	6	7
B1	0.80	0.89	2.67	1.25	1.18	1.33	1.33
B2	4.42	5.18	13.60	7.13	6.46	7.81	6.67
B3	9.54	10.22	31.09	14.60	13.90	14.97	15.48



	8	9	10	11	12	13	14
B1	1.25	1.60	1.71	1.71	1.71	1.71	2.15
B2	6.71	8.84	9.33	9.64	8.80	8.80	12.00
B3	14.57	18.99	19.86	19.83	19.89	19.89	25.36



	15	16	17	18	19	20	21
B1	2.15	2.15	2.15	1.00	2.67	4.00	3.27
B2	11.77	11.47	10.93	6.29	14.13	20.00	16.80
B3	25.24	25.19	25.05	10.52	31.39	46.44	38.32

Figure 3.8: Each motif is displayed with its corresponding complexity values B1, B2 and B3. According to all three measures, motif1 is the motif with least complexity and motif20 with the highest complexity. We see that B1 has many degenerate values for instance for motifs 10, 11, 12 and 13. B2 displays less degeneracy but B3 is the best for distinguishing between motifs.

4

Quantifying Tolerance of Proteins to Mutations by the Mutation-Minimization Method

We introduce Mutation-Minimization Method (MuMi) to study the local response of proteins to point mutations. We study parameters used in quantifying the properties of residue networks to elucidate what functional roles may be distinguished by each. We use the heat shock protein Hsp70 as the test system since it displays features that have been studied in great detail: It has many conserved residues, serves several different functions on each of its domains, and displays interdomain allostery. For the analysis of spatial arrangement of residues within the protein, we investigate the network properties of the wild type (WT) protein as well as its all single alanine residue mutants using MuMi. We propose measures to express the amount of change from the WT structure upon mutation and compare these deviations to find potential critical sites. We then map the functional significance of the potential sites to the parameter that uncovers them. We find that sites directly involved in binding are sensitive to mutations and are characterized by large displacements. On the other hand, sites that steer large conformational changes typically have increased reachability upon hydrophobic mutation occurring elsewhere in the protein. Finally, residues that control communication within and between domains reside on the largest number of paths connecting pairs of residues in the protein.

4.1 The MuMi Method

4.1.1 Protein Selection and Alanine Mutation Scan Strategy

In this study, we use three different Protein Data Bank (PDB) [31] structures representing, (i) the isolated ATP binding domain (chain D in PDB code 1DKG, residues 3-383) in nucleotide free state [32], (ii) the substrate binding domain (PDB code 1DKZ, residues 389-603) in complex with a synthetically constructed seven-residue long peptide [33], and (iii) the full length protein (chain A in PDB code 4B9Q, residues 2-602) [34]. For the latter, ATP is bound and both domains are engaged in the open configuration. Subdomains in the NBD are defined as follows [35] IA (residues 1 to 38 and 124 to 170), IB (residues 39 to 123), IIA (residues 181 to 227 and 302 to 367), IIB (residues 228 to 301), the N-terminal crossing α-helix(residues 171 to 180), and the C-terminal crossing α-helix (residues 368 to 381). The SBD is made up of the substrate binding sub-domain (residues 393-507) and the lid-domain (residues 508 to 605). The NBD and SBD are joined by a linker. We use each of the above structures for residue network construction and analysis. While 1DKZ and 4B9Q have no missing residues, those of 1DKG (M1, G2, G184, V210, D211, G212, and E213) are completed using the Accelrys Discovery Studio 3.1 program (Accelrys Inc., San Diego, CA). Each structure is then energy minimized in water. Solvation in a water box with at least 10 Å water from any given residue is constructed via VMD 1.9.1 1 [36]. Na^+ and Cl^- ions are added for a neutralized system of 150 mM ionic strength. TIP3P water model is used and the system is energy minimized under the CharmM22 force field [37] using the NAMD package. Minimization is carried out with 10000 conjugate gradient steps [38]. With this choice of minimization stopping criterion, we find that the energy difference between consecutive minimization steps is less than 0.2 kcal/mol. For point mutation scanning, we utilize the full length protein only. The well-minimized structure not only forms the basis for comparison for all the mutants, it is also the starting structure of the point mutants. Therefore, all shifts in the atomic positions are relative to the minimized wild type structure in the water environment. We generate 601 point mutants of 4B9Q whereby each residue is mutated into Ala, followed by solvation and ionization at 150 mM strength, concluded by energy minimization to the

same level of precision as the WT system described previously. This procedure includes 64 Ala to Ala mutants in positions where Ala appears in the WT structure. These cases are used to construct a baseline for the observed changes in the network parameters. We do not observe shifts in the calculated network parameters for these cases, corroborating the stability of the WT structure we use to generate all other structures. We thus obtain 601 minimized mutant structures for further analysis in comparison to the minimized WT structure. We note that the average RMSD between the mutated structures is 2.7 Å, and the largest change is 3.3 Å for the I438A mutation. We note that alanine mutations are selected over other residue types, following other work that suggests using only one alternative residue – alanine – has sufficient information [39]. Single mutation studies performed on PDZ domain in which every position is mutated one by one to every other amino acid points out minor dependence on the number of alternative substitutions tested for each position [4]. Similarly, a mutagenesis study on the voltage-sensing domain of the drk1 voltage-gated K1 channel discusses that an alanine scan is conceptually similar to but more gentle than a tryptophan scan [40].

As a result of mutation and minimization process, 601 binary adjacency matrices are computed for mutants and one for the WT. An adjacency matrix, $\mathbf{A}$ is constructed as a symmetric $N \times N$ matrix ($N = 601$ in this case) whereby the $i - j$th element is 1 if residues i and j are linked and zero otherwise. As a sample mutation, in figure 4.1, the structure of the mutant T428A is superposed with the minimized WT structure.

4.1.2 Thermal Fluctuations

The resolved crystal structure of a protein is reported by the (x, y, z) coordinates in three dimensional space along with a temperature factor for each atom. Temperature factor is a measure for the uncertainty in the atom's position which quantifies the isotropic displacement of an atom from its mean position. Temperature factor is also termed as B factor or Debye-Waller factor and given in equation 4.1 where $< u_i^2 >$ is the mean square displacement of atom i:

$$B_i = 8\pi^2 < u_i^2 > \tag{4.1}$$

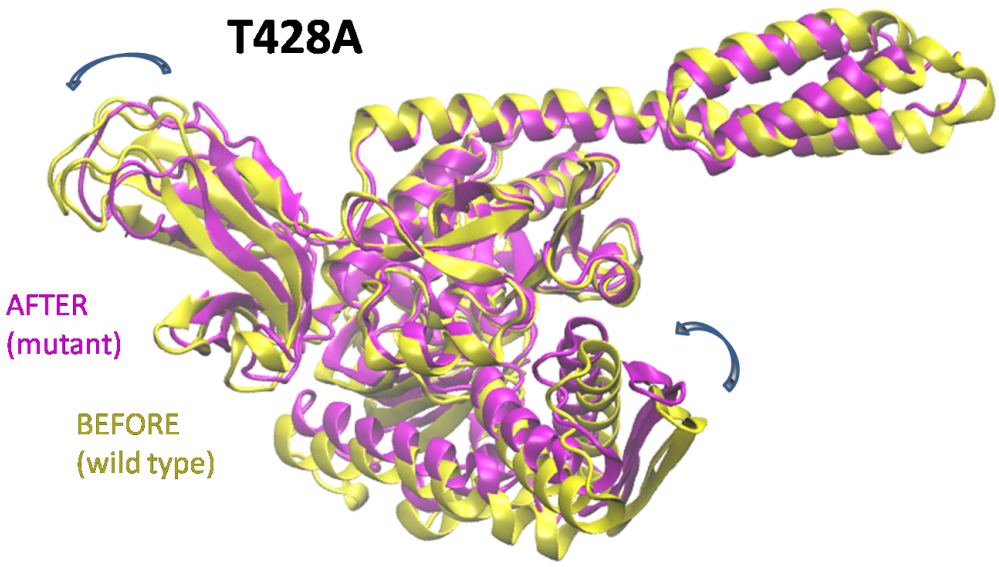


Figure 4.1: The structure of full-length HSP70 (PDB:4B9Q) is drawn in yellow. The T428A introduces the mutation. Structural differences displayed on the superposed structure.

As thermal fluctuations increase, the average displacement of an atom increases, reflected in the B-factor.

Packing is an important feature that has advanced our understanding of proteins. It inherits correlations between atomic fluctuations [41]. As studied in Gaussian network model (GNM) [42] and Anisotropic Network Model (ANM) [43], B factors are directly related to residue auto correlations as can be shown by a simple statistical mechanical treatment.

4.1.3 Measures for Structural Change

We define measures to quantify structural change after the MuMi procedure. The decision process on which measures to use is crucial in this study. For example, the average number of neighbors, k_i, is distributed in a narrow range (with $\min(k_i)$=2, $\max(k_i)$=15, $\text{mean}(k_i)\approx$6) with fluctuations in their values. Likewise, clustering coefficient, C_i is observed to be very sensitive to local changes with large variance in C_i values. Therefore, we conclude k_i and C_i are not sensitive to predicting functional sites resulting from the

MuMi scheme and do not display those results.

To measure induced changes occurring in proteins due to point mutations, we monitor the average displacement from the WT structure. For a given residue i, the displacement vector from its position in the WT structure $\mathbf{R_i^{WT}}$, upon mutating residue m is

$$\mathbf{\Delta R_i^m = R_i^m - R_i^{WT}} \tag{4.2}$$

Here $\mathbf{R_i^m}$ is the position vector of residue i after residue m is mutated and the overall structure is minimized followed by best fitting to the WT structure. The displacements may be organized into a $N \times 3N$ displacement matrix

$$\mathbf{\Delta R} = \begin{bmatrix} \mathbf{\Delta R_1^1} & \mathbf{\Delta R_1^2} & \dots & \mathbf{\Delta R_1^N} \\ \mathbf{\Delta R_2^1} & \mathbf{\Delta R_2^2} & \dots & \mathbf{\Delta R_2^N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Delta R_N^1} & \mathbf{\Delta R_N^2} & \dots & \mathbf{\Delta R_N^N} \end{bmatrix} \tag{4.3}$$

Here, each of the N rows corresponds to a perturbed (Ala mutated) residue, while the columns correspond to the response of a given residue due to mutations on other residues. We may now construct a perturbation-response matrix, $\mathbf{D}$, which is computed from the magnitudes of the displacement vectors resulting from each mutation:

$$\mathbf{D} = \begin{bmatrix} \Delta R_1^1 & \Delta R_1^2 & \dots & \Delta R_1^N \\ \Delta R_2^1 & \Delta R_2^2 & \dots & \Delta R_2^N \\ \vdots & \vdots & \ddots & \vdots \\ \Delta R_N^1 & \Delta R_N^2 & \dots & \Delta R_N^N \end{bmatrix} \tag{4.4}$$

For a protein of N residues, $\mathbf{D}$ has dimensions of $N \times N$ and is asymmetric. Finally average fluctuation vector, $\mathbf{D}$, is

$$\mathbf{D} = \frac{1}{N} \sum_{m=1}^{N} \mathbf{D_i^m} \tag{4.5}$$

$\mathbf{D}$ quantifies how much a residue would deviate from its original position on average, due to all possible alanine point mutations. Displacements of the positions D_i is a measure of local change. Calculation in equation 4.4 yields correlations between displacements due to mutations.

$$\mathbf{\Gamma^{-1}} \approx < \mathbf{\Delta R_i \cdot \Delta R_j} > \tag{4.6}$$

Thus, we first check if the Ala mutation scan yields any information complementary to the auto- and cross-correlations between residue pairs in the absence of any perturbations (equation 4.6). Note that the Gaussian network model (GNM) already provides residue cross-correlation information [42]. Alternatively, the product yields the symmetric $3N \times 3N$ second moment matrix which carries the average effect of all the perturbations. This product may also be viewed as an $N \times N$ matrix, whose jk^{th} element is the 3×3 second moment matrix of correlations between the $x-$, $y-$, and $z-$ components of the fluctuations of residues j and k:

$$(\mathbf{\Delta R^T \Delta R})_{jk} = \begin{bmatrix} <\Delta X_j \Delta X_k> & <\Delta X_j \Delta Y_k> & <\Delta X_j \Delta Z_k> \\ <\Delta Y_j \Delta X_k> & <\Delta Y_j \Delta Y_k> & <\Delta Y_j \Delta Z_k> \\ <\Delta Z_j \Delta X_k> & <\Delta Z_j \Delta Y_k> & <\Delta Z_j \Delta Z_k> \end{bmatrix} \tag{4.7}$$

Finally, the cross-correlations C_{ij} between residues i and j in response to the inserted perturbations is given by the trace of the above submatrix averaged over the N mutations:

$$C_{ij} = tr[(\mathbf{\Delta R}^T \mathbf{\Delta R})_{jk}] \tag{4.8}$$

In this study we also monitor ΔL_i, a measure of shift in average reachability with respect to the WT structure, rather than average reachability, L_i.

$$\Delta L_i = \frac{1}{N} \sum_{m=1}^{N} L_i^m - L_i^{WT} \tag{4.9}$$

where the superscript refers to the average path length of the i^{th} residue in the WT or in the m^{th} Ala mutation. ΔL_i is a measure of global structural changes unlike D_i. Betweenness centrality of WT and mutants are computed as explained in Section 2.1. We also compute ΔBC in equation 4.10 which quantifies the average change in BC values with respect to WT structure.

$$\Delta BC = \frac{1}{N} \sum_{m=1}^{N} BC_i^m - BC_i^{WT} \tag{4.10}$$

4.2 Heat Shock Protein 70 kDa: A Case Study

We perform a thorough scan of the protein with point mutations to every site, followed by energy minimization in explicit water and analyze the consequences on the global struc-

ture relative to the wild type (WT) protein. Hsp70 displays several features that have been confirmed by biochemical studies: It has many conserved residues, serves several different functions on each of its domains [44, 45], and displays interdomain allostery [46]. Hsp70 chaperone acts as a protein folding agent through a mechanism which relies on inter-domain communication between its ATPase domain and substrate binding domain [47]. Upon nucleotide binding, a docked conformation of the two regions occurs [47, 35]. The communication between distant functional sites relies on interdomain allostery which is thought to arise from a set of coevolved residues [48] E.coli Hsp70 (denoted DnaK) has a large number of conserved residues corresponding to approximately one third of its full length [49]. In addition, there is a wide spectrum of experimental studies on the Hsp70 family. Yet, because of the complex nature of the Hsp70 structure and dynamics, the mechanisms of action have not yet been fully discovered [47]. Among the Hsp70 family, the prokaryotic form DnaK acquired from E.coli shares a sequence similarity of 60% with eukaryotic forms [50]. Despite the high conservation score, there are some diversifications causing a ternary classification of the family where DnaK acts as a model for one of the classes [44, 51]. On the nucleotide binding domain (NBD) the major differences can be grouped into two: (i) DnaK class members share a distinct sequence of a loop in the IIB subdomain with different characteristics than the other classes (corresponding residues are from A276 to R302 for DnaK) and (ii) there are several structural differences around the nucleotide binding cleft; DnaK class has a hydrophobic patch and salt bridges in this region. Interestingly, these variations are observed at the nucleotide binding sites of proteins from different classes which led to the idea of building a connection between structure variations and the wide range of nucleotide association/disassociation rates between family members [44]. On the substrate binding domain (SBD), the interaction with the substrate is fundamental for chaperone activity of DnaK. In a three dimensional structure of the SBD [Protein Data Bank (PDB) code 1DKZ], the peptide NRLLLTG is enclosed by loops and a helix acting as a lid. Three elements of the architecture at this site are crucial for substrate binding mechanism: (i) the hydrophobic pocket, (ii) the hydrophobic arch, and (iii) the helical lid [52]. R536, N537, Q538 form a hinge for the helical lid that controls the rigid body rotation of the C-terminal helical subdomain [33]. The allosteric

communication between the NBD and the SBD has been well-studied [44]. The binding of ATP causes the SBD to shift from a closed to open conformation to allow substrate entry while the open lid provides space for substrates for binding; in its closed state the substrate is enclosed in the cavity by the lid [52]. We first follow the structural changes caused by the point mutations, and then quantify how these changes are reflected to the local and global network properties. Experimentally, the effect of point mutations may be revealed for those cases where the mutant is expressed and analyzed for its functional consequences, e.g. changes in binding affinity [52, 49]. While these studies have led to enhancing our understanding of how local challenges to the protein structure propagates, in most cases they are limited to the mutations near the active site.

4.2.1 Beyond Thermal Fluctuations

We compare in figure 4.2, the residue correlations obtained from GNM (i.e. the $\mathbf{\Gamma^{-1}}$ matrix) and the MuMi scheme (i.e. the $\mathbf{D}$ matrix). While the displacements are in agreement with the theoretical calculations of atomic fluctuations (figure 4.2a), there is additional information in some regions resulting from the mutations. This hints that residue displacements emerging from the mutation process are not dominated by packing only. Further, we analyze the correlation matrices of residue fluctuations. Figure 4.2b displays a comparison of two correlation matrices $\mathbf{C}$ and $\mathbf{\Gamma^{-1}}$ which are computed by the two different approaches. $\mathbf{\Gamma^{-1}}$ matrix (the lower diagonal) highlights the regions of large fluctuations. Residues 228-310 that make up the IIB subdomain of the NBD, 400-420 and 457-502 in the SBD lining the substrate binding interface and 525-600 in the lid domain display large fluctuations, but none of these regions are cross-correlated. Following the MuMi scheme (the upper diagonal in figure 4.2b), we find that all these regions are connected through an intricate network of interactions which are actuated by the mutations. Thus, structural changes due to mutations propagate through residues which are spatially far away from the mutated residue. The above observation is quantified by the joint histogram of distance from mutated residues to all others and their displacements (figure 4.2c). We find that while the largest number of displacements on the order of 2-3 Å occurs at residues ca. 12-25 Å away from the mutated site, there are many instances

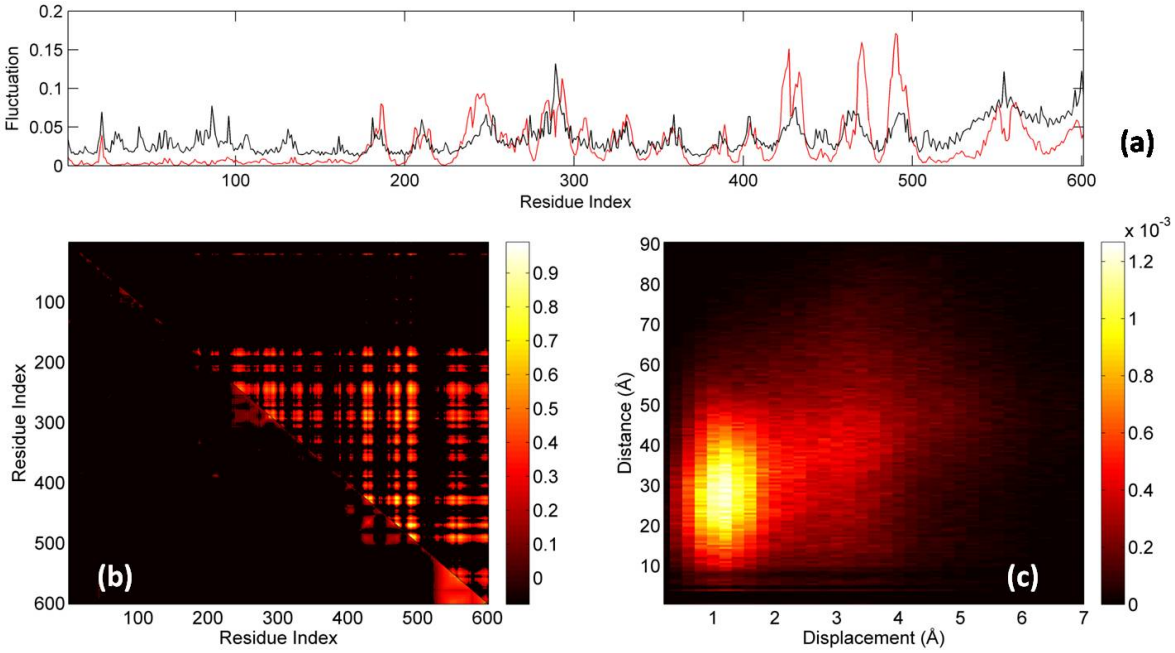


Figure 4.2: (a) The diagonal elements of Γ^{-1} are superposed with resulting $\mathbf{D}$ vector from our calculations. $\mathbf{D}$ is the square root of the diagonal of $\mathbf{C}$. Data are normalized by the total area under each curve for proper comparison.(b) Correlation matrices from two different methods are displayed as a single matrix containing $\mathbf{C}$ at the upper triangle and Γ^{-1} at the lower triangle. $\mathbf{C}$ and Γ^{-1} are thresholded by the summation of their mean and twice the standard deviation to simplify the view. (c) Joint histogram of distance from mutated residue to all others and their displacement upon mutation.

where a displacement on the order of 6 Å is observed at distances larger than 40 Å. We therefore seek to understand the information content provided by the MuMi scheme in the next subsection by projecting the results onto various network parameters described in Section 4.1.3.

4.2.2 Structure-Function Relation

Proteins can be very tolerant to mutations [53], but physical insight lacks on what makes a mutation endurable while others catastrophic [39]. Information on evolutionary conservation of amino acids has nevertheless advanced our understanding of the problem [28]. There is plenty of experimental evidence that modifications in the conserved residues are more likely to disturb the functionality of the protein. Coevolution data on residues occupying distant locations has led to attaching a functional link between regions of the protein [48, 54, 55]. Complicating the problem further is that some mutations in non-conserved sites might also damage the biological activity [39]. What makes those residues vulnerable to changes is unclear [4]. In addition, although conservation provides insights into residues of significance in the protein structure/function, it does not distinguish between stabilizing, folding or functional role that may have been taken on by a conserved site. With our method, we aim to classify the biological significance of amino acids by inspecting the defined measures.

Point Mutation Induced Large Local Rearrangements are Clustered on Sites Directly Involved in Binding

The average effect of the Ala-mutation sweep on residue displacements, quantified by D_i, is less than 3 Å for two thirds of the residues while it is larger than 6 Å for 10 of them. The distribution of D_i for the full length protein is displayed in figure 4.3a; the identities of the latter are listed in 4.1. The large local changes are confined to specific regions of the protein and the residues are highly conserved. We note that, while we use a cutoff of 6 Å for D_i to display residues in 4.1, including a less stringent value only includes additional residues in the same region as those already listed. On the ATPase domain, K294 on sheet I of the NBD is displaced by 6.5 ± 4.0 Å on average, during the alanine sweep. It is also

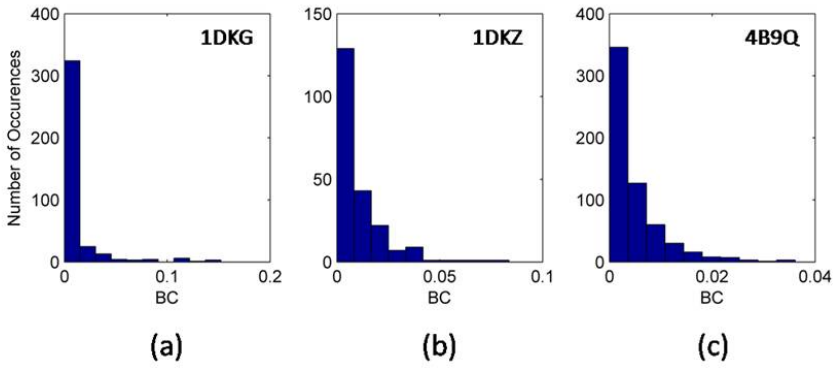


Figure 4.3: (a) Histogram of the average displacements of residues due to mutations in he MuMi analysisvector (b) Histogram of ΔL values from MuMi analysis using Eq. 4.9

highly conserved CONSURF score of 9 for both) and is part of the so-called GrpE signature motif [51]. While the mode of action of this motif is not known, altering residues by loop substitution with those from inactive Hsp70 forms affects ADP dissociation rates from the cochaperone GrpE acting as a nucleotide exchange factor by 5000-fold. This implies that the mutation sensitive loop is essential for the physical GrpE-DnaK interaction, despite having no direct contact in the x-ray structure. Based on this loop being solvent-exposed (figure 4.4a, loop containing K294 displayed in red), it was speculated that it might act as a latch to direct the binding [51, 44]. Such a role necessitates extreme flexibility while being

Table 4.1: Residues displaying significant position deviations (D_{ii}, equation 4.4) upon mutation

Residue Index	Significance	CONSURF Score
K294	Conserved structural site [44], part of the so-called GrpE signature motif [51]	9
S427	Substrate binding cavity [56, 57], S427P mutant effective in allosteric communication with the NBD [58]	8
T428	Substrate binding cavity [45]	9
M469,P470,Q471,I472	Substrate binding cavity [56, 57]	5, 9, 9, 9
D490, K491, N492	Substrate binding cavity [46]	8, 7, 5

sensitive to changes in the environment of the protein as manifested in the displacement of this residue, D_{294}. On the SBD, the highly conserved residues S427 and T428 on β sheet B and directly contacting the ligand have high D_i. Supporting these on the same β sheet, the highly conserved stretch of residues spanning M469 to I472as well as D490, K491,

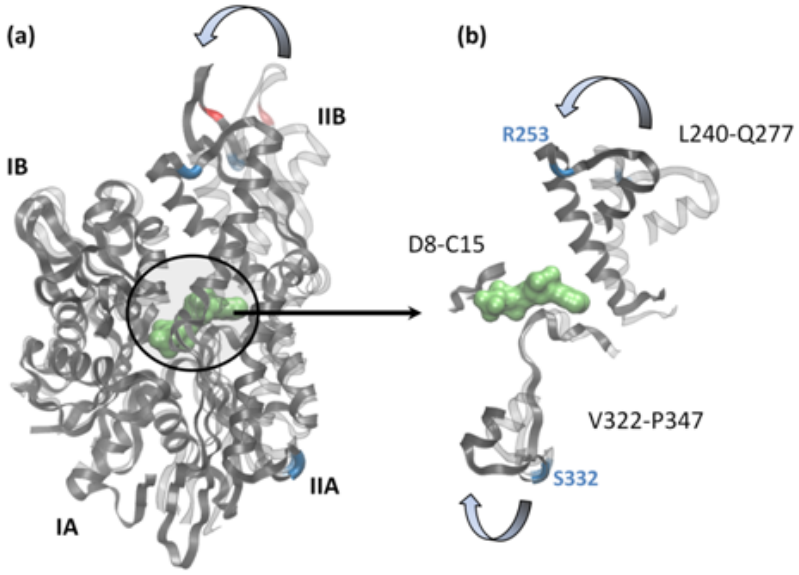


Figure 4.4: Highlighted sites on the NBD domain emerging in D and L analysis (red and blue, respectively) as well as BC (orange). (a) The NBD aligned in the nucleotide free (1DKG; transparent) and bound (4B9Q; opaque) form. Peptide is shown in green surface representation. Residues that appear in the L analysis only are shown in blue. K294 is shown in red. The four domains of the NBD are labeled. (b) A closer examination of the structure supporting ATP which is held by, (i) the loop containing residues D8-C15, (ii) the helix spanning L240-Q277, and (iii) the loop spanning V322-P347. While the structure of the first loop is intact in ATP bound – free forms, the helix and the latter loop move upon ATP binding. S332 and R253 are positioned at the base of these structures (shown in blue) and redirect the movement while their first neighbors remain intact. In particular, R253 is responsible for controlling the large closing motion of domain IIB upon ATP binding, highlighted by the arrow in part (a).

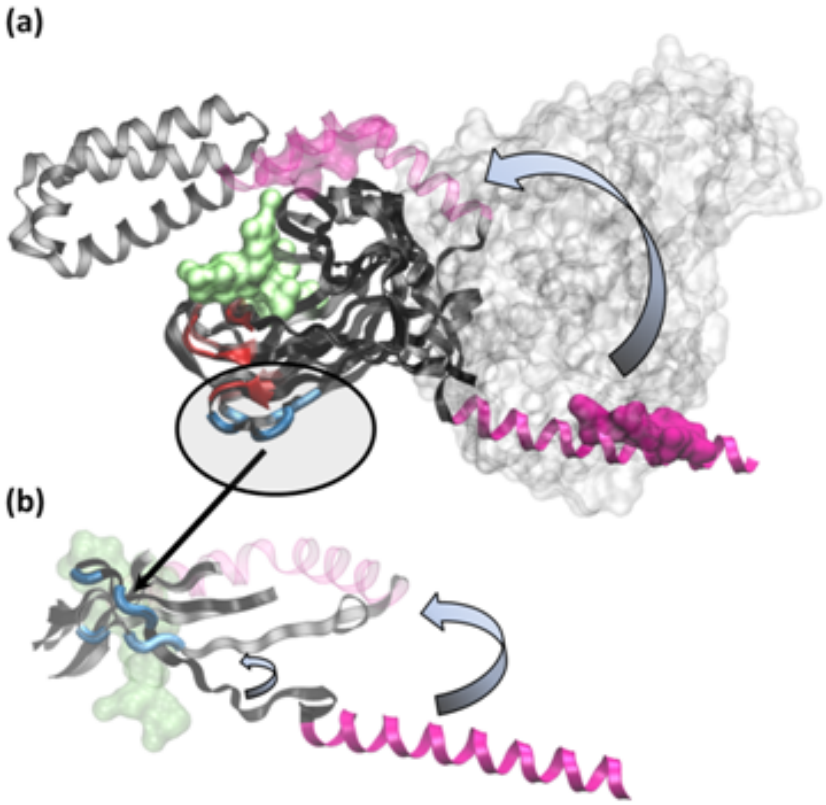


Figure 4.5: Highlighted sites on the SBD domain emerging in D(red), ΔL (blue) and BC analyses of the full structure. (a) The SBD aligned in the peptide bound (1DKZ; transparent) and unbound (4B9Q; opaque) form. Peptide is shown as green surface; the substrate binding region is tightened with a grip over the peptide. In the peptide bound (apo) form, the linker is extended; residues beyond 535 are not shown for this. Part of the linker that is displayed for the apo structure is colored in magenta on both forms (residues 510 – 535). The residues that appear in the D analysis are shown in red; they support the peptide via beta sheet B. Those that appear in L analysis only are shown in blue. Finally, residues displaying large BC are displayed as magenta surfaces. (b) Displayed from below, the part of the beta sheet which shows large L variations (blue) is displaced such that only the directionality of the following strand is different from the rest of the beta sheet in the apo form, having lost its hydrogen bonding pattern.

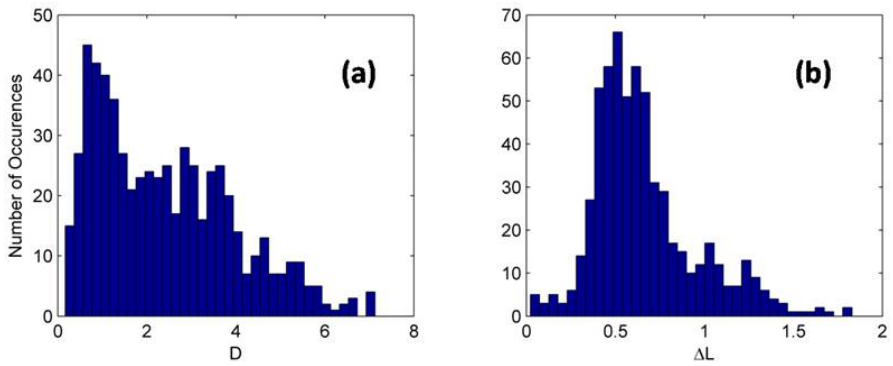


Figure 4.6: Histograms of BC values from (a) NBD, (b) SBD and (c) the full protein

N492 are affected by the mutations. Evidently, this functional region is structurally much less tolerant to mutations occurring anywhere in the protein. These are shown by the red stretch of residues on figure 4.5a. Thus, the lack of tolerance to mutations in this protein is directly related to the stability of regions on the protein that are directly involved in specific binding of substrates.

Point Mutation Induced Changes in Long Range Residue Reachability (ΔL) Highlight Sites Steering Large Subunits.

The effect of the Ala-mutation sweep on average connectedness of a given residue, quantified by change in the reachability of the residue ΔL_i, is less than one and a half steps change for most of the residues. However, the average reachability is shortened by more than 1.5 steps for a subset of residues (see figure 4.6b). ΔL_i are displayed in figure 4.7b and while they overlap with regions implicated by the change in D_i (compare figures 4.7a and 4.7b), others also appear. In fact, it is expected that those residues that are displaced by point mutations occurring all over the protein shall also have large changes in their reachability due to the rearrangements in their local network structure. K294 on the GrpE signature motif [51] and residues on β sheet B contacting the substrate (T428 and D490) appear as having extreme values with this global measure as well as emerging from D_i (compare Tables 4.1 and 4.2). In addition to those detected by simple amino acid displacements, R253 and S332 on the NBD and S493, K495, E496 on the SBD are also

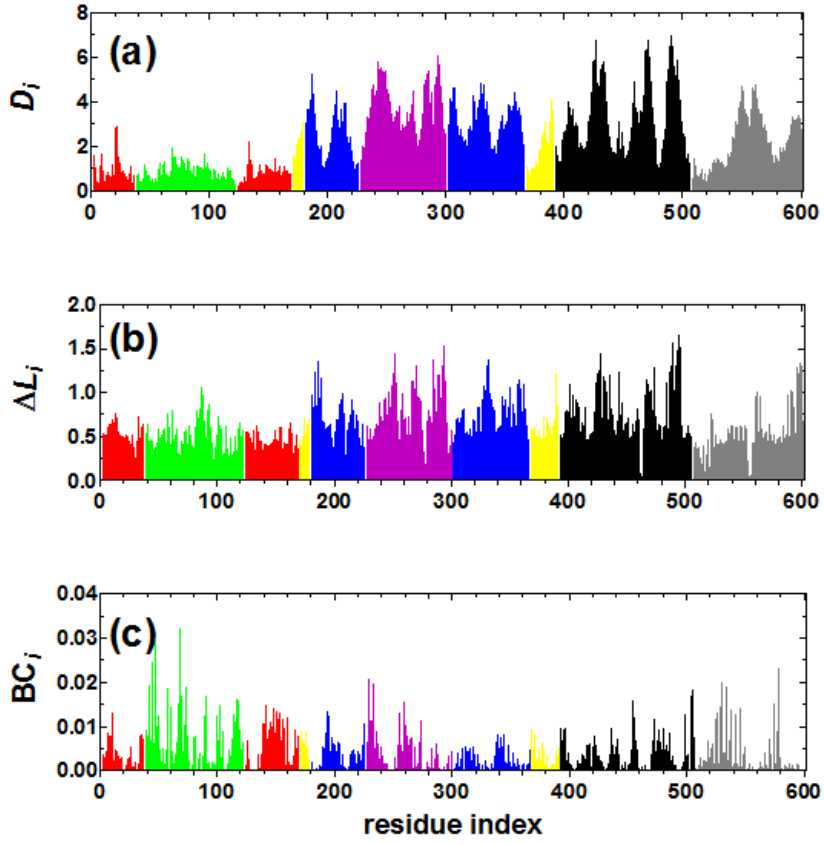


Figure 4.7: MuMi results for DnaK (a) residue displacements (D_{ii}, equation 4.3), (b) change in the average reachability of a residue upon mutation (ΔL, equation 4.9), and (c) betweenness centrality (BC) of the residues in WT structure. Residues with maximum values are listed in Tables 4.1, 4.2 and 4.3. along with possible roles in their structure. Spikes are colored according to subdomains in the NBD (IA: red, IB: green, IIA: blue, IIB: magenta, all others: yellow) and in the SBD (lid domain: gray, and the rest in black).

Table 4.2: Residues displaying significant deviations in reachability (ΔL_i, equation 4.9) upon mutation

Residue Index	Significance	CONSURF Score
R253	Distal switch at the end of helix on IIB for nucleotide binding/release [35, 59, 60],	1
K294	Conserved structural site [44], part of the so-called GrpE signature motif [51]	9
S332	Distal switch on IIA for nucleotide binding/release [this work]	1
T428	Substrate binding cavity [45]	9
D490,S493,K495,E496	Substrate binding cavity [46]	8, 8, 7, 3

uncovered by the global network analysis. Unlike the GrpE signature motif on the NBD and the β sheet B on the SBD, the local environment of these residues is less disturbed by point mutations, making them locally robust to perturbations of the protein. However, that their reachability is hampered by this procedure implies considerable rearrangements in the rest of the protein so as to keep the related neighborhood intact. For example, on the NBD, R253 and S332 are positioned in locations that may induce the motion of the structures that hold the bound ATP in place (figure 4.4b). While they do not directly contact the substrate, they act as a lever to steer the supporting helix/loop. On the SBD, residues on three strands of β sheet B appear as displaying large Di upon mutation (figure 4.5a). Their role is to support and stabilize the substrate. However, the loop connecting into the fourth strand, although having relatively smaller D_i, displays large ΔL_i for residues S493, K495, E496. This part of the sheet is connected to the strand that steers the lid domain (figure 4.5b). In peptide free form, this strand is not hydrogen bonded to the rest of the structure and the lid is open. In peptide bound form, it is steered into a position that locks the hydrogen bonding network and hence guides the lid domain towards the closed position. Note that, the residues that emerge solely from the ΔL_i analysis as being largely affected by alanine scan are not typically conserved. They occupy structural sites that do not require specificity. This is contrary to residues with large D_i whose contact structure is largely disturbed. Since these also hold locations that require specificity.

Residues Controlling Communication Within and Between Domains are Identified by Betweenness Centrality.

Insofar as average path lengths determine sites responsible for inducing inter-unit communication, we next study individual paths connecting pairs of residues to determine those that are key in controlling the communication. Our previous work has shown that a protein may be considered as an essential network of interactions overlaid by a large set of redundancies [22]. For all three structures studied in this work (4B9Q, 1DKG, 1DKZ), shortest paths are constructed from the homogeneous networks and the statistics of the residues that lie on all paths are made. The distribution of the BC values for the separate

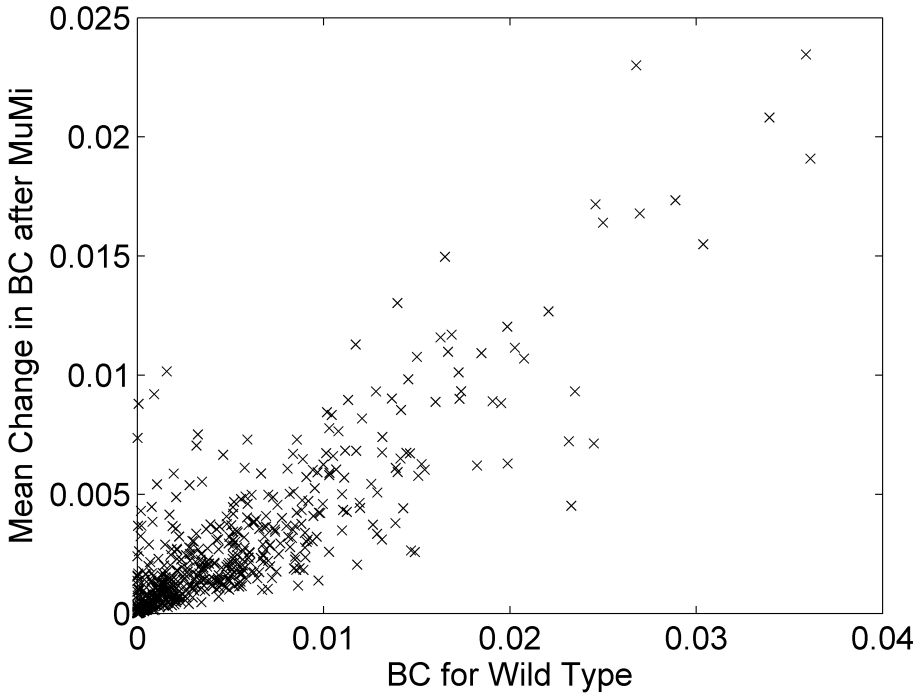


Figure 4.8: The linear relationship between BC values computed using WT structure and average amount of change in BC computed using all mutants after MuMi analysis. Residues that are displaying largest variation are identical with residues with highest BC in the WT.

domains and the full structure are displayed in figure 4.6b. Those sites that are crossed most often (having high BC) reveal locations involved in the control of communication within and between the domains as we outline in detail below (see also 4.3). We note that residues displaying largest variation in BC upon MuMi are identical with those which have the highest BC in the full WT structure (see figure 4.8). For this reason, our further analysis regards the significance of residues that exhibit highest BC for protein structure and function. These are displayed in figure 4.7 for DnaK, as calculated from the full structure. The residues displaying extreme BC values are listed in Table 4.3 for the full structure and the separate domains along with their known specific roles, if any. We find that the residues with high BC are distributed throughout the protein and their values are significantly affected by the presence of the other domain. In addition to BC analysis

from the WT structure, we compute BC values of the mutant structures. We observe a linear relationship between BC values in WT structure and the amount of change they undergo after MuMi analysis. The linearity implies that residues with highest BC values (for full structure, listed in 4.3) also display largest variation. For all five cases, the BC values decrease heavily upon any single mutation causing those to lose their characteristic of being most central nodes in the network.

Interdomain communication from BC of full structure. The full DnaK structure with a direct interaction between the NBD and SBD has been determined in the open conformation (4B9Q). Therein, ATP is bound while the substrate binding cavity is empty and the lid is open, ready for the binding of substrate. In the BC analysis, residues that appear to be most visited in the paths of the full structure are T48, I69, D233, D526, E530 (4.3). On the NBD, all three listed sites are involved in interdomain communication under different scenarios. For example, Hsp70 partners with ClpB to rescue stress-damaged proteins trapped in an aggregate. T48 is positioned on the binding interface of ClpB, but not GrpE which competes with ClpB for DnaK [61]. DnaK acts during the initial stage of rescue by exposing protein segments from the aggregate. Direct interaction between DnaK NBD and ClpB is thought to bring the exposed chain ends to ClpB for unfolding and threading of the chain [61]. The large BC of T48 for the full length DnaK highlights its role in the interdomain communication process at the initial stage of this mechanistic model of protein disaggregation. D233A mutant displays the largest increase in intrinsic ATPase activity and DnaJ stimulation (3.4 and 4.1 fold increase compared to the WT, respectively) amongst the 29 mutants on the NBD designed to test the impact of stimulated ATPase rate on the folding process [62]. Along with R71A, this mutant also displays the largest, albeit modest (10-13%), increase in luciferace refolding activity. In addition to being in close proximity to R71 mentioned above, I69 is a site coupled to the binding of the VLLL sequence that is known to be necessary and sufficient to impose the allosteric control of ATPase activity by the SBD [46]. With their large BC values, these sites are thought to hold central positions in orchestrating the coupling in DnaK. On the SBD, D526 and E530 emerge as having the largest BC when the full length protein is used in the analysis. Both these residues lie on the lid, interfacing the NBD in the peptide

Table 4.3: Residues displaying significant deviations in betweenness centrality (ΔBC_i)

Residue Index	Significance	CONSURF Score
T48	Hsp-100 (ClpB) binding [61]	8
I69	Site coupled to linker docking [46]	9
G184,T185	Positioned on exposed loop, having multiple (j,y) positions on the Ramachandran map [63]; they co-evolve with the b-sandwich core in the SBD [64].	1,1
D233	Affects intrinsic ATPase rate and DnaJ stimulation [62]. Resides in the ATP binding cavity	1
V337	V337F is a directed evolution product for efficient refolding of CAT_Cd9 [65].	6
G405	Substrate binding site [66, 48]	9
V407	Substrate binding site [57]	8
D526	Substrate affinity effects [66, 67], Lid domain [20, 46-48]	9
E530	Lid domain [68, 45, 69, 70]	2
V533	Lid domain [20, 46-48]	3

free form and the SBD in the peptide bound form (figure 4.5a, magenta surfaces). In both cases, they facilitate the communication between different regions: NBD and SBD domains in the former scenario, and the lid and the bound peptide in the latter. Besides being one of the three key elements of the substrate binding domain, the lid is involved in the ATP controlled substrate release, essential for chaperone activities of DnaK. Moro and coworkers studied the alpha helices in SBD and showed that the lid is very important in controlling the stability of protein-substrate complex and functioning of the complex [69, 68, 70]. They carried a set of truncation experiments including DnaK 1-537 and DnaK 1-507 mutants. The changes that occur upon nucleotide binding was observed in DnaK 1-537, but not in DnaK 1-507, thus revealing the importance of residues between 507 and 537. On the other hand, DnaK 1-507 mutant is found to have approximately the same values for peptide affinity with WT, similar loss of peptide binding affinity in the ATP bound state and can stimulate ATPase activity upon substrate binding [49]. One of the candidate sites for interaction of DnaJ with DnaK involves residue D526 which has the highest BC value in the open conformation. D526N mutant was suggested to have an increased on-rate for substrate by affecting the lid opening, thus mimicking the effect of ATP [67]. Strikingly, D526N mutant is found to alter kinetics of interaction with the substrate; changing substrate on/off rates without changing the KD of the reaction attributes a highly specific role to D526. Despite the communication between the SBD and the lid not being necessary for some functions, it enables SBD to trap substrates by closing upon

the substrate binding cavity. To conclude, the lid seems to be unimportant for several functions of DnaK such as interdomain allostery, but α helical parts of the lid are also essential for peptide kinetics and peptide-SBD interactions [45, 68].

Intradomain communication from BC of separate domain structures. We next treat the NBD and SBD separately to decipher key residues responsible for intradomain communication through BC analysis (4.3). For the nucleotide-free state of the NBD (PDB code 1DKG), residues G184 and T185 are observed to have the highest BC, despite them being located on an exposed loop with few contacts. The loop is flexible to the extent that in some chains of the 1DKG and 4B9Q x-ray structures, G184, and residues 181-187 are not resolved, respectively. In certain X-ray structures of the NBD of Hsp70s, the loop is positioned so that the VLLL motif in the linker region might dock through this loop (see, e.g. PDB structure 2QWO). The motif is known to both mediate the interaction with J-domain (from Hsp40 partner proteins) and to couple the NBD and SBD functions [71]. Binding of J-domain to Hsp70 disengages SBD from the NBD, rendering it conformational freedom for capturing substrates [72]. Furthermore, the loop containing T185 is crucial in that when it is substituted into eukaryotic Hsp70, it loses its interaction with the chaperon in CCT. When the reverse is done (eukaryotic loop is substituted into DnaK), CCT binding is observed [73]. The current BC results demonstrate that even in the absence of substrate binding, the NBD domain communicates through the flexible loop containing G184 and T185. Another residue with high BC that is critical in intradomain communication in the NBD is V337 (4.3). While this residue is moderately conserved (with a CONSURF score of 6, 4.3), it is one of the six residues on the NBD that is affected by directed evolution through the use of a folding-deficient C-terminal truncated choloramphenicol acetyl transferase (CAT_Cd9) for enhanced chaperone function [65]. Moreover, it is the most stable of the mutants as quantified by the two well-defined thermal unfolding temperatures in the absence of nucleotide. Nevertheless, this mutant is deficient in refolding luciferace, unlike other directed evolution products on the SBD lid domain. Thus, V337 must be effective in the selectivity towards substrates. Results for the SBD are also remarkable, because every residue having extreme BC is located at crucial functional regions. Kinetics measurements imply the opening rate of the substrate

binding cavity, bordered by the lid and the arch, limits the substrate association to the ADP bound state of DnaK [52]. Thus, residues located near the cavity, arch and the lid have kinetic control over the biological activities of DnaK. The BC findings from the peptide bound SBD are clustered in two regions, one on the lid domain and the other along the substrate binding cavity. Moreover, these two regions are in direct contact with each other (the top four BC residues are displayed in magenta in figure 4.5a). The hinge like structural element formed by the helix is crucial for substrate affinity and substrate release mechanism as verified experimentally where the function of truncated mutant DnaK (2-538) has been investigated [52]. R536 and V533 corroborate their unique positioning in controlling the communication between residues beyond Q538 with the rest of the SBD. These two residues are bridged over to V407 and G405 positioned around the substrate binding site. Direct contact between the two regions occurs via a hydrophobic patch involving V533, G405 and V407. G405S and M408I mutants were experimentally shown to have some diminishing effect on peptide binding [66]. M408 has the key side chain which forms the hydrophobic core between loops 4 and 5 and it was proposed that stabilizing contacts in the region would be disrupted upon changing its side chain [66]. One of the possible interaction sites between DnaJ which catalyzes the nucleotide hydrolysis step and DnaK is positioned near the substrate binding pocket [67]. This idea has been put forth by point mutation experiments, whereby G400D and G539D mutants are defective in peptide and DnaJ binding affinity. These residues, lining the substrate binding cavity and participating in lid domain interaction, respectively, are crucial for completing the biological cycle of DnaK. We note that, in the same study, the D526N mutant is shown to be defective in the rate of substrate/DnaJ binding. This residue has enhanced BC only in the full structure and not in the SBD analysis. Results are displayed in figure 4.7 with respect to mutated residue index, and those that display significant deviations are also listed in Tables 4.1, 4.2 and 4.3.

4.3 PDZ Domain: Another Case Study

We have applied MuMi to another relatively simple structure when compared to HSP70 to see whether the main information content provided by the method holds. This is the

third PDZ domain from the synaptic protein PSD-95. In particular, we seek to see if we may acquire information about the structure, beyond that provided by residue auto-correlations (i.e. B factors) and cross-correlations (provided by ANM or GNM methods).

4.3.1 Third PDZ Domain from the Synaptic Protein PSD-95

The third PDZ domain from the synaptic protein PSD-95 structure is determined in complex with its peptide ligand at 2.3 Å resolution by x-ray crystallography [74]. Corresponding residue numbers for the third domain are between 301 and 415 (PDB code 1BE9). PDZ domains are crucial for their role in mediating the clustering of membrane ion channels by binding to their C-terminus [74].

After the ALA mutation scan, the pairwise RMSD values between the WT structure and the mutant structures is in the interval of 0.79-1.05Å with an average of 0.89Å. Residues with index of R309, G319, S320, D332-G335, N381 appear to be significant in both methods. In addition, residue Q384 found to be highly fluctuating by GNM while residues Q391, A378 and S409 are found to have elevated correlation values with many other residues in the protein by MuMi method. The residue correlations obtained from GNM (i.e. the $\mathbf{\Gamma^{-1}}$ matrix) and the MuMi scheme (i.e. the $\mathbf{D}$ matrix) are compared in figure 4.9.

Results of full single-mutation study on PDZ domain

McLaughlin et al. generated all possible single point mutations that can occur in PDZ domain. Mutating each amino acid position to other 19 types for a protein with size 83, ends up with $83 \times 19 = 1577$ mutations. There are 20 positional mutations that cause residues to lose their functional roles (see figure 4.10 caption for full list of 20 positions). None of the mutations is found to cause residues to gain any function. These 20 residues can be considered as *significant* residues since upon their mutation loss-of-function appears. This extensive study can form a baseline to validate computational mutation studies with these experimental findings. Thus, we utilize the results by applying MuMi analysis on PDZ domain to test our measures and findings.

Starting with the WT structure, the 20 residues are mapped on several measures as

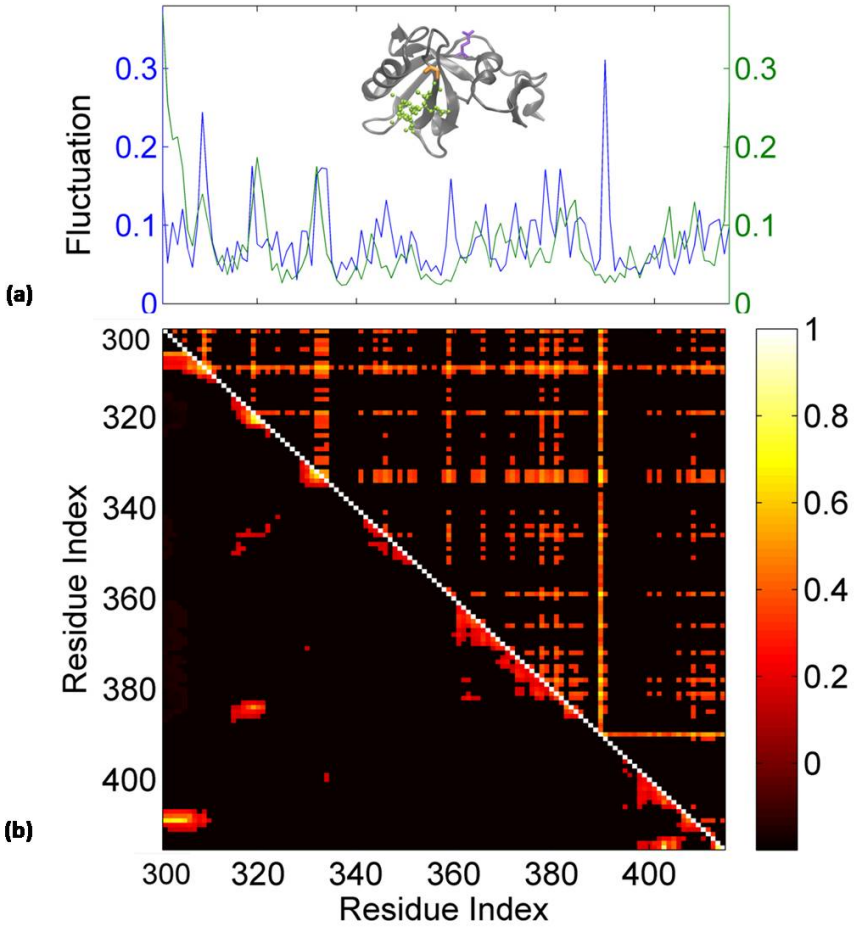


Figure 4.9: (a) Diagonal elements of $\mathbf{C}$ and $\mathbf{\Gamma^{-1}}$ for 1BE9. At the inset, PDZ domain structure and its peptide (in green) are displayed with residues having the highest (in purple) and lowest (in orange) fluctuations (Q391 and G329, respectively). (b) Comparison of two the correlation matrices: $\mathbf{C}$, computed with MuMi, is displayed in the upper triangle and $\mathbf{\Gamma^{-1}}$, computed with GNM, is displayed in the lower triangle. Diagonal is deliberately shown in white for clear visualization of the distinction between the off-diagonal terms in the two matrices. Matrices are thresholded for a clear view. Threshold value is computed as the sum of the mean value and the standard deviation of matrices.

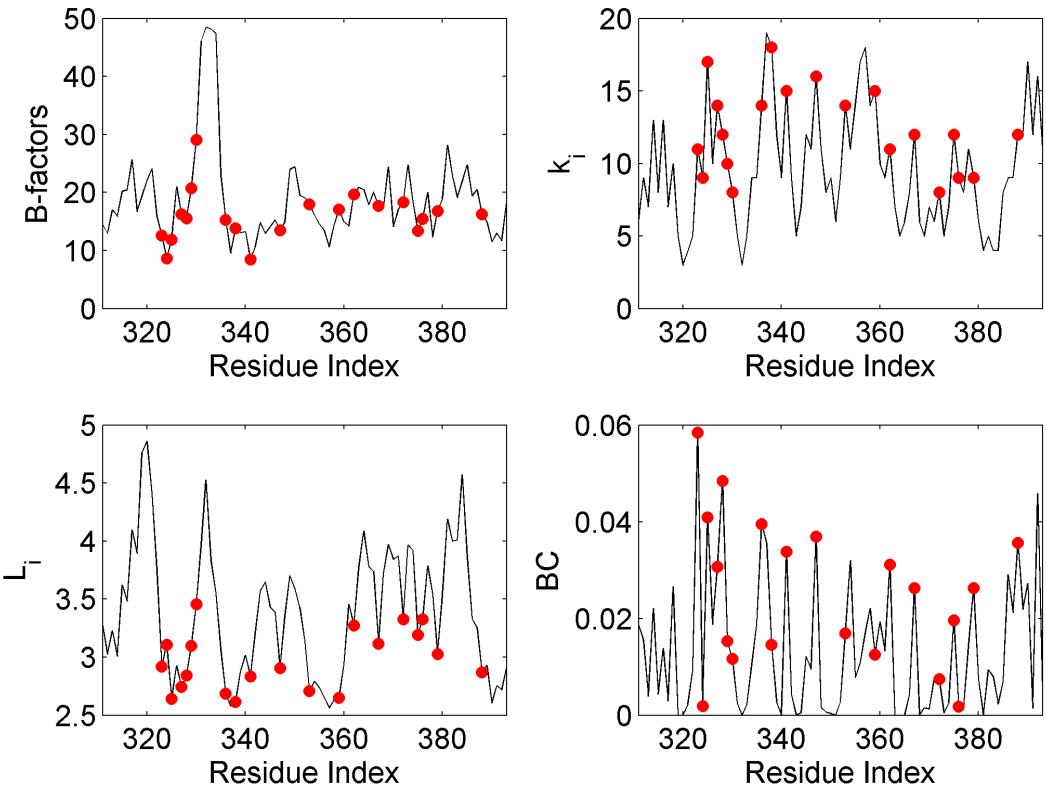


Figure 4.10: The 20 residues which are found experimentally [4] to cause loss-of-function are mapped as red dots on B-factors (from PDB file), degrees (k_i), average path length (L_i) and betweenness centrality (BC). The latest three are computed using the graph of the native structure (PDB code:1BE9). The complete list of 20 positions: 323-355, 327-330, 336, 338, 341, 347, 353, 359, 362, 367, 372, 375-376, 379 and 388.

displayed in figures 4.10 and 4.11. B-factors that are reported in PDB file does not seem a good identifier for the significant residues. The reason in that the 20 residues can have various values of B-factors and are not condensed in any specific intervals. We have previously discussed (in Section 3.3) that high connectivity of a node is a signal for its significance. We see there are 8 residues that have $k_i > 15$, which are 337, 338, 357, 325, 356, 390, 347 and 392, in the WT structure. Among them, only 338, 325 and 347 are found to affect functionality of the protein upon their mutation. The L_i graph shows that lowest values of L_i harbor many significant residues such as 325, 338 and 359. Thus, the residues that have smallest L_i can be considered as crucial for the biological activity. An interesting observation is made for the region between 360-380 where 7 *significant* residues (362, 367, 372, 375, 376 and 379) found to occupy lowest L_i in that region (but not globally lowest). Finally, all residues (323, 328, 325, 336, 347) that have the maximum BC values are found to be essential; a function is lost upon their mutation. Next, we perform MuMi analysis on PSD95$^{\text{pdz3}}$ structure and again mark the 20 positions on several measures as displayed in figure 4.11. Unlike HSP70, ΔD performs relative poorly for PDZ domain; residues (390, 319, 334, 378 and 381) that display maximum displacement upon mutations are not in full agreement with the experimental findings. They may have other functional significance which cannot be resolved by mutational studies. The residues (379, 376, 375, 325 and 323) experienced maximum change in average shortest path length,ΔL, and they are all pointed out by the mutagenesis study. As we observed earlier in our case study with HSP70, ΔBC and BC are correlated as displayed in figure 4.12. Among top five residues that have the maximum BC, 328, 392 and 325 also display maximum ΔBC and all are significant for the protein functionality. To evaluate the performance of our approach, the 20 residues that displayed maximum feature values (in case of L, minimum 20) are proposed to be significant. The number 20 is selected since the full mutagenesis study shows experimentally that in total 20 residues are significant. By using our measures, we can select top 20 residues that displayed extreme values in features and compare them with the 20 from the mutagenesis study. In table 4.4, the results are displayed for features from the native structure and the mutants (after the MuMi method). The best performing feature is ΔBC, followed by BC and the third best features are equally good: L and ΔL.

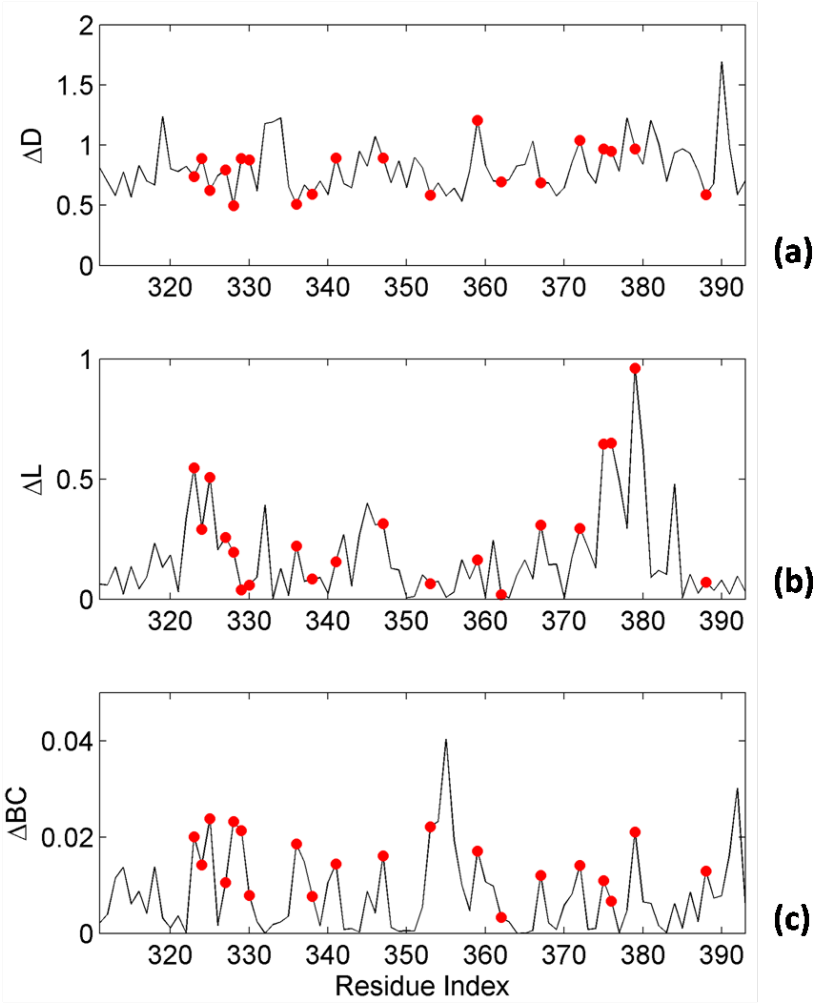


Figure 4.11: The 20 residues pointed out are mapped on the measures used in the MuMi analysis. (a) ΔD results are displayed where the residues that display maximum displacements are 390, 319, 334, 378, 381. (b) ΔL results are displayed where the residues that have maximum values are 379, 376, 375, 325 and 323. (c) ΔBC results are displayed where the residues that have maximum values are 328, 392 and 325. The importance of residues that display largest ΔD is still unclear. However, for ΔL and ΔBC measures, the significance of all top residues are verified by the complete mutagenesis study.

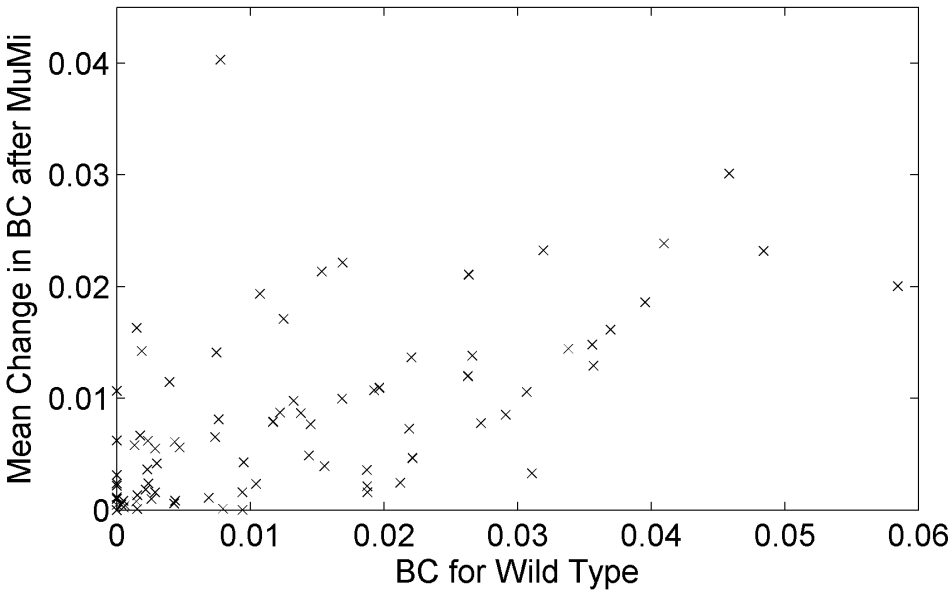


Figure 4.12: BC values of the WT structure are plotted against ΔBC. Although more scattered, the correlation between these values is significant with $R^2 = 0.48$. Thus, the residues that have higher BC also display largest deviation from their WT values after MuMi.

Table 4.4: The performance of features, as illustrated in figures 4.12 and 4.11, is given in detail. The abbreviations stand for, TP: true positive, TN: true negative, FP: false positive, FN: false negative.

	TP	TN	FP	FN
k	9	52	11	11
L	10	53	10	10
BC	11	54	9	9
B-factor	4	47	16	16
ΔD	5	48	15	15
ΔL	10	53	10	10
ΔBC	12	55	8	8

5

Conclusion

Developing simplified models of atomic systems solely using their known three dimensional structures provides a deeper understanding of their local and global properties. With the motivation of strong structure-function relation in proteins, we studied the spatial organization of amino acids to identify some interaction patterns that are not expected to be formed by chance. We believe such patterns are peculiar to proteins. We introduced a scheme for comparing proteins to non-protein structures. We developed a computational method to perform an alanine mutation scan.

We think residue networks resemble more to random graphs than they do to regular graphs. To test our idea, we generated synthetic graphs that have the same network properties with proteins: (i) a Poisson distributed degrees with mean 6 and (ii) an average clustering coefficient about 0.35. Then we analyze the differences between proteins and protein-like-synthetic networks. We find that proteins are indeed more similar to random networks in terms of clustering. However, we observe they inherit longer pathways that would be expected from their random counterparts. We think the average shortest path lengths increase due to the decrease in randomness of spatial organization of amino acid.

A thorough computational investigation of the three dimensional structures of Heat Shock Protein and third PDZ domain from the synaptic protein PSD-95 using a systematic Ala mutation scan reveals key sites that are essential in biological activities. The atomic fluctuations do not bring in valuable information additional to what one might obtain from a careful examination of the three-dimensional structure. We thus utilize residue

networks and put forward change in the reachability of residues upon mutation and their betweenness centrality in the original network structure as the network measures that are useful in distinguishing functional sites. Perturbations randomly arriving at the protein can have an influence in two ways: (i) The local neighborhood of the residue is significantly changed. Residue networks are Poisson distributed [24], but they have backlinks [75]. Although their paths are longer [22], the main influences are from the first and the second neighbors. (ii) The local neighborhood is relatively intact; however, there are global rearrangements in the rest of the protein that affect the average path length.

Bibliography

[1] D. Easley and J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[2] T. Ceska, M. Lamers, P. Monaci, A. Nicosia, R. Cortese, and D. Suck, "The x-ray structure of an atypical homeodomain present in the rat liver transcription factor lfb1/hnf1 and implications for dna binding." *The EMBO journal*, vol. 12, no. 5, p. 1805, 1993.

[3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[4] R. N. McLaughlin Jr, F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, "The spatial architecture of protein function and adaptation," *Nature*, vol. 491, no. 7422, pp. 138–142, 2012.

[5] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47–97, 2002.

[6] M. E. Newman, "The structure and function of networks," *Computer Physics Communications*, vol. 147, no. 1, pp. 40–45, 2002.

[7] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[8] P. Erdos, "Graphs with prescribed degrees of vertices (hungarian)," *Mat. Lapok*, vol. 11, pp. 264–274, 1960.

[9] R. Henson and L. Cetto, "The matlab bioinformatics toolbox," *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, 2004.

[10] M. E. Newman, "A measure of betweenness centrality based on random walks," *Social networks*, vol. 27, no. 1, pp. 39–54, 2005.

[11] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[12] A. Rapoport, "Spread of information through a population with socio-structural bias: I. assumption of transitivity," *The bulletin of mathematical biophysics*, vol. 15, no. 4, pp. 523–533, 1953.

[13] P. Erdös and A. Rényi, "{On the evolution of random graphs}," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, pp. 17–61, 1960.

[14] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[15] M. E. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, no. 2, p. 025102, 2001.

[16] ——, "Random graphs with clustering," *Physical review letters*, vol. 103, no. 5, p. 058701, 2009.

[17] S. Dorogovtsev, J. Mendes, and A. Samukhin, "How to generate a random growing network," *arXiv preprint cond-mat/0206132*, 2002.

[18] M. E. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Physical review E*, vol. 64, no. 2, p. 026118, 2001.

[19] R. Milo, N. Kashtan, S. Itzkovitz, M. E. Newman, and U. Alon, "Uniform generation of random graphs with arbitrary degree sequences," *arXiv preprint cond-mat/0312028*, vol. 106, pp. 1–4, 2003.

[20] J.-L. Guillaume and M. Latapy, "A realistic model for complex networks," *arXiv preprint cond-mat/0307095*, 2003.

[21] E. Volz, "Random networks with tunable degree distribution and clustering," *Physical Review E*, vol. 70, no. 5, p. 056115, 2004.

[22] A. R. Atilgan, D. Turgut, and C. Atilgan, "Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication," *Biophysical journal*, vol. 92, no. 9, pp. 3052–3062, 2007.

[23] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[24] A. R. Atilgan, P. Akan, and C. Baysal, "Small-world communication of residues and significance for protein dynamics," *Biophysical journal*, vol. 86, no. 1, pp. 85–91, 2004.

[25] C. Klein, A. Marino, M.-F. Sagot, P. V. Milreu, and M. Brilli, "Structural and dynamical analysis of biological networks," *Briefings in functional genomics*, p. els030, 2012.

[26] A. R. Atilgan and C. Atilgan, "Local motifs in proteins combine to generate global functional moves," *Briefings in functional genomics*, p. els027, 2012.

[27] M. Vijayabaskar and S. Vishveshwara, "Insights into the fold organization of tim barrel from interaction energy based structure networks," *PLoS Comput Biol*, vol. 8, no. 5, p. e1002505, 2012.

[28] G. Celniker, G. Nimrod, H. Ashkenazy, F. Glaser, E. Martz, I. Mayrose, T. Pupko, and N. Ben-Tal, "Consurf: using evolutionary data to raise testable hypotheses about protein function," *Israel Journal of Chemistry*, vol. 53, no. 3-4, pp. 199–206, 2013.

[29] G. Bounova and O. de Weck, "Overview of metrics and their correlation patterns for multiple-metric topology analysis on heterogeneous graph ensembles," *Physical Review E*, vol. 85, no. 1, p. 016117, 2012.

[30] D. Bonchev, "A simple integrated approach to network complexity and node centrality," *Analysis of Complex Networks: From Biology to Linguistics*, pp. 47–53, 2009.

[31] H. M. Berman, T. Battistuz, T. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain *et al.*, "The protein data bank," *Biological Crystallography*, vol. 58, no. 6, pp. 899–907, 2002.

[32] C. J. Harrison, M. Hayer-Hartl, M. Di Liberto, F.-U. Hartl, and J. Kuriyan, "Crystal structure of the nucleotide exchange factor grpe bound to the atpase domain of the molecular chaperone dnak," *Science*, vol. 276, no. 5311, pp. 431–435, 1997.

[33] X. Zhu, X. Zhao, W. F. Burkholder, A. Gragerov, C. M. Ogata, M. E. Gottesman, and W. A. Hendrickson, "Structural analysis of substrate binding by the molecular chaperone dnak," *Science*, vol. 272, no. 5268, pp. 1606–1614, 1996.

[34] R. Kityk, J. Kopp, I. Sinning, and M. P. Mayer, "Structure and dynamics of the atp-bound open conformation of hsp70 chaperones," *Molecular cell*, vol. 48, no. 6, pp. 863–874, 2012.

[35] A. Zhuravleva and L. M. Gierasch, "Allosteric signal transmission in the nucleotide-binding domain of 70-kda heat shock protein (hsp70) molecular chaperones," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 6987–6992, 2011.

[36] W. Humphrey, A. Dalke, and K. Schulten, "Vmd: visual molecular dynamics," *Journal of molecular graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[37] M. Karplus *et al.*, "Charmm: A program for macromolecular energy, minimization, and dynamics calculations," *J Comput Chem*, vol. 4, p. 187217, 1983.

[38] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, "Scalable molecular dynamics with namd," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1781–1802, 2005.

[39] T. Teşileanu, L. J. Colwell, and S. Leibler, "Protein sectors: statistical coupling analysis versus conservation," *PLoS computational biology*, vol. 11, no. 2, pp. e1 004 091–e1 004 091, 2015.

[40] Y. Li-Smerin, D. H. Hackos, and K. J. Swartz, "α-helical structural elements within the voltage-sensing domains of a k+ channel," *The Journal of general physiology*, vol. 115, no. 1, pp. 33–50, 2000.

[41] M. C. Demirel, A. R. Atilgan, I. Bahar, R. L. Jernigan, and B. Erman, "Identification of kinetically hot residues in proteins," *Protein Science*, vol. 7, no. 12, pp. 2522–2532, 1998.

[42] I. Bahar, A. R. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential," *Folding and Design*, vol. 2, no. 3, pp. 173–181, 1997.

[43] A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophysical journal*, vol. 80, no. 1, pp. 505–515, 2001.

[44] M. Mayer and B. Bukau, "Hsp70 chaperones: cellular functions and molecular mechanism," *Cellular and molecular life sciences*, vol. 62, no. 6, pp. 670–684, 2005.

[45] M. Pellecchia, D. L. Montgomery, S. Y. Stevens, C. W. Vander Kooi, H.-p. Feng, L. M. Gierasch, and E. R. Zuiderweg, "Structural insights into substrate binding by the molecular chaperone dnak," *Nature Structural & Molecular Biology*, vol. 7, no. 4, pp. 298–303, 2000.

[46] J. F. Swain, G. Dinler, R. Sivendran, D. L. Montgomery, M. Stotz, and L. M. Gierasch, "Hsp70 chaperone ligands control domain association via an allosteric mechanism mediated by the interdomain linker," *Molecular cell*, vol. 26, no. 1, pp. 27–39, 2007.

[47] M. P. Mayer, "Hsp70 chaperone dynamics and molecular mechanism," *Trends in biochemical sciences*, vol. 38, no. 10, pp. 507–514, 2013.

[48] R. G. Smock, O. Rivoire, W. P. Russ, J. F. Swain, S. Leibler, R. Ranganathan, and L. M. Gierasch, "An interdomain sector mediating allostery in hsp70 molecular chaperones," *Molecular systems biology*, vol. 6, no. 1, p. 414, 2010.

[49] D. Sharma and D. C. Masison, "Hsp70 structure, function, regulation and influence on yeast prions," *Protein and peptide letters*, vol. 16, no. 6, p. 571, 2009.

[50] K. Richter, M. Haslbeck, and J. Buchner, "The heat shock response: life on the verge of death," *Molecular cell*, vol. 40, no. 2, pp. 253–266, 2010.

[51] D. Brehmer, S. Rüdiger, C. S. Gässler, D. Klostermeier, L. Packschies, J. Reinstein, M. P. Mayer, and B. Bukau, "Tuning of chaperone activity of hsp70 proteins by modulation of nucleotide exchange," *Nature Structural & Molecular Biology*, vol. 8, no. 5, pp. 427–432, 2001.

[52] M. P. Mayer, H. Schröder, S. Rüdiger, K. Paal, T. Laufen, and B. Bukau, "Multistep mechanism of substrate binding determines chaperone activity of hsp70," *Nature Structural & Molecular Biology*, vol. 7, no. 7, pp. 586–593, 2000.

[53] W. Baase, N. Gassner, X. Zhang, R. Kuroki, L. Weaver, D. Tronrud, and B. Matthews, "How much sequence variation can the functions of biological molecules tolerate," *Simplicity and Complexity in Proteins and Nucleic Acid. H. Frauenfelder, J. Deisenhofer, and PG Wolynes, editors. Dahlem University Press, Berlin, Germany*, 1999.

[54] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: evolutionary units of three-dimensional structure," *Cell*, vol. 138, no. 4, pp. 774–786, 2009.

[55] K. A. Reynolds, R. N. McLaughlin, and R. Ranganathan, "Hot spots for allosteric regulation on protein surfaces," *Cell*, vol. 147, no. 7, pp. 1564–1575, 2011.

[56] B. Bukau and A. L. Horwich, "The hsp70 and hsp60 chaperone machines," *Cell*, vol. 92, no. 3, pp. 351–366, 1998.

[57] S. Rüdiger, A. Buchberger, and B. Bukau, "Interaction of hsp70 chaperones with substrates," *Nature Structural & Molecular Biology*, vol. 4, no. 5, pp. 342–349, 1997.

[58] D. L. Montgomery, R. I. Morimoto, and L. M. Gierasch, "Mutations in the substrate binding domain of the escherichia coli 70 kda molecular chaperone, dnak, which alter

substrate affinity or interdomain coupling," *Journal of molecular biology*, vol. 286, no. 3, pp. 915–932, 1999.

[59] Y. Liu, L. M. Gierasch, and I. Bahar, "Role of hsp70 atpase domain intrinsic dynamics and sequence evolution in enabling its functional interactions with nefs," 2010.

[60] A. Nicolaï, P. Delarue, and P. Senet, "Decipher the mechanisms of protein conformational changes induced by nucleotide binding through free-energy landscape analysis: Atp binding to hsp70," *PLoS computational biology*, vol. 9, no. 12, p. e1003379, 2013.

[61] R. Rosenzweig, S. Moradi, A. Zarrine-Afsar, J. R. Glover, and L. E. Kay, "Unraveling the mechanism of protein disaggregation through a clpb-dnak interaction," *Science*, vol. 339, no. 6123, pp. 1080–1083, 2013.

[62] L. Chang, A. D. Thompson, P. Ung, H. A. Carlson, and J. E. Gestwicki, "Mutagenesis reveals the complex relationships between atpase rate and the chaperone activities of escherichia coli heat shock protein 70 (hsp70/dnak)," *Journal of Biological Chemistry*, vol. 285, no. 28, pp. 21 282–21 291, 2010.

[63] P. M.-U. Ung, A. D. Thompson, L. Chang, J. E. Gestwicki, and H. A. Carlson, "Identification of key hinge residues important for nucleotide-dependent allostery in e. coli hsp70/dnak," 2013.

[64] I. J. General, Y. Liu, M. E. Blackburn, W. Mao, L. M. Gierasch, and I. Bahar, "Atpase subdomain ia is a mediator of interdomain allostery in hsp70 molecular chaperones," 2014.

[65] R. A. Aponte, S. Zimmermann, and J. Reinstein, "Directed evolution of the dnak chaperone: mutations in the lid domain result in enhanced chaperone activity," *Journal of molecular biology*, vol. 399, no. 1, pp. 154–167, 2010.

[66] W. F. Burkholder, X. Zhao, X. Zhu, W. A. Hendrickson, A. Gragerov, and M. E. Gottesman, "Mutations in the c-terminal fragment of dnak affecting peptide binding," *Proceedings of the National Academy of Sciences*, vol. 93, no. 20, pp. 10 632–10 637, 1996.

[67] W.-C. Suh, W. F. Burkholder, C. Z. Lu, X. Zhao, M. E. Gottesman, and C. A. Gross, "Interaction of the hsp70 molecular chaperone, dnak, with its cochaperone dnaj," *Proceedings of the National Academy of Sciences*, vol. 95, no. 26, pp. 15 223–15 228, 1998.

[68] S. V. Slepenkov, B. Patchen, K. M. Peterson, and S. N. Witt, "Importance of the d and e helices of the molecular chaperone dnak for atp binding and substrate release," *Biochemistry*, vol. 42, no. 19, pp. 5867–5876, 2003.

[69] F. Moro, V. Fernández-Sáiz, and A. Muga, "The lid subdomain of dnak is required for the stabilization of the substrate-binding site," *Journal of Biological Chemistry*, vol. 279, no. 19, pp. 19 600–19 606, 2004.

[70] F. Moro, V. Fernández, and A. Muga, "Interdomain interaction through helices a and b of dnak peptide binding domain," *FEBS letters*, vol. 533, pp. 119–123, 2003.

[71] D. P. Kumar, C. Vorvis, E. B. Sarbeng, V. C. C. Ledesma, J. E. Willis, and Q. Liu, "The four hydrophobic residues on the hsp70 inter-domain linker have two distinct roles," *Journal of molecular biology*, vol. 411, no. 5, pp. 1099–1113, 2011.

[72] J. Jiang, E. G. Maes, A. B. Taylor, L. Wang, A. P. Hinck, E. M. Lafer, and R. Sousa, "Structural basis of j cochaperone binding and regulation of hsp70," *Molecular cell*, vol. 28, no. 3, pp. 422–433, 2007.

[73] J. Cuéllar, J. Martín-Benito, S. H. Scheres, R. Sousa, F. Moro, E. López-Vinas, P. Gómez-Puertas, A. Muga, J. L. Carrascosa, and J. M. Valpuesta, "The structure of cct–hsc70nbd suggests a mechanism for hsp70 delivery of substrates to the chaperonin," *Nature structural & molecular biology*, vol. 15, no. 8, pp. 858–864, 2008.

[74] D. A. Doyle, A. Lee, J. Lewis, E. Kim, M. Sheng, and R. MacKinnon, "Crystal structures of a complexed and peptide-free membrane protein–binding domain: molecular basis of peptide recognition by pdz," *Cell*, vol. 85, no. 7, pp. 1067–1076, 1996.

[75] D. Turgut, A. R. Atilgan, and C. Atilgan, "Assortative mixing in close-packed spatial networks," *PloS one*, vol. 5, no. 12, 2010.