NETWORK CHARACTERIZATION OF PACKING ARCHITECTURE FOR CONDENSED MATTER SYSTEMS

by

DENİZ TURGUT

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

Sabancı University
January 2011

NETWORK CHARACTERIZATION OF PACKING ARCHITECTURE

FOR CONDENSED MATTER SYSTEMS

APPROVED BY:

Prof. Canan Atılgan Canan Atilgan

(Thesis Supervisor)

Prof. Ali Rana Atılgan

(Thesis Co-Supervisor)

Prof. Uluğ Çapar

Prof. Ayşe Erzan

Assist. Prof. Cleva Ow-Yang

DATE OF APPROVAL: 17.01.2011

Abstract

Networks have currently been used to model real life complex systems and they have provided additional understanding for characterizing structure-function-dynamics relationships of these complex architectures. Here we investigate statistical and spectral properties and the connections between local motifs and global behavior of networks that are formed from condensed matter systems, particularly proteins, as well as micelles, polymeric melts and Lennard-Jones clusters.

Proteins are considered as interacting residue networks. Pathways for information transfer manifested in the average path lengths are analyzed, where the energy of residue-residue interactions are imposed as edge weights in networks. Systematic removal of “low energy” interactions reveals that the network contains significant number of redundancies that provide high local clustering. The information transfer is achieved by a small number of highly clustered groups of residues, which makes the hub architecture different from that of scale-free networks. This result is then extended to protein complexes, where two proteins (ligand and receptor) interact, in order to identify essential pair-wise interactions between two proteins.

In the presence of local clustering, establishing a relationship between local structure and global properties is far from trivial. But for certain cases, applying a bottom-up approach, a relation between nearest neighbors and next-to-nearest neighbors is obtained and this relation is observed in different networks formed from condensed matter systems, as well as perfect lattice models.

To further investigate the association between local order and global structure, residue networks are considered in further detail. To outline local order, we compared residue networks to perfect lattice systems by creating self-avoiding chains on chains via Metropolis Monte Carlo method that capture three dimensional structure of protein chains as much as possible. Results show that, proteins conform to close packed ordered structures with significant voids irrespective of the underlying lattice bases.

Finally, we analyzed the spectral properties of networks used throughout the thesis. Spectral changes while breaking and rewiring the edges revealed the importance and roles of short and long-ranged contacts in determining the network structure. Comparison of spectra distributions of different networks constructed from condensed matter systems supported the result from statistical parameters that these systems have structural similarities.

Özet

Ağlar, son zamanlarda gerçek hayatta karşılaşılan karmaşık sistemleri modellemek için kullanılmaya başladı ve bu ağlar bu karmaşık yapılardaki, yapı-işlev-dinamik ilişkilerinin nitelendirilmesinde önemli katkılar sağladı. Biz burada proteinler, miseller, polimer eriyikler ve Lennard-Jones öbekleri gibi yoğun madde sistemlerinden oluşturulan ağlardaki istatistiksel ve spektral özelliklerle birlikte yerel motifler ve genel davranış arasındaki ilişkileri incelemekteyiz.

Proteinler, etkileşen rezidü ağları olarak göz önüne alınırlar. Bilgi iletimi için kullanılan yollar, rezidüler arası etkileşim enerjilerinin bağlantı ağırlığı olarak modellendiği ağ yapılarında ortalama yol uzunluğu ile incelendi. "Düşük enerjili" etkileşimlerin sistematik olarak koparılması, ağ yapılarında yerel öbeklenmenin yüksek olmasını sağlayan yedek bağlantıların oldukça fazla sayıda olduğunu ortaya çıkardı. Ölçeksiz ağlardan farklı olarak, bilgi iletimi büyük oranda öbeklenmiş az sayıda grup arasındaki etkileşimler ile sağlanmakta. Bu sonuç iki proteinin (ligand ve reseptör) etkileşimi ile oluşan protein komplekslerine genişletilerek iki protein arası önemli etkileşim çiftlerinin tanımlanmasında kullanıldı.

Yerel öbeklenmenin mevcut olduğu durumlarda yerel yapı ile genel özellikler arası ilişki bariz değildir. Fakat, bazı özel durumlar için, tabandan başlayan bir yaklaşım ile ilk komşular ile bir sonraki komşular arasında bir ilişki türetildi ve bu ilişki yoğun madde sistemleri ve mükemmel kafes modellerinden elde edilen ağlarda gözlendi.

Yerel düzen ve genel yapı arasındaki ilişkinin daha fazla irdelenmesi için rezidü ağları detaylı olarak ele alındı. Yerel düzeni ortaya koymak için, rezidü ağları mükemmel kafes yapılarından Metropolis Monte Carlo metodu kullanılarak elde edilen kendi üzerine dönmeyen zinciler ile kıyaslandı. Sonuçlar proteinlerin kullanılan kafes yapısından bağımsız olarak önemli miktarda boşluk içeren yoğun düzenli yapılara uyduğunu gösterdi.

Son olarak tez boyunca kullanılan ağ yapıları spektral özellikler bakımından analiz edildi. Bağlantıların koparılması ve rastgele bağlanması sırasında gözlenen spektral değişimler kısa ve uzun menzilli bağlantıların ağ yapısını belirlemedeki önemlerini ortaya koydu. Yoğun madde sistemlerinden elde edilen ağların spektrum dağılımı kıyaslamaları, istatistiksel değişkenlerde elde edilen sonuçla paralel olarak, bu sistemlerin yapısal benzerlikler barındırdıklarını gösterdi.

Acknowledgements

Along the years, my advisor Canan Atılgan was the main driving force of this thesis. I can't appreciate her contributions enough. She somehow believed in me when I couldn't, and always pushed me for the better. I admire her patience, and I am grateful for all the effort she put in this work.

Vision of my co-advisor, Ali Rana Atılgan, guided this thesis. Ideas he propose and the discussions he raised nourished my inspirations that made this work possible. I thank him for his endless guidance since my years as an undergraduate student.

Also, I am thankful for the discussions and critisms of my thesis jury, Ayşe Erzan, Cleva Ow-Yang, Müjdat Çetin and Uluğ Çapar. Points and issues they have brought up provided much appriciated contributions for my thesis and increased the scale of the whole work.

Realizing this thesis would not be possible if it was not with the support and companionship of my friends. Although it would be next to impossible to list all the names, I couldn't go through without mentioning my dear friends: Işıl, Eren, Osman Burak, Kerem, Sinan, Burcu, Özge, Emre and Irmak. The simple word "friend" fails to capture my relationship with you, and I feel the Turkish word "dost" is more appropriate.

In all these years, my dear friends Özlem, İbrahim, Gökhan and Murat, whom I shared an office with, deserve more than my simple grace. You turned the office space and my time in it into a much better state. I am also thankful to many more friends who made my life in Sabancı more fun and enjoyable despite my usual dark mood. I consider myself lucky to have met you.

Simple page like this is far from enough to express my gratitute but in the long process that resulted with this thesis, most of the credit belongs to my family. Dedication of this work to you would be meaningles, since this mere collection of words is nowhere near the support you provided to me throughout the years. I could only hope to be worthy of your support and belief in me.

Table of Contents

List of Figures

List of Tables

Before to dust you shall return
There is one thing that you must learn
Sorrow and pain your soul shall burn
Joy and bliss to light shall turn

∼

Dünya dediğin bir bakışımızdır bizim
Ceyhun nehri kanlı gözyaşımızdır bizim
Cehennem, boşuna dert çektiğimiz günler
Cennetse gün ettiğimiz günlerdir bizim

– Ömer Hayyam

Chapter 1

Introduction

1.1 Background

For the last two decades, complex network study has gained a lot of importance in a wide range of areas. Understanding the structure of the World Wide Web [2, 3] is crucial to categorize and catalog the web pages to utilize efficient search mechanisms. Social scientists investigate social networks to understand information flow and relationships in large social systems such as movie actors [1], scientific. co-authorship [4] and sexual contacts [5]. Understanding and preventing the spread of epidemic diseases requires careful analysis of the underlying relationships in complex networks [6, 7, 8]. All these problems from different realms of science share a common area of study called complex networks.

Networks were known in mathematics since Euler's famous Königsberg problem led to a new area of study called graph theory. From a mathematical perspective, much of the work in this area is on random graphs [9], which deals with graphs obtained by random processes. Although random graphs were extensively studied in the mathematics community, particularly by Erdös and Rényi [10], realization that real life systems may be represented by network structures accelerated complex network studies, in the 1990s.

One of the earliest results of real life networks was obtained by Stanley Milgram in the 1960s [11]. He took about 60 letters, which were all addressed to the same person in Boston, and distributed these letters to randomly selected people in Nebraska. The aim was to get these letters to their destination in Boston, but each person could only send the letter to another whom s/he knows on a first-name basis. Although only a fraction of the letters reached their destination through a chain of people, Milgram found that on the average it required about six steps to get a letter to its destination. This result provided basis for the famous phrase, six degrees of separation, and the result that two randomly selected persons can be connected with a small number of links is generally known as the small-world phenomenon.

Networks have been used extensively in many fields of study, in last decade [2, 4, 3, 5, 1]. In this study, a procedure to obtain subgraphs that would imitate certain aspects of the whole graph has been developed. Most cases in the literature studied vulnerability in the case of node removal [12, 13]; here, edge removal is studied and generalized as a subgraph deduction method. Further, certain relations between local and global measures of a network that are suggested by our numerical studies are sought. Finally these methodologies and network theory will be applied to some topics of materials science to understand and differentiate structures of various materials.

1.2 Models of Networks

There are different models for networks that try to capture the structure of real world networks. The simplest one is random graphs. These graphs are obtained by distributing a fixed number of connections between nodes. If there are N nodes and each node has k connections on average, one has to randomly distribute $Nk/2$ connections between these N nodes to form a random graph. This type of graphs was studied by Erdös and Rényi.

One can easily show that a random network has a logarithmically scaled shortest path in the limit for large N. A certain node will have k first neighbors on the average, k^2 second neighbors, k^3 third neighbors and so on. In general, the diameter, D, can be aproximated by equating number of D^{th} neighbors to the network size N. Thus, the diameter of a random network will be $D = \ln N / \ln k$. Logarithmic scaling of the largest distance with network size is one sign of small-world behavior.

In real life, there is considerable overlap of neighbors, a property that lacks in random networks, and leading to their failure to explain most of the real world networks. In other words, a node's neighbors have a significant tendency to be interconnected, i.e. a person's friends are probably also friends with each-other. This property is called clustering in general. Clustering coefficient (C) is defined as a measure for this property, where it is the ratio of the number of connections among a node's neighbors to the number of total possible pairs among its neighbors averaged throughout the network. It can be shown that, for a random graph $C = k/N$, which becomes quite small for a large network. It has been observed that, for various real life networks, the value of C is significantly larger than that of random graphs [14, 2, 15, 4, 16, 8, 1]. A network with a high clustering coefficient and small average shortest path is called a "small-world network".

1.2.1 Watts-Strogatz Model

In order to capture the high clustering coefficient, as well as the logarithmically scaling average path length between any two nodes, Watts and Strogatz [17, 1] proposed a model for generating networks. Their aim was to obtain an underlying regular lattice with some random long range connections to provide shorter pathways on the average. They started with a one dimensional regular lattice that is closed on to itself so as to form a ring. Every node in the network has initially the same number of connections, k, so that each node is connected to its $k/2$ neighbors (see

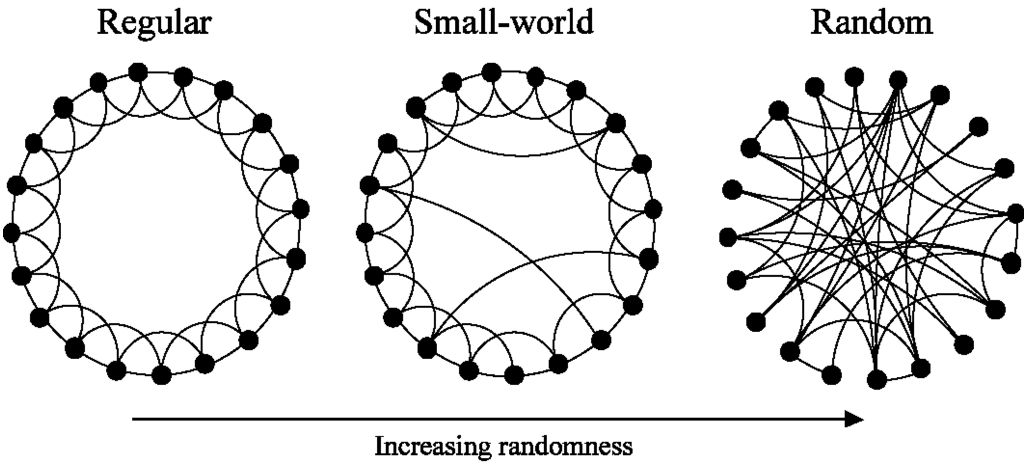


Figure 1.1. Description of Watts-Strogatz model [1].

Figure 1.1). They then consider each connection in the graph and rewire it with a probability β. For small β, this gives a mostly regular graph with few random shortcuts. For $\beta = 1$, the resulting graph will be completely random. The value of k will be preserved.

For small values of β, the clustering coefficient of the resulting network will be close to the ordered counterpart, which is considerably high. Conversely, the addition of several long range shortcuts has a dramatic effect on the characteristic path length. They reduce the average path length to values comparable to those of random graphs.

Inspired by this model, different variations of the model have been proposed. Newman and Watts [18] suggested a model that adds shortcuts, instead of rewiring links. This model provides a better basis for analysis, because it eliminates the possibility of a disconnected network, which is a risk in the original model. Another model employs addition of new nodes that are randomly connected to the original nodes [19, 20]. Both of these models show small-world behavior and result in similar networks to the original model.

1.2.2 Decentralization and other models for Small-World

Kleinberg [21, 22] suggested that, Watts-Strogatz model is not a good representation of real networks. His argument was based on Milgram's experiment. In Milgram's experiment each person on the chain is unaware of the overall structure of the network. They only use the local information to choose the next person on the chain. Yet, on the average, they manage to get to the target in a few steps. Kleinberg argued that a decentralized algorithm that only uses local data to decide on the next node could not always find the shorter paths in Watts-Strogatz model. In fact, he showed that, only a certain random connecting scheme would allow a decentralized algorithm to find the shortest paths. His model starts out with a two-dimensional regular lattice. He then adds random long range shortcuts between i and j with a probability that is proportional to $d_{i,j}^{-r}$, where $d_{i,j}$ is the Euclidian distance between nodes i and j. Kleinberg showed that for $r = 2$, there exists a simple decentralized algorithm for finding the short paths. For any other value of r, finding these short paths are much harder.

Another alternative model for small world was proposed by Albert and Barabasi [23]. Their objective was to recover the structure of the World Wide Web, where there are a small number of nodes with a lot of connections and a lot of nodes with very small connections. The model starts with a number of nodes and at each time step a new node with fixed number of edges is added to the network. These edges are connected to the existing network with a procedure called preferential attachment, where the probability that a new node will be connected to an existing node is proportional to the number of connections of the existing node.

1.2.3 Weighted Networks

Although networks provide useful tools for analysis, pure topology of the structure is only a first approximation to represent the underlying system. For example, mapping the internet backbone in a network structure could be useful, but to get a meaningful analysis, one has to incorporate the traffic and capacity data to the network. One simple way is to differentiate the connections from each other by assigning each a weight that represents the data. In other words, one introduces heterogeneity into the network.

The study of weighted networks is relatively new, because one tends to thoroughly understand the limitations of the simpler problem first. In certain cases, weighted networks can be considered as a special case of homogenous networks. Newman [24] showed that a weighted network with positive integer weights could be replaced with a homogenous network having multiple edges so that the adjacency matrices, which is a matrix that defines the interactions between nodes, are identical. For most cases, these two networks behave similarly, but in general one has to work with the weighted network.

In order to characterize the weighted networks, several new parameters are defined. It has been observed that individual edge weights themselves do not provide enough information [25]. As connectivity distribution is a defining parameter for homogeneous networks, weight distribution is also crucial in the structure of a weighted network. Any correlations between these distributions may affect network behavior. In the presence of weights, one can modify the usual network descriptors. For example, similar to the degree of a node, one can define the strength of a node by simply adding the weights of connections that emerge from the node [25, 26]. One can also modify the clustering coefficient so that it will reflect the weight structure [25]. Furthermore, in the presence of weights two useful optimal path definitions can be utilized. The first one is called "strong' path", which minimizes the maximum weight along a path over all possible paths. The second one is called "weak path" that minimizes the total weight along a path over all possible paths [27, 28, 29].

1.3 Motivation

The main goal of this thesis is to investigate network properties and particularly analyze the relationship between local and global parameters of a network by selecting the residue networks as the main case study. By "local" we refer to network properties only stem from the local neighborhood of a given node. By contrast, "global" refers to how the same node relates to the overall features of the whole network. For example, how the neighbors of a node are distributed provide local information about the network structure, whereas paths traversing between nodes would provide information about the global behavior of a network.

1.3.1 Information pathways in residue networks

Interactions, delay, and feedback are the three key characteristics of complex systems. Using these features, entities at different time and length scales communicate with great accuracy, efficiency and speed [30]. Self-assembling molecular systems are complex fluids with robust and adaptable architectures. Proteins, whose internal motions are decisive on their folding, stability, and function, are exquisite examples of these. Proteins are under constant bombardment in their environment e.g. in the cell where other small and large molecules are densely and heterogeneously distributed, or in the test tube with only water around, displaying ceaseless fluctuations around their folded structure. Since proteins function efficiently, accurately and rapidly in the crowded environment of the cell, they are expected to be effective information transmitters by design. The fact of the protein being functional or not depends on the size of these fluctuations and how they are instilled, making use of the concerted action of residues located at different regions of the protein [31, 32, 33, 34]. It is, therefore, of utmost interest to investigate how proteins respond to changes in the environment under physiological or extreme conditions.

The response of any structure to perturbations depends on its general architecture. For proteins, local, regular packing geometries [35] cannot provide short distances between highly separated residues for fast information transmission. In fact, it has been shown that random packing of hard spheres similar to soft condensed matter is observed in a set of representative proteins [36]. Consistent with the concurrent requirement of order and randomness in the protein structure, we [15] and others [37, 38, 39], have recently shown that proteins are organized within the small-world network topology. A network is referred to as small-world if the average shortest path between any two vertices scales logarithmically with the total number of vertices, provided that a high local clustering is observed [1]. Such properties are common in many real-world complex networks [20, 40], and there are examples from a diverse pool of applications such as WWW [41], the internet [42], math co-authorship [4], power grid [1] and residue networks [15].

In recent years, proteins are modeled as networks of interacting amino acid pairs to determine their network structure and to identify the adaptive mechanisms in response to perturbations [15, 43, 44]. Also, similar network treatments of proteins predict collective domain motions, hot spots, and conserved sites [45, 46, 47, 33, 48]. For these networks term residue networks is used [15] to distinguish them from protein networks which are used to describe systems of interacting proteins [49]. Statistical analysis within these works show that proteins may be treated within the small-world network topology. In the past few years, the network treatment of residues in proteins have been adopted to study their various features such as conserved long-range interactions [50], functional residues [51, 52], protein-protein association [53], and detection of structural elements [54].

In all these treatments, which have been successful in describing many important properties of proteins and provide insight as to how they function, the identities of individual amino acids are omitted in the calculations. In other words, specificity is taken into account in an indirect manner, by assuming that the locations of the different amino acid types along the contour of the polymeric chain have been operational in determining the particular average three-dimensional structure. In this viewpoint, the interactions between different pairs, triplets, etc. of amino acids are assumed to be smeared out, and the observed behavior once the protein is folded, is driven by the overall structure. In fact, it has been noted that the residue non-specific interactions contribute more to the overall stability of proteins by a factor of about five, compared to distinct residue-residue interactions [55]. Recent studies considered residue specific properties in networks and by assigning weights depending on the interactions between amino acids, it is suggested that the residue networks conform to random networks graphs with ascociated percolation behaviors [56].The question remains, however, as to the extent to which such a coarsened description of the folded protein may be used to determine other crucial properties, especially those pertaining to dynamics.

In this thesis, we elaborate on the paths between residue pairs, which we term information pathways, to understand how they relate to dynamic phenomena in proteins. In particular, it is of interest to understand allosteric interactions mediated through the changes in the dynamic fluctuations around the average structure, both in the presence and absence of conformational changes, the latter having very recently been shown to exist in proteins through a series of NMR experiments [57]. To this end, we attribute weights to the links between residue pairs using knowledge-based potentials [58, 59], and discuss the relationship between dynamic phenomena occurring in proteins and the optimal path lengths obtained from these weighted networks. We show that it is possible to extract minimal sub-graphs from the fully connected networks of residues, where a few designed-in interactions overlaying the backbone are sufficient to display communication path lengths of residue networks of interactions. We also demonstrate an application of these ideas using a non-redundant data set of interacting proteins, and extract residue pairs on the interface of the receptor/ligand that frequently appear along information pathways.

1.3.2 Local statistics of condensed matter networks

For a completely random network where the effect of local clustering is negligible, it is possible to analyze the emergence of global parameters from local distributions. In the presence of high local clustering, redundancies are introduced to a system in terms of global behavior and incorporating these effects in estimation of global parameters becomes rapidly complicated. In the path from local to global, intermediate steps require additional investigation. It can be derived that for certain networks number of neighbors of a node is proportional to the average number of neighbors of its neighbors, where this value is closely related to number of second neighbors of a node. Several real life spatial networks, including the residue networks fall under this category.

The study of real life networks, such as the world-wide web [16], internet [42], power-grids [1] and math co-authorship [4], has put forth properties that distinguish them from classical Erdös-Rnyi random networks [60]. The variety of degree distributions and other statistical measures that emerge has heightened the interest in complex networks. With the proposition of algorithms by Watts-Strogatz [1] and Barabsi-Albert [23] to generate real life-like networks, this area has been investigated extensively [22, 61]. The classification of networks is mostly based on measures such as degree distributions, average clustering, and average path length [14, 62].

In recent years, proteins were investigated as networks, by taking the amino-acids as nodes. Termed as residue networks (RN), edges between neighboring nodes are represented by their bonded and non-bonded interactions [15, 63, 64, 65]. Several studies have shown that residue networks have small-world topology [15, 37, 38, 39], characterized by their logarithmically scaling average path lengths with network size, despite displaying high clustering. Further studies also utilized network models for protein structures to predict hot spots [46, 45, 47, 48], conserved sites [46, 45, 47, 48, 50, 66, 67], domain motions [68, 46, 45, 47, 69, 48], functional residues [51, 33, 52, 70] and protein-protein interactions [53]. The small-world topology of residue networks is established, and various network properties such as the clustering coefficient, path length, and degree distribution are used to account for, e.g. the different fold-types in proteins [50], interfacial recognition sites of RNA [66], and bridging interactions along the interface of interacting proteins [63]. In light of these studies, we expect other self-organized molecular systems of synthetic origin to display similar topology.

In fact, a hierarchical arrangement of the nodes is expected to occur in self organization of atoms and molecules under the influence of free energetic driving forces. In graph theory, hierarchies have been quantified by the presence of (dis)assortative mixing of their degrees, defined as nodes with high degrees having a tendency to interact with other nodes of (low)high degrees [71]. Analytical and computational models for generating assortatively mixed networks were proposed [72, 73]. Newman has shown that assortatively mixed networks percolate more easily and they are more robust towards vertex removal [72, 74]; most social networks are examples of these. In this work, we find RN of proteins to also have assortative mixing, although many biological networks such as protein-protein interactions and food webs were found to display disassortative behavior.

It is expected that in networks displaying any degree of correlations, local properties of the constructed graphs will have an effect on the global features. However, a connection between the local and global network properties and the underlying structure of molecular systems has yet to be established. In this study, we derive a relationship relating the nearest neighbor degree correlation of nodes, their degree, and clustering coefficient. We next show that a linear relationship is valid for two types of self-organized molecular systems: (i) Folded proteins and (ii) block cooligomers in a solvent that encourages micelle formation. Furthermore, simulated configurations of Lennard-Jones clusters also approximate the findings as well as a simple polymeric system forced into a close-packed structure under extremely high pressure. We also show that model hexagonal close packed (HCP) structures may be used to reproduce many of the graph properties of the above-mentioned systems. A brief description of the model systems are summarized under the Methods section. This study is a first step towards the design principles of complex molecular networks.

1.3.3 Packing of proteins

Local clustering in these networks is a direct result of their three dimensional structure. Therefore, it is imperative to understand the effect of structure to network parameters. Focusing on residue networks, we look for local ordering in protein structures by generating lattice based self-avoiding chains that would approximate the real protein chain.

Research on lattice representation of proteins dwells on two problems. The first problem is to accelerate modeling efforts by confining conformational moves restricted in conformational space. In this setting, the fundamental use of the underlying lattice is to provide a basic grid for realizing and updating conformations [75]. There are many folding algorithms based on these ideas and they are widely used in the computational biology community [76]. These algorithms make use of various lattice types [77]. Notably, Covell and Jernigan uses a face-centered cubic (FCC) lattice; they suggest a way to identify a lattice walk that approximates the native state [78]. Similarly, a lattice model based on the diamond cubic lattice (equivalent to a FCC lattice with a two point basis) has been introduced for predicting folded conformations at low spatial resolution, without reference to a native state [79]. The use of closed pack structures has been suggested in studies on local packing of residues [80], and hydrophobic-hydrophilic interactions [81].

The chain fitting problem onto a crystal lattice in $\Re^3$ using root mean square deviation metric has been shown to be NP complete [82], if self-avoidance criteria is strictly and rigorously enforced. Therefore various heuristic approaches have been developed for attacking the problem. The simpler problem which does not entail the self-avoiding property can be solved in polynomial time and two such chain-fittin algorithms have been developed so far [83, 84].

Covell and Jernigan attempt to create all conformations on an FCC lattice and choose the optimal conformation based on non-bonded pairwise potential energy minimization [78]. The use of dynamic programming algorithms in finding an optimal lattice fit to a template chain [85, 86] has been suggested as an alternate approximate solution which iterates by minimizing a global error function. Alternatively, a greedy algorithm has been proposed and this attracted considerable attention [87]. Yet another method is the self-consistent mean field theory approach which finds the optimal fit starting from a set of lattice points through an iterative procedure to minimize an energy function with a lattice probability weight matrix [88]. Although the large majority of research focuses on representing backbone fitting, side chain atoms can also be accounted for without much difficulty [89].

Many studies favor crystal prototypes FCC or HCP for realistic representation of protein energetics. Yet alternative closed pack structures which possess different stacking patterns has never been accounted for in treating in them [77]. This is important because altering stacking of triangular layers does not disturb closed-packedness [90].

In our approach we introduce a Metropolis Monte Carlo scheme where the random conformational variations on lattice sites are evaluated by structural alignment of resulting self-avoiding lattice chains onto real protein chains by use of quaternion based alignment algorithm [91]. Acceptance of new conformations are then based on the root mean square deviations of aligned sequences. We then analyze resulting self-avoiding chains and compare them to their protein counterparts by looking at the spacial and network properties.

1.3.4 Spectral properties of networks

Spectral analysis of systems provide valuable information about their dynamic properties. For proteins, normal mode analysis was used to analyze coupled motions in low frequency modes and helped classification of protein motions, e.g., hinge bending and shear [92]. It has further been shown that the predominant contributions to these motions may be described by a single, most collective mode for some proteins, whereas it may be obtained from a superposition of several modes for others

[93]. With the advent of coarse graining of biomolecular structures through residue-based network models [43, 45, 92], it has been possible to study a large number of protein structures. These anisotropic network models (ANMs) take into account the three-dimensional geometry of interacting pairs of residues to study the modal behavior of proteins. Using such information, it is possible to morph between the apo and holo structures to gain insight into the intermediates that lead to the final structure [94, 95, 96]. Eigenvectors corresponding to the lowest eigenvalues provides information regarding the confromational changes during binding [97, 95, 93].

In terms of networks, spectral properties gained attention since the distribution of eigenvalues of normalized Laplacian [98] characterize several aspects of the network such as algebraic connectivity, motif replication and bipartiteness [98, 99, 100, 101]. An extention of normalized Laplacian to three dimensions was recently applied to the analysis of local arrangements in residue networks [102].

Here we employ spectral analysis of normalized Laplacian to networks obtained from condensed matter system in order to characterize structural properties. Although the spectra of normalized Laplacian is not unique, i.e. different networks with identical eigenvalues may be formed, these isospectral systems behave similarly in terms of monitored network parameters [98] and can be considered a family of systems with similar properties.

Chapter 2

Network descriptors

Networks are modeled with mathematical constructs called graphs. A graph G, consists of a set of vertices $V(G)$ (also called as nodes) and a set of edges $E(G)$ where an edge is an unordered pair of vertices in $V(G)$. An edge between x and y can be denoted in short form as xy. If an edge xy is present in the graph, x and y are called adjacent vertices and y is denoted as a neighbor of x.

Although equality between two graphs requires that they have the same vertex and edge sets, simple reordering of the vertices in the vertex set does not alter the relationship in a graph. Therefore instead of equality, it is generally more convenient to define isomorphism between graphs. Two graphs A and B are said to be isomorphic if there is a bijection f from $V(A)$ to $V(B)$ such that $f(x)$ and $f(y)$ are adjacent if and only if x and y are adjacent. Isomorphic graphs can be treated as equal graphs without loss of generality.

A graph is called complete if every pair of its vertices are adjacent, and it is called empty if the edge set is an empty set. The above definition of a graph assumes a symmetric relationship between edges, i.e. if x is a neighbor of y, then y is also a neighbor of x, and these graphs are called simple graphs. Although it is possible to define asymmetric relations between vertices via directed edges, this work utilizes undirected networks therefore directed graphs are not discussed. Depending on the model, values can be associated with edges to differentiate variations in relative importance within the edges. These values are called weights of edges.

Subgraphs deduced from graphs usually provide important properties. A subgraph of a graph is a graph with vertex set and edge set that are subsets of the parent graph. A clique is a subgraph that is complete. A path of length l from x to y is a sequence of $l + 1$ distinct vertices starting with x and ending with y such that every consecutive vertices are adjacent. A graph is called connected if there is a path between any two vertices in the graph. A cycle is a subgraph where every vertex has exactly two neighbors. Minimum possible cycle is a three vertex subgraph, which is also a clique and often called a triangle. At the other extreme graph without any cycles is called a tree. In a connected graph with cycles, there is more than one path between any two vertices. Therefore, the shortest path length between two vertices is generally described by the path with the smallest length. In the presence of weights, it is also useful to use alternative optimal path length definitions by the length of the path that minimizes a function of the edge weights along the path.

2.1 Matrix representations

A graph is usually represented with a matrix called the adjacency matrix $\mathbf{A}$. Rows and columns of the adjacency matrix correspond to the vertices and A_{ij} entry is the number of edges between vertices i and j. Since all the graphs in this work does not contain multiple edges between two vertices, adjacency matrices are binary (i.e. A_{ij} entry of the adjacency matrix is either 1 or 0 depending whether or not vertices i and j are adjacent). It should be noted that the adjacency matrix fully defines a graph and the parameters that are often used to classify networks, can be computed directly from the adjacency matrix.

The most common parameter that is of importance is the degree k_i of vertex i. Degree is basically the number of neighbors of a given vertex and it can be calculated as;

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{2.1}$$

where N is the number of vertices in the graph. Higher order degree correlations are also of importance and may be utilized to identify more distinguishing features of the network. For instance, average nearest neighbor degree of a node i, denoted by $k_{nn,i}$, is the average degree of its neighbors and may be written in terms of the adjacency matrix.

$$k_{nn,i} = \sum_{j=1}^{N} \sum_{m=1}^{N} A_{ij} A_{jm} = \sum_{j=1}^{N} A_{ij} k_j \tag{2.2}$$

Normalized third degree correlations (C_i), known also as the clustering coefficient, is widely used to characterize the distinctness of networks. It is defined as the ratio of the number of interconnections between a node's neighbors to the number of all possible connections. C_i is closely related to the number of triangles involving the vertex i and can be considered as a measure of local cliqueness around a vertex.

$$C_i = \frac{\frac{1}{2} \sum_{j=1}^{N} \sum_{m=1}^{N} A_{ij} A_{jm} A_{im}}{\frac{k(k-1)}{2}} \tag{2.3}$$

While k_i, $k_{nn,i}$, and C_i are descriptors of local structure, another common parameter used to classify the global structure of graphs is the average shortest path length, L_i of a node. Given that the shortest path length from i to j is L_{ij}, it is the average number of steps that are traversed from all other nodes to node i:

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

Another matrix that is associated with graphs is the Laplacian (also known as the Kirchoff) matrix. The Laplacian of a graph L is used extensively in the graph theory literature and bears some important aspects of a graph. It is defined as

$\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = k_i$. The Laplacian is a positive-semidefinite matrix and its spectrum may be used to diagnose certain underlying features of the graph. For instance, the second lowest eigenvalue is associated with the algebraic connectivity of the graph and it denotes how well connected the graph is. In this study, we use the normalized Laplacian, $\mathbf{L}^*$.

$$\mathbf{L}^\star = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{A})\mathbf{D}^{-\frac{1}{2}} \tag{2.5}$$

The spectrum of the normalized Laplacian is also used to categorize networks, i.e. the presence of an eigenvalue at $\lambda = 2$ implies the network is bipartite, the multiplicity of the eigenvalues at $\lambda = 1$ is a measure of motif duplication in the network, and the second eigenvalue indicates how well the network is connected [9, 98, 100].

Chapter 3

Optimal paths in residue networks

One of the ways that a graph can be used to analyze is the information transfer in the network through the connections. Here, we consider proteins as a network of interacting residues and we elaborate on the paths between residue pairs, which we term information pathways, to understand how they relate to dynamic phenomena in proteins. In particular, it is of interest to understand allosteric interactions mediated through the changes in the dynamic fluctuations around the average structure, both in the presence and absence of conformational changes, the latter having very recently been shown to exist in proteins through a series of NMR experiment [57]. To this end, we attribute weights to the links between residue pairs using knowledge-based potential [103, 59], and discuss the relationship between dynamic phenomena occurring in proteins and the optimal path lengths obtained from these weighted networks. We show that it is possible to extract minimal sub-graphs from the fully connected networks of residues, where a few designed-in interactions overlaying the backbone are sufficient to display communication path lengths similar to that of the full residue network. We also demonstrate an application of these ideas using a non-redundant data set of interacting proteins, and extract residue pairs on the interface of the receptor/ligand that frequently appear along information pathways.

3.1 Model

3.1.1 Spatial residue networks

For the single protein calculations, we utilize 595 proteins with sequence homology less than 25% [104] and sizes spanning ca. 50 to 1000 residues. For the receptor-ligand complexes, on the other hand, we use the non-redundant benchmark set of Weng and collaborators developed for testing docking algorithms that contains an overall of 59 pairs of proteins with 22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, 11 other complexes, and seven difficult test cases [105]. We form spatial residue networks from each of these proteins using their Cartesian coordinates reported in the protein data bank (PDB) [106]. In these networks, each residue is represented as a single point, centered on the C_β atoms; the C_α atoms are used for Glycine residues. Given the C_β coordinates of a protein with N residues, a contact map can be formed for a selected cut-off radius, r_c, an upper limit for the separation between two residues in contact. This contact map also describes a network which is generated such that if two residues are in contact, then there is a connection (edge) between these two residues (nodes). Thus, the elements of the adjacency matrix, $\mathbf{A}$, are given by

$$A_{ij} = \begin{cases} H(r_c - r_{ij}) & i \neq j \\ 0 & i = j \end{cases} \tag{3.1}$$

Here, r_{ij} is the distance between the i^{th} and j^{th} nodes, $H(x)$ is the Heaviside step function given by $H(x) = 1$ for $x > 0$ and $H(x) = 0$ for $x \leq 0$. We adopt the value for the cutoff distance $r_c = 6.7$Å that includes all neighbors within the first coordination shell around a central residue.

In the case of the weighted residue networks, we assign weights to the edges according to the inter-residue interaction potentials of Miyazawa and Jernigan [103] and Thomas and Dill [59]. These are statistical potentials extracted from a protein data base. Both potentials have been extensively tested in threading algorithms [107, 58], protein stability and designability studies [108], folding and binding energetics, as well as amino acid classification [109]. The Miyazawa-Jernigan (MJ) potential is based on a set of protein subunit structures exceeding 1600 in number [103]. In their treatment of the problem, the system is taken as an equilibrium mixture of unconnected residues and effective solvent atoms. The Bethe approximation is employed to estimate the contact energies from the numbers of contacts that arise in the sample. Excluded volume is taken into account by the inclusion of a hard-core repulsion between the residues and a repulsive packing-density-dependent term. The Thomas-Dill potential, on the other hand, utilizes a much smaller data set of 37 proteins [59]. The authors use the folded chain conformation as the reference state, instead of a collection of randomly mixed particles of residues and solvent molecules [in treatments using the Bethe approximation, the problem of reference states has been addressed and corrections have been proposed[110]]. Thomas and Dill employ an iterative method which extracts pair potentials that incrementally drive the system towards a lowest energy structure that corresponds to the native structure. The main discrepancies in the statistical potentials that result from the approximate treatment or neglect of excluded volume, chain connectivity and interdependence of pairing frequencies are therefore intrinsically taken care of.

Here, we have repeated all the calculations using both the Miyazawa-Jernigan and the Thomas-Dill knowledge-based potentials. Despite differences in details, the main results and conclusions reached do not change with the choice of potential. In what follows, we therefore report only results from the Thomas-Dill potentials. We assign e_{ij}, value of the connection between the i^{th} and j^{th} residue, according to the inter-residue interaction potential between the i^{th} and j^{th} residue types. Thus, the links connecting the residue pairs with the least favorable interaction energy have the lowest weight, i.e. the highest value.

3.1.2 Network descriptors

The networks are classified by local and global parameters, all of which can be derived from the adjacency matrix. In the absence of edge weights, the most general descriptors of the network structure are average degree of a node (equation 2.1), and the average shortest path length (equation 2.4) through the network. The average degree of the network is thus $z = \langle k_i \rangle$, where the brackets denote the average over all nodes. The degree of the residue networks follow the Poisson distribution [15].

The shortest path length, L_{ij}^h, of a homogeneous network, where the links have no weights, is the minimum number of connections that must be traversed to connect residue pair i and j. In computing the shortest path between a pair of nodes, we make use of the fact that the number of different paths connecting a pair of nodes i and j in n steps is given by $(\mathbf{A}^n)_{ij}$. Thus, the shortest path between nodes i and j is given by the minimum power, m, of $\mathbf{A}$ for which $(\mathbf{A}^m)_{ij}$ is non-zero.

In the presence of weights, it is possible to define additional path lengths so as to take into account the skewing effects of the weights. Weights may be factored into the path lengths using different optimality criteria. We define two criteria for paths between two residues [27, 28, 29], weak disorder and strong disorder. In the former one, the optimal path connecting residues i and j is the length of the path, L_{ij}^w, that minimizes the sum of the weights along the path and it can be written as;

$$L_{ij}^w = \text{length}(\underset{p_{ij}}{\text{argmin}}(\sum_{e \in p_{ij}} w(e))) \tag{3.2}$$

where p_{ij} is a path from node i to node j, e is an edge in p_{ij} and $w(e)$ is the weight for edge e. We employ Dijkstra's algorithm to compute the optimal paths in the weak disorder case. In the latter (strong disorder) case, L_{ij}^s is the length of the shortest path that minimizes the maximum weight along the path.

$$L_{ij}^{s} = \text{length}(\underset{p_{ij}}{\text{argmin}}(\max_{e \in p_{ij}} w(e))) \tag{3.3}$$

To obtain L_{ij}^{s}, we sort the links in descending order and sequentially remove the links beginning with the highest weight (lowest energy). We continue to remove the links until we find the bottleneck link which will cause the connectivity between vertices i and j to be lost. We then compute the length of this remaining path in terms of the number of intervening links. Note that once the optimal path connecting residues i and j is determined, the path length is simply the number of connections along the path; i.e. the step lengths themselves are not weighted.

The characteristic path length of the network is then the average over all possible node pairs,

$$L^{\dagger} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} L_{ij}^{\dagger} \tag{3.4}$$

where the dagger symbol, †, represents the homogeneous, weak or strong paths, L^{h}, L^{w}, and L^{s}, respectively. Here, N is the number of residues in the protein. Note that $L^{\dagger}$ is a measure of the global properties, reflecting the overall efficiency of the network, under the imposed constraints; i.e. the lower $L^{\dagger}$ is, the faster information is communicated through the network.

3.2 Results

3.2.1 Random coils as a basis for comparison

Proteins may be modeled as networks where a special set of interactions are imposed on chain connectivity and the extent to which such interactions are specially designed are of interest here. In this study we generate a variety of networks based on selected proteins. A firm basis for comparing the various networks that may be formed from a given chain with a known contact number is a chain of the same length and the same number of connections for each of its nodes, but a randomized set of links between the nodes. To generate such networks, we rewire every residue (node) randomly to another residue chosen from a uniform distribution such that each residue has the same number of neighbors (contact number, k_i), while the contact order changes; chain connectivity is preserved by keeping the $(i,\ i+1)$ contacts intact. Such a network corresponds to the random coil conformation of a polymer chain at an arbitrary point in time. Degree distribution of residue networks is Poisson [15]. It is also known from network theory that a completely random, Poisson distributed network has the shortest path length,

$$L_{random} = \frac{\ln N}{\ln z} \tag{3.5}$$

where z is the average degree of the network.

Figure 3.1 shows different shortest path lengths for residue networks. Results are presented for the non-redundant set of 595 proteins whereby values for proteins of size $(m \pm 1)10; m = 3, 5, \ldots$ are averaged. Protein path lengths computed with the weak disorder limit are not distinguishable from those of shortest paths on homogeneous networks; both may be best-fitted by a line of slope 5.2. Optimization with the strong criterion results in networks with significantly longer path lengths (best-fitting line through the data is shown by the dashed line; slope is 9.0). For

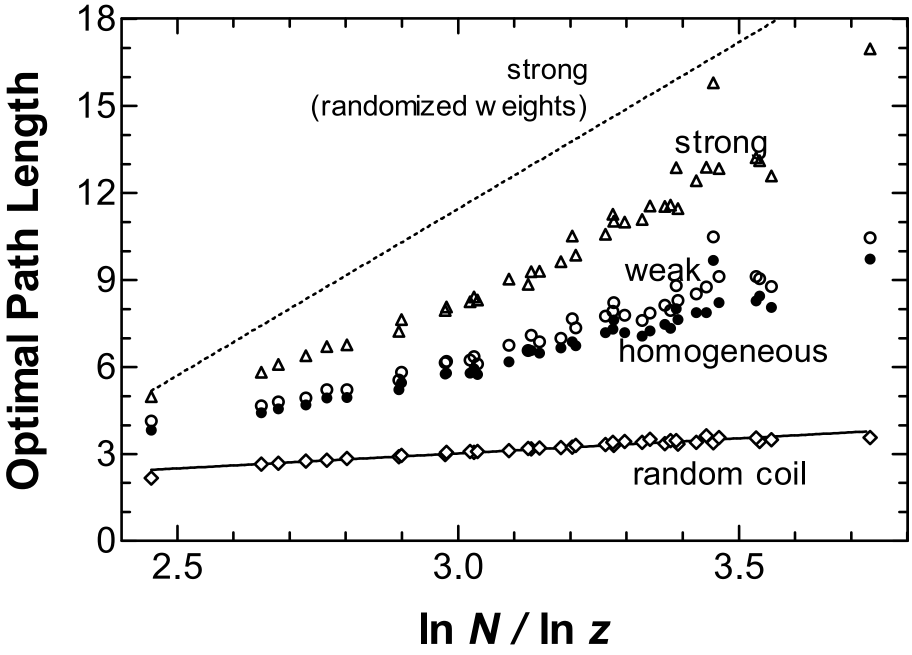


Figure 3.1. Optimal path lengths, $L^h(\bullet)$, L^w (solid line), $L^s(\circ)$, of the protein networks in comparison to those of the theoretical value of Poisson distributed random networks of the same size and number of neighbors (L_{random}, eq. 3.5).

comparison, random coils have also been generated by random rewiring of the residue networks while preserving connectivity (see text). These networks provide the same result as a totally randomized network (no chain connectivity) of the same size (slope is 1.0). At the other extreme, randomized weights have been imposed on the original residue networks (dotted line). L^s for these are longer by a factor of ca. 1.3, indicating that the weights in a protein are specifically distributed.

As shown in figure 3.1 bottom curve, it is verified that the randomized chains behave exactly as expected from a completely random collection of nodes. Average path lengths on the residue networks, L^h, on the other hand, are significantly higher than the randomized networks while still preserving the approximately logarithmic dependence on number of residues, as shown with the filled circles in figure 3.1. The loss of high optimality (i.e. a two-fold increase in the shortest path lengths compared to a random network) must be compensated by the emergence of functionality in the self-organized structure. This exchange is achieved along the scaffold of the non-random networks formed by the residues of the proteins.

3.2.2 Optimal paths in the presence of weights

In the absence of weight information of the links (i.e. for a homogeneous network), L^h is the only parameter we can use as a measure of the distance between nodes in the network with N vertices. In the presence of weights, the heterogeneity of the medium is taken into account; hence different types of optimality criteria can be defined. In the case of weak disorder, the sum of the potentials along the optimal path is minimized to obtain L^w. This can be interpreted as the path that causes minimum possible total disturbance to the residues along the path. The links with lower potentials are more likely to tolerate the disturbances. In Figure 3.1 we display a comparison of shortest paths of homogeneous and weak disordered networks, L^h (symbols) and L^w (line), respectively, with that of the random coil. The correlation between the two data sets is excellent, showing that the weighted network in the weak disorder limit behaves similar to the homogeneous network. The optimal path in the strong disorder, on the other hand, is the path that minimizes the maximum of the potentials along the path, which can be interpreted as the shortest path that causes minimal maximum disturbance along the path. As exhibited in Figure 3.1 for the strong disorder case (see the open circles and the overlaying best-fitting dashed line), L^s is significantly larger than L^w by an average factor of 1.3.

3.2.3 Are weights imposed on the links significant for the protein?

To answer this question, we randomly reassign the potentials attributed to pairs of residues. This is achieved by redistributing the 210 different types of pair potentials in the Thomas-Dill potential matrix, so that the same residue type pair always has the same value. As such, the underlying network structure remains unchanged, while the optimal paths that are preferred will be affected. The results based on these networks are obtained from five realizations of this randomization.

Two major observations are made for such networks: In the weak disorder limit, the optimal path lengths increase (data not shown), signifying that the residue pairs are specially distributed in the protein network so as to have similar allotment of weights around a given node, although the values themselves have a large span [-1.8 ...1.5]. Moreover, the strong paths in the weight-randomized networks are longer (shown by the dashed line in figure 3.1), further corroborating this finding with the more stringent constraint that key links minimizing the maximum weight along given paths exist in the folded protein.

3.2.4 Identifying redundancies in the protein communication pathways by extracting sub-networks

We deduce sub-networks from the original residue networks of each of the 595 proteins utilized in this work by systematically removing links that have values higher than a given cut-off value, e_{cut}. Chain connectivity is preserved regardless of the residue types flunking a given bond. We rely on the fact that, a protein under external disturbance will have a higher tendency to lose communication through high energy contacts, while the low energy ones will be more cohesive. The shortest path lengths of each of the remaining networks are subsequently computed.

Several important cases are presented in figure 3.2, as a function of the random coil of the same size, N, and the same original number of neighbors, z. The distribution of the links is shown in the inset to this figure, and the chosen cut-off values are marked on the distribution. Sub-networks from the original residue networks are deduced using the edge values, whose distribution for the 210 possible residue pair interactions are shown in the inset. Edges with values higher than a given cut-off, e_{cut}, are removed and the new shortest path lengths of these sub-networks are computed; connectivity is preserved. The redundancy in the proteins is such that, when ca. half of the long-range contacts are removed, the system still has the same path length. Upon further removal of contacts, the paths get longer, and

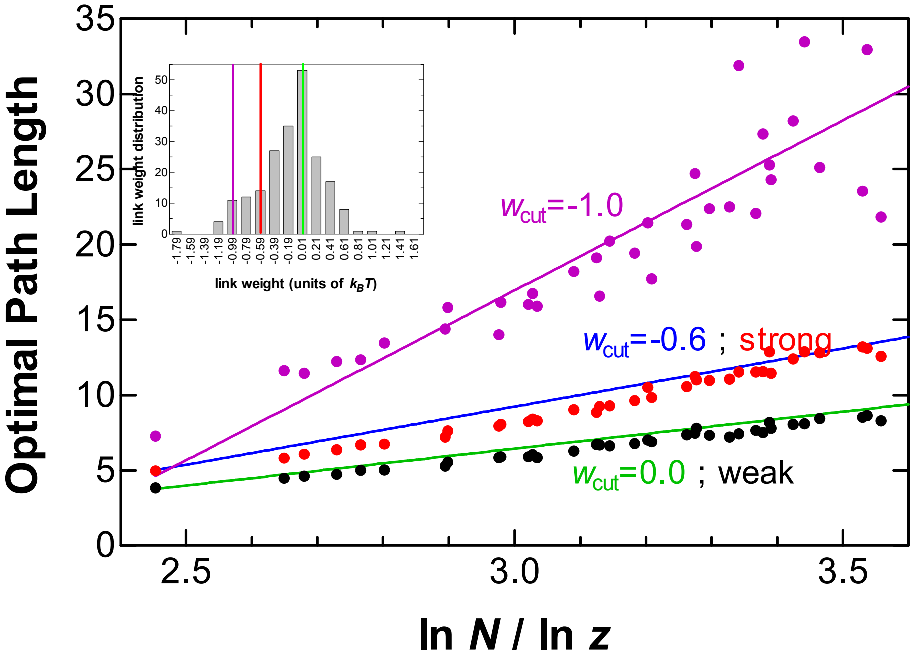


Figure 3.2. Optimal path lengths of the protein networks constructed with various schemes as a function of the randomized counterparts of the original networks (eq. 3.5).

they overlap with L^s at $e_{cut} = -0.6\mathrm{k_B T}$ (only ca. 20% of the long-range contacts remaining). Further removal of contacts results in a sudden increase in the shortest path lengths, exemplified by the case of $e_{cut} = -1.0\mathrm{k_B T}$ (slope = 22.6).

The redundancy in the proteins is such that, when ca. half of the non-bonded contacts are disregarded, $e_{cut} = 0$, the system still has the same shortest path length as the full protein that preserves all of its contacts (compare the green line and the black data points). Upon further removal of links, the paths get longer, and they overlap with L^s at $e_{cut} = -0.6\mathrm{k_B T}$ (compare the blue line and the red data points). At this point, only ca. 20% of the long-range contacts remain in the sub-networks. Further removal of contacts results in a sudden increase in the shortest path lengths, exemplified by the case of $e_{cut} = -1.0\mathrm{k_B T}$. In figure 3.1, this data set is shown in purple, along with the best fitting line (slope = 22.6, in comparison to the random networks where the slope is one). Note also that the scatter in the data is extreme, signifying that the logarithmic dependence of path lengths on number of residues is lost.

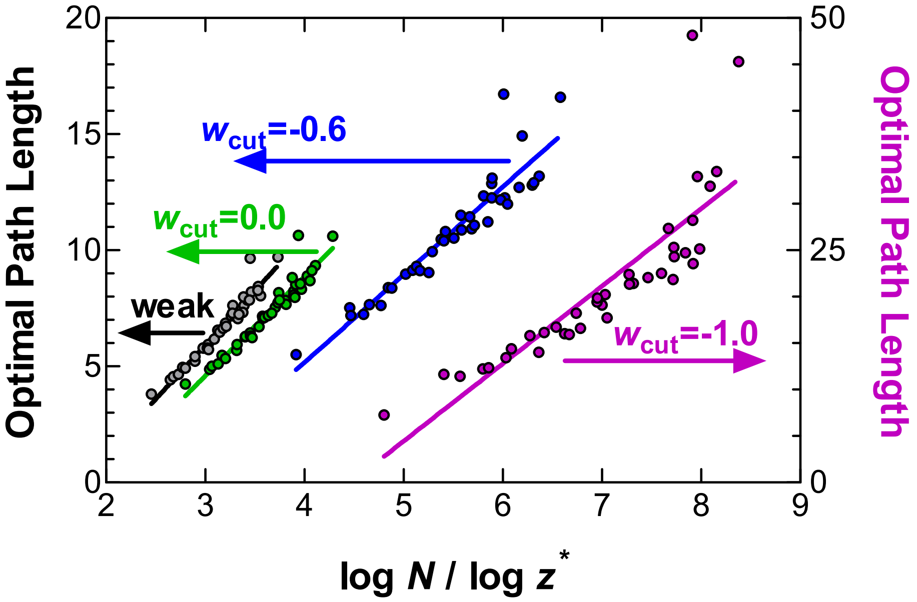


Figure 3.3. Optimal path lengths of the protein networks constructed with various schemes as a function of the randomized counterparts of the newly constructed networks, $L_{random} = \ln N / \ln z^*$.

Another way to observe this data is by plotting the shortest path lengths of the sub-networks as a function of the random coil of the same size, N, and the modified (reduced) number of neighbors, z^* (figure 3.3). Although the path length increases as networks with less contacts are formed, as expected, the slope of the best-fitting line remains constant until $e_{cut} = -0.6\mathrm{k_BT}$, i.e. coincides with the original, fully connected network that utilizes the strong paths as was shown in figure 3.2. Further removal of links results in a dramatic increase in the shortest paths, as exemplified by the $e_{cut} = -1.0\mathrm{k_BT}$ case (purple; values on the right y-axis). Again, it is observed that the scatter in the data increases as the sub-networks approach a linear chain ($e_{cut} = 1.8\mathrm{k_BT}$, i.e. only connectivity remains).

3.3 Discussion

A folded protein needs to perform its function under the constraints that the overall shape is suitable for the task it undertakes, while it is not energetically penalized. As a molecular machine, it needs to optimize the time it takes to communicate the incoming information, which, to a first approximation, may be assumed to be linearly dependent on the shortest path length in its residue network. Excluded volume imposes another limit on the size of the molecule. As incoming information, we refer to perturbations that are imparted on one or several of the residues. Changes in the environmental conditions that are reflected on thermodynamic parameters, such as the temperature, will affect the whole system. The latter are not of concern in this study, since these may potentially change the overall network structure.

In the previous section we have displayed results that introduce several different perspectives to evaluate how folded proteins are organized so as to manage their redundancies under sub-optimal conditions. Our basis for comparison is the random coil, whereby a Poisson distributed arrangement of residues will always lead to the most optimal path length, given by the analytical relationship of equation 3.5. The random networks constructed for figure 3.1 have the same average number of neighbors as their folded network counterparts [$z = 6.9$]. They may be thought of as compact chains that constantly change their partners at different points in time. They, therefore, represent an average over many significantly different configurations, in direct opposition to the case of a folded protein, where residues always keep the same neighbors while they fluctuate in space. For a given amount of excluded volume, decided upon by chain connectivity and the number of long-range contacts, the random coils give a limiting value for how fast information may be spread through the system.

On the other hand, information spreading will take on different forms in a protein depending on the type of local perturbation that is received. Two limiting situations may be distinguished: (i) Proteins experience constant random fluctuations from the environment under the usual conditions they function; e.g. random collisions with solvent molecules, formation of local hot spots, etc. We classify these perturbations, extensive in number but small in the size of fluctuation they invoke, as everyday events. (ii) At other times, there will be large perturbations that will be targeted on specific regions, such as those occurring during binding, or approach of a large cellular body to unspecified regions of the protein. We classify these perturbations as extreme events. The modes of response from the protein are expected to be different for the two types of events. In other biological systems, such modified reactions to different types of input (global vs. pathway specific noise) were also observed and quantified; e.g. for the variation in the behavior of genetically identical cells [111, 112].

In folded proteins, the network structure, equivalent to a coarse graining obtained from the average conformation of the folded structure, is expected to remain nearly the same under both conditions. However, the way the energy will be transmitted throughout the network will differ according to the type of perturbation. Noting that the network is mostly made up of residues held together by non-bonded interactions, the proximity of pairs of residues will not differ; e.g., in many cases, the structure of the bound and unbound forms of a ligand protein to its receptor is less than the experimental uncertainty as in the case of chymotrypsin inhibitor II [33]. However, the transfer of information (energy) along the residue network will only occur if the fluctuations in neighboring residues are correlated along any chosen pathway [as conformational variability increases, the communication of a signal in a molecule, e.g. conductance, occurs with less strength and over a broader range of values, as was recently demonstrated through unique experiments in a series of diphenyl containing small molecule systems[113]]. For small perturbations caused by random fluctuations, the correlations between neighboring residues are expected not to be affected, and the most probable pathway for information transmission is the lowest energy one i.e. L^w. For large impacts (extreme events), although the overall network structure will be preserved due to the pressure exerted by the compact structure of the molecule, the correlations between pairs of residues that are weakly connected to each other will be lost. For the purpose of information propagation, those pathways may be assumed to be non-existent; i.e. those network connections will be lost.

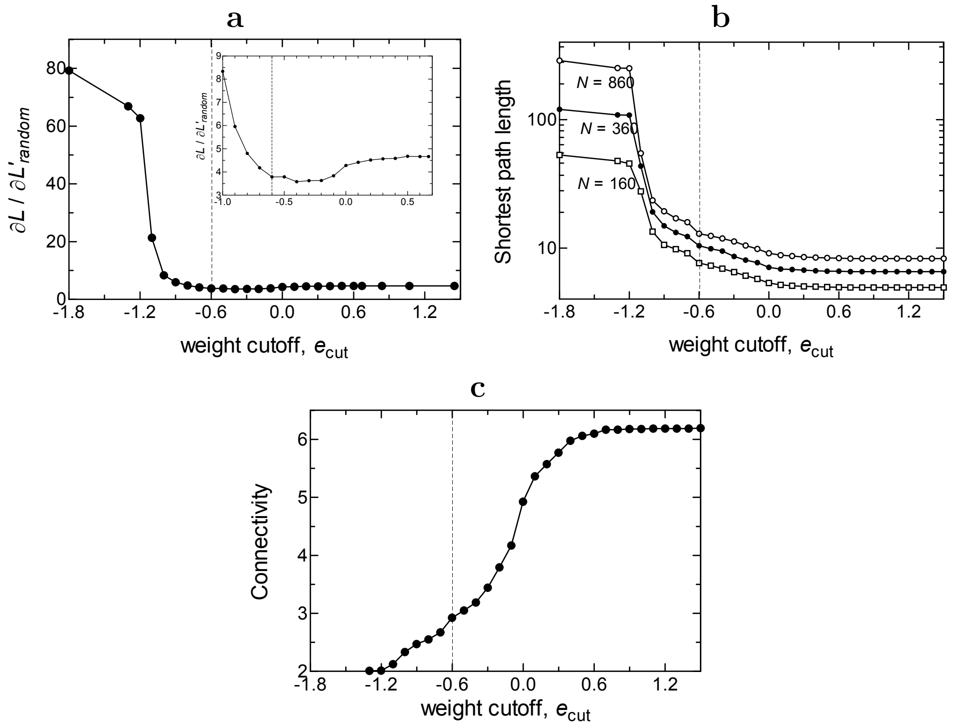


Figure 3.4. Change in network parameters of the sub-networks.

3.3.1 Properties of the residue network under varying degrees of external perturbations.

Usually, the impacts imparted on the protein in its usual environment will be intermediate between the two extremes of small perturbations and large impacts. Our analysis in figure 3.3 shows the operational limits of these molecular machines: We may classify those perturbations that delete nearly half the non-bonded contacts from being functional (i.e. $e_{cut} = 0.0\mathrm{k_B T}$) as everyday events. The change in the average path length of the protein relative to the change in that of the randomly rewired counterpart ($\delta L'/\delta L'_{random}$, where L' refers to path length on the sub-networks with the lower average connectivity, z^*) remains fixed for that range (figure 3.3). The latter quantity is shown for the whole range of values of e_{cut} in figure 3.4a. In the same range of values, the average shortest path length, a size dependent quantity, is also constant (figure 2.4b). The change in the average number of neighbors of a node is also relatively small, decreasing from 6.2 to 5 (figure 3.4c). Noting that two of these neighbors are located along the chain, at $e_{cut} = 0.0\mathrm{k_B T}$ an average node has lost one of its four non-bonded neighbors.

Further removal of the links signifies even larger perturbations to the protein. Up to ca. $e_{cut} = -0.6\mathrm{k_B T}$, where the shortest path lengths on the sub-networks coincide with the strong paths of the original weighted residue networks (marked by the dashed lines in figures 3.4a-c), the quantity $\delta L'/\delta L'_{random}$ shows a decreasing trend (inset to figure 3.4a). In the range of $e_{cut} = -0.60.0\mathrm{k_B T}$, the increase in L is less than a factor of two for all sizes of proteins, whereas its value increases logarithmically beyond that cut-off ($e_{cut} < -0.7\mathrm{k_B T}$; see figure 3.4b). The logarithmic dependence of the path length on chain size is also preserved in this range (see figures 3.2 and 3.3). Note that at this critical value of the cut-off, only about one non-bonded contact per average node remains (figure 3.4c).

Representative proteins of $\alpha, \beta, \alpha/\beta$ types are shown in figure 3.5; ribbon diagrams of the structures deposited in the Protein Data Bank [106] are shown in the first column. All non-bonded contacts (thin lines) superimposed on the backbone (thick lines) are shown in the second column. The strongest links that form the underlying structure and that give the polymeric chain its protein-like path lengths are shown in the third column. 14, 21, 13, and 18 % of the non-bonded contacts remain in these proteins Any other interactions added to these create redundancies that contribute to the robustness of the structure so that the protein is able to function under the harsh conditions of the cell. In reality, depending on the size and direction of the impact, some of the weaker links that are located far from that site may be preserved; i.e. we do not expect the links to be lost hierarchically. Nevertheless, the proteins reaction to the perturbation, as measured by the average path lengths of the effectively remaining contacts, is relatively insensitive to size and direction, as long as the most cohesive of the interactions remains intact.

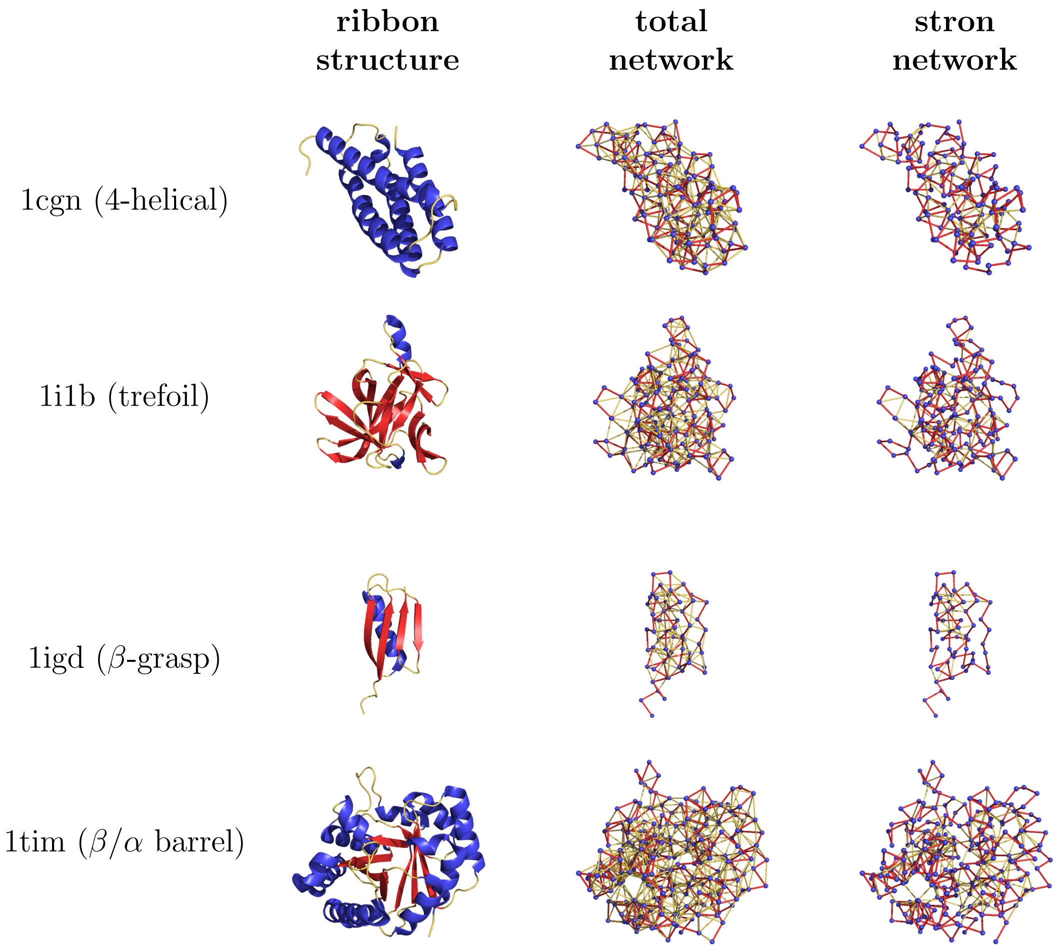


Figure 3.5. Example networks from proteins with common folds.

3.3.2 Practical application: Optimal paths in interacting proteins

We postulate that residues, frequently found along the paths connecting a receptor ligand pair, control the communication between the two proteins. Since binding is an event that requires exchange of large amounts of energy, in this treatment, we use the optimal paths with strong disorder which emphasize the largest barriers to be crossed along the way. Using the benchmark set of 59 receptor-ligand complexes [105], we seek the pairs of residues that are most significant in determining key interactions.

We first record the pairs that form bridges between receptor and ligand for every path that originates in the receptor and ends in the ligand; i.e. residue i is located on the receptor and residue j is located on the ligand and they are connected within the network formed by the protein protein complex. We then take into account the fact that the propensity of a selected amino acid type being located along the interaction surface significantly varies, as reported by Ma et al. [114]; e.g. TRP, ARG and GLN are the residues that are found most frequently on the interface. Therefore, we normalize the probability of finding a residue pair along the strong pathways, $p_{i\leftrightarrow j}$. Thus, the conditional probability, $p(i \leftrightarrow j \mid i, j)$, can be computed by relating the probability that the pair actually appears along the selected paths, to the probability of each of the residues in the pair being located on the interface, q_i and q_j:

$$p(i \leftrightarrow j \mid i, j) = \frac{p_{i\leftrightarrow j}/q_i q_j}{\sum p_{i\leftrightarrow j}/q_i q_j} \tag{3.6}$$

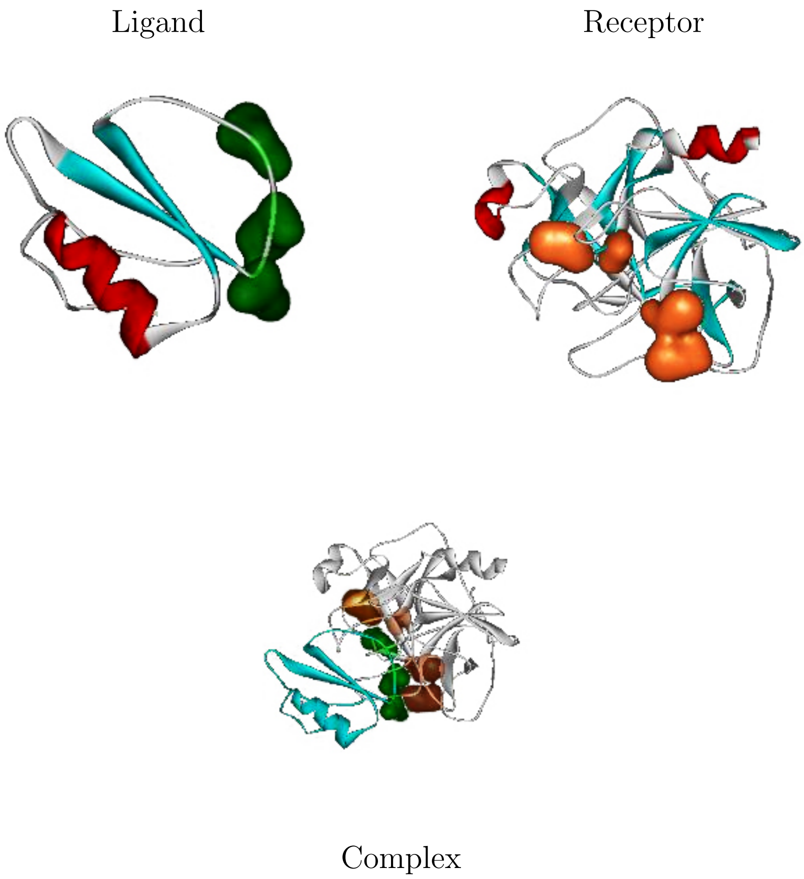
$p_{i\leftrightarrow j}$ is assumed to be proportional to the frequencies that these pairs are observed in the interface along the strong paths determined in this study. q_i and q_j are taken to be proportional to the propensity of the residue to be found in the interface of either the ligand or the receptor, as reported in the literature [114]. The

Table 3.1. Residue pairs that appear in the interface with significantly enhanced probabilities.

Residue Pair			**Propensity-normalized**	**Contact Potential**
(Receptor	$\rightarrow$	**Ligand)**	**probability** $p(i \leftrightarrow j \mid i, j)$	**(units of** $\mathrm{k_B T}$**)**
ILE	$\rightarrow$	VAL	0.130	-0.98
ALA	$\rightarrow$	ILE	0.041	-0.64
ILE	$\rightarrow$	ILE	0.039	-0.71
ILE	$\rightarrow$	LEU	0.036	-1.04
GLU	$\rightarrow$	LYS	0.032	-0.09
LEU	$\rightarrow$	ILE	0.030	-1.04
VAL	$\rightarrow$	VAL	0.027	-1.15

resulting conditional probabilities of the most significant pairs are listed in Table 3.1, along with the value of the TD contact potential.

Note that the pairs that are used in the paths consist mostly of the hydrophobic-hydrophobic interaction types, though not necessarily appearing in the order of cohesive energy. In fact, if all amino acids are grouped in the broadest sense of hydrophobic, polar, charged, and GLY, over 42% of all pairs that appear along the interface and that are on the strong paths make hydrophobic-hydrophobic contacts. Furthermore, the interactions need not be symmetric; in fact, the most significant pairs have ILE on the receptor and VAL on the ligand (normalized probability is 0.13). The reverse arrangement does not appear to be significant. A similar observation is also made for the ALA - ILE pair. In contrast, ILE and LEU pairs appear to be involved in specific interactions, though not with a significant preference for the ligand or the receptor. One example ligand-receptor system of α-chymotrypsin in complex with eglin c is shown in figure 3.6. Bridging residue pairs that are on the largest number of pathways between the receptor and the ligand are shown in orange and green, respectively. The interacting pairs are (enzyme - inhibitor): PHE^{39} - TYR^{49}, PHE^{41} - LEU^{47}, VAL^{213} - LEU^{45}, TRP^{215} - LEU^{45}; note that LEU^{45} interacts with two residues. Note that in the large interaction surface of the protein pairs, it is possible to identify four key interactions utilizing three residues on one protein and four on the other.



Ligand

Receptor

Complex

Figure 3.6. Example receptor-ligand system of the enzyme eglin c in complex with the inhibitor α-chymotrypsin; PDB code: 1acb.

Chapter 4

Relationship between k_{nn} and k for complex networks

A connection between the local and global network properties and the underlying structure of these self-organized molecular systems has yet to be established. The problem becomes increasingly complicated, when there are deviations from a random network, where interactions between neighbors, their neighbors, *etc.* are not negligible. However, it is possible to tackle this problem with a bottom-up approach. To address this, we derive a relationship that relates the degree of a node to the average degree of its neighbors in the presence of clustering. We show that this relation holds for various spatial networks that are obtained from self organized systems, condensed soft matter and regular crystalline structes.

We finally show that the correlations extend into farther neighbors of a node for the subset of spatial networks studied here. Thus, it is possible to extend local information to global information for some network structures.

4.1 Relationship between k_{nn} and k in graphs with uniform clustering

The generating function, $G_0(x)$, for the probability distribution of vertex degrees k is given by [115],

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k \tag{4.1}$$

where $|x| \leq 1$, p_k is the probability that a randomly chosen vertex on the graph has degree k, and its distribution is normalized with $G_0(1) = 1$. The $G_0(x)$ function generates the probability distribution of degrees, capturing all the discrete probability values through the derivatives property,

$$p_k = \frac{1}{k!} \frac{d^k G_0}{dx^k}\bigg|_{x=0} \tag{4.2}$$

The n^{th} moment of the distribution can thus be calculated from

$$\langle k^n \rangle = \sum_k k^n p_k = \left[\left(x \frac{d}{dx} \right)^n G_0(x) \right]_{x=1} \tag{4.3}$$

In particular, the average degree of a vertex is $\langle k \rangle = z = \sum_k k p_k = G_0'(1)$.

If one randomly chooses m vertices from a graph, then the powers property of the generating function provides a route to generating the distribution of the sum of the degrees of those vertices by $[G_0(x)]^m$.

We define outgoing edges from the first neighbors of a randomly chosen vertex as those that connect to vertices that are different from the first neighbors of the originally chosen vertex. It is first necessary to define the generating function for the distribution of the degree of the vertices one arrives at, along a randomly chosen edge. That vertex will be reached with probability proportional to its degree, kp_k, so that the normalized distribution is generated by

$$\frac{\sum_k k p_k x^k}{\sum_k k p_k} = \frac{G_0'(x)}{G_0'(1)} \tag{4.4}$$

Starting from a randomly chosen vertex and following each of its edges to arrive at the k nearest neighbors, each of the vertices arrived at will have outgoing edges that is given by the degree of that vertex less the edge that one arrives along and the edges that interconnect these nearest neighbors, or backlinks, b. Thus, the generating function for the outgoing edges from each vertex is,

$$G_1(x) = \frac{\sum_k kp_k x^{k-1-b}}{\sum_k kp_k} \tag{4.5}$$

Note that b itself depends on k.

The number of backlinks, b, is given in terms of the clustering coefficient, C, around a given node with degree k. Using the definition of C, with the number of interconnections, I, between its first neighbors, $C = I/[k(k-1)/2]$, the average number of backlinks for each of the k neighboring nodes is, $b = 2I/k = C(k-1)$. This will lead to the generating function for outgoing edges as:

$$G_1(x) = \frac{\sum_k kp_k x^{(k-1)(1-C)}}{z} \tag{4.6}$$

The generating function for the distribution of all outgoing links from the k neighbors of the original node is then obtained from the powers property:

$$G_k(x) = G_1(x)^k = \left[\frac{\sum_k kp_k x^{(k-1)(1-C)}}{z}\right]^k \tag{4.7}$$

The average number of outgoing links is computed from the first moment of the generating function

$$G'_k(1) = \frac{k(1-C)(\langle k^2 \rangle - z)}{z} \tag{4.8}$$

k_{nn} is the nearest neighbor correlations, defined as the total number of neighbors of a given node which emanates from a selected node of k neighbors. Thus, it is given by the sum of the number of outgoing links, the backlinks per k neighbor and the k

links that connect the original node to the first neighbors:

$$k_{nn} = \frac{G_k'(1) + 2I + k}{k} = Ck + \frac{\langle k^2 \rangle (1 - C)}{z} \tag{4.9}$$

Note that, for a finite and constant clustering coefficient, k_{nn} is always expected to be linear in k_n, with slope C. The intercept, on the other hand, depends on the degree distribution. For example, for a Poisson distributed network, such as residue networks as was shown in, $p_k = z^k e^{-z}/k!$, the relation takes the form

$$k_{nn} = Ck + (1 + z)(1 - C) \tag{4.10}$$

Further note that this development is not only true for constant C, but also for all networks where C is independent of k. This is because the summation in equations 4.6 and 4.7 are again directly evaluated for $\langle C \rangle$.

4.2 Model Systems

4.2.1 Self-organized molecular structures

In this subsection we describe how the networks are constructed for the self-organized molecular structures studied in this work.

Residue Networks (RN)

These networks are formed from experimentally determined protein structures obtained from the Protein Data Bank (PDB) [106]. For the RN calculations we utilize a set of 595 single-chain proteins with sizes between 54-1021 and having a sequence homology less than %25 [104].

Given a protein, each amino-acid is represented by a node that is centered at the position of C_β atoms, or the C_α atom in the case of glycine. Edges are added between two nodes, if they are closer than a selected cutoff, r_c. We call these residue networks. We use $r_c = 6.7$Å as in our previous work [15, 63, 102], which is the distance where the first coordination shell ends, as computed from the radial distribution function.

Micellar Networks (MN)

Unlike proteins, there is no experimentally available atomistic structure data for self-organized synthetic molecules. We therefore generate such data using Dissipative Particle Dynamics (DPD) simulations. DPD is a coarse grained simulation methodology. The equilibrium morphology of a group of beads is obtained by integrating out the fast motion of atoms. In addition to the random and dissipative forces, the net forces on the beads are soft and repulsive conservative forces. Then, the simulation is carried out by integrating Newton's law of motion. DPD simulations allow for reaching much larger length and time scales for macromolecular systems. Thus, self-organization of systems of large sizes can be observed. Here, we simulate the micelle formation by ABC type oligomers of styrene-co-perfluoroalkylethylacrylate in tetrahydrofuran (F beads). The co-oligomer consists of ten styrene monomers (A beads), seven perfluoroheptane monomers (C beads) and a linker monomer (B bead). The styrene monomers in the co-oligomer have a tendency to interact with the solvent, whereas the fluorinated parts prefer to segregate, thus resulting in micelle formation. The equilibrium morphology depends on the concentration of oligomer in the solution [116]. A general overview of the method and parameterization is given below:

Flory-Huggins mean-field theory of polymers explains the miscibility of polymer and a given solvent by comparing the free energy of the mixture before and after mixing. Similarly, with some modification, the DPD method can be employed to describe the thermodynamics of polymer blends, diblock copolymers and their blends with homopolymers. Then, the mixing energy can be related to the dimensionless Flory-Huggins interaction parameters, χ, which is the energy difference for taking a polymer (A) from its own environment and putting it into a solvent or another polymer (B), normalized to $\mathrm{k_B T}$. χ is defined as $\chi = z[\varepsilon_{AB} - \frac{1}{2}(\varepsilon_{AA} + \varepsilon_{BB})]/k_B T$, where z is the coordination number, ε_{AA},ε_{BB},and ε_{AB} correspond to the energies for AA, BB and AB interactions, respectively. For soft sphere interactions in DPD, the Flory-Huggins parameter χ can be written as, $\chi = 2\alpha(\alpha_{AB} - \alpha_{AA})(\rho_A + \rho_B)/k_B T$. α is related to the pair correlation function, and a_{AB} is the repulsion parameter between two corresponding beads. For a density of $\rho = 3$ DPD units, Groot and Warren developed the empirical relationship to calculate a_{ii} and a_{ij} parameters, $a_{ii} = 25k_B T$ and $a_{ij} \approx a_{ii} + 3.27\chi_{ij}$. The parameters used, and the forces involved are given with the details of these DPD simulations in [117].

We report results from systems, in which the volume fraction, ν, of the oligomers is 0.3, 0.6 and 0.9, respectively. We find that at these concentrations, the triblock co-oligomers self-organize into spherical, cylindrical and lamellar morphologies as the concentration is increased. Once the organized structures are obtained, we focus on one substructure from the simulated system; *e.g.*, the set of oligomers that form a complete sphere are taken as the structure, whose network will be formed. Thus, the spherical structure is made up of 50 chains, the cylindrical structure has 100 chains, and the lamellar structure has 150 chains. Finally, we concentrate on the fluorinated segments of these segments, which have self-organized, due to the driving forces inherent to the system parts. By computing the radial distribution functions around these beads, we find that the first coordination shell ends at 1.1 DPD units. We use this cutoff distance to form the network, whose properties are studied. Chain connectivity of a copolymer is preserved, regardless of the particle separation; *i.e.* $(i, i+1)$ connections are always present.

4.2.2 Other atomic/molecular structures

It is important to investigate the differences between the network properties of self-organized molecular structures, and other systems of atomic/molecular origin; in particular the effects of excluded volume and chain connectivity on the observed behavior must be investigated. To this end, we also study the structure of networks obtained from Lennard-Jones clusters (excluded volume) and polybutadiene melts (excluded volume and chain connectivity). The coordinate data are obtained as described below.

Lennard-Jones Clusters (LJC)

The structure of clusters of atoms is an area of intense scientific research, since the properties of materials become size dependent, when systems are small enough. By clusters, we refer to groups of atoms from tens to thousands of atoms. Lennard-Jones clusters (LJC) are a group of atoms that contain purely Lennard-Jones interactions between pairs of atoms. Geometric optimization of these clusters requires developing efficient search algorithms, since the conformational space available to a cluster of atoms increases explosively. The atomic coordinates of LJC for sizes 3-1000 are deposited on the Cambridge Cluster Database [118]. Many of them are described by icosahedral motifs with an incomplete core [119]. Here we examine clusters of sizes 350-550, in intervals of 50 atoms. The cutoff distance for adjacency matrix construction is 1.6Å [120].

Polybutadiene Melts (PBD)

We investigate networks constructed from PBD melts that have been obtained from molecular dynamics (MD) simulations. The system consists of monodisperse *cis*-1,4-PB of 32-chains, each with 32 repeat units (C_{128}). The initial coordinates of the system studied was prepared in the Amorphous Construction Module of the Accelerys Material Studio 4.4 [121] at a density of 0.92 gr/cm^3, which occupies a cubic box of 47 Å on each side. Minimization, pre-equilibration and integration of the equations of motions were done with the NAMD program [122]. The interaction potentials for PBD chains reported in [123] are adopted. For all simulations, 1 fs integration time step was used. Temperature and pressure were maintained constant in the MD simulations at their prescribed values by employing the Langevin thermostat-barostat. For the non-bonding interactions, the cut-off distance of 10 Å was used with a switching function activated at 8 Å.

To obtain well-equilibrated samples of PBD chains with the correct chain statistics, the initial structure, which is energy minimized for 10000 steps, is depressurized by placing the chains into a larger cubic box of 300 Å on each side. NVT simulations of this low-density system is carried out for 10 ns at 430^o K. We then cool the system to 300^{o}K by equilibrating for an additional 20 ns. Consequently, we compress it with NPT simulations at 1 atm at 430^{o}K for 1 ns. We check that the conformational properties (as measured by the characteristic ratio) and the thermodynamic measurable (e.g. thermal expansion coefficient and compressibility) are compatible with the values in reference [124]. The data used in the current calculations are obtained from highly pressurized PBD melts via NPT simulations at 100 GPa and 430^o K. We collect data for 50 ns. PBD melts are coarse grained by using the coordinates for the center of mass of carbon atoms in the butadiene repeat units. The cut-off distance for network construction is chosen at 5 Å, the ending point of the first coordination shell.

4.2.3 Lattice-based Network Models

To interpret the results obtained for the molecular structures, we attempt to find base-models that best describe their statistical and spectral properties. Here, we describe the regular lattices that are used as model systems for this purpose. Their generating functions and sample basic connections based on these lattices are shown in table 4.1.

Ring Lattice (RL)

This is a one-dimensional lattice of nodes residing on a ring. Each node is connected to $z/2$ predecessor and $z/2$ succeeding nodes, therefore having a total of z connections leading to constant values of network parameters given by $C = [3(z-1)]/[4(z-2)]$, $k_{nn} = z$, and $L = [N(N+z-2)]/[2z(N-1)]$.

Simple Cubic (SC)

Ths is a basic cubic crystalline structure, where the nodes are placed in the corners of a cubic lattice. Connection to only the nearest-neighbors would lead to $C = 0$. Therefore, each node is connected to its first and second nearest-neighbors.

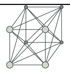
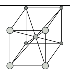
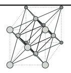
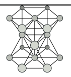
Body-Centered Cubic (BCC)

This also has a cubic unit cell, with the difference from the SC being an additional node in the center of each cube. Again, its first and second nearest-neighbors are connected to a node.

Face-Centered Cubic (FCC)

This is one of the close-packed structures. There are nodes at the corners and at the centers of the faces of a cube. The network is formed by connecting the nearest neighbors. This network is also known as second nearest neighbor diamond (2nnd) lattice. It was shown that residue coordination in proteins can be modeled

Table 4.1. Network models used and the generating functions for degree distributions.

Network type	Size, N	Generating function, $G_0(x)$
SC 	343	$0.024x^6 + 0.175x^9 + 0.437x^{13} + 0.364x^{18}$
BCC 	432	$0.005x^4 + 0.014x^5 + 0.056x^6 + 0.014x^7 + 0.111x^8 + 0.222x^9 +$ $0.005x^{11} + 0.056x^{12} + 0.222x^{13} + 0.296x^{16}$
FCC 	500	$0.008x^3 + 0.096x^5 + 0.384x^8 + 0.512x^12$
HCP 	500	$0.004x^3 + 0.032x^4 + 0.076x^5 + 0.028x^6 + 0.060x^7 + 0.080x^8 +$ $0.224x^9 + 0.048x^{10} + 0.064x^{11} + 0.384x^{12}$

as a distorted model of a 2nnd lattice [80].

Hexagonal Close Packed (HCP)

This is the other possible close-packed structure that can be formed with identical spheres. Nodes are arranged on a plane in a hexagonal formation, and stacked on top of each other with alternating order. The first nearest neighbors are connected to obtain the network.

4.3 Linear relationship between the first and second degree correlations

4.3.1 Residue networks (RN)

We apply the results derived in section 4.1 to RN constructed from folded protein structures. Previous studies on RN showed that these networks have high clustering, as opposed to their random counterparts, and have comparable shortest path lengths as the random networks; therefore, they can be considered as having

Table 4.2. Network parameters $\langle C \rangle$ and $\langle k^2 \rangle / z$ computed from the generated graphs and predicted from the least squares linear fit to k_{nn} vs. k curves.

		Calculated[c]		**Predicted**[d]	
		$\langle C \rangle$	$\langle k^2 \rangle / z$	$\langle C \rangle$	$\langle k^2 \rangle / z$
Residue Networks[a]	595 Proteins; $\langle N \rangle = 254$	0.38(0.02)	6.2(0.5)	0.35±0.01	5.8±0.2
	$N = 140 - 160$	0.38(0.02)	6.1(0.4)	0.32±0.01	5.7±0.2
	$N = 190 - 210$	0.39(0.02)	6.2(0.4)	0.32±0.02	5.8±0.4
	$N = 290 - 310$	0.37(0.01)	6.6(0.3)	0.36±0.01	6.2±0.2
Micellar Networks[a]	$\nu = 0.3$	0.45(0.14)	10.3(3.5)	0.40±0.02	10.5±0.8
	$\nu = 0.6$	0.43(0.15)	9.9(3.7)	0.51±0.02	10.2±0.8
	$\nu = 0.9$	0.41(0.15)	9.4(3.5)	0.51±0.02	9.6±0.6
Lennard-Jones Clusters[b]	$N = 350$	0.47(0.08)	15.1(7.2)	0.33±0.07	14.4±1.4
	$N = 400$	0.47(0.08)	15.3(7.1)	0.31±0.06	14.5±1.1
	$N = 450$	0.46(0.08)	15.4(7.0)	0.33±0.07	14.6±1.3
	$N = 500$	0.46(0.08)	15.5(6.9)	0.33±0.07	14.6±1.4
	$N = 550$	0.47(0.08)	15.6(6.9)	0.37±0.12	15.3±2.6
Polymeric melts[b]	T = 300 K, P = 1 atm	0.14(0.20)	3.8(2.7)	0.33±0.04	3.7±0.3
	T = 430 K, P = 106 atm	0.45(0.10)	12.8(6.2)	0.52±0.03	12.4±0.7

[a] Degree distribution is well-described by Poisson; therefore predictions by eq. 4.9 and 4.10 lead to the same result. Also $z = \langle k \rangle = \langle k^2 \rangle / z$ for these systems.

[b] Degree distributions are not well-described by Poisson. Predictions are made through eq. 4.9.

[c] Standard deviations calculated from the data are reported in parentheses.

[d] Error margins on the predicted values are reported.

small-world topology. In these studies, comparisons were performed for the average properties throughout the network between the RN and its randomly rewired counterparts. Although average values do confirm that RNs have small-world properties, detailed analyses of the individual parameters are needed to assess similarity with artificially generated networks.

In reference [15] it was shown that the degree distributions of RN are Poissonian; the mean is 6.2. Therein, it was also shown that the residues in the core have a mean clustering coefficient of approximately 1/3, whereas this value approaches 0.5 for the nodes that reside along the surface. Averaged over the set of 595 proteins, the clustering coefficient of RN has the value 0.38. In Figure 4.1, we display k_{nn} versus k data from three sets of proteins, $N = 140 - 160$ (48 proteins), $N = 190 - 210$ (29 proteins), and $N = 290 - 310$ (31 proteins), as well as the whole set of 595 proteins. The linearity between k_{nn} and k holds for all sizes of proteins, despite the size differences, as well as the deviation of the clustering coefficient distribution from Dirac delta function. We adopt Equation 4.10 to analyze the relationship between

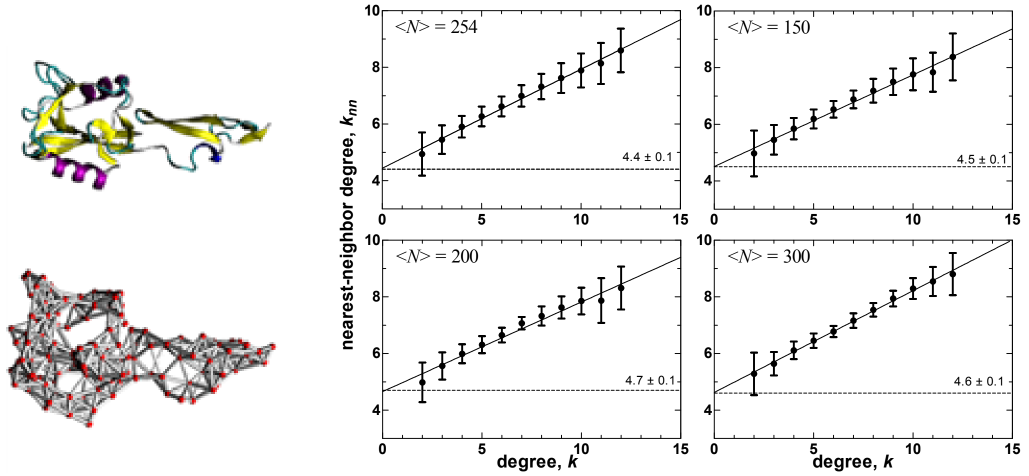


Figure 4.1. An example residue network (RN) where the sample protein (PDB code 1ESL), its network construction and averaged k_{nn} vs. k plots for proteins for four cases

k_{nn} and k in RN and we find that the slope can be characterized by the average clustering coefficient of the network. The values of $\langle C\rangle$ and z, calculated directly from the network and predicted via Equation 4.10, are listed in Table 4.2. Nodes with degree 1, 13, 14 and 15 are omitted since there are relatively small number of nodes with such degrees (¡ 25) to provide meaningful statistics. Within the error bounds, the predictions of theory are valid; the only slight deviation occurs as an underestimation of $\langle C\rangle$ for the smaller proteins where the surface effects (and the variance in C) are more pronounced. We shall elaborate further on the surface effects.

4.3.2 Micellar Networks (MN)

We expect other self-organized molecular structures to display network properties similar to the RN obtained from proteins, provided that they are thermodynamically stable and have a given average structure around which fluctuations are observed. Similar to the proteins, these structures follow certain organization rules due to the (in)compatibility of their chemical units with the solvent. Other environmental factors, such as the temperature or the concentration, play a role on the type of organization observed. As example systems, we choose micelles of different morphologies formed by the ABC type co-oligomers, whose coordinates were obtained from DPD simulations.

At low concentrations, these oligomers organize to form spherical micelles. As the concentration increases, adjacent spheres begin to merge and attain a cylindrical morphology. Further increase in the concentration results in the formation of lamellae. In Figure 4.2a, we display the spherical, cylindrical and the lamellar formations excerpted from oligomer concentrations of $\nu = 0.3$, 0.6, and 0.9, respectively. Styrene monomers, the linker beads and the perfluoroheptane monomers are represented as black, red and white spheres, respectively. Note that it is the core region (i.e. the fluorinated regions shown as white spheres) that maintains the stable morphology, while the corona formed by the red and gray beads shows large fluctuations in conformation. Thus, we use the coordinates of the white blobs to generate the MN. The degree and clustering coefficient distributions of three sample networks are shown in Figure 4.2b. It is important to note that, regardless of the type of self organization, these network parameters show a similar pattern. The degree distribution may be approximated by a Poisson distribution.

Similar to RN, analysis of the k vs. k_{nn} relationship for MN reveals a positive linear correlation, regardless of morphology (Figure 4.2c). The values of $\langle C \rangle$ and z, calculated directly from the network and predicted via Equation 4.10, are also listed in Table 4.2. Nodes with less than five and more than 15 connections are omitted, due to the lack of statistics of blobs with so few or so many neighbors. The slope of the best-fitting line is close to the average clustering coefficient. Thus, theoretical predictions from the slope and intercept of the k vs. k_{nn} relation show a good correlation with the numerical results.

The linear relationship between k_{nn} and k also predicts the increase in z with size in RN, as well as the decrease in z with concentration (and morphology change) in MN. The theory slightly underestimates the clustering coefficient of RN, whereas it overestimates that of MN. This is due to surface effects: in proteins, nodes along

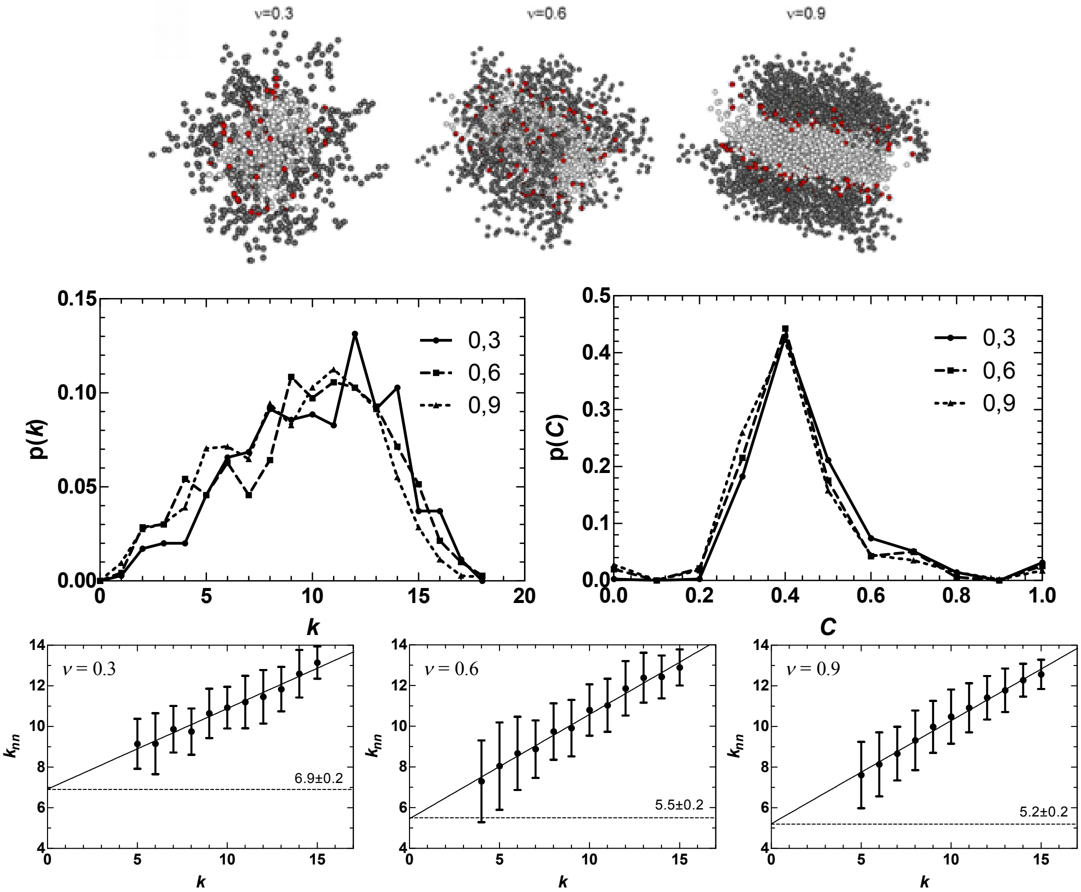


Figure 4.2. Self-organized micellar structures studied in this work at three different concentrations.

the surface have high clustering coefficients, as shown in reference [15], because these nodes have few links that are interconnected, increasing the average clustering coefficient. Conversely, in MN surface nodes along the core are connected to the solvo-phillic arms of the chains. These connections, which are omitted in the calculations, have the reverse effect on the average value of the clustering coefficient. Thus, the value of C predicted for both RN and MN reflects the network structure below the surface.

4.3.3 Lennard-Jones Clusters (LJC)

Atoms occupy a specified volume in space, and as a result, there is an upper bound on the number of neighbors that may be within the direct interaction range of a given node. Furthermore, since our nodes comprise of coarse-grained clusters of atoms that are not arranged in a spherically symmetric manner, the number of neighbors may be as large as 16 for some nodes. This is in contrast to the maximum coordination of 12 expected of regular lattices of spherical particles. All of the networks studied here have this property. However, the extent to which this excluded volume effect influences the predictions of the previous subsection is unclear. To further investigate this point, we study LJC, which are clusters of atoms of minimized energy that interact purely via Lennard-Jones interactions. We confine our attention to those within the size range up to 550, which is compatible with the network sizes of RN and MN studied in previous subsections. A sample three dimensional visualization of Lennard-Jones cluster is plotted in Figure 4.3.

We find a linear relationship between k_{nn} and k, as in the previous self-organized systems. We observe that the degree of these systems cannot be described by Poisson distributions. The clustering distributions, on the other hand, are identical to those of MN. We therefore utilize Equation 4.9 instead of Equation 4.10, which provides a prediction of the average clustering from the slope, but the ratio $\langle k^2 \rangle / z$ from the intercept. These results are also presented in Table 4.2. In all the LJC, we find $\langle C \rangle$ to be consistently underestimated by the theory, while the $\langle k^2 \rangle / z$

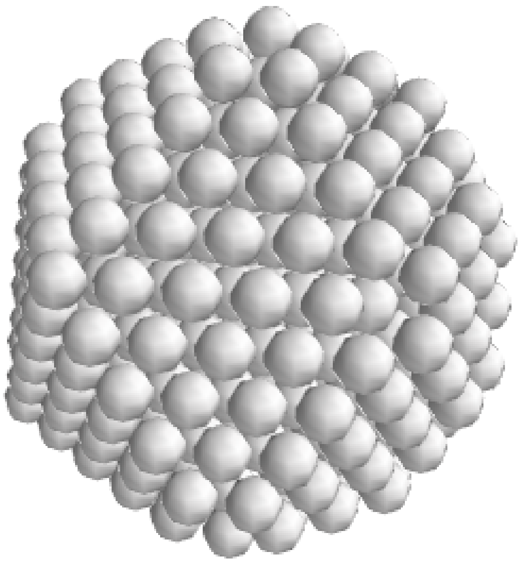


Figure 4.3. Three dimensional visualization of Lennard-Jones cluster with $N = 500$

values are well-predicted. Thus, although the excluded volume imposes restrictions on the degree distributions, in particular it leads to assortative mixing in the graph structure, it does not have a direct effect on the local clustering.

4.3.4 Polybutadiene Melts (PBD)

Finally, we study polymeric melts to discern the effect of connectivity on the statistical properties of the networks. The linear relationship between k_{nn} and k is also observed for this system at both moderate and high density. Degree distributions of both systems deviate from Poisson. At very high compressions, represented here by the system at 430 K and 106 atm, clustering distribution is very similar to those obtained for MN and LJC. On the other hand, at moderate densities, exemplified here by PBD at 300 K and 1 atm, the clustering has two maxima whereby nearly half the nodes have $C \approx 0$, while the rest have a peakish distribution centered at $C = 0.3$. This leads to $\langle C \rangle = 0.14$ with a standard deviation of 0.20. Predictions via Equation 4.9 largely overestimate $\langle C \rangle$, while the $\langle k^2 \rangle / z$ is well-predicted at moderate density. On the other hand, at very high compressions, both quantities are predicted via the theoretical fit, with a slight overestimation of $\langle C \rangle$.

Putting together the results obtained thus far, we conclude that the excluded volume leads to the assortative mixing of the local structure, described by the positive slope of between k_{nn} and k curves. Furthermore, the extrapolation of the curves to low connectivity ($k > 0$) leads to a prediction of the $\langle k^2 \rangle / z$ values. We observe this behavior regardless of the type of system studied. Additional constraints on the local organization of the beads would lead to further local structuring which is measurable by the slope of these curves converging to $\langle C \rangle$. We find that chain connectivity does not bring about such local organization of the beads, as shown by the PBD system at moderate density. However, systems attaining dense core structures do converge to this limit. Such high densities may be attained by imposing external factors such as the high pressure on PBD; alternatively, the core regions of self-organized systems prefer to realize such an arrangement due to the free energetic requirements of arranging chains with both solvo-phobic and solvo-phillic regions in a solvent that creates the driving force for the formation of the densely packed core.

4.3.5 Model networks

We utilize ordered networks to test the linear relationship between k_{nn} and k of Equation 4.9. Ring lattice (RL) is homogeneous, therefore it shows a delta distribution for k_i. In the other network models, derived from three-dimensional regular structures, there is a wider distribution due to the finite size of nodes and the resulting edge effects.

The key to using Equation 4.9 to describe the relationship between k_{nn} and k is the uniform distribution of backlinks, which necessitates either a narrow distribution of the clustering coefficient, or the distribution of clustering coefficient is constant with respect to the degree (equations 4.6 and 4.7). For the homogeneous RL, this is a spike centered at $C = 2/3$. For all the other model networks, the value is 0.4 ± 0.1 (Table 4.3). Clustering distributions for these networks are not delta, but the invariant clustering distribution with respect to degree would still result in a linear k_{nn} versus k. Figure 4.4 shows mean clustering values and standard deviations

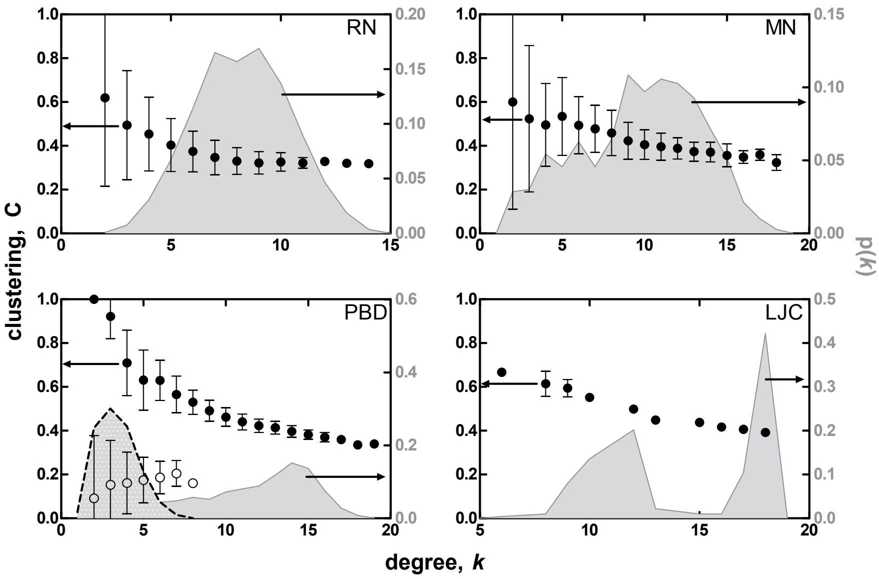


Figure 4.4. Averaged clustering vs. k plots for RN ($N = 190-210$), MN ($\nu = 0.60$), PBD (filled circles for high pressure and empty circles for low pressure cases) and LJC ($N = 500$) superposed onto degree distributions for corresponding networks.

for these model networks superposed onto degree distributions for a corresponding network. For RN and MN, average clustering shows little variations especially for the degrees that have statistically significant samples. Two cases of PBD show quite different characteristics for C versus k relation. The low pressure system is loosely packed, which results in a low degree network, and the clustering values are close to zero. Therefore, average clustering values with respect to degree vary considerably from the network average clustering coefficient. On the other hand, for a high pressure, the resulting system is much denser and has a higher degree of clustering. For degrees that are observed more frequently, average clustering remains relatively constant. For LJC, average clustering follows a linear trend with a negative slope, yet the slope is small (-0.02), and relation could be assumed independent with respect to degree.

Although the parameters for model networks deviate from the assumptions slightly, the k_{nn} versus k relation may still be investigated to see the extent to which the variances are tolerated. Using linear fits to the k_{nn} versus k curves for SC, BCC, FCC and HCP models (in RL, k_{nn} values are identical for each node), and applying

Table 4.3. Network parameters $\langle C \rangle$ and $\langle k^2 \rangle / z$ computed from the generated graphs and predicted from the least squares linear fit to k_{nn} vs. k curves.

	Calculated		**Predicted**	
	$\langle C \rangle$	$\langle k^2 \rangle / z$	$\langle C \rangle$	$\langle k^2 \rangle / z$
SC	0.44(0.10)	14.8(6.8)	0.45±0.02	14.9±0.4
BCC	0.43(0.06)	11.9(5.3)	0.46±0.09	10.7±1.6
FCC	0.41(0.08)	10.4(4.7)	0.34±0.03	10.4±0.3
HCP	0.41(0.10)	10.2(4.6)	0.38±0.06	9.9±0.8

equation 4.9 to obtain the slope and the intercept values, we predict the average clustering coefficients of these networks and the ratio $\langle k^2 \rangle / z$. These are compared to the actual values from the model networks in Table 4.3. There is a high degree of accuracy with all models, despite the fact that the clustering coefficient distributions are not delta functions in these networks.

4.4 Local motifs and higher order relations

Relating global parameters, such as the shortest path length, to local parameters, such as the degree distribution, relies on the higher order degree correlations. If one has the number of r^{th}-nearest neighbors for a node i, $n_{i,r}$, calculation of the average shortest path is a simple summation;

$$L_i = \frac{1}{\sum_r n_{i,r}} \sum_r r * n_{i,r} \tag{4.11}$$

For random networks with known degree distributions, higher order nearest neighbors can be calculated in the limit of large number of nodes. Then the average shortest path can be estimated as a function of the number of first and second nearest neighbors [115]. However for an arbitrary network, this calculation is not straightforward. Although correlations within higher degrees will affect the results, the linear relation between k and k_{nn} obtained for certain systems (Section 4.1)

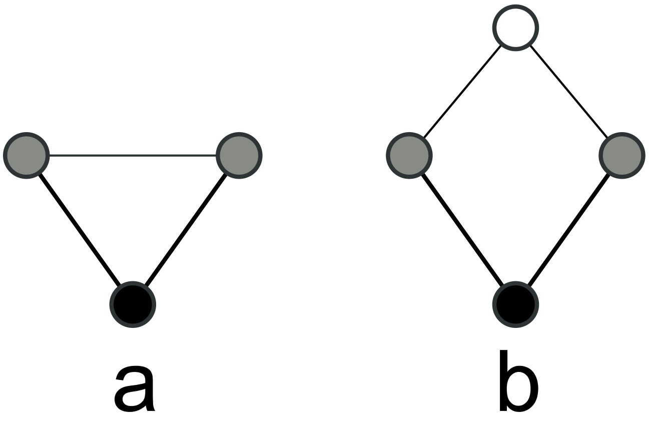


Figure 4.5. Possible cases where the neighbors of a node's neighbors does not result in second-nearest neighbors.

would help in estimating the second nearest neighbors.

k_{nn} accounts for the average number of neighbors that a node's neighbor has. Although this definition covers the number second neighbors, there are cases that need extra attention. For a node (black) with non-zero clustering coefficient, some of the neighbors (grey) are inter-connected, thus forming a triangle (Figure 4.5a). Although this connection between neighbors is accounted for in k_{nn}, it should be discarded from the number of second-nearest neigbors, as we have don in Equations 4.6-4.9.

There may also be cases where the neighbors (grey) of a node (black) share a common neighbor (white) and they will form a diamond shape (Figure 4.5b). Here the new node is a second-nearest neighbor, but it is counted twice.

Here, the problem lies with the counting of triangles and diamonds in a network. The triangle count is directly related to the clustering coefficient, and it can be deduced from that, while the diamond count is not trivial. Figure 4.6 shows the dependence of number of diamonds to triangles per node with respect to each other for residue networks of three different sets of varying sizes, between 140-160, 190-210 and 290-310. There is a clear linear relation between these two geometric motifs in residue networks. Therefore, the actual counting of diamonds can be skipped, and

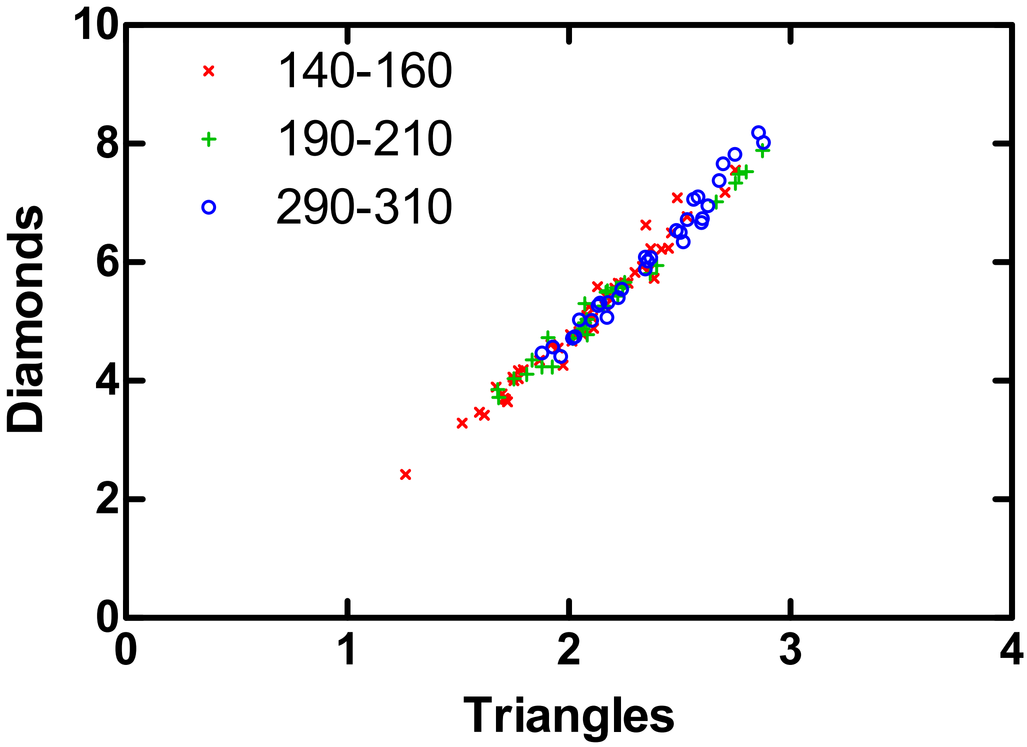


Figure 4.6. Average diamond per node vs. average triangle per node for residue networks.

the number of diamonds can be approximated using a simple linear equation. As a future study, higher order motifs, such as pentagons, heptagons *etc.*, will be investigated in relation to the number of triangles. Then, coupling this information with higher order measures like k_{nn} might lead to an accurate approximation of higher order nearest neighbors. From that point, relating local measures, such as degree and clustering coefficient, to global parameters like shortest path length, would be straigtforward.

4.5 Discussion

The study presented in this chapter is based on the premise that network structures are better classified by the distributions of their network parameters rather than the average values. One example is with approximating residue networks derived from proteins with the regular RL. Although it is relatively easy to generate a corresponding RL with a few randomly rewired links having the same average degree and clustering coefficient as the RN, neither the second degree correlations nor the global properties (e.g. path length) will be reproduced with this approach. However, a comparison of distributions of the many parameters involved is not straightforward. We further show that, in these spatial networks, such correlations may further extend into higher order neighboring relations, allowing a connection between local and global network properties.

To make the problem tractable, we derive a relationship between k_{nn} and k for networks with arbitrary degree distributions, but with narrowly distributed finite clustering (Equation 4.9). This subset of constraints is relevant to the study of complex systems, because the results directly apply to the study of self-organized molecular structures, which are characterized by Poissonian degree distributions, and narrowly distributed clustering coefficients. As such, the derived linear relationship between k_{nn} and k displayed in Equation 4.10 should apply. In randomly-packed chain systems, this relationship is expected to be lost, as is observed when the corona region of the micellar networks (*i.e.* the disorganized parts of the chains protruding into the solvent) is also included in the calculations (data not shown). We validate the derived relationship on several model networks based on three dimensional regular structures, as well as those constructed from proteins and micelles of self-organizing co-oligomers.

The close packed structures emerge as model systems that approximate the network properties of self-organized molecular structures: They yield both the local statistical averages and their distributions, in addition having to similar spectra. However, the crystal-based structures are highly regular, not only displaying non-logarithmic size dependence of path lengths, but also highly deviating from the narrow distribution of L_i that is characteristic of efficient information transfer in self-organizing molecular structures. Using these model networks as the basis, one may generate networks, by introducing a few random links, whereby the local properties are preserved, while the desired global properties are approximated. In a forthcoming study, we shall report a detailed analysis of RN and close-packed structures with such random links.

Chapter 5

Packing of proteins

For the case of spatial networks, graph tools may be used to identify certain characteristics of structures. Here we consider residue networks in detail and utilize methods from graph theory to characterize packing structure in proteins.

So far it has been found that the local ordering of residues plays an important part in the overall behavior. Furthermore, in terms of network parameters, a residue network is better modelled with a three dimensional lattice, specifically a close packed lattice structure like HCP or FCC [80]. To further investigate the lattice-protein relation, we propose a scheme for mapping the protein onto a crystalline lattice.

5.1 Model

Since the interaction of residues are our main concern, a coarse grained approach to protein structure is necessary. As in section 3.1.1, coarse graining is achived via taking the C_α atoms of residues, thus reducing the protein to a single chain. C_α atoms along the chain are evenly distrubuted with a spacing of 3.7Å. Our goal here is to find a self avoiding walk on a crystalline lattice basis that will capture the protein chain as much as possible. We selected the basic cubic lattices (simple cubic (SC), body-centered cubic (BCC), face-centered cubic (FCC)) and hexagonal closed pack (HCP) as our base lattices. Mapping is optimized by the Metropolis Monte Carlo (MMC) method [125] with the minimizing function selected as the root mean square deviation (RMSD) from the protein chain.

As the walk on the lattice traverses along first nearest neighbors, the lattice is formed, such that the first nearest neighbor distance is equal to the average C_α distance for the protein in hand. Although it is possible to start with a random self-avoiding walk, in order to improve convergence performance, we start with a conformation that is closer to the protein. This starting conformation is obtained by first aligning the first two residues of the protein chain with two points on the lattice. Then the next point is selected from the nearest neighbors on the lattice that is closer to the next residue, as well as preserving self-avoidance. This procedure is repeated for all the remaining residues. Although this starting conformation is fairly optimal, it just aligns the first two residues, and the rest is formed accordingly. Therefore considering the overall alignment of the chains, there might still be room for optimization. This optimization is done by the Metropolis Monte Carlo method coupled with a quaternion based orientational alignment, in order to compare the original and generated chains. In this algorithm, we choose a site at random and flip it to a position, which will preserve the backbone connectivity as well as self-avoidance. The acceptance is carried out with a MMC scheme, wherein the energy function is simply set as RMSD.

The algorithm is schematically explained in Figure 5.1. First, the protein (red line) is positioned on the lattice with two consecutive residues aligned with lattice points. Then a best-fitting, self-avoiding chain (green line) is estimated by selecting the lattice sites that are closest to the residue sites. After the initial self-avoiding chain is constructed, a random transformation (either a site flip or a bond flip) is applied to the green chain, and a new conformation is obtained at blue line. Then, the protein is aligned to this new chain with a quaternion-based fitting algorithm. Root mean square deviation (RMSD) of the new structure is calculated

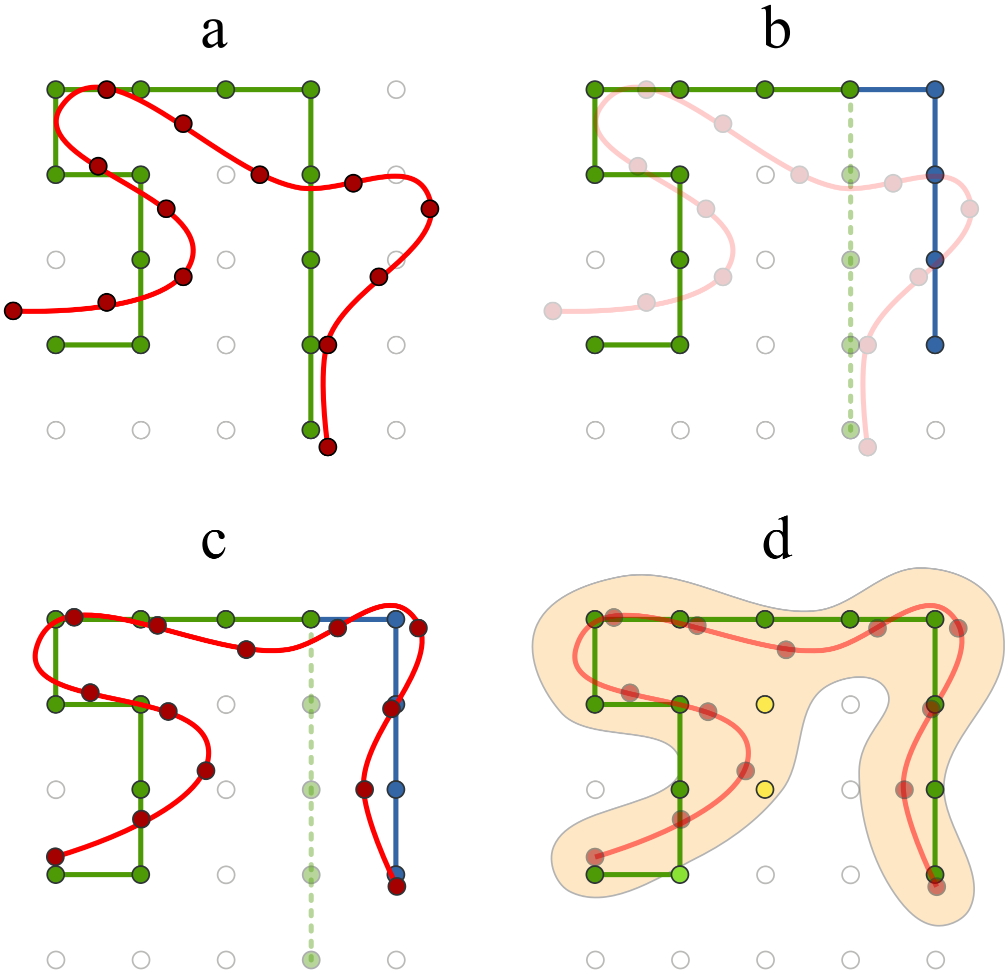


Figure 5.1. Schematic representation of the fitting algorithm.

and compared with the previous RMSD. The new conformation is accepted using a Metropolis Monte Carlo scheme. Steps b and c in the figure together conforms to a single Monte Carlo step. After the final chain is obtained, the packing ratio for lattice estimation is calculated with the use of the solvent-accessible surface (red area) of the protein. Chain sides that are inside (dark green dots) or outside (light green dots) the surface are identified, as well as the lattice points that are inside the surface but not used (yellow dots) for the protein fitting.

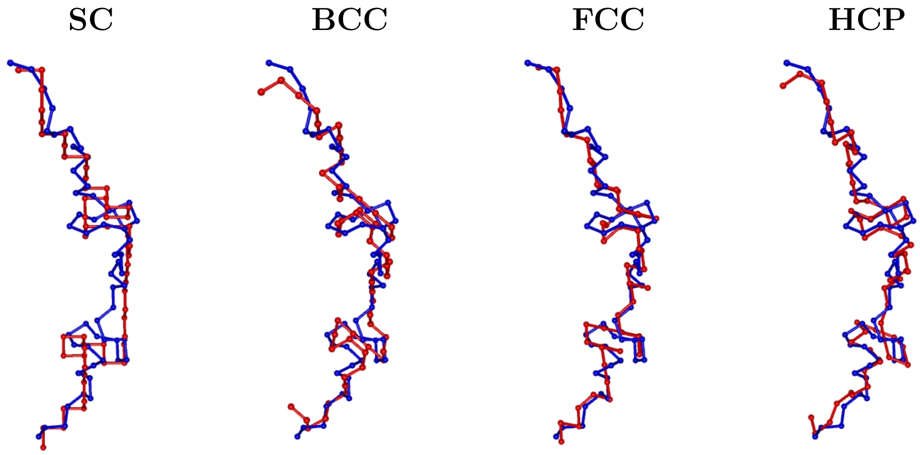


Figure 5.2. Protein (blue) and predicted lattice chains (red) for 1aaf.

Table 5.1. Average RMSD values of self avoiding chains.

	140-160	**190-210**	**290-310**
SC	3.89±0.11	4.11±0.11	4.19±0.13
BCC	4.04±0.14	3.14±0.08	3.07±0.07
FCC	2.13±0.04	2.20±0.04	2.22±0.04
HCP	2.14±0.03	2.20±0.04	2.17±0.03
RCP	2.15±0.04	2.22±0.03	2.18±0.03

5.2 Results

MMC simulations were run for three different sets of proteins with sizes between 140-160, 190-210 and 290-310. For each protein, five simulations were performed. At each simulation 5000 MMC steps of bond flipping followed by 20000 MMC steps of site flipping. Bond flipping captures the overall conformation, whereas site flipping provides local optimization. Figure 5.2 shows the three dimensional structure for the protein and the fitted chains for 1aaf.

RMSD values for the final predicted chains are shown in Table 5.1. For larger proteins the SC model performs worse than the other lattice bases. On the other hand, BCC leads to better results, as the chain length increases. SC imposes stricter constraints and has a lower coordination order than BCC. Therefore, as the size increased, it becomes more difficult to conform to these constraints in SC, impairing its performance.

To further quantify the quality of fit of various lattice types, we incorporate the relative orientations of the residues. Bond-orientational order parameter defined by Steinhardt *et al.*,[126] is a well-established metric in the study of packed spheres. This parameter is a measure of the distribution of a residue's neighbors in three dimensional space. Neighbors are defined as residues that are closer than a given cut-off value, r_{cut}, for the residue of interest. Therefore, we can use this measure to analyze the quality of fit for lattice self-avoiding chains. Bond-orientational order parameter for residue i is defined as;

$$Q_l(i) = \left(\frac{4\pi}{2l+1} \sum_{m=-l}^{l} \left| \frac{1}{N_b(i)} \sum_{n=1}^{N_b(i)} Y_{lm}[\theta(\vec{r}_n - \vec{r}_i), \phi(\vec{r}_n - \vec{r}_i)] \right|^2 \right)^{1/2} \tag{5.1}$$

where $Y_{lm}[\theta(\vec{r}_n - \vec{r}_i), \phi(\vec{r}_n - \vec{r}_i)]$ are the spherical harmonic functions for a bond vector from residue i to n, θ and ϕ are the polar angles of this bond. $N_b(i)$ is the total number of such contacts of residue i with distances closer than r_{cut}. Since this definition of the order parameter only depends on the vector between residues, for even values of l, it is independent of the selection of origin and orientations of coordinate axes. For a single protein, we average $Q_l(i)$ over all residues to obtain the order parameter for the protein:

$$\langle Q_l \rangle = \frac{1}{N} \sum_{i=1}^{N} Q_l(i) \tag{5.2}$$

Among the different choices for l, Q_6 is commonly employed as the bond orientational parameter, because it concurrently yields non-zero values for hexagonal close packed, cubic (simple, body centered, and face centered) and icosahedral configurations [126]. Note that, in order to obtain a full description of a system, all Q_l with different even l values should be calculated. Thus for assigning a system to a given lattice type, all these Q_l parameters should match.

For a system with cubic packing symmetries, Q_2 is zero, therefore it will not provide any non-trivial information for comparing SC, BCC and FCC systems studied in this work. Figure 5.3 shows orientational order parameters for $l = 4$, $l = 6$ and $l = 8$ for various r_{cut} values averaged over 48 proteins with sizes in the range 140-160. Lattice fits (lines) overlaid onto orientational order parameters obtained from the original protein coordinates (grey shaded area). For small r_{cut} values, Q_l values for lattice fits show small variations from the protein Q_l. This is expected, since lattice point approximations are not perfect due to the constraints from lattice structures. In Q_8, the discrepancy between SC and protein persists to higher values of r_{cut}, indicating that SC fails to capture longer range order in proteins. For Q_6, the difference between BCC and protein is larger for short range order. Although underlying symmetries for FCC and HCP are quite different, it is important to note that for all three orientational order parameters presented in Figure 5.3, chains approximating proteins using these underlying lattice points show similar ordering.

Among all regular lattices, FCC and BCC result in the best possible fits. It is also important to note that results of FCC and BCC are consistent for different protein sizes. It can be argued that close-packed structures have higher coordination numbers; therefore they have a higher degree of freedom in choosing neighboring lattice sites. We note that all the lattice structures have distinct symmetries, and one lattice system can not be replicated with the other. For example, if the protein itself had a conformation compatible with, say BCC, the corresponding lattice fit would have outperformed all the others.

5.3 Which closed packing?

Almost identical performance of FCC and HCP bears the question that whether it is possible to distinguish these two models. A common property within these structures is that they are both maximally packed models. To see if the packing is the driving property, we next introduce a random close packed model. FCC and

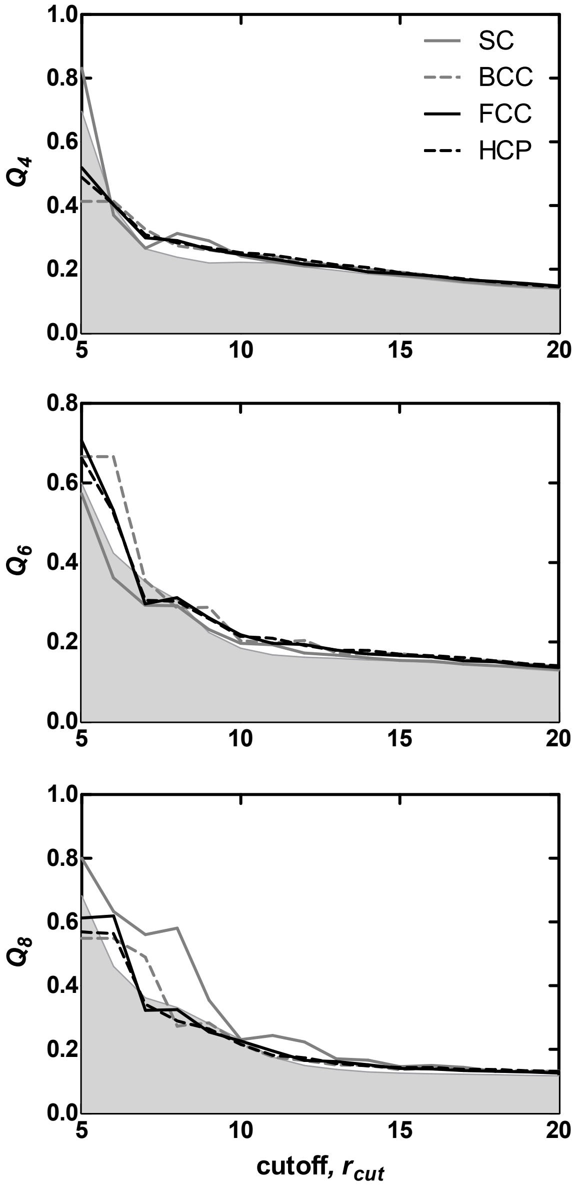


Figure 5.3. Comparison of Q_4, Q_6 and Q_8 for lattice fits (lines) and protein chains (grey shaded area) for sizes in the range 140-160

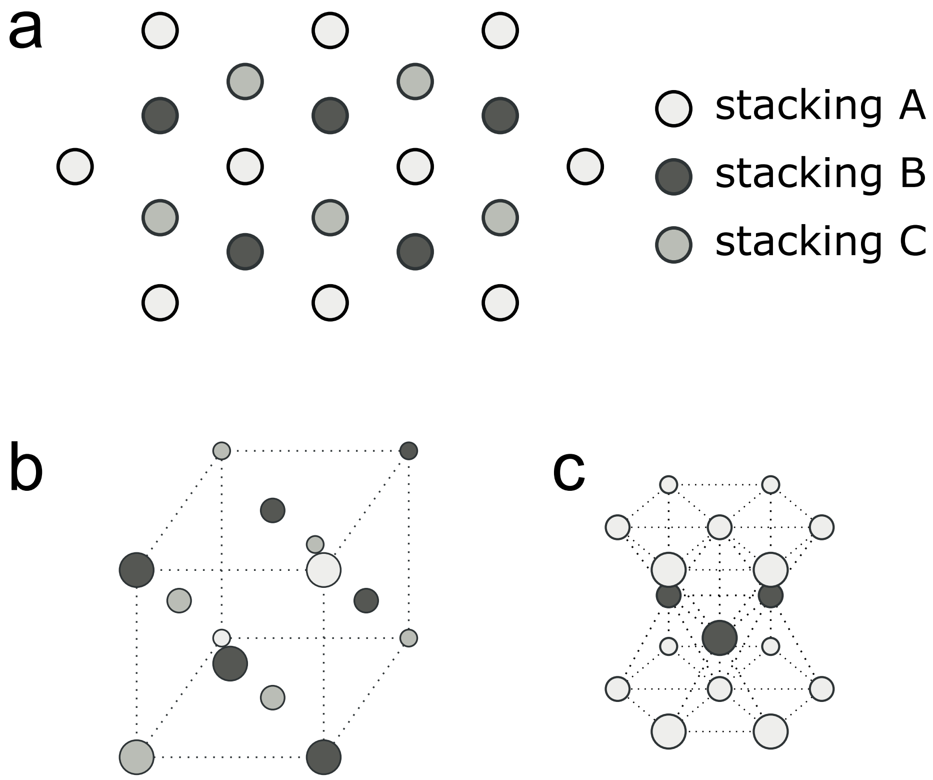


Figure 5.4. Hexagonally packed systems. a) Three possible hexagonal stackings that would result in a close packed lattice. b) FCC lattice with A, B and C layers shown. c) HCP lattice with A and B layers shown.

HCP both are composed of stacked hexagonal packed layers. While HCP is formed by the doubly alternating sequences such as ABABAB.., FCC has a triple ordered sequence such as ABCABC..., where B and C are obtained by two possible shiftings of hexagonal layer that would conform maximum packing (shown in figure 5.4).

Any random sequence of A, B and C, with the condition that no two consecutive layers are the same, would result in the same packing density as FCC and HCP. Thus, it is easy to construct a lattice that will not have the overall symmetries of FCC and HCP, but is still packed with the densest conformation. This provides an ideal basis for testing whether a particular close packed system is preferable to the other. Therefore we applied MMC fitting of proteins to randomly ordered stacking sequences that are denoted as random closed pack (RCP).

Average RMSD values for RCP lattice are also presented in table 5.1. Within error bounds, the performance of RCP is identical to that of FCC and HCP. Thus, the proteins may be modeled as a lattice as long as it is maximally packed.

5.4 Network parameters

Besides the spatial deviation, the fitting procedure may be analyzed with network methodology. Here, we construct the networks from proteins and predicted lattice chains with the same methodology mentioned in section 3.1.1. A cut-off of 6.7 Å is selected as before. We compare the distributions for network parameters below (see figure 5.5. These are the degree distributions (a), average next-nearest neighbor distributions (b), clustering coefficient distributions (c) and shortest path length distributions (d) for lattice fits (lines) plotted together with the protein distributions (grey shaded area). From the degree distribution, with the exception of BCC, it may be concluded that the lattice chains would result in nodes with higher degree. This can be attributed to the fact that lattice structure constrains the chain to a more dense conformation thus giving rise to a higher number of contacts. This is also reflected in k_{nn} distribution. Although, individual degrees of nodes vary the overall C and L distributions remains same more or less. Despite the slight differences in the distributions, the predicted lattice chains capture the network specific properties of proteins considerably well.

5.5 Packing fraction

We are then faced with the question of what differentiates proteins from close packed structures. Insofar as the network approachs of the present thesis is concerned, the adjacency matrix of a perfect close packed lattice is modified by removing links. This corresponds to voids within the packed protein structure similar to the voids occuring in the close packed optimized structures of LJC discussed

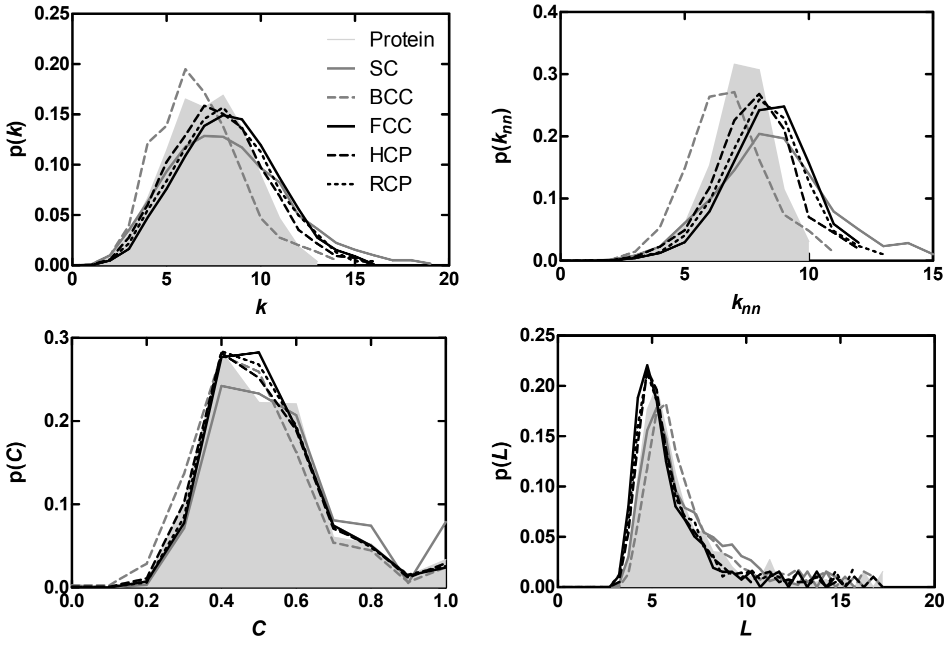


Figure 5.5. Network parameters for original proteins and best fit lattices for sizes $N = 190 - 210$.

in section 4.2.2. We therefore analyze the final structure following the best-fitting procedure. We consider the protein occupying a volume that is determined by the solvent accessible surface area of the backbone only [127]. Then we characterize the lattice points with respect to the volume occupied by the protein. It is important to note that all the lattice types pack proteins into these volumes with considerable voids. More than one third of the lattice sites (35% for SC, 41% for BCC, 37% for FCC and 37% for HCP) that reside in the solvent accessible surface was not used for residue fitting (yellow points in figure 5.1d). The fraction of lattice sites of the self avoiding chain that arere outside the covered volume is 40% for SC, 28% for BCC, 17% for FCC and 17% for HCP. These results are in aggreement with the RMSD values. They also suggest the number of eliminated adjacency matrix elements of an original close packed structure.

Chapter 6

Spectral properties of networks

In the previous chapters we have discussed the spatial networks we have studied mainly by statistical quantitites such as degree, clustering coefficient, path lengths, as well as their distributions. However one may also obtain a plathora of information from a spectral characterization of the network structure, which we investigate next.

The Laplacian of a network is used extensively in the literature. Networks are characterized using the spectra of the Laplacian. One variant of the Laplacian is the normalied Laplacian, $\mathbf{L}^*$, as described in equation 2.5 which is used in this study to identify characteristics of the networks.

The advantage of using the normalized Laplacian spectra is that the eigenvalues are in the range 0 and 2. This makes it ideal to compare networks with different sizes. Moreover, normalized Laplacian also contains significant information about the network, such as motif replication, bipartiteness and connectivity [98, 99, 100, 101]. Although there are isospectral networks with different adjacency matrices, these networks with identical spectra share similar properties in terms of network statistics. Therefore, the normalized Laplacian spectra may be used to distinguish networks with similar properties, that otherwise have similar statistical properties.

6.1 Edge cutting in residue networks

In terms of networks statistics, the procedure of edge cutting described in chapter 3 reveales that there is considerable redundancy in residue networks. Therein, networks are constructed with connecting residues that are closer than a predefined cutoff value (6.7 Å) and weights are added to the edges depending on the interacting aminoacid types. Weights are selected from well known residue-residue potentials of Thomas-Dill [59] and Miyazawa-Jernigan [103].

After network formation, edges are cut systematically. A cutoff potential, e_{cut} is selected and edges with interaction potentials larger than e_{cut} are removed while keeping the chain connectivity, regardless of the interaction potentials of bonded aminoacids. This is done to mimic the changes in the protein when a large disturbance such as bonding to a ligand occurs. Under such circumstances, information pathways that are not stable (i.e. high energy interactions) would be severed and no longer be used. In chapter 3, we monitored the average path length in reduced networks with changing weight cutoff to capture the effects of these events on communication pathways in proteins.

Here, we repeat the procedure of removing edges and monitor the normalized Laplacian spectra distribution for 48 proteins with sizes between 140-160. Figure 6.1a shows the fraction of edges with weights larger than e_{cut}. Edges that conform to chain connectivity are excluded, since they are not removed in the deletion process. There are only a couple of residue-residue interactions with potentials greater than $0.6\mathrm{k_B T}$, so it is expected that significant changes in network structure would occur below this cutoff. At $e_{cut} = -0.1\mathrm{k_B T}$, approximately half of the non-bonded edges are deleted and at $e_{cut} = -1\mathrm{k_B T}$, 90% of the non-bonded edges are removed from the network.

Then we characterize the removed edges according to their distances along the chain, which we call "contour distance". The contour distance of an edge between

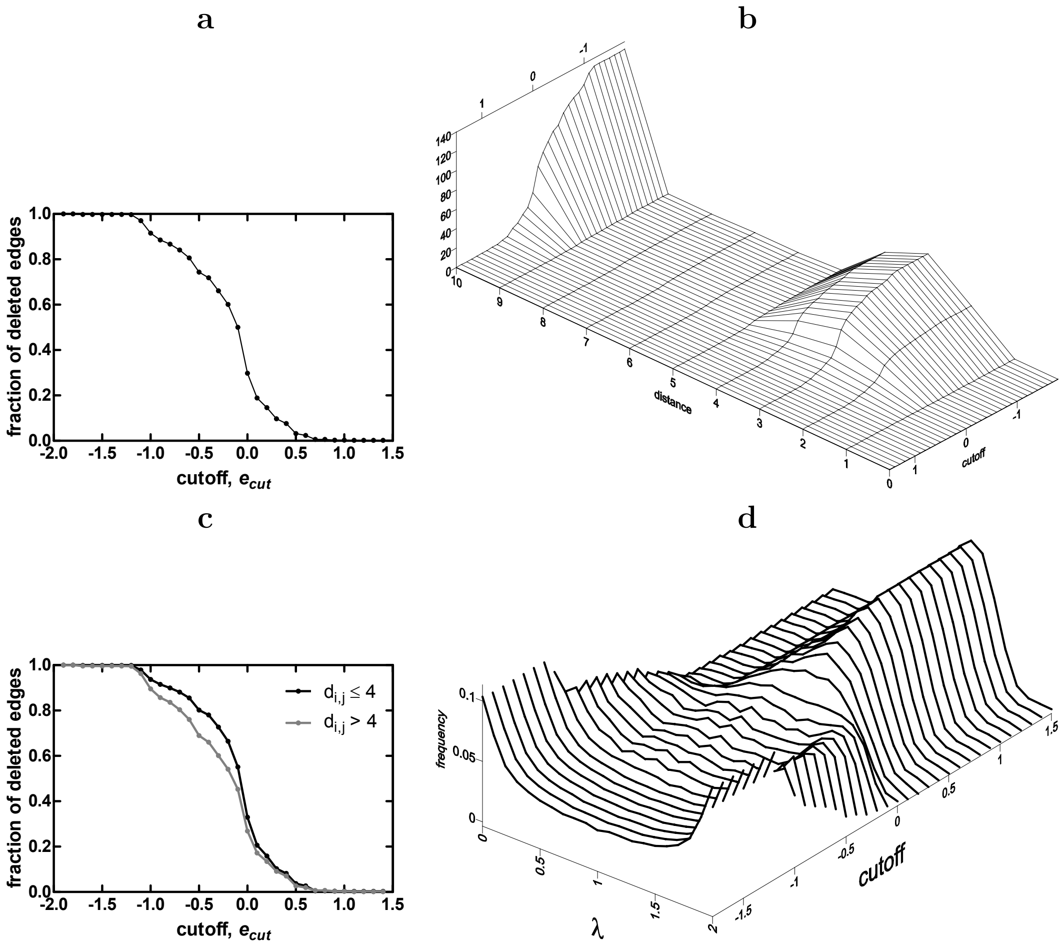


Figure 6.1. Change in a) fraction of deleted edges, b) sequential distance of deleted edges, c) fraction of short and long range edge deletions and d)normalized Laplacian spectra distribution with weight cutoff, e_{cut} in units of $\mathrm{k_B T}$.

two residues, $d_{i,j}$, is defined as the number of bonds between those along the chain. So, if residues are numbered as consecutive integeres along the chain, the distance for an edge that connects i^{th} and j^{th} residues is $d_{i,j} = |i - j|$. Figure 6.1b shows the number of edges with potentials greater than e_{cut} grouped by the distance along the chain, where edges with $d_{i,j} \geq 10$ are combined and shown as 10 in the figure. For all the non-bonded edges about 50% are in the region $d_{i,j} \leq 4$, with $d_{i,j} = 3$ as maximum. These would mostly correspond to the interactions within alpha-helical regions and turns. The other haf of the long range interactions occur between residue pairs that have $d_{i,j} > 4$ combined.

Figure 6.1c displays the fraction of edges with interaction potentials greater than e_{cut} where edges are grouped as short range contacts ($d_{i,j} \leq 4$) and long range contacts ($d_{i,j} > 4$). This shows the difference between short range contacts and long range contacts. Long range contacts are formed with aminoacids with more cohesive interactions (more negative values), so at a given weight cutoff a larger fraction of the long range edges remain intact. This supports the results of chapter 3 that the shortest path length does not change significantly up to $e_{cut} = -0.6\mathrm{k_B T}$.

Figure 6.1d shows the change in spectra distribution with weight cutoff. As a descriptor of global network properties, eigenvalue spectra distribution is more sensitive to changes in the network structure than shortest path length. For even a very small number of edge removals ($e_{cut} = 0.5\mathrm{k_B T}$), the eigenvalue spectrum reflects this change with minor variations. Still, the overall profile remains constant until $e_{cut} = -0.1\mathrm{k_B T}$. After that, the peak around $\lambda = 1$, which is associated with local motifs, starts to decrease and reaches to a minumum at $e_{cut} = -0.6\mathrm{k_B T}$. Further removing edges, connectivity of the graph decreases and this is reflected with the increased number of eigenvales near zero. Moreover, as a linear chain is a bipartite graph, the eigenvalue spectrum is symmetric around 1 and $\lambda = 2$ is present, thus leading to the increase in eigenvalues around 2. Considering the contour distance information provided by figures 6.1b&c and 3.2, we conclude that the essential information for constructing a protein-like network depends on proper long-range contacts.

6.2 Edge rewiring in residue networks

The previous section indicated the differences in edges in terms of the contour distance. Long range edges in proteins provide fast information transfer, whereas short range edges mostly take part in the local structure of the network. In order to investigate the contribution of these different edge types systematically, we propose a selective rewiring of edges according to their chain distances.

The network for a given protein is constructed, as before, by taking the C_β atoms of residues (C_α for Glycine) as representative points for the nodes and adding connections between nodes that are closer than a cutoff value that is obtained from the radial distribution function of residues in space. After constructing the networks, edges are randomized selectively according to their contour distances. The connections in the network are rewired randomly for two complimentary cases. In one case, connections with contour distances smaller than a predefined value are kept and the rest is rewired, so that short order interactions remain intact whereas long range interactions are randomized. In the other case, connections with contour distances larger than a value are held and connections closer along the chain are rewired; i.e. the long range interactions are conserved while the short range effects are randomly distributed.

In order to eliminate differences resulting from variations in degree distributions, rewiring algorithm works under the constraint that initial degree for a node remains constant. The algorithm starts by finding and breaking the edges that need to be rewired under given conditions which leads to a partial network with nodes that have free edges to be connected. Then a random pair of nodes that were not neighbors in the original structure is selected and connected with an edge. This process is continued until there are no nodes with free edges. It is possible to arrive at a point where there are free edges, yet there are no possible non-neighbor pairings available. In such cases, a random edge that was rewired previously is broken and new possible pairings are searched randomly. This breaking and connecting loop

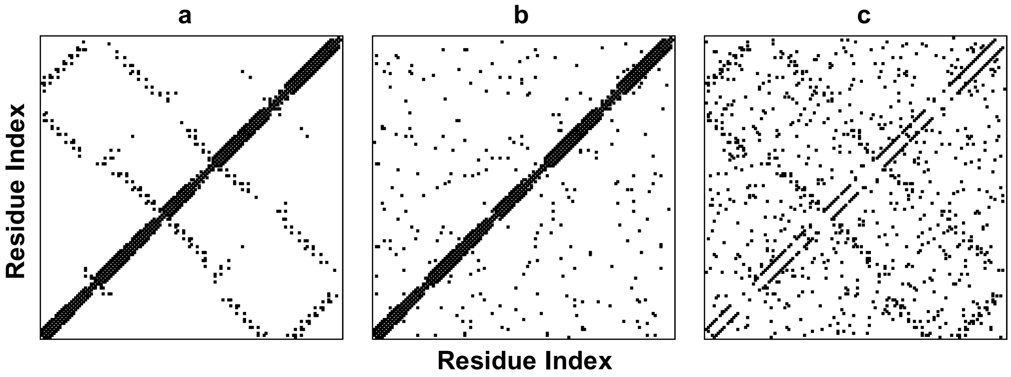


Figure 6.2. Sample adjacency matrices for protein 1AEP.

is repeated until no free connections remain. Figure 6.2 shows sample adjacency matrices of protein 1AEP before and after rewiring. Black dots on point (i, j) represents the presence of an edge between nodes i and j. Figure displays adjacency matrices of the original protein (a) and rewired networks with keeping the short range contacts (b), where all edges with $d_{i,j} > 4$ are rewired and with keeping long range contacts (c), where all edges with $d_{i,j} < 4$ are reqired.

We applied the rewiring algorithm to residue networks constructed from folded protein structures with sizes between 140-160 residues. Average C distribution, k_{nn} distribution and k vs k_{nn} plots for a these proteins for various rewiring scenarios are presented in figure 6.3. Figure 6.3a (upper row) shows these plots for rewiring networks where short range contacts are kept. Various scenarios are generated whereby only the chain connectivity is preserved (all contacts except $d_{i,j} = 1$ reqired), as well as all short range ones are preserved ($d_{i,j} > 4$ rewired) or part of the long range contacts are preserved ($d_{i,j} \geq 16$ rewired). It is important to note that k_{nn} distribution and k vs k_{nn} plots do not change dramatically with different randomization levels. However, clustering coefficients are effected noticeably. As more edges are rewired (i.e. keeping only $d_{ij} = 1$) the behavior of networks resemble that of a random network, where local clustering diminishes to zero.

For Figure 6.3b (lower row), similar figures are plotted by keeping long range interactions and rewiring local contacts. Clustering coefficient shows similar behavior to that of the counterpart in a. As more contacts are rewired, the network will

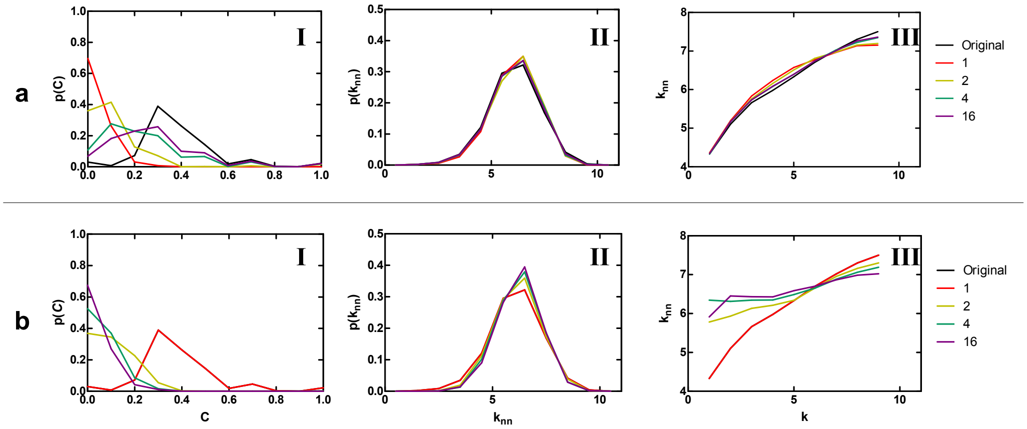


Figure 6.3. Network parameters for randomly rewired proteins by various (a) short and (b) long-range contacts

approach to a random network. k_{nn} distribution again, does not depend on rewiring, but the individual k_{nn} values are affected. This can be better seen in the k vs. k_{nn} plot of this set, where the slope of the curve decreases as more edges are rewired. Thus, low degree nodes are mostly peripheral nodes and they tend to have only local contacts, which are also closer to the bounds, thus having smaller degrees. So in the original structure, low degree nodes have low degree neighbors which results in a lower k_{nn}. When local contacts are rewired, most of the connections of these peripheral nodes will be rewired, and since the degree distributions are Poisson, on the average these will connect to higher degree nodes which results in an increase from the original values.

Furthermore, we monitor the spectra of the normalized Laplacian as the rewiring advances. In Figure 6.4a, local contacts are preserved and long range contacts are rewired. When keeping only $d_{ij} = 1$ (i.e. chain connectivity) and rewiring the rest of the connections, the spectrum of the network approaches to that of random network, i.e. a semicircle [128]. The overall properties of the spectra and the network is mostly captured by keeping further local contacts, for example when first 4 contacts are kept and the rest is rewired. Conversely by preserving long range connections and shuffling local contacts, the spectra resembles that of a random even for shuffling the first contacts only (i.e. keeping $d_{ij} \geq 2$). Thus, although chain connectivity is crucial for capturing overall spectra, it is not sufficient. Several local contacts should also be included along with random long-range contacts.

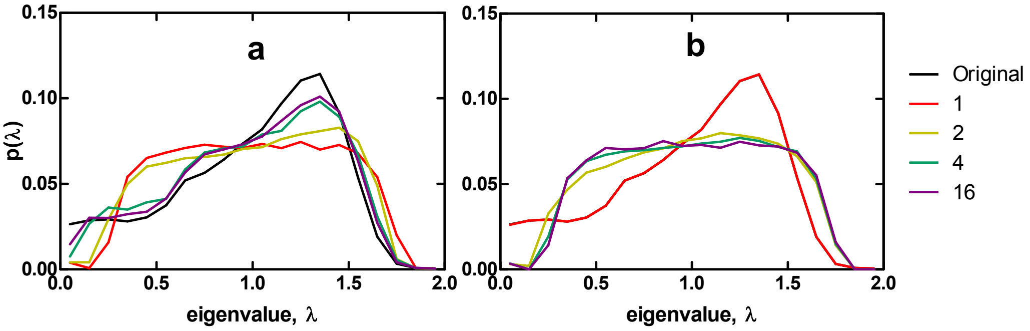


Figure 6.4. Normalized Laplacian spectrum for randomly rewired with preserving (a) short and (b) long-range contacts.

6.3 Lattice packings of proteins

Results of chapter 3 showed that residue networks and regular lattice structures share common properties in terms of graph parameters. In chapter 5, we used lattice based templates to generate protein chains. Although self-avoiding chains produced on regular lattice structures can approximate protein structures, the resulting chains occupied a volume with voids. Therefore perfect lattice models fall short of obtaining protein models in terms of network parameters.

Here, we investigate the best fitting chains obtained by Metropolis Monte Carlo simulations described in chapter 5 by considering the normalized Laplacian spectra distributions (see figure 6.5 for comparison of lattice chains (lines) to proteins (grey shaded area) with sizes between 140-160). Network parameters shown in figure 5.5 are not sufficiently distinguishing the lattice models in terms of their fit performances. However, normalized Laplacian spectra distributions present a better tool for comparison. Although the overal profile is obtained, SC and BCC deviates from the target spectra significantly. On the other hand, close packed structures (FCC, HCP and RCP) capture the structure of residue networks quite well. Slight increase of eigenvalues around $\lambda = 1$ in close packed structures compared to residue networks may be attributed to the increase in local order via additional constraints imposed on the self-avoiding chains due to the underlying lattice structure.

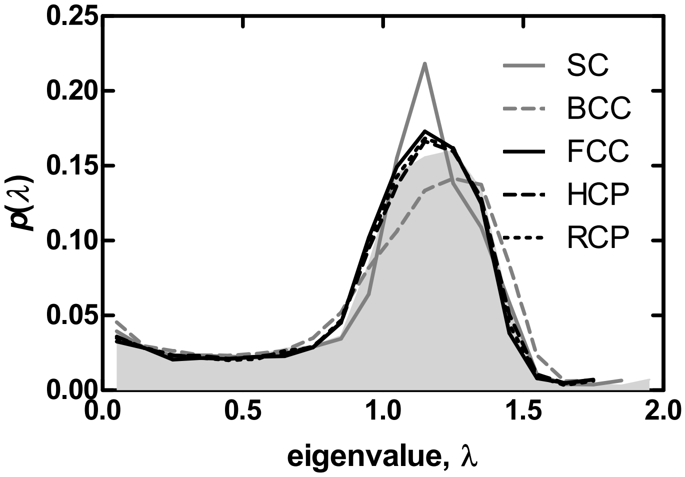


Figure 6.5. Normalized Laplacian spectra distributions for self-avoiding chains obtained by Metropolis Monte Carlo simulations outlined in chapter 5. Results from simple cubic (SC), body centered cubic (BCC), face centered cubic (FCC), hexagonal close packed (HCP) and random close packed (RCP) with the actual values from protein networks (grey shaded area). Distributions are averaged over 58 proteins with sizes between 140-160.

6.4 Other Model Networks

To further investigate how the spectra define other close packed structures, in addition to proteins, we analyzed normalized Laplacian spectra distribution for the systems that are investigated in chapter 3. Banerjee and Jost previously studied the spectra of many different model and real-life networks [129] and classified them according to the peak locations. The multiplicity of the eigenvalue $\lambda = 1$ is of particular interest, indicating the amount of motif duplications in the network. Also of importance is the closeness of the smallest non-zero eigenvalue to 0, which is a measure of the difficulty to disconnect the graph [98]. From another perspective, the density of the eigenvalues near $\lambda = 0$ is a measure of the collectivity in the network.

The spectra of RN ($N = 190 - 210$), MN ($\nu = 0.60$), LJC ($N = 500$), HCP ($N = 500$), PBD, HCP ($N = 500$) are shown in figure 6.6. On the lower bottom corner HCP with 3% randomly rewired edges is superposed on RN. Addition of a few long range edges affects the overall structure greatly. Rewired HCP spectra is closer to that of RN. The effect of changing morphology in MN is shown in the inset in the region $0.9 < \lambda < 1.5$. The spectra of the three morphologies are the same elsewhere. All spectra are characterized by a wide peak in the region $1 < \lambda < 1.5$, overlaid by a long tail at $0 < \lambda < 1$. They are truncated at $\lambda > 1.5$. There is evidence that the skewness of the eigenvalue distributions are directly related to the high clustering property of these systems, e.g., compared with the spectra presented in ref. [129] for systems of low $\langle C \rangle$ values. The lowest eigenvalue region is associated with a global connectivity in the system, a property only present in RN which comprise of single folded chains. Conversely, the distributions are peaked for systems lacking chain connectivity (kurtosis is positive for LJC and HCP, and negative for the others). Thus, although chain connectivity is not encoded in the network construction process; i.e. nodes are connected if they are within a cutoff distance of each other, regardless of residing on the same chain or not, the spectra distinguish between the many chain systems (MN, PBD) and collection of non-bonded particles (LJC, HCP) as opposed to the single chain systems of RN.

The overall spectrum of HCP is typical of the molecular structure networks. The spectra of MN and PBD are particularly well-reproduced by that of HCP (Pearson correlation coefficients between HCP/RN, HCP/MN, HCP/LJC and HCP/PBD are 0.82, 0.96, 0.82, 0.93, respectively.) The differences between the spectra of HCP and MN are most prominent in the region $\lambda > 1$ which contains information on the details of the local motifs, whereas the long-range effects (collectivity) in these structures are well described by the HCP scaffold. The larger discrepancy between the spectra of RN and HCP in the region $\lambda > 1$ is due to the wealth of local motifs occurring in the different arrangements of secondary structural elements of proteins. In fact, the situation may be rectified by randomly rewiring a few edges in the latter. We will thoroughly investigate these properties of RN in a forthcoming paper.

The spectra may be qualitatively used to identify some of the subtleties of these networks. For example, LJC spatially forms a packed structure due to the energy minimization. Although, there are variations from a perfectly ordered structure and missing lattice points exist, the core region shows considerable homogeneity. This can be observed via the peak at $\lambda = 1$ and the overall clustering of frequencies in the range $\lambda \geq 1$.

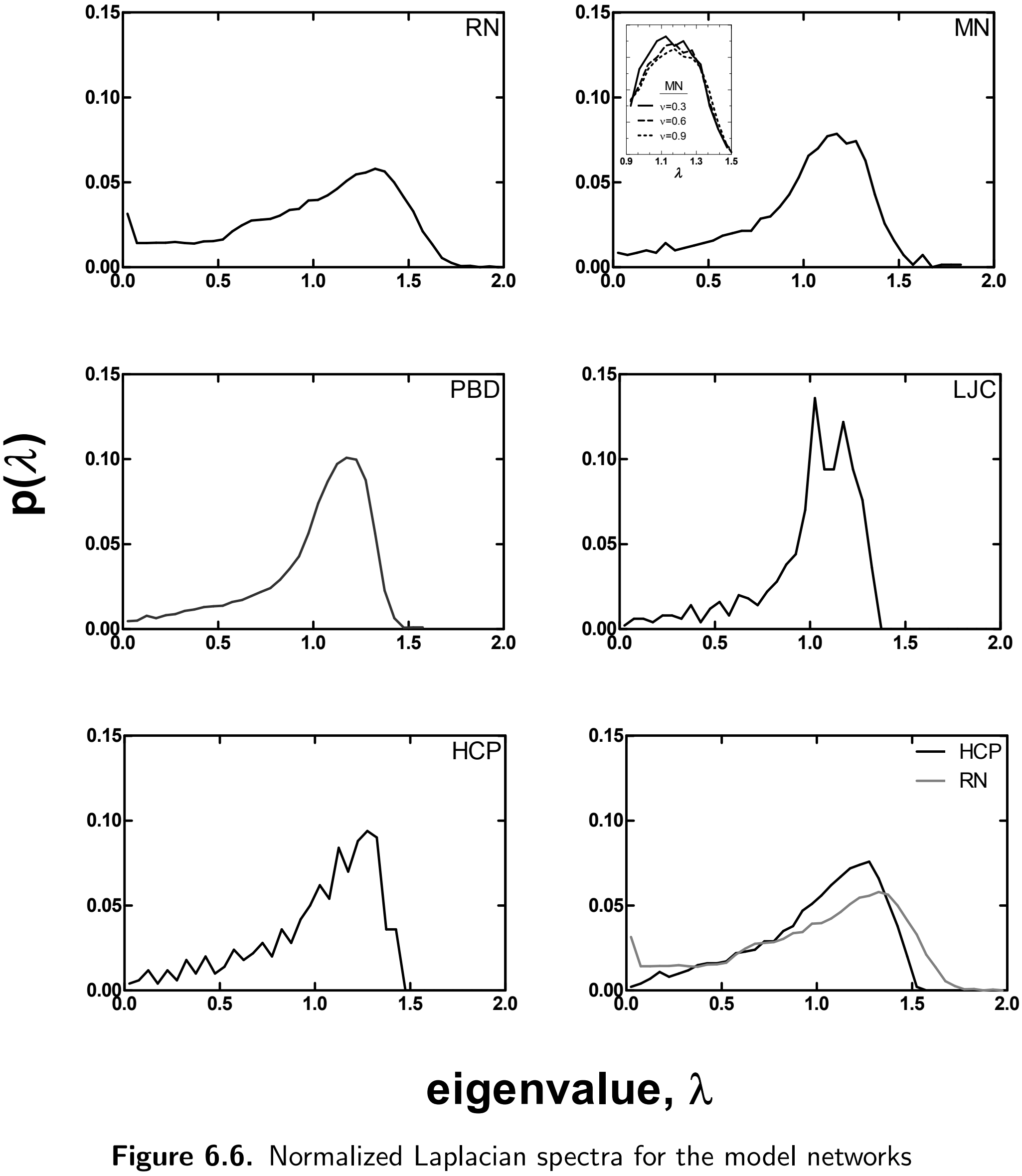


Figure 6.6. Normalized Laplacian spectra for the model networks

The spectra may also be used to identify the different morphologies of MN by focusing on the region in the vicinity of $\lambda = 1$. The region of $0.9 < \lambda < 1.5$ in the spectra of the different MN is magnified as an inset in figure 6.6, noting that the rest of the spectra are exactly overlaid in these systems. As the concentration ν is decreased from 0.9 to 0.6 to 0.3 (see figure 4.2), the morphology shifts from lamellar to cylindrical to spherical, accompanied by increased symmetry. Each structure is expected to contain a larger amount of the repetitive local motifs that make up the corresponding morphology. Indeed, the spectra contain this information with a shift towards the $\lambda = 1$ value as the spherical structure is approached.

6.5 Statistical or spectral characterization?

Throughout the study outlined in this thesis, we analyzed real structures by looking at their contact networks. Analysis of networks based on statistical measures such as degree, nearest neighbor degree, clustering coefficient and shortest path length, as well as the distributions of normalized Laplacian spectra.

The question then rises as to which measure (statistical or spectral) is important in classification and characterization of networks. Edge cutting method described in chapter 3, revealed that although certain statistical measures, such as shortest path length remained constant up to a certain deletion of edges, other parameters such as degree changed drastically (see figure 3.4). Conversely, the spectral distributions are quite sensitive to edge deletion and variations can be seen even for a small amount of disturbance in the network (see figure 6.1).

Similarly, for the self-avoiding chains generated by the algorithm outlined in section 5.1, statistical properties are insufficient in differentiating structural differences in networks (see figure 5.5). However, spectral distributions presented figure 6.5 outline small differences among these networks.

On the other hand, in the residue network rewiring scheme (section 6.2), rewiring long-range interactions have quite small difference in terms of spectral distributions (figure 6.4). But the change in networks can be observed more clearly in the clustering coefficient distribution (figure 6.3).

Concluding from these facts, statistical and spectral properties do not provide alternative to one another. In contrast, they give complemantary information about the network and one should consider both in order to characterize any network.

Chapter 7

Conclusions and Future Work

In this study, we utilized network models for condensed matter systems to analyze key properties that will aid in the classification of condensed matter systems. Networks are constructed from three dimensional stuctures with radial distribution functions of nodes to define the conditions for constructing contact maps. The ultimate goal of such sudies is not only to distinguish and classify different packing architectures in condensed matter systems, but also to derive coarse-grained potentials for constructing networks conforming to a given set of constraints [59].

- Systematic removal of edges in residue networks depending on the interaction potential of contacting residues revealed that they contain significant number of redundancies.
- By identifying information pathways in proteins using optimal path lengths, we find that in addition to chain connectivity, small number of long range contacts govern the information flow and conformational changes in case of extreme events.
- Furthermore, using this approach, we have been able to define key residues that form bridges between interacting proteins. The few key contact pairs may be used as primary links in identifying the interaction geometry, overlaid

by the energy lowering contributions from the rest of the pairs in solving protein-protein interaction problems.

- We derived a relationship between nearest neighbor and next-nearest neighbor correlations in networks. For networks with arbitrary degree distributions where the clustering coefficient is independent of degree distributions, k_{nn} linearly depends on k with a slope defined by clustering coefficient and intercept as a function of clustering coefficient and degree distributions.

- We showed that, the above-mentioned relation holds for various networks constructed from condensed matter systems, such as proteins, flurinated block co-polymer systems that form micellar structures, polybutadiene simulations under high pressure, clusters of identical particles obtained by minimizing Lennard-Jones potentials and perfect lattice structures.

- Combining the fact that geometric motifs such as triangles and diamonds are closely correlated in these systems, the relation between higher order neighbors and immeadiate neighbors may provide insight into the scaling between local and global parameters.

- Concentrating further on proteins, we analyze the spacial packing of residues in the folded conformation by comparing the protein chains to self-avoiding walks on regular lattice structures. We proposed a Metropolis Monte Carlo scheme coupled with quaternion-based fitting step to obtain lattice approximations of proteins.

- Comparison within different lattice systems revealed that close packed lattices such as face-centered cubic, hexagonal close packed and random close packed capture the spatial distribution of amino acids in proteins, indicating that the folded structure of the protein tends to pack maximally in order to obtain maximum coordination within the chain whereas the overall packing contains significant void fractions.

- The normalized Laplacian spectra distributions are used to characterize structural properties. Spectral distribution changes in residue networks confirmed the results of statistical measures. Further looking at the contacts that remains in the presence of large perturbations revealed that long range hydrophobic contacts play an important role in defining global information pathways in proteins. This is in paralel with the findings that hydrophobic interactions is key defining measure in protein folding and further research of these reduced systems may reveal important dynamics in folding mechanisms of proteins.

- To understand the role of short range and long range contacts, we apply a selective rewiring method to residue networks depending on the sequential distance of contacts. This reveals that the chain and the contacts that are closer along the chain define the local structure whereas contacts that are farther in terms of sequential distance govern the fast information pathways.
- Analysis of condensed matter systems considered in chapter 4 in terms of their spectral properties showed that overall structures of these systems behave similarly indicating that they conform to a family of systems.

References

[1] Watts, D.J. and Strogatz, S.H. [1998]. “Collective dynamics of ’small-world’ networks.” *Nature*, 393(6684):pp. 440.

[2] Albert, R., Jeong, H. and Barabasi, A. [1999]. “Internet - diameter of the world-wide web.” *Nature*, 401(6749):pp. 130.

[3] Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. [1999]. “Trawling the web for emerging cyber-communities.” *Computer Networks-The International Journal Of Computer And Telecommunications Networking*, 31(11-16):pp. 1481.

[4] Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A. and Vicsek, T. [2002]. “Evolution of the social network of scientific collaborations.” *Physica A*, 311(3-4):pp. 590.

[5] Liljeros, F., Edling, C., Amaral, L., Stanley, H. and Aberg, Y. [2001]. “The web of human sexual contacts.” *Nature*, 411(6840):pp. 907.

[6] Balcan, D., Gonçalves, B., Hu, H., Ramasco, J., Colizza, V. and Vespignani, A. [2010]. “Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model.” *Journal of Computational Science*.

[7] Colizza, V., Barrat, A., Barthélemy, M. and Vespignani, A. [2006]. “The role of the airline transportation network in the prediction and predictability of global epidemics.” *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):p. 2015.

[8] Pastor-Satorras, R. and Vespignani, A. [2001]. “Epidemic spreading in scale-free networks.” *Physical Review Letters*, 86(14):pp. 3200.

[9] Bollobas, B. [1985]. *Random Graphs.* Academic Press, New York.

[10] Erdös, P. and Rényi, A. [1959]. "On random graphs." *Publicationes Mathematicae*, 6:pp. 290.

[11] Milgram, S. [1967]. "The small world problem." *Psychology Today*, 2:pp. 60.

[12] Callaway, D., Newman, M., Strogatz, S. and Watts, D. [2000]. "Network robustness and fragility: Percolation on random graphs." *Physical Review Letters*, 85(25):pp. 5468.

[13] Cohen, R., Erez, K., ben Avraham, D. and Havlin, S. [2000]. "Resilience of the internet to random breakdowns." *Physical Review Letters*, 85(21):pp. 4626.

[14] Albert, R. and Barabasi, A.L. [2002]. "Statistical mechanics of complex networks." *Reviews of Modern Physics*, 74(1):pp. 47.

[15] Atilgan, A.R., Akan, P. and Baysal, C. [2004]. "Small-world communication of residues and significance for protein dynamics." *Biophysical Journal*, 86(1):pp. 85.

[16] Huberman, B.A. and Adamic, L.A. [1999]. "Internet - growth dynamics of the world-wide web." *Nature*, 401(6749):pp. 131.

[17] Watts, D. [1999]. *Small Worlds: The Dynamics of Networks between Order and Randomness.* Princeton University, Princeton, NJ.

[18] Newman, M. and Watts, D. [1999]. "Renormalization group analysis of the small-world network model." *Physics Letters A*, 263(4-6):pp. 341.

[19] Kasturirangan, R. [1999]. "Multiple scales in small-world graphs." *cond-mat/9904055.*

[20] Newman, M. [2000]. "Models of the small world." *Journal Of Statistical Physics*, 101(3-4):pp. 819.

[21] Kleinberg, J. [2000]. "Navigation in a small world - it is easier to find short chains between points in some networks than others." *Nature*, 406(6798):pp. 845.

[22] Kleinberg, J. [2000]. "The small-world phenomenon: An algorithmic perspective." *Proc. 32ND ACM Symposium on the Theory of Computing.*

[23] Barabasi, A.L. and Albert, R. [1999]. "Emergence of scaling in random networks." *Science*, 286(5439):pp. 509.

[24] Newman, M. [2004]. "Analysis of weighted networks." *Physical Review E*, 70(5):p. 056131.

[25] Barrat, A., Barthelemy, M., Pastor-Satorras, R. and Vespignani, A. [2004]. "The architecture of complex weighted networks." *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 101(11):pp. 3747.

[26] Yook, S., Jeong, H., Barabasi, A. and Tu, Y. [2001]. "Weighted evolving networks." *Physical Review Letters*, 86(25):pp. 5835.

[27] Braunstein, L., Buldyrev, S., Cohen, R., Havlin, S. and Stanley, H. [2003]. "Optimal paths in disordered complex networks." *Physical Review Letters*, 91(16):p. 168701.

[28] Chen, Y., Lopez, E., Havlin, S. and Stanley, H. [2006]. "Universal behavior of optimal paths in weighted networks with general disorder." *Physical Review Letters*, 96(6):p. 068702.

[29] Cieplak, M., Maritan, A. and Banavar, J. [1994]. "Optimal paths and domain-walls in sthe strong disorder limit." *Physical Review Letters*, 72(15):pp. 2320.

[30] Kitano, H. [2002]. "Systems biology: A brief overview." *Science*, 295:pp. 1662.

[31] Baysal, C. and Atilgan, A.R. [2005]. "Relaxation kinetics and the glassiness of native proteins: Coupling of timescales." *Biophys. J.*, 88:pp. 1570.

[32] Baysal, C. and Atilgan, A.R. [2002]. "Relaxation kinetics and the glassiness of proteins: The case of bovine pancreatic trypsin inhibitor." *Biophys. J.*, 83:pp. 699.

[33] Baysal, C. and Atilgan, A.R. [2001]. "Coordination topology and stability for the native and binding conformers of chymotrypsin inhibitor 2." *Proteins-Structure Function and Genetics*, 45(1):pp. 62.

[34] Zaccai, G. [2000]. "How soft is a protein? a protein dynamics force constant measured by neutron scattering." *Science*, 288:pp. 1604.

[35] Raghunathan, G. and Jernigan, R. [1997]. "Ideal architecture of residue packing and its observation in protein structures." *Prot. Sci.*, 6:pp. 2072.

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

[37] Bagler, G. and Sinha, S. [2005]. "Network properties of protein structures." *Physica a-Statistical Mechanics and Its Applications*, 346(1-2):pp. 27.

[38] Greene, L.H. and Higman, V.A. [2003]. "Uncovering network systems within protein structures." *Journal of Molecular Biology*, 334(4):pp. 781.

[39] Vendruscolo, M., Dokholyan, N.V., Paci, E. and Karplus, M. [2002]. "Small-world view of the amino acids that play a key role in protein folding." *Physical Review E*, 65(6).

[40] Strogatz, S.H. [2001]. "Exploring complex networks." *Nature*, 410:pp. 268.

[41] Adamic, L. and Huberman, B. [1999]. "Growth dynamics of the world-wide web." *Nature*, 401:p. 131.

[42] Vázquez, A., Pastor-Satorras, R. and Vespignani, A. [2002]. "Large-scale topological and dynamical properties of the internet." *Physical Review E*, 65(6).

[43] Atilgan, A., Durell, S., Jernigan, R., Demirel, M., Keskin, O. and Bahar, I. [2001]. "Anisotropy of fluctuation dynamics of proteins with an elastic network model." *Biophysical Journal*, 80(1):pp. 505.

[44] Yilmaz, L.S. and Atilgan, A.R. [2000]. "Identifying the adaptive mechanism in globular proteins: Fluctuations in densely packed regions manipulate flexible parts." *J. Chem. Phys.*, 113:pp. 4454.

[45] Bahar, I., Atilgan, A.R. and Erman, B. [1997]. "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential." *Folding & Design*, 2(3):pp. 173.

[46] Bahar, I., Atilgan, A.R., Demirel, M.C. and Erman, B. [1998]. "Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability." *Physical Review Letters*, 80(12):pp. 2733.

[47] Bahar, I., Erman, B., Jernigan, R.L., Atilgan, A.R. and Covell, D.G. [1999]. "Collective motions in hiv-1 reverse transcriptase: Examination of flexibility and enzyme function." *Journal of Molecular Biology*, 285(3):pp. 1023.

[48] Demirel, M.C., Atilgan, A.R., Jernigan, R.L., Erman, B. and Bahar, I. [1998]. "Identification of kinetically hot residues in proteins." *Protein Science*, 7(12):pp. 2522.

[49] Jeong, H., Mason, S., Barabasi, A. and Oltvai, Z. [2001]. "Lethality and centrality in protein networks." *Nature*, 411(6833):pp. 41.

[50] Higman, V. and Greene, L. [2006]. "Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology." *Physica A-Statistical Mechanics And Its Applications*, 368(2):pp. 595.

[51] Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. and Pietrokovski, S. [2004]. "Network analysis of protein structures identifies functional residues." *Journal of Molecular Biology*, 344(4):pp. 1135.

[52] del Sol, A., Fujihashi, H., Amoros, D. and Nussinov, R. [2006]. "Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families." *Protein Science*, 15(9):pp. 2120.

[53] Brinda, K.V. and Vishveshwara, S. [2005]. "Oligomeric protein structure networks: insights into protein-protein interactions." *Bmc Bioinformatics*, 6.

[54] Taylor, T. and Vaisman, I. [2006]. “Graph theoretic properties of networks formed by the delaunay tessellation of protein structures.” *Physical Review E*, 73(4):p. 041925.

[55] Bahar, I. and Jernigan, R.L. [1997]. “Inter-residue potentials in globular proteins and the dominance of highly specific hydrophillic interactions at close separation.” *J. Mol. Biol.*, 266:pp. 195.

[56] Brinda, K., Vishveshwara, S. and Vishveshwara, S. [2010]. “Random network behaviour of protein structures.” *Mol. BioSyst*, 6(2):pp. 391.

[57] Popovych, N., Sun, S., Ebright, R.H. and Kolodimos, C.G. [2006]. “Dynamically driven protein allostery.” *Nature Structural and Molecular Biology*, 13:pp. 831.

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

[59] Thomas, P. and Dill, K. [1996]. “An iterative method for extracting energy-like quantities from protein structures.” *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 93(21):pp. 11628.

[60] Erdös, P. and Rényi, A. [1961]. “On the evolution of random graphs.” *Bulletin of the Institute of International Statistics.*

[61] Paperin, G., Green, D. and Leishman, T. [2008]. “Dual phase evolution and self-organisation in networks.” *Simulated Evolution and Learning*:pp. 575.

[62] Newman, M.E.J. [2003]. “The structure and function of complex networks.” *Siam Review*, 45(2):pp. 167.

[63] Atilgan, A.R., Turgut, D. and Atilgan, C. [2007]. “Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication.” *Biophysical Journal*, 92(9):pp. 3052.

[64] Brinda, K.V. and Vishveshwara, S. [2005]. “A network representation of protein structures: Implications for protein stability.” *Biophysical Journal*, 89(6):pp. 4159.

[65] Sathyapriya, R. and Vishveshwara, S. [2007]. "Structure networks of e-coli glutaminyl-trna synthetase: Effects of ligand binding." *Proteins-Structure Function and Bioinformatics*, 68(2):pp. 541.

[66] Lee, C.Y., Lee, J.C. and Gutell, R.R. [2007]. "Networks of interactions in the secondary and tertiary structure of ribosomal rna." *Physica a-Statistical Mechanics and Its Applications*, 386(1):pp. 564.

[67] Muppirala, U.K. and Li, Z.J. [2006]. "A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues." *Protein Engineering Design & Selection*, 19(6):pp. 265.

[68] Aftabuddin, M. and Kundu, S. [2007]. "Hydrophobic, hydrophilic, and charged amino acid networks within protein." *Biophysical Journal*, 93(1):pp. 225.

[69] Bode, C., Kovacs, I.A., Szalay, M.S., Palotai, R., Korcsmaros, T. and Csermely, P. [2007]. "Network analysis of protein dynamics." *Febs Letters*, 581(15):pp. 2776.

[70] del Sol, A., Fujihashi, H., Amoros, D. and Nussinov, R. [2006]. "Residues crucial for maintaining short paths in network communication mediate signaling in proteins." *Molecular Systems Biology*.

[71] Pastor-Satorras, R., Vazquez, A. and Vespignani, A. [2001]. "Dynamical and correlation properties of the internet." *Physical Review Letters*, 87(25).

[72] Newman, M.E.J. [2003]. "Mixing patterns in networks." *Physical Review E*, 67(2).

[73] Xulvi-Brunet, R. and Sokolov, I.M. [2004]. "Reshuffling scale-free networks: From random to assortative." *Physical Review E*, 70(6).

[74] Newman, M.E.J. [2002]. "Assortative mixing in networks." *Physical Review Letters*, 89(20).

[75] Lau, K.F. and Dill, K.A. [1989]. "A lattice statistical mechanics model of the conformational and sequence spaces of proteins." *Macromolecules*, 22:pp. 3986.

[76] Hart, W.E. and Newman, A. [2001]. "Protein structure prediction with lattice models." In S. Aluru, editor, "Handbook of Computational Molecular Biology," Chapman-Hall/CRC Press, Boca Raton, pp. 30.1–21.

[77] Godzik, A., Kolinski, A. and Skolnick, J. [1993]. "Lattice representations of globular proteins: How good are they?" *Journal of Computational Chemistry*, 14(10):pp. 1194.

[78] Covell, D.G. and Jernigan, R.L. [1990]. "Conformations of folded proteins in restricted spaces." *Biochemistry*, 29(13):pp. 3287.

[79] Hinds, D.A. and Levitt, M. [1992]. "A lattice model for protein structure prediction at low resolution." *Proceedings of the National Academy of Sciences of USA*, 89:pp. 2536.

[80] Bagci, Z., Jernigan, R.L. and Bahar, I. [2002]. "Residue coordination in proteins conforms to the closest packing of spheres." *Polymer*, 43(2):pp. 451.

[81] Agarwala, R., Batzogloa, S., Dancik, V., Decatur, S.E., Hannenhalli, S., Farach, M., Muthukrishnan, S. and S, S. [1997]. "Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the hp model." *Journal of Computational Biology*, 4(3):pp. 276.

[82] Manuch, J. and Ram, G.D. [2008]. "Fitting protein chains to cubic lattice is np-complete." *Journal of Bioinformatics and Computational Biology*, 6(1):pp. 93.

[83] Huang, X. [2007]. "Fitting protein lattice to integer programming approach, msc. thesis." Technical report, Simon Fraser University.

[84] Thomas, D. [2006]. "Algorithms and experiments for the protein chain lattice fitting problem, msc. thesis." Technical report, University of Alberta.

[85] Rykunov, D., Reva, B.A. and Finkelstein, A.V. [1994]. "A rapid and precise method for lattice approximation of the course of a protein chain based on a dynamic programming algorithm." *Molekuliarnaia biologiia*, 28(4):pp. 855.

[86] Rykunov, D.S., Reva, B.A. and Finkelstein, A.V. [1995]. "Accurate general method for lattice approximation of three dimensional structure of a chain molecule." *Proteins*, 22(2):pp. 100.

[87] Park, B. and Levitt, M. [1995]. "The complexity and accuracy of discrete state models of protein structure." *Journal of Molecular Biology*, 249(2):pp. 493.

[88] Koehl, P. and Delarue, M. [1998]. "Building protein lattice models using self-consistent mean field theory." *Journal of Chemical Physics*, 108(22):pp. 9540.

[89] Reva, B.A., Rykunov, D.S., Olson, A.J. and Finkelstein, A.V. [1995]. "Constructing lattice models of protein chains with side groups." *Journal of Computational Biology*, 2(4):pp. 527.

[90] De Graef, M. and McHenry, M.E. [2007]. *Structure of Materials, An Introduction to Crystallography, Diffraction and Symmetry.* Cambridge University Press.

[91] Kearsley, S.K. [1989]. "On the orthogonal transformation used for structural comparisons." *Acta Crystallographica Section A*, 45(2):pp. 208.

[92] Hinsen, K. [1998]. "Analysis of domain motions by approximate normal mode calculations." *Proteins: Structure, Function, and Bioinformatics*, 33(3):pp. 417.

[93] Tama, F. and Sanejouand, Y. [2001]. "Conformational change of proteins arising from normal mode calculations." *Protein Engineering*, 14(1):p. 1.

[94] Gerstein, M. and Krebs, W. [1998]. "A database of macromolecular motions." *Nucleic acids research*, 26(18):p. 4280.

[95] Keskin, O. [2007]. "Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: a case study of antibodies." *BMC Structural Biology*, 7(1):p. 31.

[96] Kim, M., Jernigan, R. and Chirikjian, G. [2002]. "Efficient generation of feasible pathways for protein conformational transitions." *Biophysical Journal*, 83(3):pp. 1620.

[97] Atilgan, C. and Atilgan, A. [2009]. "Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein." *PLoS Computational Biology*, 5(10):pp. 527.

[98] Chung, F. [1997]. *Spectral graph theory.* AMS.

[99] Bollóbas, B. [1998]. *Modern graph theory.* Springer.

[100] Godsil, C. and Royle, G. [2001]. *Algebraic graph theory.* Springer.

[101] Mohar, B. [1991]. "The laplacian spectrum of graphs." *Graph Theory, Combinatorics, and Applications*, 2.

[102] Atilgan, C., Okan, O.B. and Atilgan, A.R. [2010]. "How orientational order governs collectivity of folded proteins." *Proteins: Structure, Function, and Bioinformatics*, 78(16):pp. 3363.

[103] Miyazawa, S. and Jernigan, R. [1996]. "Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading." *Journal Of Molecular Biology*, 256(3):pp. 623.

[104] Fariselli, P. and Casadio, R. [1999]. "A neural network based predictor of residue contacts in proteins." *Protein Engineering*, 12(1):pp. 15.

[105] Chen, R., Mintseris, J., Janin, J. and Weng, Z. [2003]. "A protein-protein docking benchmark." *Proteins*, 52:pp. 88.

[106] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. [2000]. "The protein data bank." *Nucl. Acids Res.*, 28:pp. 235.

[107] Cao, H., Ihm, Y., Wang, C.Z., Morris, J.R., Su, M., Dobbs, D. and Ho, K.M. [2004]. "Three-dimensional threading approach to protein structure recognition." *Polymer*, 45:pp. 687.

[108] Li, H., Tang, C. and Wingreen, N.S. [2002]. "Designability of protein structures: A lattice-model study using the miyazawa-jernigan matrix." *Proteins*, 49:pp. 403.

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

[110] Betancourt, M.R. and Thirumalai, D. [1999]. “Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes.” *Prot. Sci.*, 8:pp. 361.

[111] Colman-Lerner, A., Gordon, A., Serra, E., Chin, T., Resnekov, O., Endy, D., Pesce, C.G. and Brent, R. [2005]. “Regulated cell-to-cell variation in a cell-fate decision system.” *Nature*, 437(7059):pp. 699.

[112] Eldar, A. and Elowitz, M. [2005]. “Systems biology - deviations in mating.” *Nature*, 437(7059):pp. 631.

[113] Venkataraman, L., Klare, J.E., Nuckolls, C., Hybertsen, M.S. and Steigerwald, M.L. [2006]. “Dependence of single-molecule junction conductance on molecular conformation.” *Nature*, 442:pp. 904.

[114] Ma, B., Elkayam, T., Wolfson, H. and Nussinov, R. [2003]. “Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces.” *Proc. Natl. Acad. Sci. USA*, 100:pp. 5772.

[115] Newman, M.E.J., Strogatz, S.H. and Watts, D.J. [2001]. “Random graphs with arbitrary degree distributions and their applications.” *Physical Review E*, 6402(2).

[116] Ozen, A.S., Sen, U. and Atilgan, C. [2006]. “Complete mapping of the morphologies of some linear and graft fluorinated co-oligomers in an aprotic solvent by dissipative particle dynamics.” *Journal Of Chemical Physics*, 124(6):p. 064905.

[117] Kacar, G., Atilgan, C. and Ozen, A.S. [2010]. “Mapping and reverse-mapping of the morphologies for a molecular understanding of the self-assembly of fluorinated block copolymers.” *Journal of Physical Chemistry C*, 114(1):pp. 370.

[118] Wales, D.J. and Doye, J.P.K. [1997]. “Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms.” *The Journal of Physical Chemistry A*, 101(28):pp. 5111. doi:10.1021/jp970984n. URL `http://pubs.acs.org/doi/abs/10.1021/jp970984n`.

[119] Xiang, Y.H., Jiang, H.Y., Cai, W.S. and Shao, X.G. [2004]. "An efficient method based on lattice construction and the genetic algorithm for optimization of large lennard-jones clusters." *Journal of Physical Chemistry A*, 108(16):pp. 3586.

[120] Xiang, Y.H., Cheng, L.J., Cai, W.S. and Shao, X.G. [2004]. "Structural distribution of lennard-jones clusters containing 562 to 1000 atoms." *Journal of Physical Chemistry A*, 108(44):pp. 9516.

[121] Accelrys Inc. [2008]. "Materials studio, release 4.4."

[122] Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L. and Schulten, K. [2005]. "Scalable molecular dynamics with namd." *Journal of Computational Chemistry*, 26(16):pp. 1781.

[123] Can, H., Kacar, G. and Atilgan, C. [2009]. "Surfactant formation efficiency of fluorocarbon-hydrocarbon oligomers in supercritical co2." *Journal of Chemical Physics*, 131(12).

[124] Tsolou, G., Harmandaris, V.A. and Mavrantzas, V.G. [2006]. "Temperature and pressure effects on local structure and chain packing in cis-1,4-polybutadiene from detailed molecular dynamics simulations." *Macromolecular Theory and Simulations*, 15(5):pp. 381.

[125] Teller, E., Metropolis, N. and Rosenbluth, A. [1953]. "Equation of state calculations by fast computing machines." *J. Chem. Phys*, 21(13):pp. 1087.

[126] Steinhardt, P., Nelson, D. and Ronchetti, M. [1983]. "Bond-orientational order in liquids and glasses." *Physical Review B*, 28(2):pp. 784.

[127] Connolly, M. [1983]. "Solvent-accessible surfaces of proteins and nucleic acids." *Science*, 221(4612):pp. 709.

[128] Wigner, E. [1955]. "Characteristic vectors of bordered matrices with infinite dimensions." *Annals of Mathematics*, 62(3):pp. 548.

[129] Banerjee, A. and Jost, J. [2008]. "Spectral plot properties: Towards a qualitative classification of networks." *Networks and Heterogeneous Media*, 3(2):pp. 395.

[130] Dijkstra, E. [1959]. “A note on two problems in connexion with graphs.” *Numerische mathematik*, 1(1):pp. 269.

[131] “Jmol: an open-source java viewer for chemical structures in 3d.” URL `http://www.jmol.org/`.

Appendix A

JResNets: A web-based service for calculation of strong paths in residue networks

We provide a web-based tool (JResNets) for the calculation and visualization of strong paths for residue networks. It can be reached under the **Services** link at the website: http://midst.sabanciuniv.edu

A.1 Method Overview

Input for the server is a protein structure defined by the PDB format. Additional inputs are required to form the network structure. The method starts with the formation of a network representation for the given protein structure. Every amino acid in the protein structure (or its specified chains) is represented with a node that is centered on the carbon atom of choice. The nodes are connected to each other if the distance between them is smaller than the cutoff value provided. Weights are associated with the connections according to the amino acids using well known attraction potentials obtained by Thomas-Dill [59] or Miyazawa-Jernigan [103].

The next step is the calculation of optimal path lengths and optimal path costs. An optimal path between two nodes is defined as the shortest path that minimizes the maximum weight along the path, and the cost for that path is the maximum weight. A modification of Djikstra algorithm [130] is used to calculate the strong path costs between two nodes by changing the cost term from the summation of weights to the maximum of weights. After the calculation of optimal path costs, for each pair of nodes a reduced network is constructed by removing connections with weights greater than the cost of path between these nodes. Then optimal paths can be calculated using breadth-first search algorithm on this unweighted reduced network.

A.2 Webserver

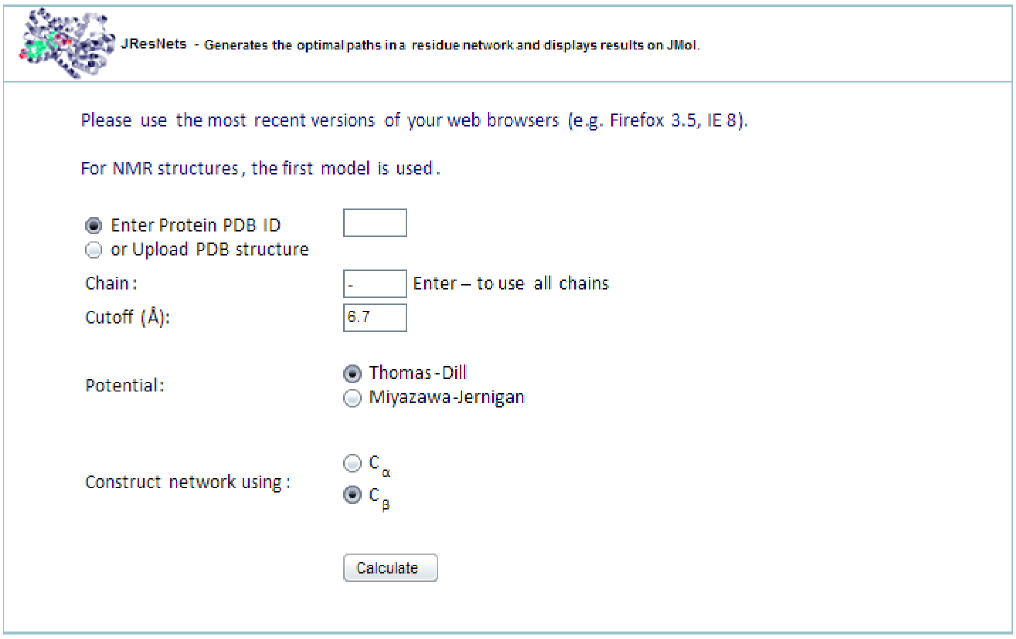
A.2.1 Input

The basic input form is simple (see figure A.1. The user is required to define the protein structure either by entering the PDB-ID or by uploading a PDB file. If PDB-ID is supplied, the server will automatically download the ascociated PDB file from the Protein Data Bank [106]. Additional parameters are required to set how the network is formed. If the user wants the results for only a specific chain, then the chain field must be filled accordingly. A cut-off distance is required to form the network with non-bonded interactions. A value of 6.7 Å is pre-entered as a default value, which corresponds to the first coordination shell [15]. A potential is selected to assign weights to connections from amongst the alternatives of TD or MJ. Finally, the user can specify which carbon atom of the amino acids will be used as a center for the nodes (default is C_β).

A.2.2 Output

A typical calculation for JresNets takes between seconds to couple of minutes depending on the size of the protein. After the calculation, server presents a link to download the zipped outputs (see figure A.2).

The main output is a text file where each optimal path between all pairs of nodes are listed in the allpathways.txt with the following format:


JResNets - Generates the optimal paths in a residue network and displays results on JMol.

Please use the most recent versions of your web browsers (e.g. Firefox 3.5, IE 8).

For NMR structures, the first model is used.

● Enter Protein PDB ID
○ or Upload PDB structure

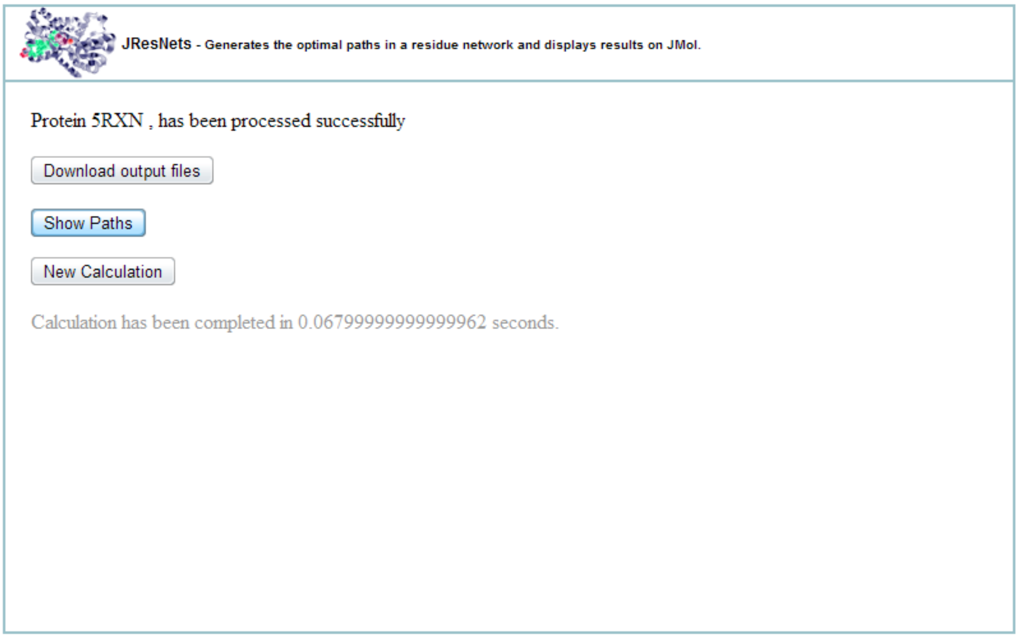
Chain: - Enter – to use all chains

Cutoff (Å): 6.7

Potential: ● Thomas-Dill ○ Miyazawa-Jernigan

Construct network using: ○ C_α ● C_β

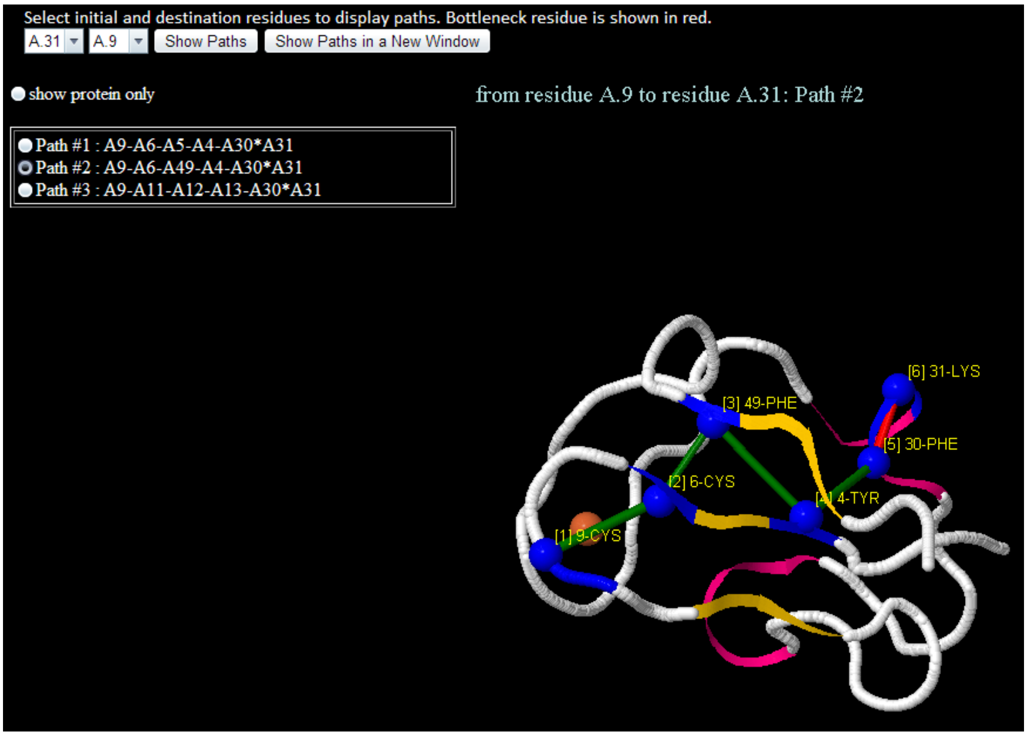
Figure A.1. Input screen for webserver.


JResNets - Generates the optimal paths in a residue network and displays results on JMol.

Protein 5RXN , has been processed successfully

Calculation has been completed in 0.0679999999999962 seconds.

Figure A.2. Output screen for webserver.



Select initial and destination residues to display paths. Bottleneck residue is shown in red.

A.31 ▾ A.9 ▾ Show Paths Show Paths in a New Window

○ show protein only

from residue A.9 to residue A.31: Path #2

○ Path #1 : A9-A6-A5-A4-A30*A31
● Path #2 : A9-A6-A49-A4-A30*A31
○ Path #3 : A9-A11-A12-A13-A30*A31

Figure A.3. Visualization screen for strong paths with JMol applet.

```
A.1 -> A.19
path #1 :  A.1(MET)   A.15(PRO) * A.16(GLU) * A.26(PRO)   A.25(ASN)   A.18(GLY)   A.19(ASP)
path #2 :  A.1(MET)   A.15(PRO) * A.16(GLU) * A.26(PRO)   A.25(ASN)   A.23(GLY)   A.19(ASP)
path #3 :  A.1(MET)   A.15(PRO) * A.16(GLU) * A.26(PRO)   A.25(ASN)   A.24(VAL)   A.19(ASP)
```

A.1 − > A.19 in the output denotes the path start and end aminoacids, where all possible optimal paths are listed in the following line as the sequences of amino acids. The stars in the sequence denote the cost defining connections for this pathway, e.g. in the example provided the PRO^{15} - GLU^{16} and GLU^{16} - PRO^{26} links have the highest weight. Two additional output matrices of size $N \times N$ are also supplied where they show the optimal (lengths_strong.txt) and shortest path lengths (lengths_homo.txt) for each pair of nodes.

At this point, the user also has the option to viualize paths in a page via the web based viewer Jmol [131]. The visualizaton page has two inputs: Starting and destination nodes. After selecting these nodes all possible optimal paths between them are listed, where the user can choose and toggle to view the different alternatives. Selected path superposed on to the protein will be shown in a Jmol applet (see figure A.3).