BINDING PROTEINS IN THE SMALL WORLD OF RESIDUE NETWORKS

by
Güngör Özer
B.S. Chem. Koç University, 2002

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

Sabancı University
July 2004

annem'e
ve
babam'a...

ABSTRACT

We analyze the network properties of a set of 59 pairs of proteins in their bound and unbound forms. We verify that these "residue networks" are in the "small world" class in their separate forms as well as in complex. We also investigate the different network properties of interface residues compared to those of other surface residues of the complexes. The results point that the average shortest paths of interface residues are in general lower than other surface residues even in their unbound forms. Moreover, the residues that are used in the shortest pathways between the receptor and the ligand are analyzed with the theory of betweenness centrality. When specific weights are assigned to the links in the network, the same pairs of residue types emerge as important hubs when complexation occurs. The calculations are further implemented to decoy structures of 15 of the bound complexes, with 10 decoys for each complex. We find the characteristic path length as the most important descriptor for differentiating between decoys and native structures.

ÖZET

Bu çalışmada 59 protein kompleksinin hem ayrık hem de bağlı durumdaki ağ özelliklerini inceledik. Ağ şeklinde ifade edilen proteinlerin “küçük dünya” özelliklerinin kompleks oluşumu sonrasında da devam ettiğini gözlemledik. Ayrıca, bağlanmada etkin rol oynayan amino asitlerin diğer yüzey amino asitlerine göre –bağsız durumda bile– daha yüksek ortalama en kısa yol değerlerine sahip olduğunu gördük. Reseptör ve ligand amino asitleri arasındaki en kısa yollarda kullanılan ve reseptör’den ligand’a geçiş noktalarını oluşturan amino asit çiftlerinin kullanılan farklı kontak enerjilerinden bağımsız olarak, çoğunlukla, benzer olduğunu gösterdik. Daha sonra, aynı hesaplamaları gerçek ve bozulmuş kompleks yapıların ağ özellikleri üzerinde gerçekleştirdik. Ortalama kısa yol değerleri yine en belirleyici özellik olarak ortaya çıktı. Ayrıca Miyazawa ve Jeringan’ın kontak enerjilerinin bu gerçek ve bozulmuş yapılar içindeki reseptor-ligand geçiş amino asitlerini daha iyi ayrımsadığını gözlemledik.

ACKNOWLEDGEMENTS

This thesis needs to be claimed as “our” thesis with all the following people.

I would like to put my supervisor, Canan Baysal, to the first position to own our thesis. Without her insightful advice and guidance I could not go one step forward in this research. I have felt her great encouragement by me in all the problems on the way.

I am grateful to Uğur Sezerman for his leading suggestions. He is the one who introduced me the great world of computational biology. I hope sometime somewhere we will again work together.

Special thanks to Cumhur Aksoy for his collaboration and friendship.

I also would like to recognize Mehmet Ali Gülgün, Yusuf Menceloğlu and Alpay Taralp for their effort to keep our team together and efficient.

I would like to specially thank to Hakan Erdoğan for accepting to read our thesis and evaluate my work.

I also want to thank my friends Serkan Baş, Burak Birkan, Özgür Bozat, Karahan Bulut, Sarp Kaya, Yener Kuru, Ünal Şen, Mesut Ünal, Fahri Başar Ulusoy, Hasan Vatansever and all my friends for sharing my two years at Sabancı University. Without them life would be so boring that I might give up working on my thesis.

Finally, I would like to express my greatest gratitude to all my family to whom I dedicate this work. Their patience to my absence while working on my thesis was admirable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Å	Angstrom
C_α	Central carbon atom
C_β	Side chain carbon atom
∞	Infinity
Ø	Empty set

LIST OF ABBREVIATIONS

$\mathbf{A}$	Adjacency matrix
C. elegans	Caenorhabditis elegans
C	Clustering coefficient
K	Degree
L	Characteristic path length
L_{weak}	Optimal path in weak disorder case
L_{strong}	Optimal path in strong disorder case
MJ	Miyazawa and Jernigan
PC	Percent centrality
RMSD	Root mean square distance
S	Strength
s.d.	Strong disorder
SCN	Scientist collaboration network
TD	Thomas and Dill
WAN	World airport network
w.d.	Weak disorder

BINDING PROTEINS IN THE SMALL WORLD OF RESIDUE NETWORKS

by
Güngör Özer
B.S. Chem. Koç University, 2002

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

Sabancı University
July 2004

annem'e
ve
babam'a...

ABSTRACT

We analyze the network properties of a set of 59 pairs of proteins in their bound and unbound forms. We verify that these "residue networks" are in the "small world" class in their separate forms as well as in complex. We also investigate the different network properties of interface residues compared to those of other surface residues of the complexes. The results point that the average shortest paths of interface residues are in general lower than other surface residues even in their unbound forms. Moreover, the residues that are used in the shortest pathways between the receptor and the ligand are analyzed with the theory of betweenness centrality. When specific weights are assigned to the links in the network, the same pairs of residue types emerge as important hubs when complexation occurs. The calculations are further implemented to decoy structures of 15 of the bound complexes, with 10 decoys for each complex. We find the characteristic path length as the most important descriptor for differentiating between decoys and native structures.

ÖZET

Bu çalışmada 59 protein kompleksinin hem ayrık hem de bağlı durumdaki ağ özelliklerini inceledik. Ağ şeklinde ifade edilen proteinlerin “küçük dünya” özelliklerinin kompleks oluşumu sonrasında da devam ettiğini gözlemledik. Ayrıca, bağlanmada etkin rol oynayan amino asitlerin diğer yüzey amino asitlerine göre –bağsız durumda bile– daha yüksek ortalama en kısa yol değerlerine sahip olduğunu gördük. Reseptör ve ligand amino asitleri arasındaki en kısa yollarda kullanılan ve reseptör’den ligand’a geçiş noktalarını oluşturan amino asit çiftlerinin kullanılan farklı kontak enerjilerinden bağımsız olarak, çoğunlukla, benzer olduğunu gösterdik. Daha sonra, aynı hesaplamaları gerçek ve bozulmuş kompleks yapıların ağ özellikleri üzerinde gerçekleştirdik. Ortalama kısa yol değerleri yine en belirleyici özellik olarak ortaya çıktı. Ayrıca Miyazawa ve Jeringan’ın kontak enerjilerinin bu gerçek ve bozulmuş yapılar içindeki reseptor-ligand geçiş amino asitlerini daha iyi ayrımsadığını gözlemledik.

ACKNOWLEDGEMENTS

This thesis needs to be claimed as “our” thesis with all the following people.

I would like to put my supervisor, Canan Baysal, to the first position to own our thesis. Without her insightful advice and guidance I could not go one step forward in this research. I have felt her great encouragement by me in all the problems on the way.

I am grateful to Uğur Sezerman for his leading suggestions. He is the one who introduced me the great world of computational biology. I hope sometime somewhere we will again work together.

Special thanks to Cumhur Aksoy for his collaboration and friendship.

I also would like to recognize Mehmet Ali Gülgün, Yusuf Menceloğlu and Alpay Taralp for their effort to keep our team together and efficient.

I would like to specially thank to Hakan Erdoğan for accepting to read our thesis and evaluate my work.

I also want to thank my friends Serkan Baş, Burak Birkan, Özgür Bozat, Karahan Bulut, Sarp Kaya, Yener Kuru, Ünal Şen, Mesut Ünal, Fahri Başar Ulusoy, Hasan Vatansever and all my friends for sharing my two years at Sabancı University. Without them life would be so boring that I might give up working on my thesis.

Finally, I would like to express my greatest gratitude to all my family to whom I dedicate this work. Their patience to my absence while working on my thesis was admirable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Å	Angstrom
C_α	Central carbon atom
C_β	Side chain carbon atom
∞	Infinity
Ø	Empty set

LIST OF ABBREVIATIONS

$\mathbf{A}$	Adjacency matrix
C. elegans	Caenorhabditis elegans
C	Clustering coefficient
K	Degree
L	Characteristic path length
L_{weak}	Optimal path in weak disorder case
L_{strong}	Optimal path in strong disorder case
MJ	Miyazawa and Jernigan
PC	Percent centrality
RMSD	Root mean square distance
S	Strength
s.d.	Strong disorder
SCN	Scientist collaboration network
TD	Thomas and Dill
WAN	World airport network
w.d.	Weak disorder

1. INTRODUCTION

Understanding and predicting the protein function computationally is one of the major scientific obstacles. Protein-protein interactions play a key role in the functioning of proteins. The problem of *Protein Docking* (i.e. finding the geometry in which two or more proteins interact under physiological conditions) arises here. Many scientists from different backgrounds try to understand the logic behind this problem using different methodologies. Although no satisfactory answer has yet been reached, every study contributes to the solution in various ways and more importantly, all research done so far, leads the researchers to new ideas, new techniques and new strategies.

Weng and coworkers have been developing a docking algorithm that, for the present, takes into account pairwise shape complementarity, desolvation and electrostatics simultaneously which gave good results [Chen *et al.*, 2003] in 2003 CAPRI[1] challenge. In our research, our purpose is to contribute to the understanding of the docking problem with a statistical analysis of interacting proteins. We treat these interacting proteins, on a coarse level, as residue networks. We use the carefully selected set of Chen *et al.*, consisting of 59 protein complexes. We also utilize decoy structures for 15 of these 59 complexes.

Small World Network (SWN) properties have recently been shown to arise in miscellaneous systems such as airport network [Guimerà *et al.*, 2003] social networks [Wattz and Strogatz, 1998], neuronal networks [Wattz and Strogatz, 1998], food webs [McCann, 1998], disease spreading [Woolhouse and Donaldson, 2001], scientific collaborations [Newman, 2001], and the World Wide Web [Barabasi, Nature, 1999; Barabasi, Science, 1999]. Also residue networks in proteins show SWN behavior [Atilgan *et al.*, 2004] and the analysis of complex structures of proteins is of great significance since it is known that in all types of networks there is a strong relation between structure and function.

Proteins perform their functions basically by binding to smaller molecules (i.e. ions, fats, sugars etc.), nucleic acids, or other proteins. Conformational changes in the unbound form of the proteins (frequently in the side chains, sometimes in the whole

[1] Critical Assesment of Prediction of Interactions

backbone) are commonly encountered and that is a main obstacle for the solution of the docking problem. It is also observed for many systems that some residues are more important than others while proteins are carrying out their functions. Here, some critical questions arise: Is it also possible to observe that kind of key roles in residue networks? How do the structural changes during binding appear while analyzing residue networks? More importantly, may the representation of protein complexes as networks give a concrete clue about the structural changes during binding, or structure-function relationships?

2. PROTEIN COMPLEXES AS NETWORK STRUCTURES

In this research, proteins are treated as networks formed using the spatial coordinates of atoms. We assume each residue as a different node. Two nodes are connected by an edge if the distance between the pairs of residues is less than a particular cut-off radius.

2.1 Overview

A network is basically any set of interconnected nodes. Here, by replacing the term “node” with a particular object (telephone, computer, people, railroad, neuron etc.) it will be possible to express any set of interacting entities as a network. In other words, any complex system can be modeled as a network by using a clear definition for the nodes and the interactions between the nodes.

Any network can be represented graphically as a diagram –called graph– with vertices (sng. Vertex, any node in that graph) and edges (links connecting these vertices) [Wilson and Watkins, 1990].

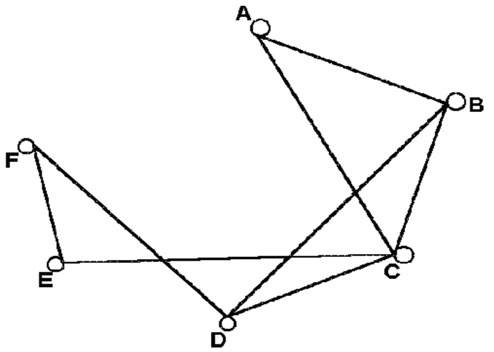


Figure 2. 1 A simple graph with 6 vertices and 8 edges[2].

For a better understanding, take the members of a population as the vertices, then the edges can be any kind of relationship between each member like friendships,

[2] http://www.geom.uiuc.edu/~zarembe/graph3.gif

business ties, etc. The brain is a network of neurons, organizations are networks of people, and global economy is a network of national economies which are networks of markets which in turn are networks of producers and consumers [SFI bulletin, vol.14 no.2]. Such networks that exist in real life have unique properties like error tolerance, speed and flexibility. These characteristics play central roles in, for example, spreading diseases through social networks or viruses through computer networks, or propagation of power failures through energy grids. Therefore, understanding the roles of these properties sufficiently will lead to better and more efficient design of networks, which would lead to better productivity in various fields [Watts, 1999].

2.2 Small World Networks

At the extremes, a network may be considered to be either completely regular or completely random; however, many biological, technological and social networks are somewhere between these complete regularity and randomness. With the random rewiring procedure shown in figure 2.2 [Wattz and Strogatz, 1998] it will be easy to interpolate the regular and random networks. Starting from a ring lattice with *n* vertices and *k* edges for each vertex, each edge is rewired at random with a probability *p*. Consequently, the graph is tuned between ultimate regularity *(p* = 0*)* and ultimate randomness *(p* = 1*)*.

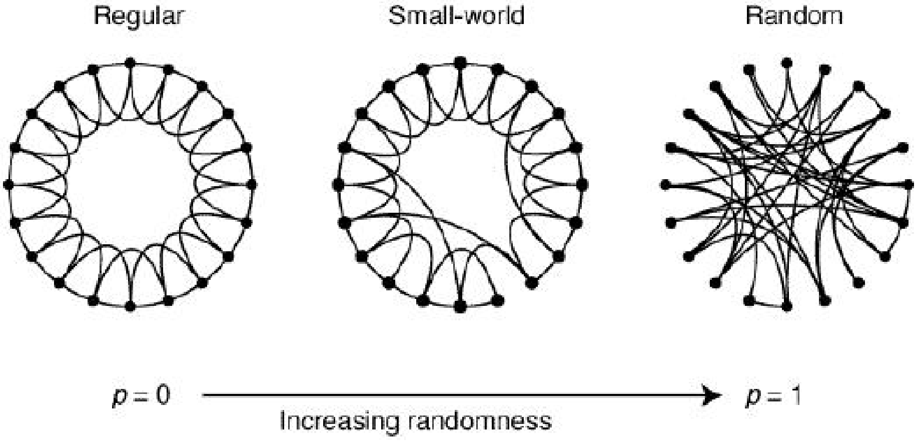


Figure 2. 2 The transition through regularity and randomness in a simple topology.

Figure 2.2 [Watts and Strogatz, 1998] is a demonstration of the random rewiring procedure for the interpolation between a regular network and a random network without altering the number of vertices (*n* = 20, there are 20 nodes) and edges (*k* = 4, each node has 4 connections to other nodes). In this process a vertex is chosen and the edge that connects the chosen vertex to its closest neighbor is reconnected to another vertex chosen uniformly at random over the entire ring. Duplication of edges is forbidden, otherwise the edges conserve their original positions. This process is repeated by moving clockwise around the lattice ring, considering each vertex in turn until a lap is completed. Then, the same process is repeated for more distant edges until all the edges are considered. So, for $p = 0$ the original lattice is preserved. As p increases, the graph become more disordered. Finally, at $p = 1$, the graph has the ultimate randomness with all the edges rewired randomly according to a Gaussian distribution of the neighbors. The clear observation with this figure is that, for the intermediate values of p the graph is highly clustered like a regular graph; on the other hand, it has small values of characteristic shortest paths –also defined as Network Diameter[3] [Barthèlèmy, 1999]– like a random graph. Table 2.1 below shows this difference in the characteristics of regular and random networks for three different systems (film actors, power grid and C. elegans);

System	L_{actual}	L_{random}	C_{actual}	C_{random}
Film Actors	3.65	2.99	0.79	0.00027
Power Grid	18.7	12.4	0.08	0.005
C. elegans	2.65	2.25	0.28	0.05

Table 2. 1 Empirical examples of SWN's; $L \approx L_{random}$, $C >> C_{random}$. [Wattz and Strogatz, 1998].

According to this table, all three systems show SWN properties. Since these systems are not hand-picked networks, it is quite logical to claim that small-world phenomenon is probably generic for many large, sparse networks found in nature [Wattz and Strogatz, 1998].

Thus, in order to deeply understand the general characteristics of SWNs, there are two significant network properties; quantifying the amount of clustering, C and quantifying the value of characteristic shortest path, L. Moreover, in real life, unlike the ideal lattice ring in figure 2.1, the number of edges from vertex to vertex may differ. Therefore, another important property quantifying the number of neighbors, k, and its distribution, $p(k)$, should also be considered.

[3] http://www.ssec.wisc.edu/~billh/gbrain0.html

2.2.1 Degree (*k*)

The number of neighbors of each node, for any complex network, has valuable information on the structure of the corresponding network. In order to quantify this characteristic, let *p(k)* be the fraction of nodes with *k* neighbors [Strogatz, 2001]; k_i, itself; on the other hand, is the number of neighbors that the i^{th} node has.

Atilgan *et al.* demonstrate the following contact distribution for residue networks.

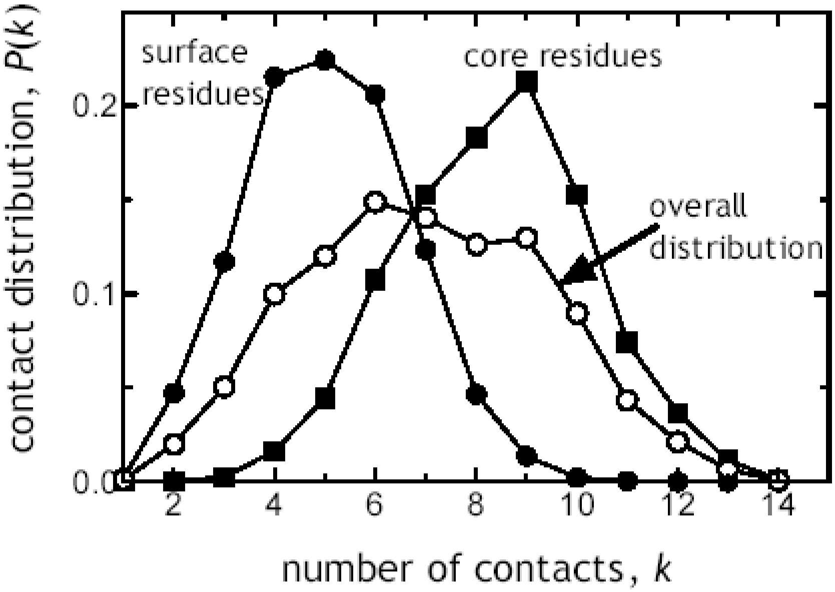


Figure 2. 3 Degree distribution of residue networks [Atilgan *et al.*, 2004].

These residue networks were constructed from proteins using a cutoff radius of 7 Å, a Gaussian distribution with a mean of 6.9 neighbors is constructed representing the connectivity distribution of residue networks. The distribution in the graph above is in a good agreement with that proposed previously for all 20 different types of amino acids [Miyazawa and Jernigan, 1996].

2.2.2 Characteristic Path Length *(L)*

One of the most important statistics of graphs is the characteristic path length (*L*) that is the average distance between every vertex and every other vertex. Distance here does not refer to any metric space between vertices. Yet, the shortest path between any two nodes is simply the minimum number of edges that must be traversed in order to reach a vertex from another vertex, and *L* is the average of these over all nodes in the system [Watts, 1999].

L itself is not indicative of the topology of a particular network. Instead, *L scaling* (the scaling of *L* with n –the size of the network– or *k* –the average number of neighbors–) display some characteristic for the system.

The different behavior of *L* in a regular network and in a SWN is the change of *L* with size; for a detailed explanation, in a regular network *L* increases linearly with the size of the network ($L \sim n$) and in a SWN *L* increases with the logarithm of the size of the network ($L \sim ln(n)$) [Barthèlèmy, 1999].

More interestingly, a very good correlation between path lengths and dynamics of the proteins were discovered [Atilgan *et al.*, 2004]. Atilgan *et al.* compared the residue fluctuations computed by the Gaussian Network Model, which many researches [Bahar, 1997; Bahar, 1999; Baysal and Atilgan, 2001; Ming, 2003] proved to be in an excellent agreement with experimentally extracted β-factors, with the average path lengths of individual residues. They found *L* to be in good agreement with residue fluctuations and therefore experimental results.

2.2.3 Clustering Coefficient *(C)*

In a network, the density of neighboring clusters is an important factor characterizing the topology of that network. Clustering coefficient (C_i) of a particular node is the probability that the neighbors of a node are neighbors of each other. *C* of a network, on the other hand, is the average of the clustering coefficient of every vertex of that network.

The clustering coefficient of a vertex with *k* neighbors is defined as follows. The maximum number of edges interconnecting these *k* neighbors is *k*(*k*-1)/2; however, the actual number is usually less than that maximum since in a spatial network of a particular cut-off radius, it is very unlikely to have all the neighbors interconnected. The ratio of this actual number to the maximum possible number of edges gives the clustering coefficient of that particular vertex.

Atilgan *et al.* (2004) proposed a relationship of depth (i.e. the shortest distance from a residue to the surface of a protein) with *C* and *L* values seen in the graph below;

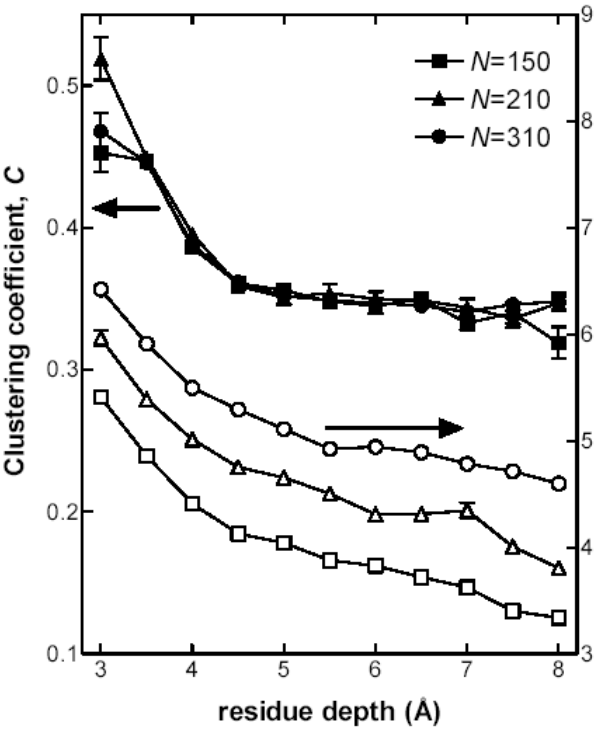


Figure 2. 4 vs. *C* and depth vs. *L* for three different system sizes [Atilgan *et al.*, 2004]. The reasoning concerning the conclusion inferred from the graph above is discussed in section 3.1.

2.2.4 Weighting Effect

Most studies concerning complex networks assume all the edges of the network to be identical. In practice, however, the weights (e.g. the quality or cost) of these links are not equal resulting in heterogeneity within a particular network. Ignoring the weights might lead to an incorrect evaluation of the parameters mentioned before, or even the whole network itself.

Recently, some researchers include this effect in their studies, and give new definitions for the network parameters discussed in the previous sections. In this thesis, two different weighting sets are used. These are the interresidue contact potentials proposed by Miyazawa and Jernigan in 1996 (MJ Potential) and by Thomas and Dill again in 1996 (TD Potential) [Miyazawa and Jernigan, 1996; Thomas and Dill, 1996].

2.2.4.1 Degree of Weighted Networks

The term obtained by extending the definition of vertex degree in terms of assigning weights is stated as *strength*, s_i. This new quantity measures the total weight of the connections of a particular residue. As an example, consider the SCN, scientist collaboration network [Newman, 2001a; Newman 2001b; Barabasi *et al.*, 2002], the strength defines the scientific productivity since it is equal to the number of publications of any given scientist.

Strength of a particular node is an important measure of the significance of that particular node in communication through the network. To quantitatively characterize the role of network elements in information flow, a new term defined as the betweenness centrality [Goh *et al.*, 2001; Barrat *et al.*, 2004] has been used.

2.2.4.2 Betweenness Centrality

Betweenness centrality simply accounts for the number of shortest paths, between all pairs in the network, passing through a given vertex. Centrality is often used in transportation networks (e.g. *WAN* – world airport network) to estimate the traffic handled by the vertices (e.g. airports).

Depending on the heterogeneity of the system, the quantity of betweenness centrality, and thus the quality, of a node may vary. The reason for that is the difference between definitions of shortest paths in weighted and non-weighted links. The level of disorder in a particular network would lead the information flow between any two nodes to have different paths than the paths in the weightless system. Thus, the residues existing in the weighted shortest paths would differ.

2.2.4.3 Characteristic Path Length of Weighted Networks

Shortest path between any two nodes was previously defined as the minimum number of edges that must be traversed in order to reach a vertex from another vertex. Note that, considering shortest path as the least number of edges between two vertices is meaningful only when all the edges are assumed to have the same weights.

When such a set of weights is applied, the network becomes disordered and there arises a need for a new definition of L. We define two new Ls for the new disordered network: *Weak Disorder* in which all links on the path contributes to the optimal path, and *Strong Disorder* in which the power of a strong link dominates the optimum path [Braunstein, 2003]; L_{weak} and L_{strong}, respectively.

2.2.4.4 Clustering Coefficient of Weighted Networks

Clustering around a particular vertex could also become more designative with the contribution of weights. Considering different links to have different values will lead a better understanding of local cohesiveness. For, not only the number of closed triplets in the neighborhood around a vertex but also the total of their relative weights is also included in the quantification.

The global cohesiveness of a weighted network is closely related to the general behavior of local clustering. If the weights of the closed triplets are more likely to have larger weights than the others within the network, then the average weighted clustering of the network will also be larger than that of a network with identical weights assigned to each edge. Similarly, with smaller weights of the edges forming the close triplets, the average weighted clustering of the network will have smaller value. Figure 2.5 demonstrates this on a small network of five nodes and seven edges;

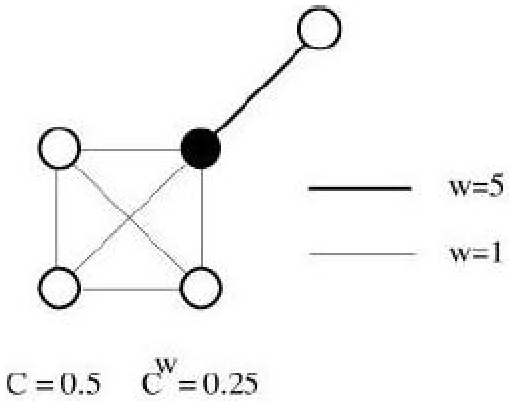


Figure 2. 5 Difference between the weighted and non-weighted clustering coefficient.

2.3 Proteins as Networks of Their Interacting Residues

In order to treat the proteins in our data set as networks, we have used the method developed by Baysal and coworkers [Atilgan, 2004]. The generation of the networks and calculation of the parameters of interest (*C*, *L*, and *K*) are implemented over all the data set.

2.3.1 Protein Network Generation

Each protein is converted into a network by taking every residue in a structure as a vertex and the interaction among them as edges [Yilmaz and Atilgan, 2000]. The position of each residue is assigned the coordinate of C_β atoms of that particular residue (C_α for Glycine). Two residues are considered as connected if they are within distance of a particular cut-off radius from each other (6.7 Å in our calculations) and they are said to be interacting/in contact. An example of a generated residue network is demonstrated in Figure 2.6 with 7 Å cut-off radius.

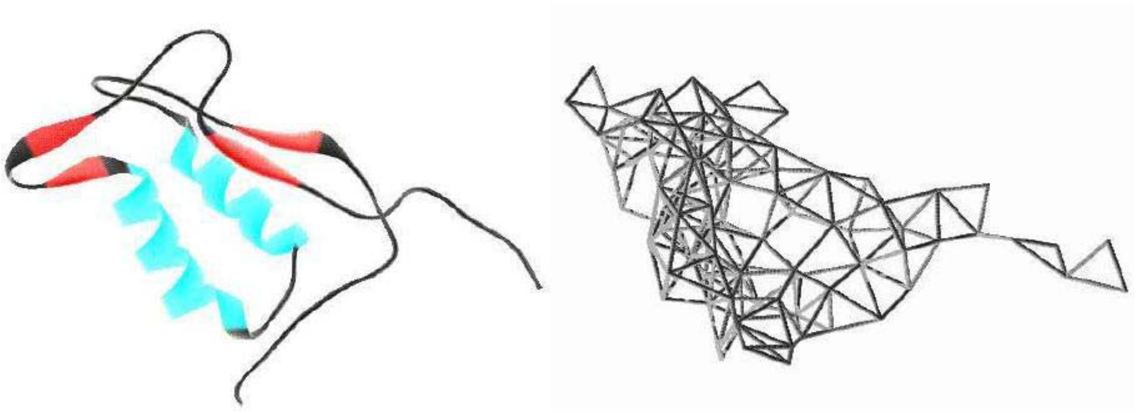


Figure 2. 6 Folded structure and network representation of 1-β converting enzyme.

For each residue, their contacts are found by calculating the distance from the corresponding residue to all the other residues. Therefore, an *NxN* matrix –where N is the number of residues in the protein is formed by assigning the matrix elements a value of 1 if the residues of interest are in contact and with 0 if not. This symmetric matrix, the so called *Adjacency Matrix*, can be mathematically expressed as [Atilgan *et al.*, 2004];

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

where r_{ij} is the distance between the i^{th} and j^{th} nodes, *H(x)* is the Heavyside step function given by $H(x) = 1$ for $x > 0$ and $H(x) = 0$ for $x \leq 0$, and r_c is the given cut-off radius.

All the properties *C*, *L* and *K* can easily be calculated using this matrix. These computations differ slightly when the weighting term w_{ij} for each pair of residues is considered.

2.3.2 Calculation of Degree

The degree of a particular residue k_i is simply obtained by counting the number of neighbors of that residue. The mathematical expression is shown below;

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{2}$$

Thus, the average connectivity of the network is just the arithmetic average over k_i;

$$K = \frac{\sum_{i=1}^{N} k_i}{N} \tag{3}$$

The strength, on the other hand, of a particular residue is calculated by simply adding up the weights of every edge belonging to that residue. As follows [Barrat *et al.*, 2004],

$$s_i = \sum_{j=1}^{N} A_{ij} * w_{ij} \tag{4}$$

Therefore, the average strength of the network is the arithmetic average over s_i;

$$S = \frac{\sum_{i=1}^{N} s_i}{N} \tag{5}$$

2.3.3 Calculation of Betweenness Centrality

The definition of betweenness centrality includes a simple term – the percent centrality (*PC*) of a particular node, which, here, is calculated by;

$$PC_i = \frac{100}{N_S * N_D} \sum_s \sum_d b_{isd} \tag{6}$$

where *S* and *D* are two different sets of nodes; *source* and *destination*, respectively and N_S and N_D are the number of nodes in both sets. On the other hand, b_{isd} is a function of which value is either 1 when the i^{th} node is in the optimal path between the s^{th} and d^{th} nodes, 0 otherwise.

When the source, *S*, is taken as the receptor residues and the destination, *D*, as the ligand residues, the last residue in the receptor and the first residue in the ligand define the contact pair used in the information flow within the complex protein [Barrat *et al.*, 2004].

2.3.4 Calculation of Characteristic Path Length

It is not possible to calculate *L* directly from the adjacency matrix; yet, the powers of this matrix lead to this information. If the shortest path between the i^{th} and j^{th} residues is *d*, then the ij^{th} entry of the d^{th} power of the adjacency matrix should become non-zero whereas it is zero for all smaller powers. Thus, to get all the shortest paths between all pairs of residues, the multiplication of the matrix with itself should run until all the elements of the resulting matrix are non-zero. The mathematical expression for the shortest path of a particular pair of residues, *i* and *j* is;

i. if no weights are assigned for the links between pairs of residues;

$$L_{ij} = \mathrm{n}$$

wehere *n* is that power of A_{ij} which is non-zero for the first time.

ii. if different weights are assigned for different pairs of residues;

   a. in weak disordered networks, the optimal path between any two vertices *i* and *j* is considered as the path with minimum weighted sum of all edges on the way, found using Dijkstra Algorithm (Appendix A);

   b. in strong disordered network, on the other hand, the optimal path is the path in which the maximum weight on the way is the minimum among all other paths.

L_{ij} will, therefore, be equal to the number of edges connecting the optimal paths found for these two cases.

The average shortest path, *L*, when no weights are assigned for the edges can be computed from [Atilgan *et al.*, 2004];

$$L = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{J=i+1}^{N} L_{ij} \tag{7}$$

2.3.5 Calculation of Clustering Coefficient

C for the i^{th} residue of a particular structure with *k* neighbors is the ratio of the number of actual connections between the neighbors of the i^{th} residue to the all possible ones [Atilgan *et al.*, 2004];

$$C_i = \frac{\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} A_{ij} A_{ik} A_{kj}}{C(k_i,2)} \qquad (8)$$

where *C(k_i,2)* is the combination relationship and counts the maximum number of possible connections between the k_i first neighbors.

When the elements of a particular network are not identical, the weights of interconnected edges should be considered [Barrat *et al.*, 2004].

$$C_i^w = \frac{1}{s_i(k_i - 1)}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{(w_{ij} + w_{jk})}{2} A_{ij} A_{jk} A_{ik} \qquad (9)$$

The average clustering coefficient of a particular network is, therefore;

$$C = \frac{\sum_{i=1}^{N} C_i}{N} \qquad (10)$$

2.3.6 The Data Sets

Recent developments in proteomics and structural genomics are resulting in a continuously increasing number of single and complex protein structures deposited in the Protein Data Bank.[Bernstein *et al.*, 1977; Abola *et al.*, 1987] Yet the number of the structures that form a complex is still limited with a few hundred deposited coordinate sets, most of which are from a highly limited variety of proteins.

The complete list of the data set is in Appendix C.

2.3.6.1 Native Complex and Unbound Structures

We have used a set[4] of 59 complexes: 22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, 11 other complexes and 7 difficult complexes (having significant conformational change for more than half of the interface backbone residues, see below for the definition of *interface*). Among them, there are 31 unbound-unbound and 28 unbound-bound cases. Among the unbound-unbound test cases, 16 are enzyme-inhibitor, 5 antibody-antigen, 5 others and 5 difficult.

There are two different definitions of *interface residues* analyzed separately in this research. (*i*) Any residue having at least an atom closer than 10 Å to any atom in the corresponding protein (if the residue of interest is a ligand residue then the corresponding protein is the receptor of the complex, or vice versa) then that residue is said to be an interface residue, (*ii*) The limiting value in this second definition is 6.7 Å instead of 10 Å, and the comparisons are made using only C_β (C_α for Glycine) coordinates instead of comparing all atomic distances.

On the other hand, *surface residues* are defined as the residues having at least an atom within 4 Å depth[5] of the protein of interest [Chakravarty&Varadarajan, 1999]. The values of atomic depths are calculated using Monte Carlo procedure outlined in their paper. The depth of a residue is assumed as the depth of the atom with the lowest value of distance to surface among all atoms of that particular residue.

2.3.6.2 Decoy Complex Structures

The other data set we have used is formed of 15 decoy complexes all of which we also have the original complex structures in our original data set. In this set, there are 10 decoy structures for each case. The RMSDs of these decoy structures are ranging from the lowest range of 9-10 Å to the highest of 40-41 Å. These decoys are selected from a

[4] http://zlab.bu.edu/zdock/

[5] Residue depths are calculated by the DEPTH program of Chakravarty1 and Varadarajan (1999).

much crowded list –of possible docked complexes generated– choosing the ones with highest surface complemantarity scores by Weng and coworkers [Chen *et al.*, 2003].

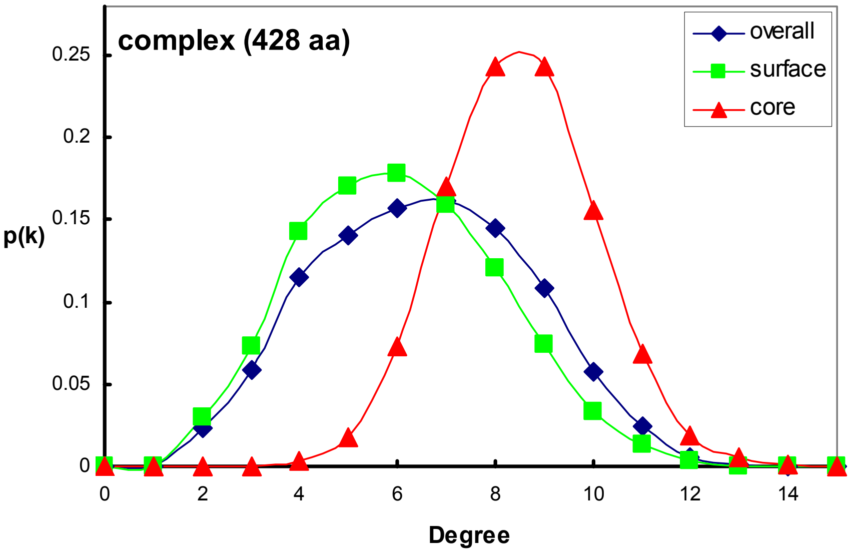
The data set of interface and surface residues used in the computational implementation for these decoys are the same as mentioned in the pervious subsection.

3. RESULTS AND DISCUSSION

We have used to different definitions for interface residues as mentioned in section 2.3.6.1: (*i*) residues having at least an atom closer than 10 Å to any atom in the other protein of the complex, (*ii*) residues having their C_β (C_α for Glycine) atom closer than 6.7 to any C_β (C_α for Glycine) in the other protein of the complex. We have found the results, with the latter definition, more considerable. Therefore, only these results are introduced in this thesis.

3.1 Basic Network Parameters

We first repeated the study of Atilgan *et al.* using a cut-off distance of 6.7 Å instead of 7 Å used in that work. We choose 6.7 Å since it is the limit of the first coordination shell. [Atilgan *et al.*, 2003; Akan, 2002]



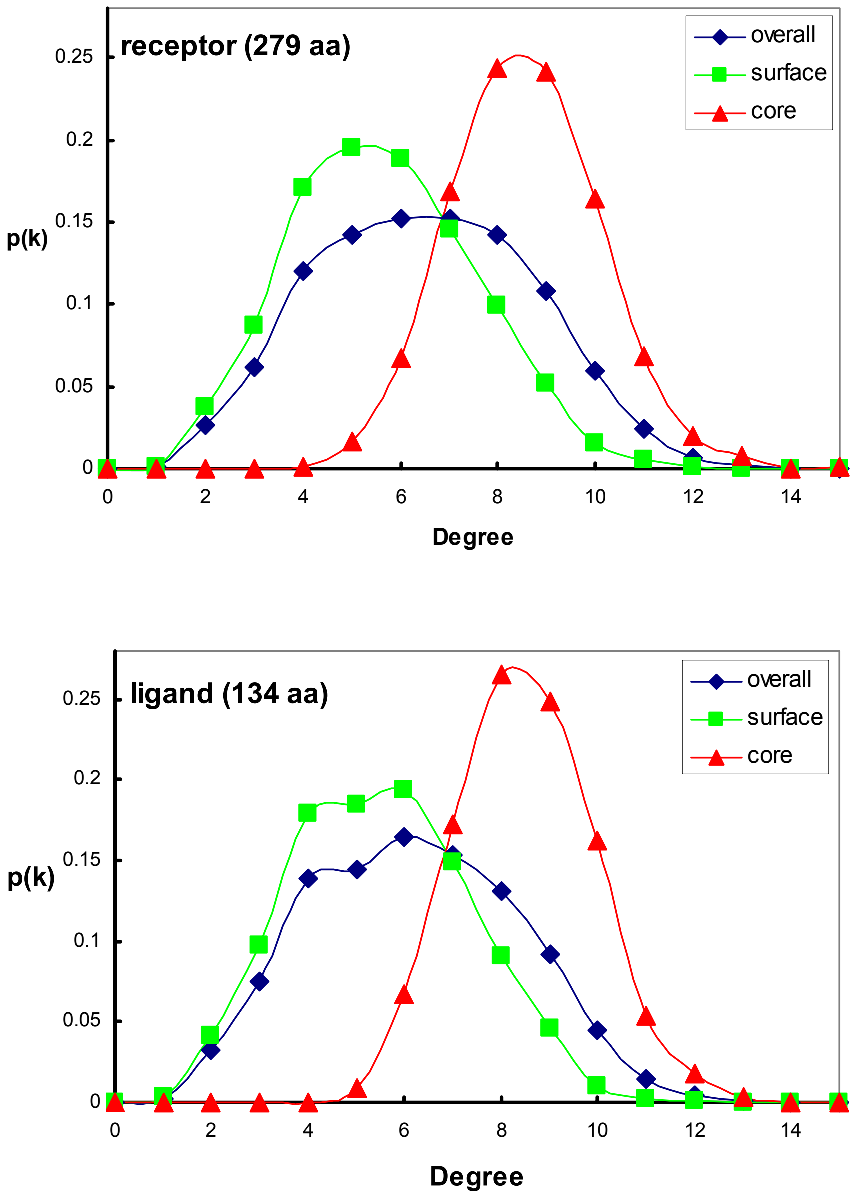
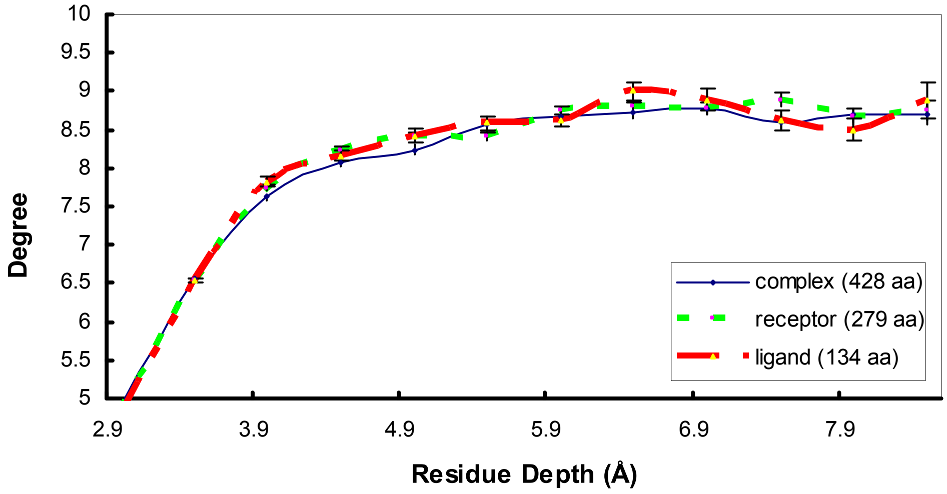
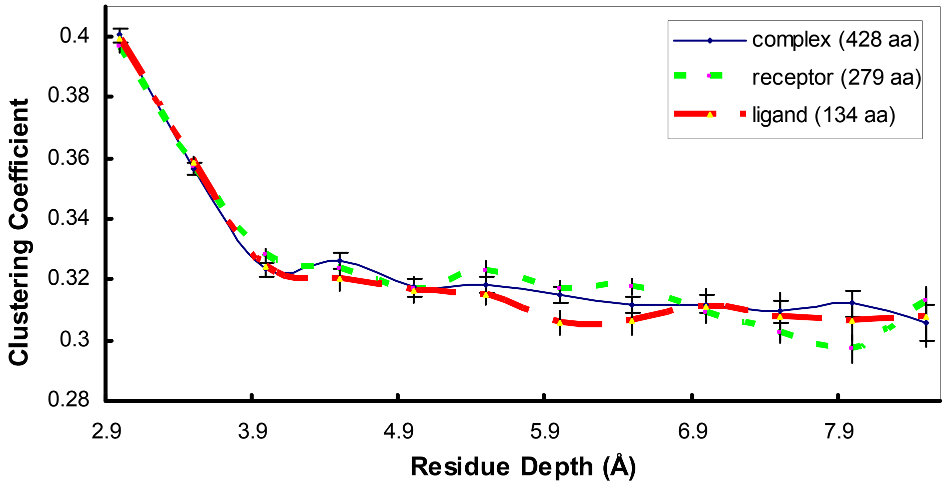


Figure 3. 1 Residue contact distribution at $r_c = 6.7$ Å for the (a) complex, (b) receptor, (c) ligand.

Previous research shows that the distribution over connectivity displays a Gaussian distribution for both the hydrophobic core and molten surface residues. Note that Atilgan *et al.* demonstrate this property of residue networks at 7 Å cut-off. A similar distribution was observed Miyazawa&Jernigan for each of the 20 amino acid types [Miyazawa&Jernigan, 1996]. The graphs above are plotted using the results of residue networks with 6.7 Å. The observation of a Gaussian distribution does not change with the size of the networks: the residue network of (a) the whole complex with an average of 428 amino acids per complex (b) only the receptor proteins with an average of 279 amino acids per protein (c) only the ligand proteins with an average of 134 amino acids per protein.

In the same study of Atilgan *et al.* the relation of *C* and *L* with *residue depth* was investigated (again at 7 Å cut-off) separately. They have shown that the clustering coefficient is independent of network size, and the value of *C* approaches a fixed value of 0.35 at depths greater than 4 Å. On the other hand, the characteristic path length decreases consistently with depth, and as the network size increases, the curve shifts to higher values. In our study, similar results with identical inferences are extracted with r_c = 6.7 Å (Figure 3.2). There is a difference observed at the converged value of *C* (0.35 → 0.31).





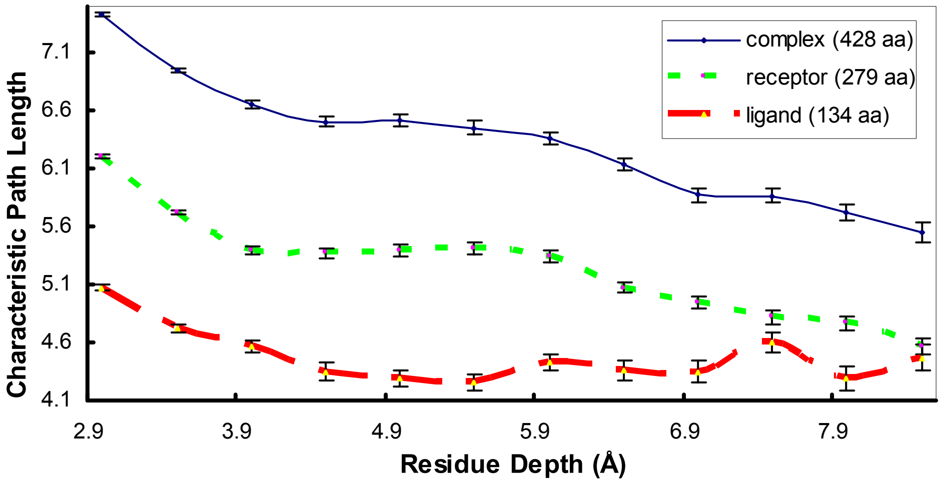


Figure 3. 2 Depth dependence of network parameters; (a) degree, (b) clustering coefficient, (c) characteristic path length.

3.2 Proteins Organize into Larger Networks by Binding

The presence of SWN properties in proteins has already been shown in previous research [Atilgan *et al.*, 2004], and in the previous parts of this thesis. Is the same status valid for protein complexes? To answer this question, we have analyzed the networks formed by the complexes of our data set. The results strongly support the idea that the bound complexes of proteins behave also as SWNs.

As mentioned in section 3.1, C –irrespective of system size– decreases from a value of 0.55 and at depths greater than 4 Å become fixed at a value of *ca.* 0.31 (for r_c = 6.7Å). Figure 3.3 shows that the average C values of interface residues decreases after the complex formation, indicating that the surface residues of unbound proteins gain core status upon complexation. Note that interface residues are thos that reside on different proteins and that are at most 6.7 Å from each other.

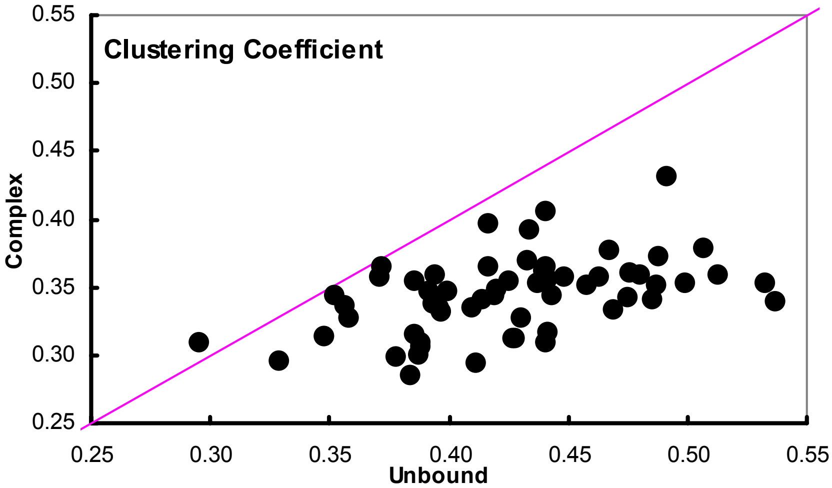


Figure 3. 3 Clustering coefficients of the interface residues.

That the bound proteins form a perfect match so that the interface resembles the inside of a single protein, is also indicated by the change in the average connectivities of the interface residues;

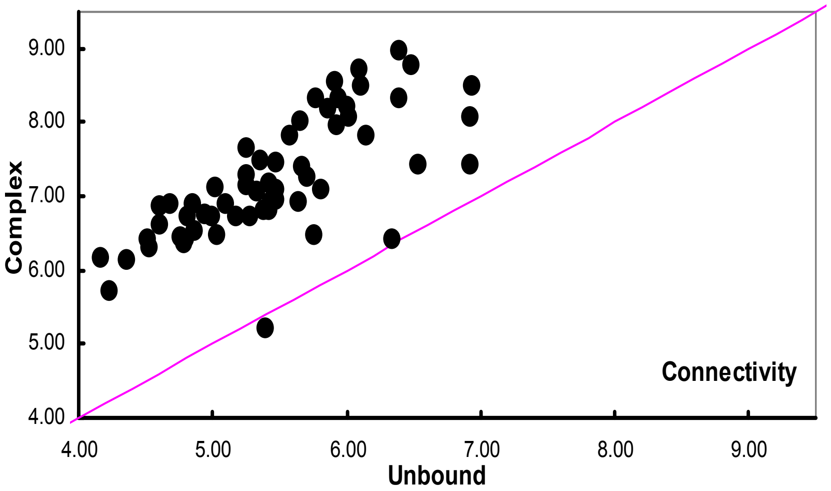


Figure 3. 4 Connectivities of the interface residues.

The depth-connectivity relation was discussed in section 3.1 (Figure 3.2a), suggesting an increase from the molten surface to the hydrophobic core approaching a value of *ca.* 9. The graph above demonstrates the increase in the average connectivity of interface residues; however, it does not approach the values of the core. The connectivity indeed never reaches 9 except for a few complexes, proposing that the depth of interface residues in the complex structure is not so high.

3.3 Identifying interface residues by comparing path lengths in unbound forms

The agreement of shortest paths with residue fluctuations, thus with protein dynamics, was discussed in section 2.2.3. Therefore, it is quite reasonable to expect some residues –interface residues in our definition– to have higher shortest paths than others since proteins function via the residues that fluctuate dynamically in space. However, the results do not support those expectations with all three calculations of *L* (i.e. weighted strong disorder and weak disorder, and weightless) as seen in figure 3.5.

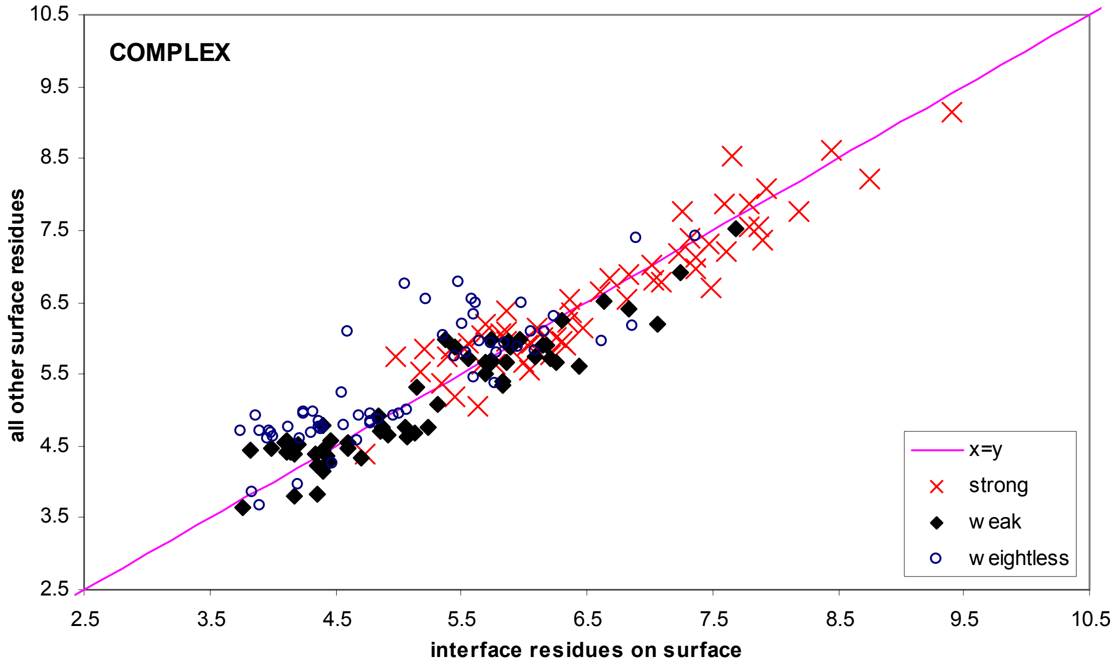


Figure 3. 5 Comparison of average shortest path lengths (MJ potential set). The average shortest paths of interface residues that are placed at a depth of 4 Å or less (i.e. in the unbound forms of the receptor and the ligand, separately) are compared with those of all other residues that reside at the same depth. The result does not reflect our expectation that the interface residue would have higher *L* than that of an arbitrary surface residue (discussed more in Appendix D.1). On the contrary, the latter has slightly higher shortest paths on average. A similar graph and same inference is made from the results of links weighted by Thomas and Dill’s set of contact potentials, as seen in Figure 3.6.

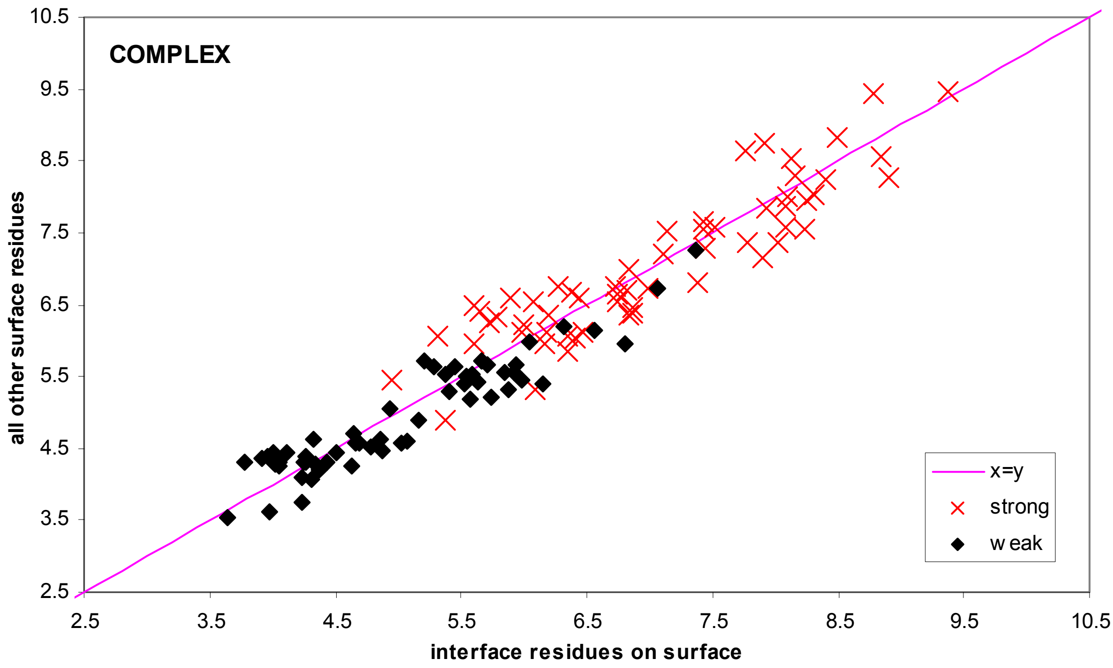


Figure 3. 6 Comparison of characteristic path lengths (TD potential set).

From this figure (note that, the weightless average shortest paths are not included) and Figure 3.5, we can propose that changing the weights of the interacting residue potentials does significantly change the quantitative values of *L*.

The quantitative change can be clearly observed from the following comparison of average shortest paths of surface residues calculated with MJ and TD contact potentials, separately. As one can see in Figure3.7a and Figure 3.7b, in the data calculated with MJ and TD potentials, L_{weak} and L_{strong} are strongly correlated. Besides, the difference between L_{strong} and L_{weak} calculated with MJ potentials is almost identical with the difference between L_{strong} and L_{weak} calculated with TD potentials.

The results for receptors and ligands (separately) are discussed in Appendix D.

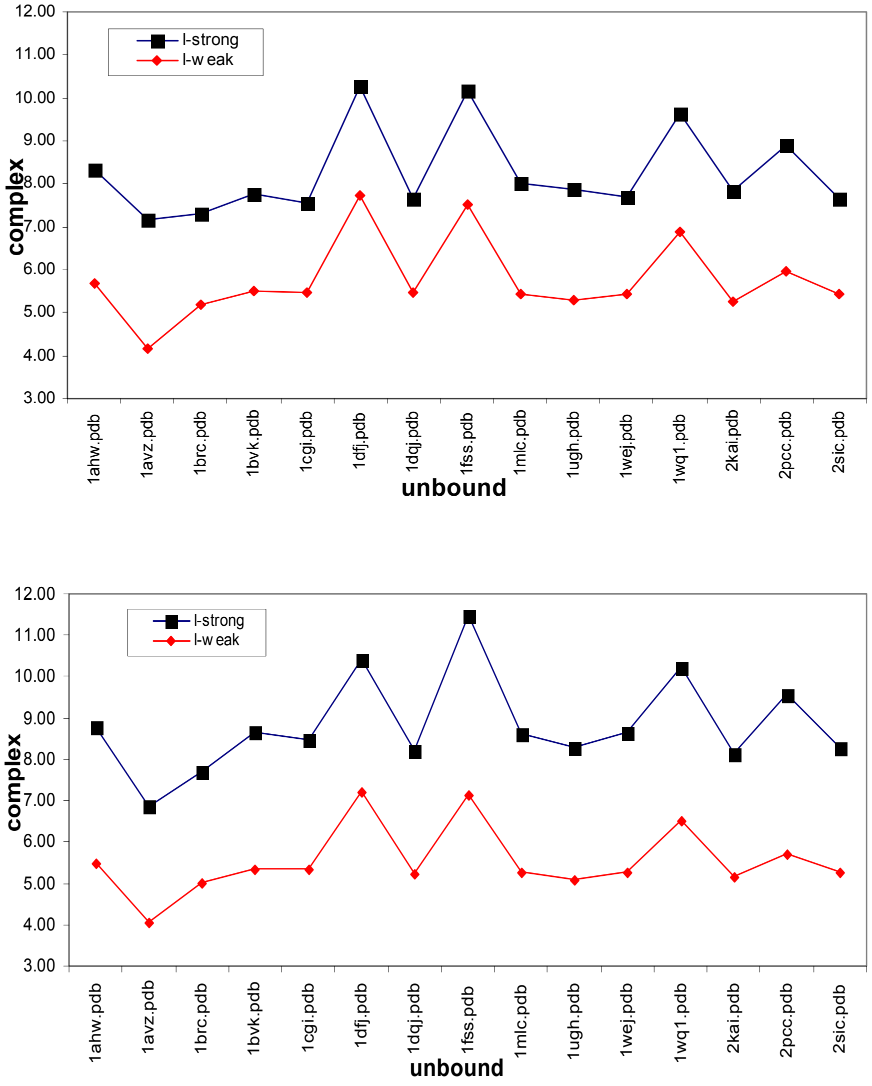


Figure 3. 7 Comparison of *L* of all surface residues calculated using (a) MJ potentials. (b) TD potentials.

3.4 Amino acid based analysis of residue networks

The characteristic network parameters are also used to compare the possible different behavior of amino acid types. The global properties *K* and *C* should obviously differentiate between residue types. For example, the inference that is made in section 3.2 –that small proteins organize into larger systems with the same properties upon binding– could be proved applying the same comparison as in that section onto different types of amino acids. The same hypothesis of depth-*K* and depth-*C* relation is used, and thus the following two graphs support our theory of the formation of a larger network with the same properties (complex) from binding of two smaller networks (receptor and ligand).

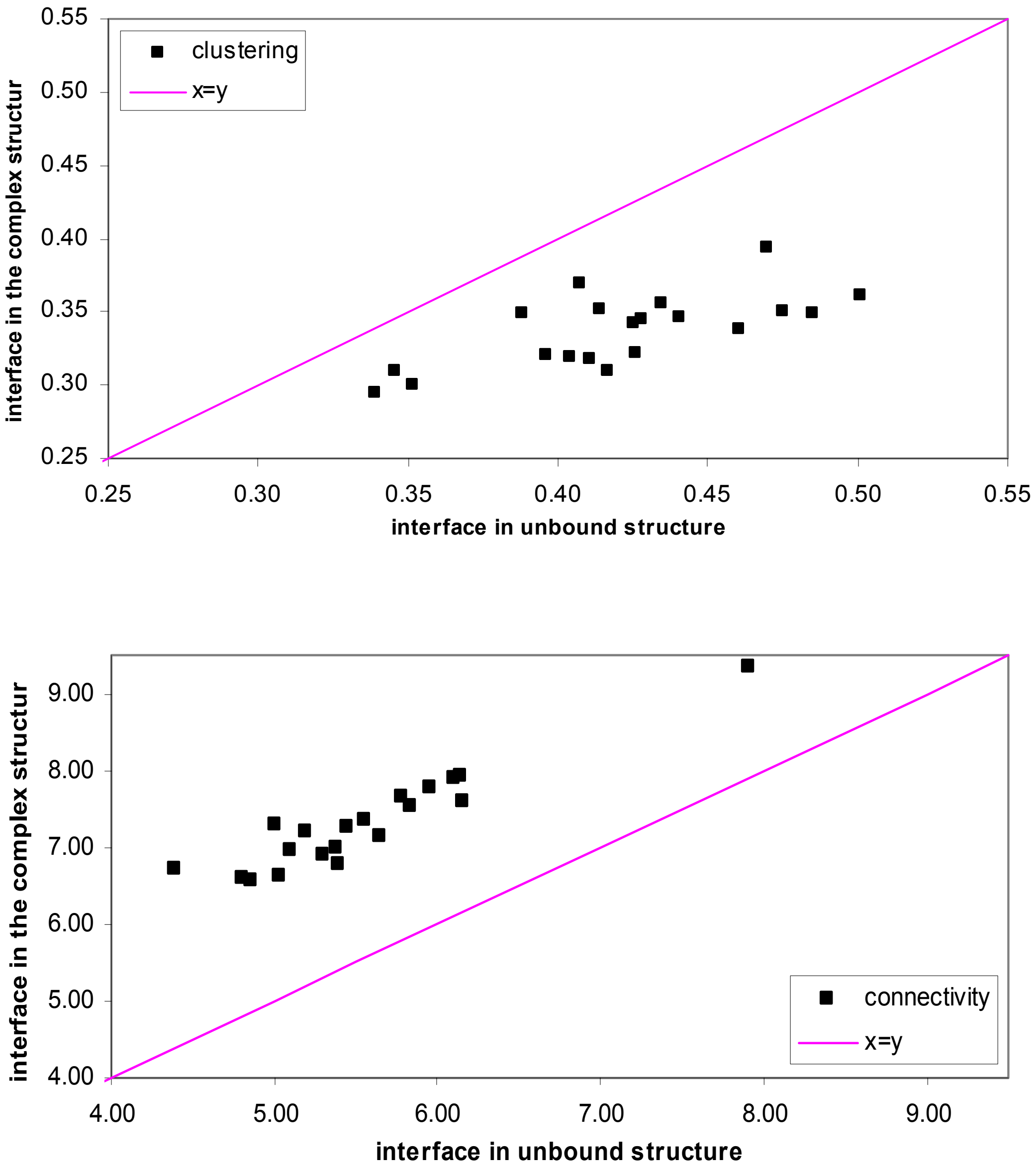


Figure 3. 8 (a) Clustering coefficient, (b) connectivity of interface residues in unbound form vs. in complex form.

From the above two figures, it can be inferred that all characteristics of small world networks are also well fitting with the averaged amino acid values of SWN parameters *K*, *C*, and *L*.

3.5 Native Structures vs. Decoy Structures

Since decoy structures of complex proteins are formed by misdocking of ligand to receptor, one can rightfully expect different behaviors from complex networks. Yet, our calculations concerning both native and decoy structures result in no clear difference between their SWN parameters. As an example, the following five figures compare *C*, *K* and *L* values of interface residues belonging to both type of structures (note that, L_{strong} and L_{weak} are calculated using MJ potentials set);

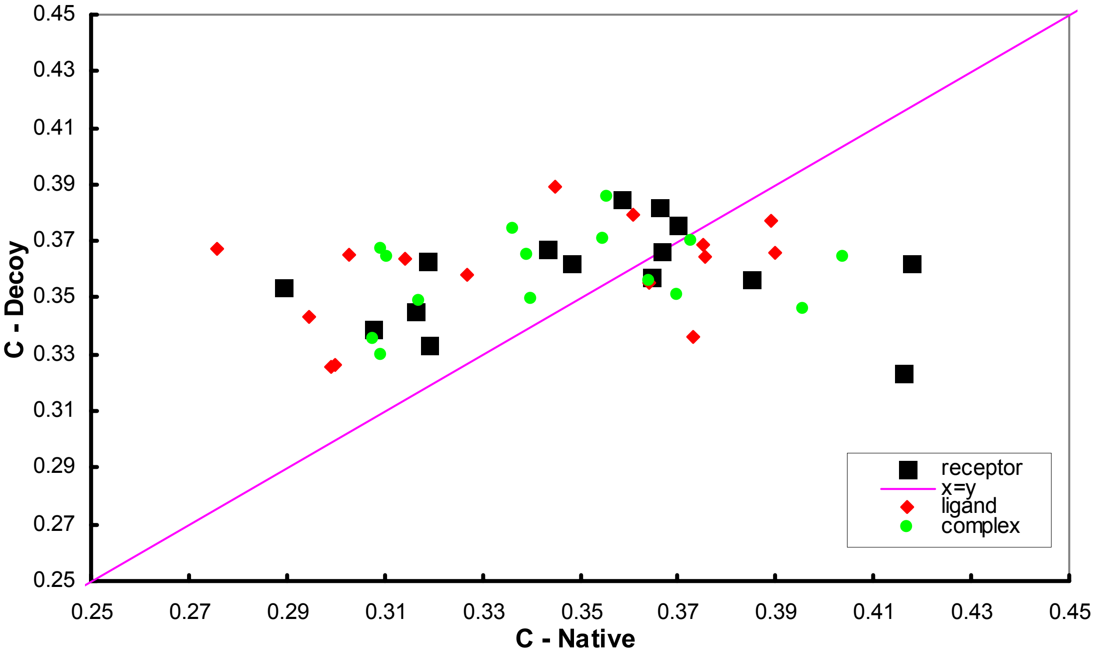


Figure 3. 9 Clustering of interface residues of native and decoy structures.

In Figure 3.9, relatively higher values of *C* of interface residues in decoy structures with respect to that in native structures is observed (i.e. *ca.* %70 of all three kinds –receptor, ligand, and complex– have higher values for decoys). That is a good demonstration of wrongly bound proteins. Since the geometric docking of a particular ligand over a particular receptor requires the perfect matching of protrusions with intrusions, decoys not fulfilling that matching would clearly result in less clustering in the interacting part of the complex. It is also observed that the clustering coefficient of the native structures spans a broad range (from 0.3 to 0.4), yet that of decoys span in relatively narrow range (from 0.27 to 0.43). Same range difference is also observed between the connectivities of decoy structures and native structures (Figure 3.10).

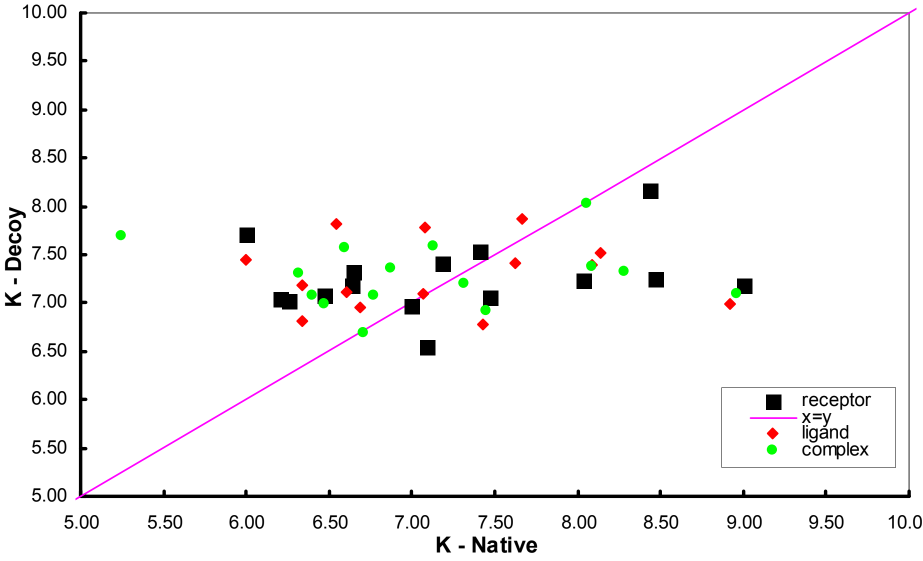


Figure 3. 10 Connectivity of interface residues of native and decoy structures.

Connectivity of the same interface residues, on the other hand, does not differ between native and decoy structures in any way, although our expectation includes relatively lower *K* values for decoys since the possibility of forming new neighbors is low because of the lack of exact geometric match.

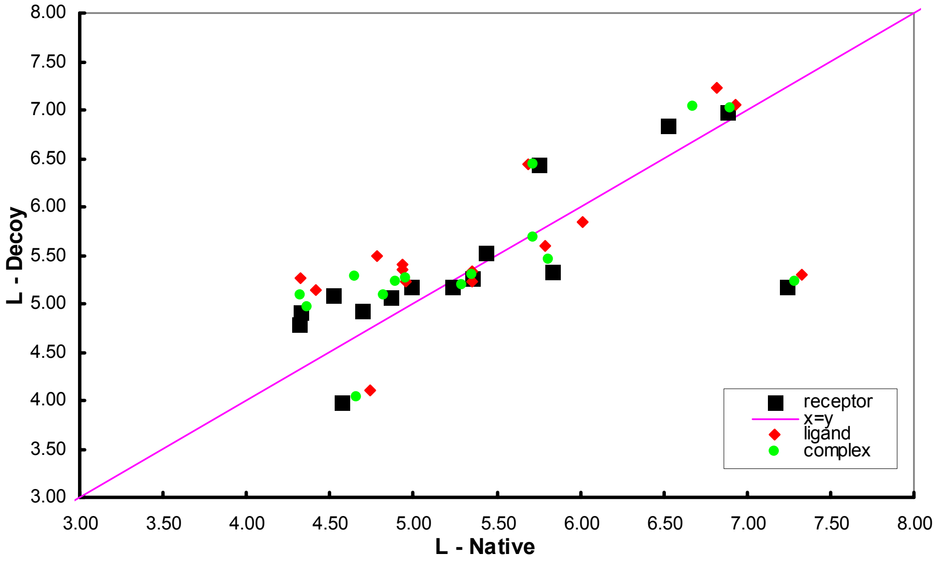


Figure 3. 11 *L* of interface residues of native and decoy structures.

Using the same justifications made for the reasoning of Figure 3.9 above, the relatively high values of *L* in Figure 3.11 is explained. The lack of all possible contacts between receptor and ligand prevents some optimal paths existing in native structure to occur in decoys, again due to non-exact geometric match. L_{weak} shows a similar behavior in the comparison between native structures and decoy structures (Figure 3.12).

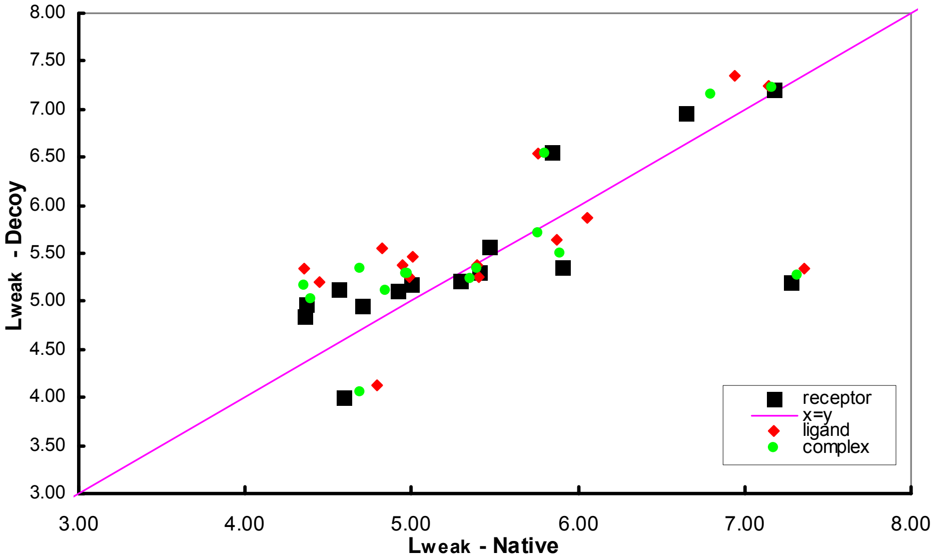


Figure 3. 12 L_{weak} of interface residues of native and decoy structures (TD).

Contrary to the fact inferred in the previous graph, Figure 3.13 –demonstrating the comparisons of L_{strong} – shows no difference between native and decoy structures.

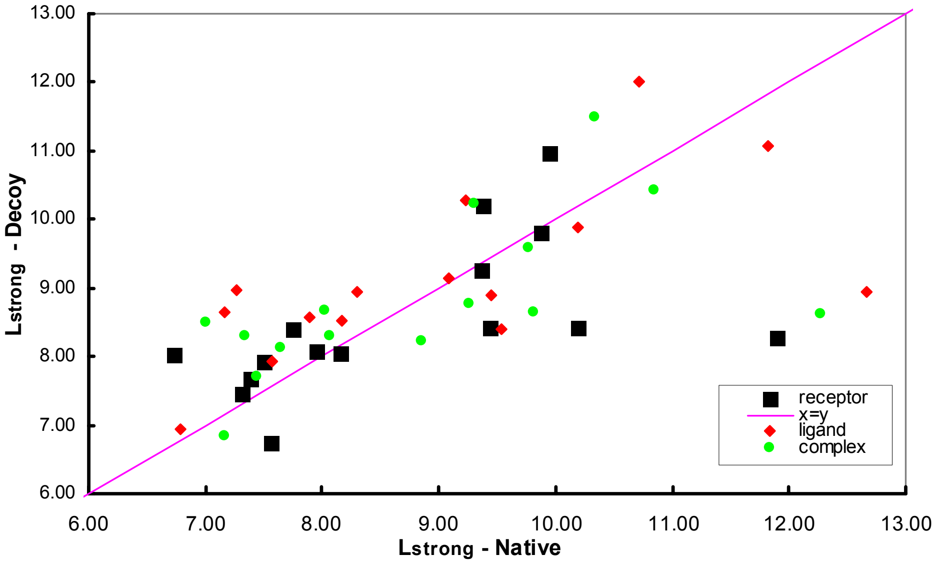


Figure 3. 13 L_{strong} of interface residues of native and decoy structures (TD).

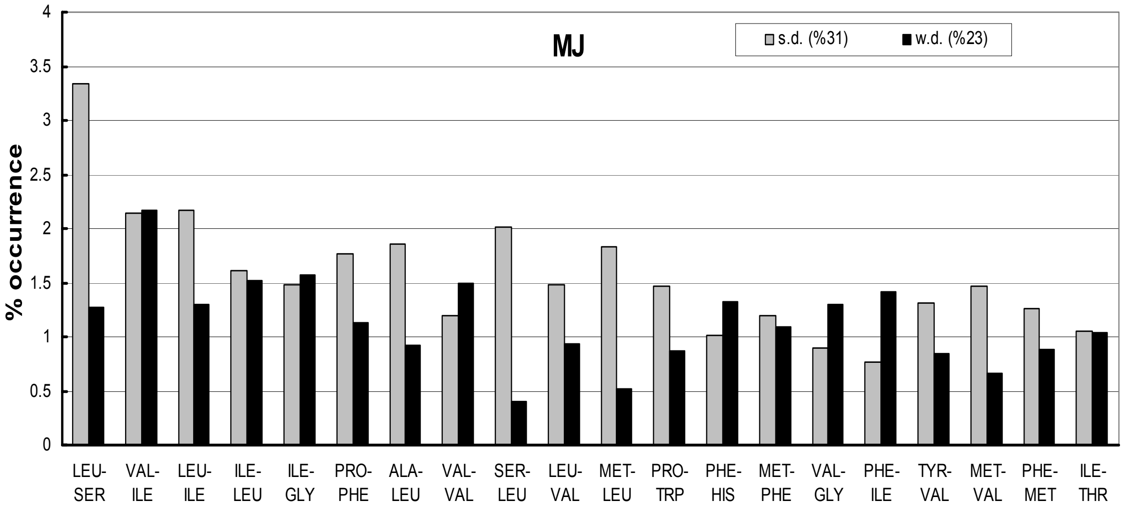
The reason for this observation hides in the definition of L_{strong} (i.e. the optimal path is dominated by a strong link on the way). In the line of this definition, the L_{strong}

does not depend on the number of contacts between a receptor and a ligand unless the mismatch of the geometry does not cause the loss of a strong link between receptor and ligand existing in the native structure. Therefore, the scatter in the graph in Figure 3.13 is quite reasonable.

Figure 3.12 and 3.13 are drawn using the data calculated by TD potentials. Similar graphs and same inferences are observed with the data calculated by MJ potentials.

3.6 Contact Residues used in the Shortest Paths

It is also of interest to determine the amino acid pairs that significantly couple in bound complexes. In this part of the thesis, the pairings existing in the shortest pathways from every node in the receptor protein to every node in the ligand protein are analyzed.



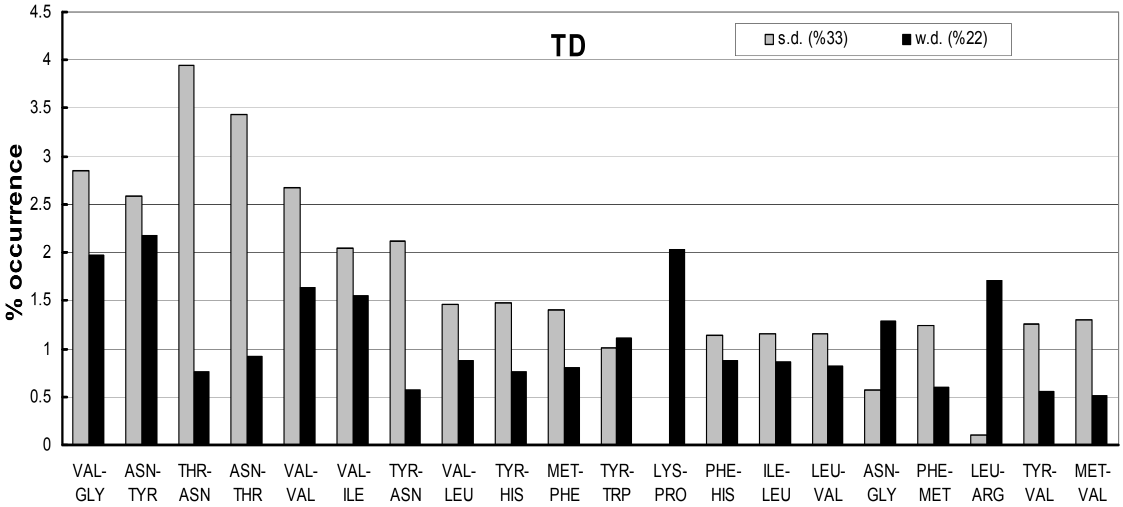


Figure 3. 14 Receptor→Ligand residue contacts with (a) MJ and (b) TD potential sets (native).

The figures above reflect the percentage of amino acid pairs used in the optimal paths from every residue in receptor proteins to every residue in ligand proteins for calculations of *L* with both strong disorder and weak disorder. While choosing the pairs, the percent occurrences of strong disorder and weak disorder cases are added up and the top 20 pairs are graphed. (Notice that the values seen in the labels equal the total percentage of these 20 pairs).

The role of weights –the TD and MJ contact potentials calculated differently– can be clearly seen. The 20 pairs, in each diagram, are chosen among all the possible 400 pairs of 20 different amino acids.

There are 10 pairs (*ILE-LEU, TYR-ASN, VAL-GLY, PHE-HIS, LYS-PRO, THR-ASN, TYR-HIS, LEU-ARG, ASN-GLY,* and *PHE-MET*) occurring in both top 20 used contact residues data (50 percent difference) extracted using TD and MJ contact energy sets. That shows the different behavior of a system with different weights.

As expected, in figure 3.14a, most of the pairs agree with the suggestion of Miyazawa and Jernigan saying that their set of contact potentials are well fitting with the hydrophobicities attained by experimental data. Only 5 of each 40 amino acids forming the 20 pairs are not hydrophobic, and it is quite sensible that hydrophobic residues prefer to cover the surroundings with other residues more than water itself. The other interesting observation in this same figure is that the total percentage of the 20 pairs existing in the graph gives similar results; *%31 and %23* for strong and weak disorder, respectively. That is also consistent when we think of their contact potential data; noting that (from Appendix B) the values of hydrophobic interactions are much higher than that of hydrohobic-polar or polar-polar interactions on average and the different hydrophobic contact energies for different hydrophobic pairs are very close to each other [Miyazawa&Jernigan, 1996].

On the other hand, in figure 3.14b the difference in the total percentages are relatively higher (total with strong disorder is 33% while it is 22% for weak disorder). The deviation of the hydrophobic interaction potentials that Thomas and Dill suggested is much higher than that of potentials suggested by Miyazawa and Jernigan, the same is also true for hydrophobic-polar and polar-polar interaction potentials [Thomas&Dill, 1996].

The two graphs above are the demonstrations for native structures of complex proteins. We now would like to see if a similar inference is possible for decoys. The following two bar type graphs reflect the results for decoy structures:

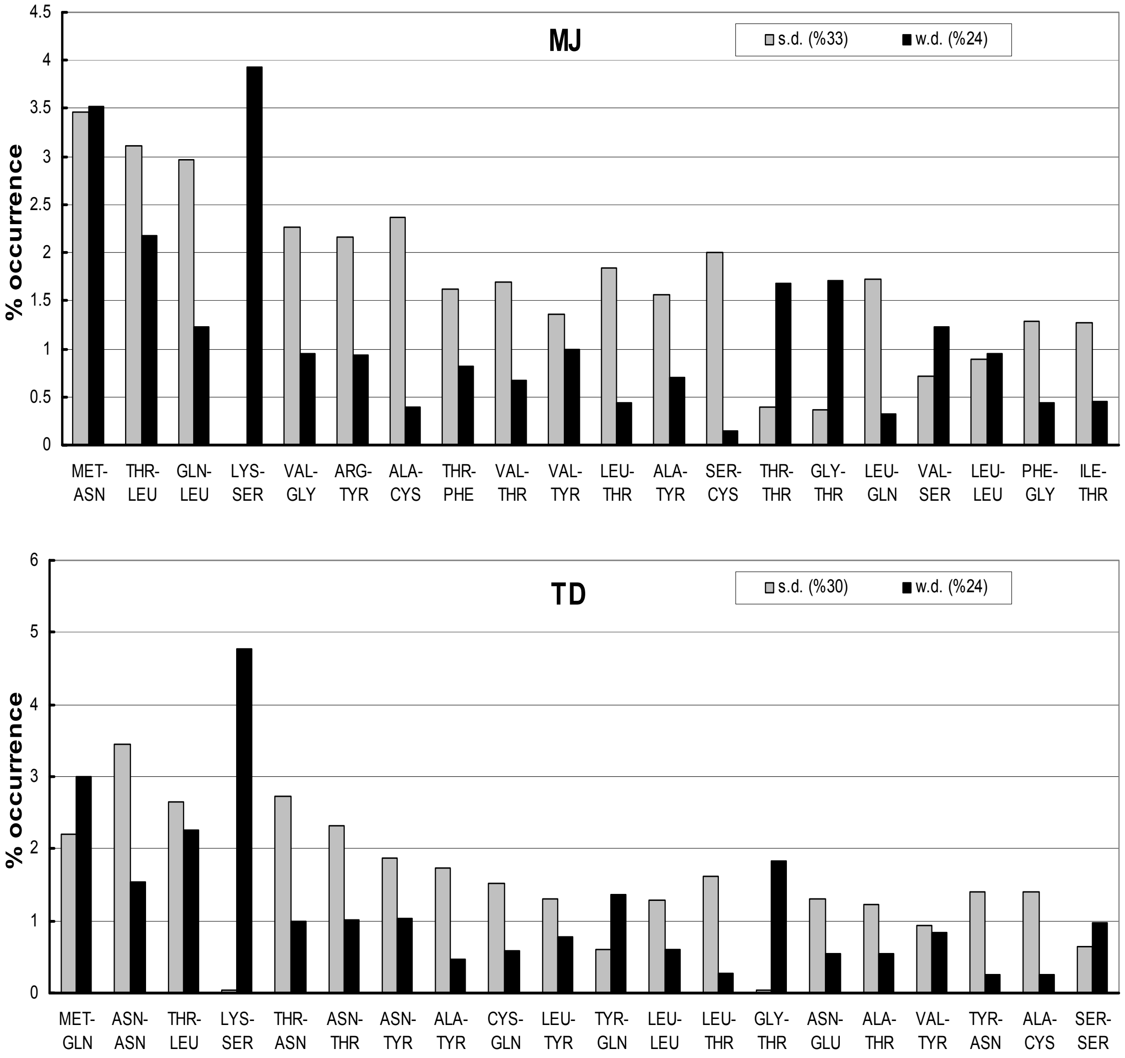


Figure 3. 15 Receptor→Ligand residue contacts with (a) MJ and (b) TD potential sets (decoy).

A first glance to the graph extracted with the MJ potentials set, the number of polar residues in the top 20 pairs increased from 13 (out of 40) in the native set to 18 (out of 40) in the decoy set. A similar increse from 16 to 26 occurs with TD potentials set.

Better explanations could be claimed by the following graphs comparing the behavioral differences between native and decoy networks.

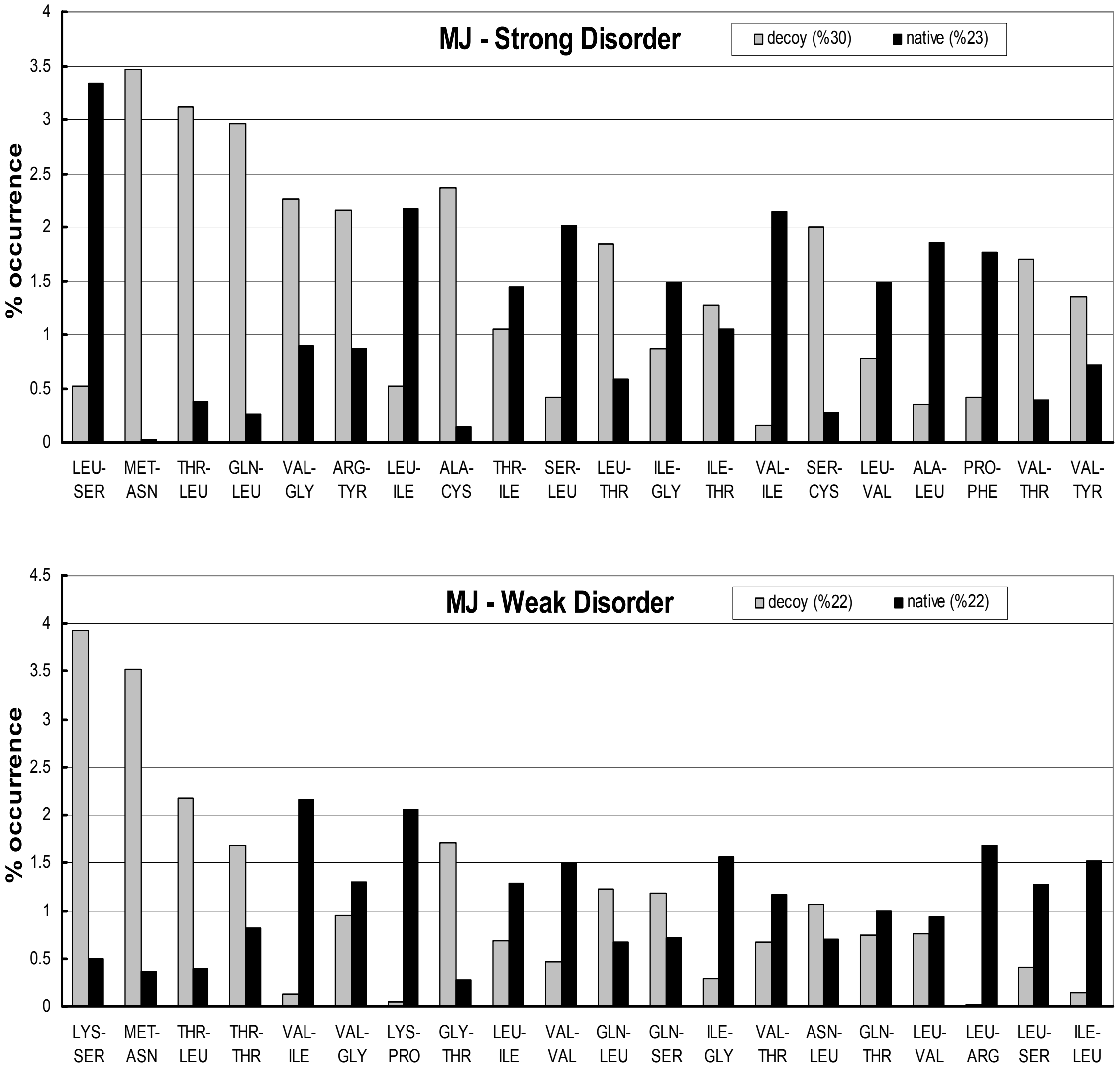
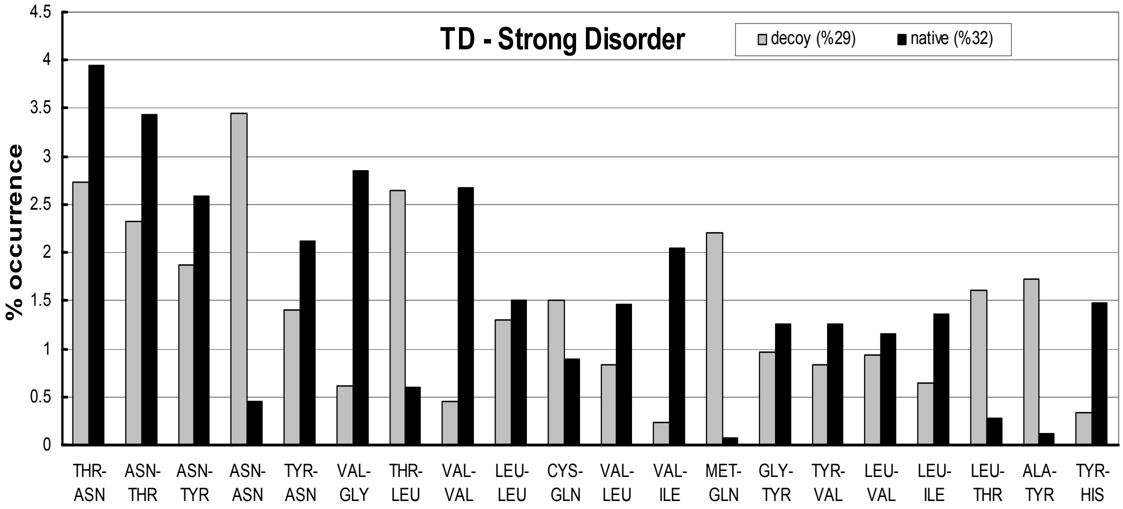


Figure 3. 16 Receptor→Ligand residue contacts (native vs. decoy structures) of (a) strong disorder and (b) weak disorder (MJ potentials).

Using the figures above, the clear and useful inference could be the dramatic difference between the total percentages of weakly disordered native and decoy networks; native structures have almost twice as large contribution in the top 20 pairs than that of decoy structures. Therefore, identifying receptor-ligand contact residues used in the weakly disordered optimal paths could be a good strategy of discrimination of native from decoy structures.

Looking at the total percentages, the difference in figure 3.16b and figure 3.17.a-b could be regarded negligible. However, a closer look would show that the values differ dramatically for individual pairs. For example, *LEU-SER* (figure 3.16a) pair has a percentage of 3.5 in native networks; however, it has only a percentage of ≈0.5 in decoy networks. Similar inferences could be made with many pairs in both strong and weak disorder data extracted both with the MJ set (above) and the TD set (below). However, MJ set obviously discriminates better between decoys and natives since the calculations with MJ potentials result in greater difference between the values of individual pairs.



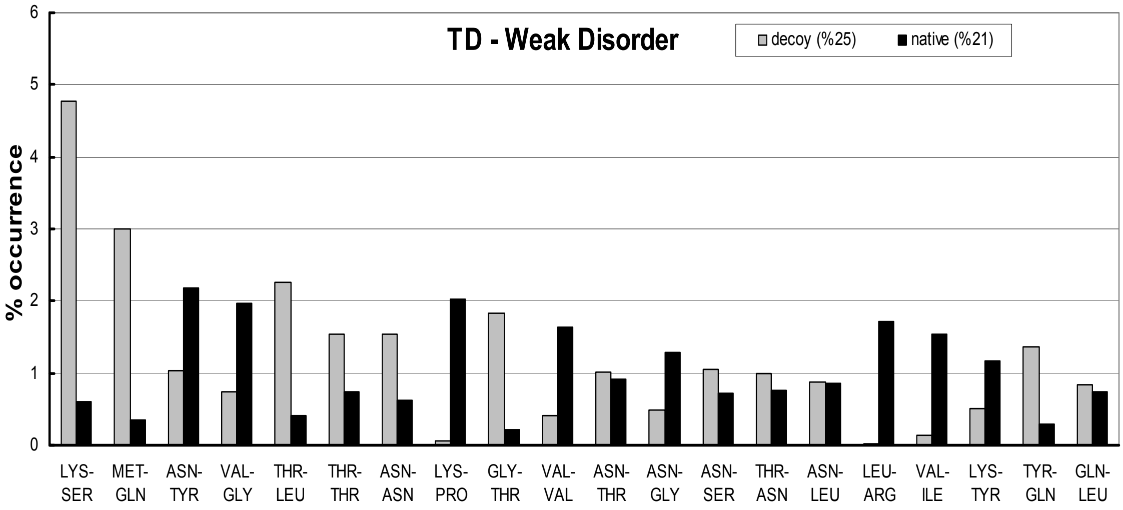


Figure 3. 17 Receptor→Ligand residue contacts (native vs. decoy structures) of (a) strong disorder and (b) weak disorder (TD potentials).

4. CONCLUSION

In this thesis, a set of 59 protein-protein complexes, and the unbound proteins that form the complexes, are converted into networks of interacting residues and their network properties are examined.

The formation of a new network that emerges by the binding of a ligand protein to its receptor is shown to carry the characteristics of a small world network (i.e. the contact distribution of the complex is the same as that of single proteins, the depth dependence of average shortest path and average clustering is in good agreement with the known relation). The same agreement is shown to exist with the calculations averaging the *K* and *C* values of different types of amino acids.

The identification of interface residues in the native structures of complexes is also tested with different definitions of shortest path (i.e. shortest path with no weights, *L*, with weak disorder L_{weak}, and with strong disorder L_{strong}). Results suggest a relatively lower value of average shortest path of interface residues that are in the molten surface compared to that of all other surface residues. The same comparison of the small world characteristics between native structures and decoy structures are also implemented. *L* and L_{mean} comparisons show that the decoys have relatively higher values than the native structures. On the other hand, L_{strong}, *C* and *K* results do not reflect any discriminating characteristics between these different structures.

Receptor to ligand residue contacts are analyzed in the final part of the thesis. Here we give importance not to all residues in the interface, but only to those residues that are most frequently used in all the possible pathways between the receptor and the ligand. These pairs mostly consist of hydrophobic residues with the calculations using both the MJ and TD potential sets. It was also observed that these pairs are different in strong disorder and weak disorder. However, we are led to different results depending on which set is used in the calculations. For example, the percentage of hydrophobic residues within the top 20 pairs is relatively higher for the MJ case. In addition, pairs are observed to differ between the results of native structures and decoy structures. It is shown that the MJ set discriminates between the decoy and native structures in the strong disorder case, whereas TD set is more discriminating in the weak disorder case. Nevertheless, irrespective of the potential set used, we find *ILE-LEU, TYR-ASN, VAL-GLY, PHE-HIS, LYS-PRO, THR-ASN, TYR-HIS, LEU-ARG, ASN-GLY,* and *PHE-MET* pairs to be frequently utilized in complexation.

In future studies, one might bring together these network features to develop a methodology to discriminate the structures closest to the native complex from amongst a large set of structures reproduced by a prediction algorithm. It should also be possible to point out an irregular network structure along the interface. Such a property will especially be useful to determine situations where there is substantial structural change upon binding.

REFERENCES

1. Abola E.E., F.C. Bernstein, S.H. Bryant, T.F. Koetzle, and J. Weng, *"Protein Data Bank" in Crystallographic Databases - Information Content, Software Systems, Scientific Applications*, eds. F.H. Allen, G. Bergerhoff and R. Sievers, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987. p. 107-132.

2. Bahar, I., B. Erman and A.R. Atilgan, *Direct Evaluation of Thermal Fluctuations in Proteins Using a Single Parameter Harmonic Potential*, Fold. Des., 1997. **2(3)**: p. 173-181.

3. Bahar, I., B. Erman, R.L. Jernigan, A.R. Atilgan and D.G. Covell, *Collective Dynamics of Hiv-1 Reverse Transcriptase: Examination of Flexibility and Enzyme Function*, J. Mol. Biol., 1999. **285**: p. 1023-1037.

4. Barabasi, A.-L, R. Albert and H Jeong, *Internet: Growth Dynamics of the World-Wide Web*, Nature, 1999. **401**: p. 130-131.

5. Barabasi, A.-L. and R. Albert, *Emergence of Scaling in Random Networks*, Science, 1999. **286**: p. 509-512

6. Barabási, A. L., H. Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vicsek, *Evolution of the social network of scientific collaborations*, Physica A, 2002. **311**: p. 590-614

7. Barrat A., M. Barthèlèmy, R. Pastor-Satorras, and A. Vespignani, *The architecture of complex weighted networks*, Proc. Natl. Acad. Sci., 2003. **101(11)**: p. 3747-3752.

8. Barthèlèmy, M. and L.A.N. Amaral, Small-World Networks:Evidence for a Crossover Picture. Phys. Rev. Lett., 1999. 82: p. 3180-3183.

9. Baysal, C. and A.R. Atilgan, *Elucidating the Structural Mechanisms for Biological Activity of the Chemokine Family*, Proteins, 2001. **43**: 150-160.

10. Baysal, C. and A.R. Atilgan, *Relaxation Kinetics and the Glassiness of Proteins: The Case of Bovine Pancreatic Trypsin Inhibitor,* Biophys. J., 2004. **83**: p. 699-705.

11. Baysal, C, P. Akan & A.R. Atilgan, *Small-World Communication of Residues and Significance for Protein Dynamics,* Biophys. J., 2004. **86**: p. 85-91.

12. Bernstein, F.C., T.F. Koetzle , G.J. Williams, E.F. Jr. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *The protein data bank: a computer based archival file for macromolecular structures.* J Mol Biol 1977. **112**: 535–542.

13. Braunstein, L.A, S.V. Buldyrev, R. Cohen, S. Havlin and H.E. Stanley, *Optimal Paths in Disordered Complex Networks*, Phys. Rev. Lett., 2003. **91**, 168701.

14. Braunstein, L.A., S. Sreenivasan, T. Kalisky, S.V. Buldyrev, S. Havlin and H.E. Stanley, *Effect of Disorder Strength on Optimal Paths in Complex Networks*, submitted to Phys. Rev. E, 2004.

15. Chakravarty, S. and Varadarajan, R., Residue Depth: A Novel Parameter for the Analysis of Protein Structure and Stability, Structure, 1999 **7**: p. 723-732.

16. Chen, R., W. Tong, J. Mintseris, L. Li and Z. Weng, *ZDOCK Predictions for the CAPRI Challenge*. Proteins, 2003. **52**: p. 68-73.

17. Goh, K.-I., B. Kahng and D. Kim, *Universal Behavior of Load Distribution in Scale-Free Networks*, Phys. Rev. Lett., 2001. **87**: 278701.

18. Guimerà, R., S. Mossa, A. Turtschi and L.A.N Amaral, Structure and Efficiency of the World-Wide Airport Network, preprint (available online at http://arxiv.org/abs/cond-mat/0312535)

19. McCann, K., A. Hatings and G. R. Huxel, *Weak Trophic Interactions and the Balance of Nature*, Nature, 1998. **395**: p. 794-798.

20. Ming, D., Y. Kong, Y. Wu and J. Ma, *Substructure Synthesis Method for Simulation Large Molecular Complexes*, Proc. Natl. Acad. Sci., 2003. **100**: p. 104-109

21. Miyazawa, S. and R.L. Jernigan, *Residue-Residue Potentials with a Favorable Contact Pair Term and an Favorable High Packing Density Term, for Simulation and Threading*. J. Mol. Biol., 1996. **256**: p. 623-644.

22. Newman, M. E. J., *The Structure of Scientific Collaboration Networks*, Proc. Natl. Acad. Sci., 2001. **98**: p. 404-409

23. Newman, M. E. J., *Scientific collaboration networks. I. Network construction and fundamental results*, Phys. Rev. E, 2001. **64**: 016131.

24. Newman, M. E. J., *Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality*, Phys. Rev. E, 2001. **64**: 016132.

25. Rosen, K. H., *Discrete Mathematics and Its Applications.* 2003, New York, NY, USA: McGrav-Hill Companies, Inc.

26. Strogatz, S.H., Exploring Complex Networks, Nature, 2001. **410**: p. 268-276.

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

28. Voet, D. and J.G. Voet, *Biochemistry (2nd Edition)*. 1995, Somerset, NJ, USA: John Wiley & Sons, Inc.

29. Watts, D.J. and S.H. Strogats, *Collective Dynamics of 'Small World' Networks*. Nature, 1998. **393**: p. 440-442.

30. Watts, D.J., Small *Worlds: the Dynamics of Networks Between Order and Randomness*. 1999, Princeton, NJ, USA: Princeton University Press.

31. Wilson, R. J., and J. J. Watkins, *Graphs: An Introductory Approach*. 1990, New York, USA: Wiley.

32. Woolhouse, M. and A. Donaldson, *Managing Foot and Mouth*, Nature, 2001. **410**: p. 515-517

APPENDIX A: DIJKSTRA's ALGORITHM

Many different algorithms have been developed to find a shortest path between two vertices in a weighted graph. The one that is discovered by the Dutch mathematician Edsger Dijkstra in 1959 is used in our calculations. The algorithm simply proceeds by making the choice that looks best at each step[6].

In a weakly disordered network the optimal path between any two vertices *i* and *j* is considered as the path with minimum weighted sum of all edges on the way and is simply found by iteratively calculating the following series.

$$L_{ij} = \min\{L_{k-1}(i,j), L_{k-1}(i,x) + w_{x\,j})\}$$

Dijkstra's algorithm proceeds by forming a distinguished set of vertices, S_k, at each step, *k*. Initially S_0 = Ø and S_k, at each iteration, is formed from S_{k-1} by adding a vertex, *x*, that is not in S_{k-1} if it results in smaller sum – $L_{k-1}(i,x) + w_{x\,j}$. The iterations are implemented until *j* is added to S_k [Rosen, 2003].

An example with a weighted network of six vertices and nine edges is demonstrated in Figure A. The purpose is to find a shortest path between *a* and *z*. Initially S_0 = Ø (no vertex is in circle) and the shortest paths to all other edges are assigned ∞. At each step a different vertex is added to S_k, finally *z* is added and the computation terminates.

[6] http://www.cs.dartmouth.edu/~chepner/cs15/notes/22_graphs.html

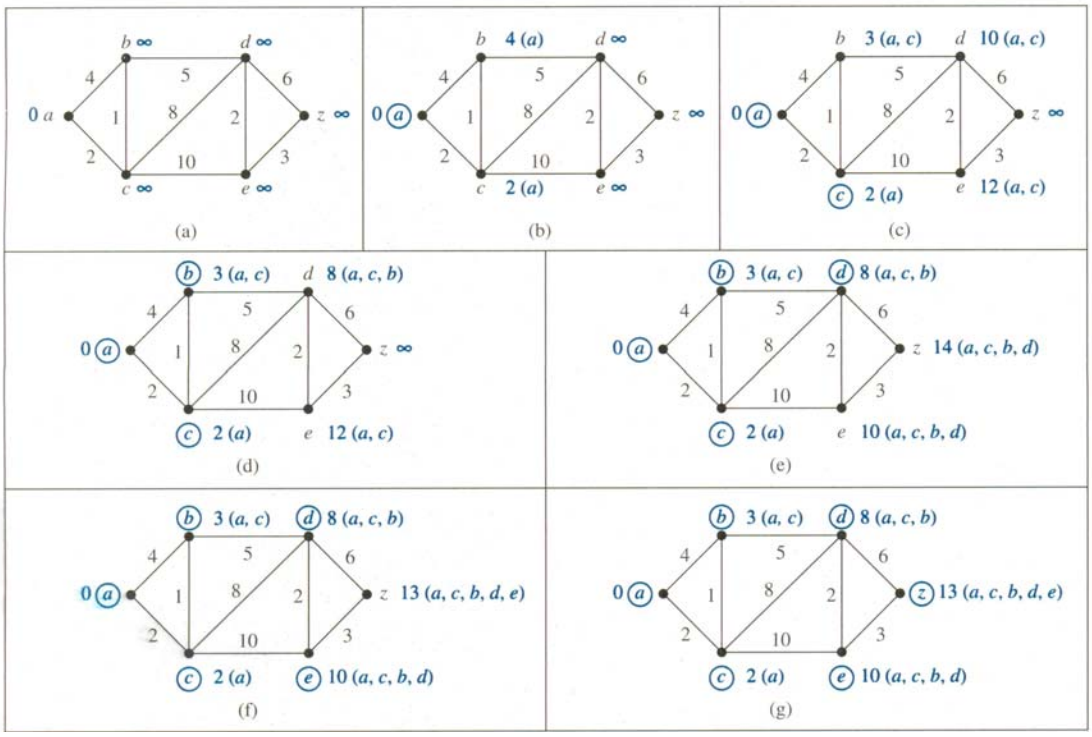


Figure A. 1 Using Dijkstra's Algorithm to find a shortest path from *a* to *z* [Rosen, 2003].

APPENDIX B: INTERRESIDUE CONTACT POTENTIALS

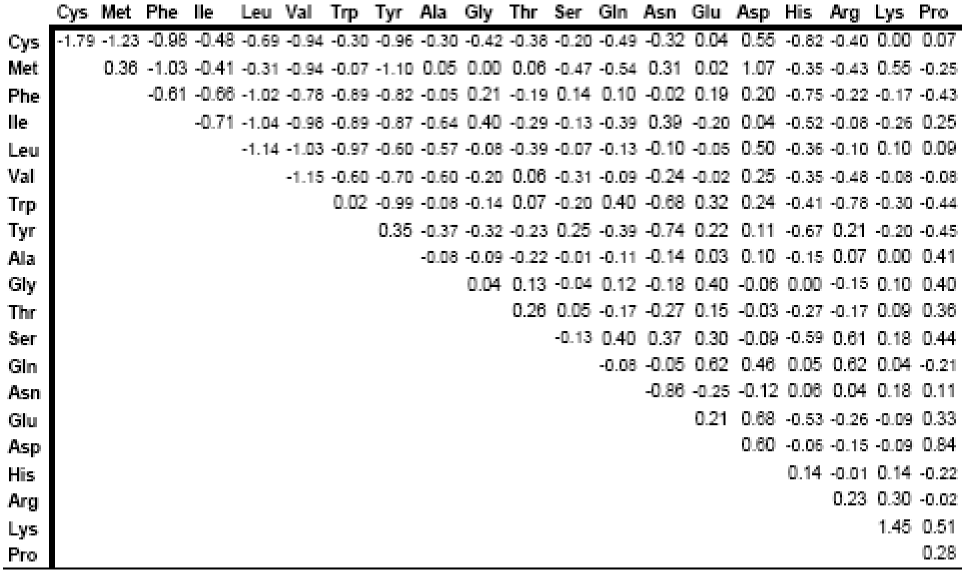
B.1 Overview

The problem, *Protein Folding*, is the most widely studied issue in the world of structural biology. The number of elucidated structures of proteins increase rapidly, and it becomes more and more possible to find a generalized answer for the problem day by day. However, current computational power is not enough for detailed molecular dynamics simulations that employ potentials for full atomic representations of proteins. Therefore, different approaches are also studied. One point of view is contributing the solution by deriving potential functions –energy like quantities– for interacting pairs of residues. Most of these statistical studies apply Boltzmann relation to the pairing frequencies of amino acids observed in known protein structures for their derivations.

The works belonging to Thomas & Dill (TD) and Miyazawa & Jernigan (MJ) give quite successful approximations. MJ contact potentials are calculated statistically over a large set of known structures; on the other hand, TD contact potentials were calculated iteratively, over a relatively smaller set of proteins, until a convergence is reached.

In our thesis, the set of contact potentials for both these two works are modified and used to assign the weights of connected residue pairs. The modification is simply like that; we first add the absolute value of the smallest potential to all potentials (i.e. the smallest, therefore, would be equal to zero), then for each different network we add the average weight of the network to all individual weights.

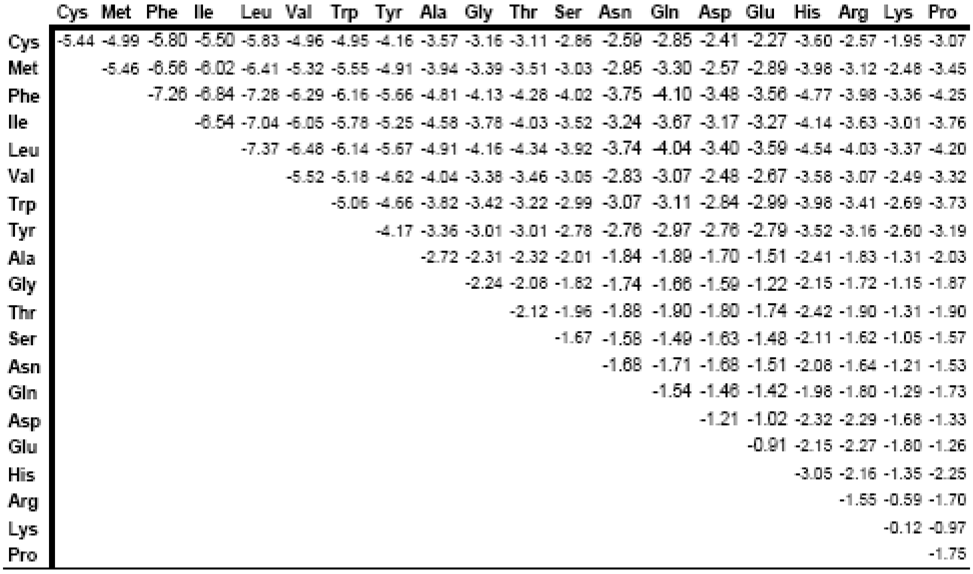
B.2a Thomas & Dill



	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Gln	Asn	Glu	Asp	His	Arg	Lys	Pro
Cys	-1.79	-1.23	-0.98	-0.48	-0.69	-0.94	-0.30	-0.96	-0.30	-0.42	-0.38	-0.20	-0.49	-0.32	0.04	0.55	-0.62	-0.40	0.00	0.07
Met		0.36	-1.03	-0.41	-0.31	-0.94	-0.07	-1.10	0.05	0.00	0.06	-0.47	-0.54	0.31	0.02	1.07	-0.35	-0.43	0.55	-0.25
Phe			-0.61	-0.66	-1.02	-0.78	-0.89	-0.82	-0.05	0.21	-0.19	0.14	0.10	-0.02	0.19	0.20	-0.75	-0.22	-0.17	-0.43
Ile				-0.71	-1.04	-0.98	-0.89	-0.87	-0.64	0.40	-0.29	-0.13	-0.39	0.39	-0.20	0.04	-0.52	-0.08	-0.26	0.25
Leu					-1.14	-1.03	-0.97	-0.60	-0.57	-0.08	-0.39	-0.07	-0.13	-0.10	-0.05	0.50	-0.36	-0.10	0.10	0.09
Val						-1.15	-0.60	-0.70	-0.60	-0.20	0.06	-0.31	-0.09	-0.24	-0.02	0.25	-0.35	-0.48	-0.08	-0.08
Trp							0.02	-0.99	-0.08	-0.14	0.07	-0.20	0.40	-0.68	0.32	0.24	-0.41	-0.78	-0.30	-0.44
Tyr								0.35	-0.37	-0.32	-0.23	0.25	-0.39	-0.74	0.22	0.11	-0.67	0.21	-0.20	-0.45
Ala									-0.08	-0.09	-0.22	-0.01	-0.11	-0.14	0.03	0.10	-0.15	0.07	0.00	0.41
Gly										0.04	0.13	-0.04	0.12	-0.18	0.40	-0.06	0.00	-0.15	0.10	0.40
Thr											0.26	0.05	-0.17	-0.27	0.15	-0.03	-0.27	-0.17	0.09	0.36
Ser												-0.13	0.40	0.37	0.30	-0.09	-0.59	0.61	0.18	0.44
Gln													-0.08	-0.05	0.62	0.46	0.05	0.62	0.04	-0.21
Asn														-0.86	-0.25	-0.12	0.06	0.04	0.18	0.11
Glu															0.21	0.68	-0.53	-0.26	-0.09	0.33
Asp																0.60	-0.06	-0.15	-0.09	0.84
His																	0.14	-0.01	0.14	-0.22
Arg																		0.23	0.30	-0.02
Lys																			1.45	0.51
Pro																				0.28

Table B. 1 TD Interresidue contact potentials [Thomas and Dill, 1996].

B.2b Miyazawa & Jernigan



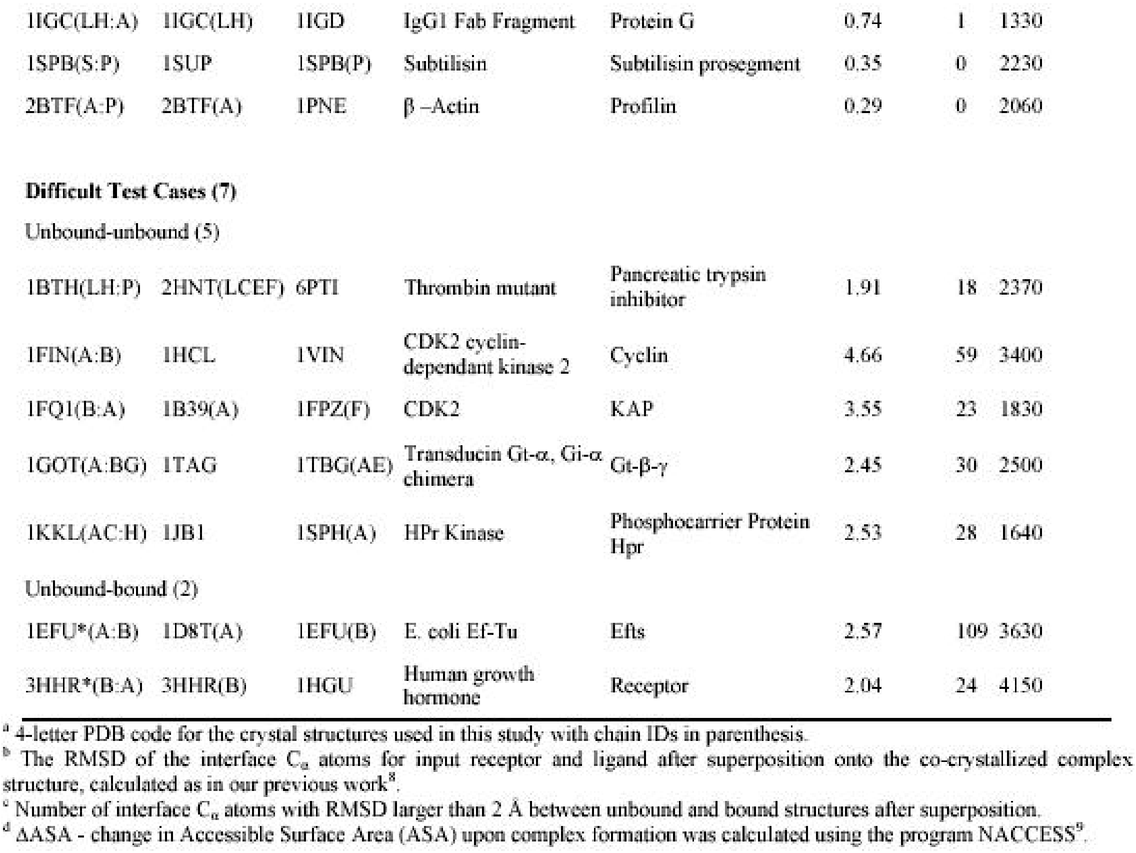
	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Asn	Gln	Asp	Glu	His	Arg	Lys	Pro
Cys	-5.44	-4.99	-5.80	-5.50	-5.83	-4.96	-4.95	-4.16	-3.57	-3.16	-3.11	-2.86	-2.59	-2.85	-2.41	-2.27	-3.60	-2.57	-1.95	-3.07
Met		-5.46	-6.56	-6.02	-6.41	-5.32	-5.55	-4.91	-3.94	-3.39	-3.51	-3.03	-2.95	-3.30	-2.57	-2.89	-3.98	-3.12	-2.48	-3.45
Phe			-7.26	-6.84	-7.28	-6.29	-6.16	-5.66	-4.81	-4.13	-4.28	-4.02	-3.75	-4.10	-3.48	-3.56	-4.77	-3.98	-3.36	-4.25
Ile				-6.54	-7.04	-6.05	-5.78	-5.25	-4.58	-3.78	-4.03	-3.52	-3.24	-3.67	-3.17	-3.27	-4.14	-3.63	-3.01	-3.76
Leu					-7.37	-6.48	-6.14	-5.67	-4.91	-4.16	-4.34	-3.92	-3.74	-4.04	-3.40	-3.59	-4.54	-4.03	-3.37	-4.20
Val						-5.52	-5.18	-4.62	-4.04	-3.38	-3.46	-3.05	-2.83	-3.07	-2.48	-2.67	-3.58	-3.07	-2.49	-3.32
Trp							-5.06	-4.66	-3.82	-3.42	-3.22	-2.99	-3.07	-3.11	-2.84	-2.99	-3.98	-3.41	-2.69	-3.73
Tyr								-4.17	-3.36	-3.01	-3.01	-2.78	-2.76	-2.97	-2.76	-2.79	-3.52	-3.16	-2.60	-3.19
Ala									-2.72	-2.31	-2.32	-2.01	-1.84	-1.89	-1.70	-1.51	-2.41	-1.83	-1.31	-2.03
Gly										-2.24	-2.08	-1.82	-1.74	-1.66	-1.59	-1.22	-2.15	-1.72	-1.15	-1.87
Thr											-2.12	-1.96	-1.88	-1.90	-1.80	-1.74	-2.42	-1.90	-1.31	-1.90
Ser												-1.67	-1.58	-1.49	-1.63	-1.48	-2.11	-1.62	-1.05	-1.57
Asn													-1.68	-1.71	-1.68	-1.51	-2.08	-1.64	-1.21	-1.53
Gln														-1.54	-1.46	-1.42	-1.98	-1.80	-1.29	-1.73
Asp															-1.21	-1.02	-2.32	-2.29	-1.68	-1.33
Glu																-0.91	-2.15	-2.27	-1.80	-1.26
His																	-3.05	-2.16	-1.35	-2.25
Arg																		-1.55	-0.59	-1.70
Lys																			-0.12	-0.97
Pro																				-1.75

Table B. 2 MJ Interresidue contact potentials [Miyazawa and Jernigan, 1996].

APPENDIX C: DATA SET

Complex[a]	Receptor[a]	Ligand[a]	Receptor Description	Ligand Description	RMSD[b](Å)	CA[c]	ΔASA[d](Å^2)
Enzyme-inhibitor (22)							
Unbound-unbound (16)							
1ACB(E:I)	5CHA(A)	1CSE(I)	α -chymotrypsin	Eglin C	0.7	1	1540
1AVW(A:B)	2PTN	1BA7(A)	Trypsin	Soybean Trypsin inhibitor	0.35	0	1740
1BRC(E:I) ✗	1BRA	1AAP(A)	Trypsin	APPI	0.44	0	1320
1BRS(A:D)	1A2P(B)	1A19(A)	Barnase	Barstar	0.47	0	1560
1CGI(E:I) ✗	1CHG	1HPT	α-chymotrypsinogen	Pancreatic secretory trypsin inhibitor	1.48	14	2050
1CHO(E:I)	5CHA(A)	2OVO	α -chymotrypsin	Ovomucoid 3rd Domain	0.59	1	1470
1CSE(E:I)	1SCD	1ACB(I)	Subtilisin Carlsberg	Eglin C	0.43	0	1490
1DFJ(I:E) ✗	2BNH	7RSA	Ribonuclease inhibitor	Ribonuclease A	1.04	13	2580
1FSS(A:B) ✗	2ACE(E)	1FSC	Snake Venom Acetylcholinesterase	Fasciculin II	0.75	1	1970
1MAH(A:F)	1MAA(B)	1FSC	Mouse Acetylcholinesterase	Fasciculin 2	0.6	0	2150
1TGS(Z:I)	2PTN	1HPT	Trypsinogen	Pancreatic secretory trypsin inhibitor	1.49	17	1720
1UGH(E:I) ✗	1AKZ	1UGI(A)	Human Uracil-DNA glycosylase	Inhibitor	0.53	1	2190
2KAI(AB:I) ✗	2PKA(XY)	6PTI	Kallikrein A	Trypsin inhibitor	0.7	2	1420
2PTC(E:I)	2PTN	6PTI	β-trypsin	Pancreatic trypsin inhibitor	0.32	0	1430
2SIC(E:I) ✗	1SUP	3SSI	Subtilisin BPN	Subtilisin inhibitor	0.4	0	1620
2SNI(E:I)	1SUP	2CI2(I)	Subtilisin Novo	Chymotrypsin inhibitor 2	0.37	0	1630
Unbound-bound (6)							
1PPE(E:I)	2PTN	1PPE(I)	Trypsin	CMT-1	0.27	0	1690
1STF(E:I)	1PPN	1STF(I)	Papain	Stefin B	0.25	0	1790
1TAB(E:I)	2PTN	1TAB(I)	Trypsin	BBI	0.27	0	1360
1UDI(E:I)	1UDH	1UDI(I)	Virus Uracil-DNA glycosylase	Inhibitor	0.36	0	2020
2TEC(E:I)	1THM	2TEC(I)	Thermitase	Eglin C	0.19	0	1560
4HTC(LH:I)	2HNT(LCEF)	4HTC(I)	A –Thrombin	Hirudin	0.56	2	3320
Antibody-antigen (19)							
Unbound-unbound (5)							
1AHW(DE:F) ✗	FGN(LH)	1BOY	Antibody Fab 5G9	Tissue factor	0.71	1	1900
1BVK(DE:F) ✗	1BVL(LH)	3LZT	Antibody Hulys11 Fv	Lysozyme	1.22	3	1400
1DQJ(AB:C) ✗	1DQQ(LH)	3LZT	Hyhel - 63 Fab	Lysozyme	0.73	3	1760

1MLC(AB:E)✗	1MLB(AB)	1LZA	IgG1 D44.1 Fab fragment	Lysozyme	0.85	3	1390
1WEJ(LH:F)✗	1QBL(LH)	1HRC	IgG1 E8 Fab fragment	Cytochrome C	0.32	0	1180
Unbound-bound (14)							
1BQL(LH:Y)	1BQL(LH)	1DKJ	Hyhel - 5 Fab	Lysozyme	0.52	2	1630
1EO8(LH:A)	1EO8(LH)	2VIU(A)	Bh151 Fab	Influenza Virus Hemagglutinin	0.28	0	1530
1FBI(LH:X)	1FBI(LH)	1HHL	IgG1 Fab fragment	Lysozyme	0.5	0	1690
1IAI(MI:LH)	1AIF(LH)	1IAI(LH)	IgG1 Idiotypic Fab	Igg2A Anti-Idiotypic Fab	0.99	12	1890
1JHL(LH:A)	1JHL(LH)	1GHL(A)	IgG1 Fv Fragment	Lysozyme	0.26	0	1240
1KXQ(D:E)	1PIF(A)	1KXQ(E)	α-amylase	Camelid AMD9 Vhh Domain	0.43	0	2140
1KXT(A:B)	1PIF(A)	1KXT(B)	α-amylase	Camelid AMB7 Vhh Domain	0.39	0	1620
1KXV(A:C)	1PIF(A)	1KXV(C)	α-amylase	Camelid AMD10 Vhh Domain	0.24	0	1620
1MEL(B:M)	1MEL(B)	1LZA	Vh Single-Domain Antibody	Lysozyme	0.65	2	1690
1NCA(LH:N)	1NCA(LH)	7NN9	Fab NC41	Influenza Virus Neuraminidase	0.24	0	1950
1NMB(LH:N)	1NMB(LH)	7NN9	Fab NC10	Influenza Virus Neuraminidase	0.21	0	1350
1QFU(LH:A)	1QFU(LH)	2VIU(A)	Igg1-k Fab	Influenza Virus Hemagglutinin	0.27	0	1840
2JEL(LH:P)	2JEL(LH)	1POH	Jel42 Fab Fragment	A06 Phosphotransferase	0.18	0	1500
2VIR(AB:C)	2VIR(AB)	2VIU(A)	Igg1-lamda Fab	Influenza Virus Hemagglutinin	0.41	1	1260
Others (11)							
Unbound-unbound (5)							
1AVZ(B:C)✗	1AVV	1SHF(A)	HIV-1 NEF	FYN tyrosin kinase SH3 domain	0.73	1	1260
1L0Y(A:B)	1BEC	1B1Z(A)	T Cell Receptor β chain	Exotoxin A1	0.83	2	1130
1WQ1(G:R)✗	1WER	5P21	RAS activating domain	RAS	0.83	9	2910
2MTA(LH:A)	2BBK(LH)	1AAN	Methylamine dehydrogenase	Amicyanin	0.34	0	1460
2PCC(A:B)✗	1CCA	1YCC	Cytochrome C Peroxidase	Iso-1-Cytochrome C	0.44	1	1140
Unbound-bound (6)							
1A0O(A:B)	1CHN	1A0O(B)	Che A	Che Y	1.59	9	1130
1ATN(A:D)	1ATN(A)	3DNI	Actin	Deoxyribonuclease I	0.31	0	1770
1GLA(G:F)	1GLA(G)	1F3G	Glycerol kinase	GSF III	0.37	0	1300



1IGC(LH:A)	1IGC(LH)	1IGD	IgG1 Fab Fragment	Protein G	0.74	1	1330
1SPB(S:P)	1SUP	1SPB(P)	Subtilisin	Subtilisin prosegment	0.35	0	2230
2BTF(A:P)	2BTF(A)	1PNE	β –Actin	Profilin	0.29	0	2060
Difficult Test Cases (7)							
Unbound-unbound (5)							
1BTH(LH:P)	2HNT(LCEF)	6PTI	Thrombin mutant	Pancreatic trypsin inhibitor	1.91	18	2370
1FIN(A:B)	1HCL	1VIN	CDK2 cyclin-dependant kinase 2	Cyclin	4.66	59	3400
1FQ1(B:A)	1B39(A)	1FPZ(F)	CDK2	KAP	3.55	23	1830
1GOT(A:BG)	1TAG	1TBG(AE)	Transducin Gt-α, Gi-α chimera	Gt-β-γ	2.45	30	2500
1KKL(AC:H)	1JB1	1SPH(A)	HPr Kinase	Phosphocarrier Protein Hpr	2.53	28	1640
Unbound-bound (2)							
1EFU*(A:B)	1D8T(A)	1EFU(B)	E. coli Ef-Tu	Efts	2.57	109	3630
3HHR*(B:A)	3HHR(B)	1HGU	Human growth hormone	Receptor	2.04	24	4150

[a] 4-letter PDB code for the crystal structures used in this study with chain IDs in parenthesis.

[b] The RMSD of the interface C_α atoms for input receptor and ligand after superposition onto the co-crystallized complex structure, calculated as in our previous work[8].

[c] Number of interface C_α atoms with RMSD larger than 2 Å between unbound and bound structures after superposition.

[d] ΔASA - change in Accessible Surface Area (ASA) upon complex formation was calculated using the program NACCESS[9].

Table C.1 The list of complex systems used in our calculations.

In Table C.1 one can find the list of all the complexes together with their unbound (or re-assembled from complex structure) receptor and ligand proteins.

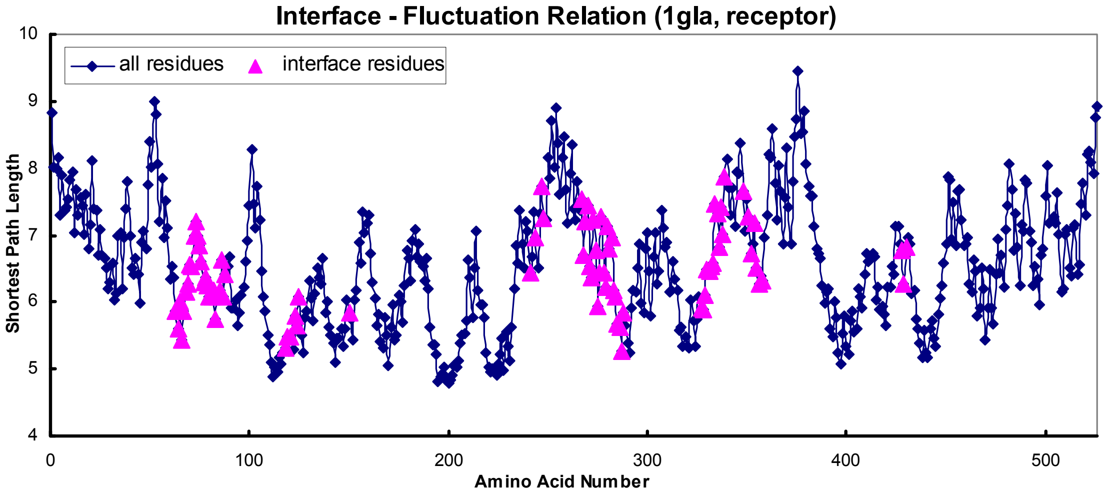
The ones with an *x* mark next to the name of the complex (also underlined) form the list of our decoy data set.

APPENDIX D: RESULTS IN DETAIL

D.1 Fluctuation and Shortest Path Relation

Previous research showed that there is a good agreement between residue fluctuations and shortest paths [Atilgan *et al.*, 2004]. It was also shown that fluctuation and protein dynamics are correlated. Therefore, we expected interface residues to have higher shortest path lengths than ordinary surface residues. However, we have observed just the opposite, in section 3.3.

The figures below (Figure D.1a and D1b) correspond the shortest path distribution of all residues and interface residues (of receptor and ligand of the complex, 1gla – Phosphotransferase) with different signs. The residues with the largest shortest paths are not necessarily the interface residues in the either receptor or the ligand.



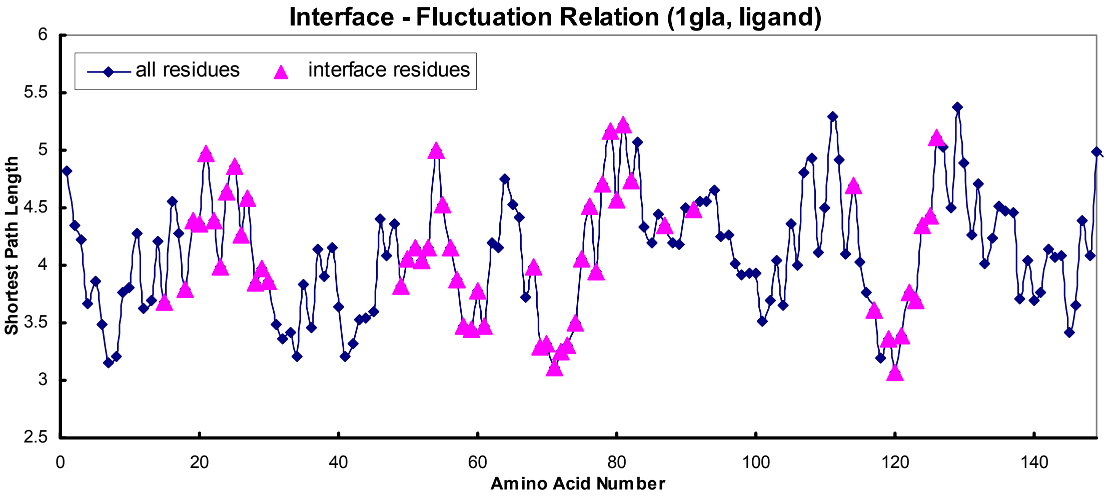
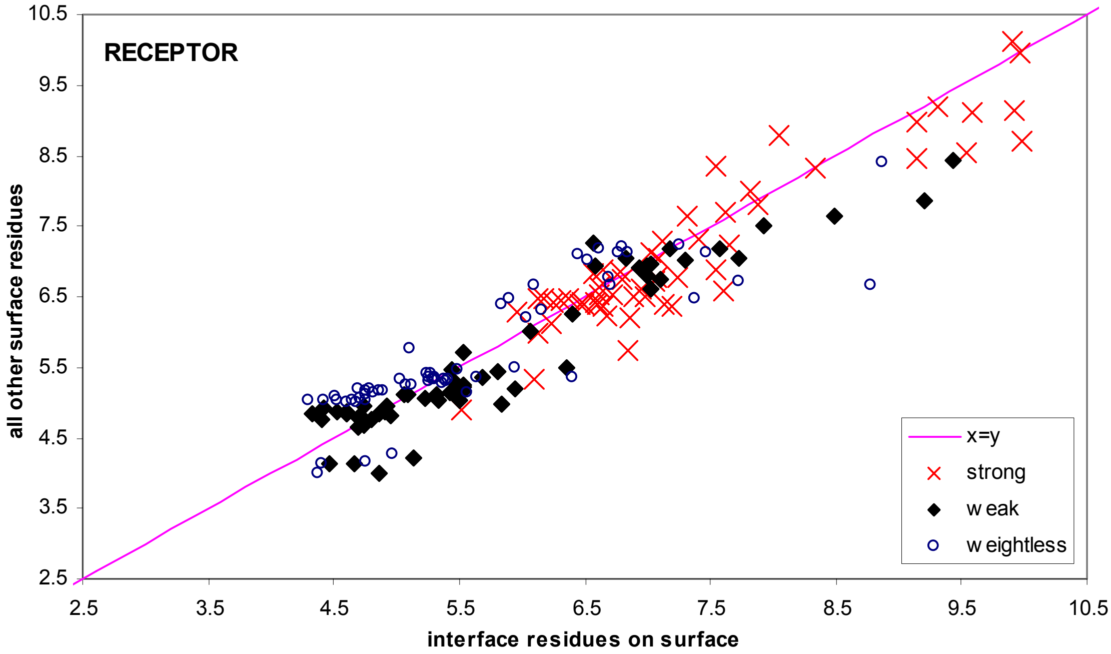


Figure D. 1 Fluctuation of inteface residues for (a) receptor, (b) ligand.

D.2 Average Shortest Paths in Receptors and Ligands Individually.

In section 3.3 it was observed that the shortest paths (L; L_{weak} and L_{strong}) best discriminate interface residues on the molten surface over all other surface residues even in the unbound form. It has been observed that the interface residues have slightly lower average shortest path values of the overall complex. How is the relation if we think of the receptors and ligands separately? Figure D.2 demonstrates this relation;



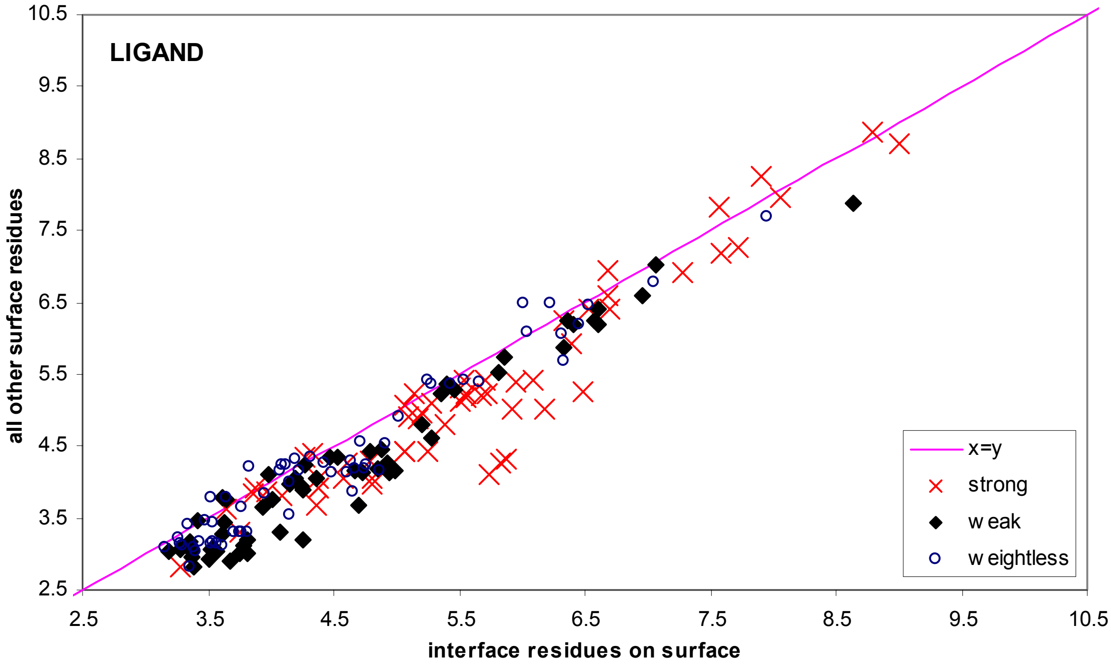


Figure D. 2 Comparison of average shortest path lengths -MJ potential set- for (a) receptors and (b) ligands.

As seen from these graphs; in ligand proteins, residues having relatively higher shortest paths are more preferably to be interface residues. On the other hand, in receptor proteins the opposite is observed. So, in ligands interface residues fluctuate more; on the contrary, in receptors interface residues fluctuate less.

D.3 Identifying Interface Residues with Clustering and Connectivity.

In section 3.3 we have concluded that shortest path discriminates between interface residues and surface residues better than other network parameters (i.e. degree, strength, clustering coefficient and weighted clustering coefficient). The comparison of the values of these parameters between interface residues on molten surface and all other surface residues can be seen in Figure D.3 and D.4.

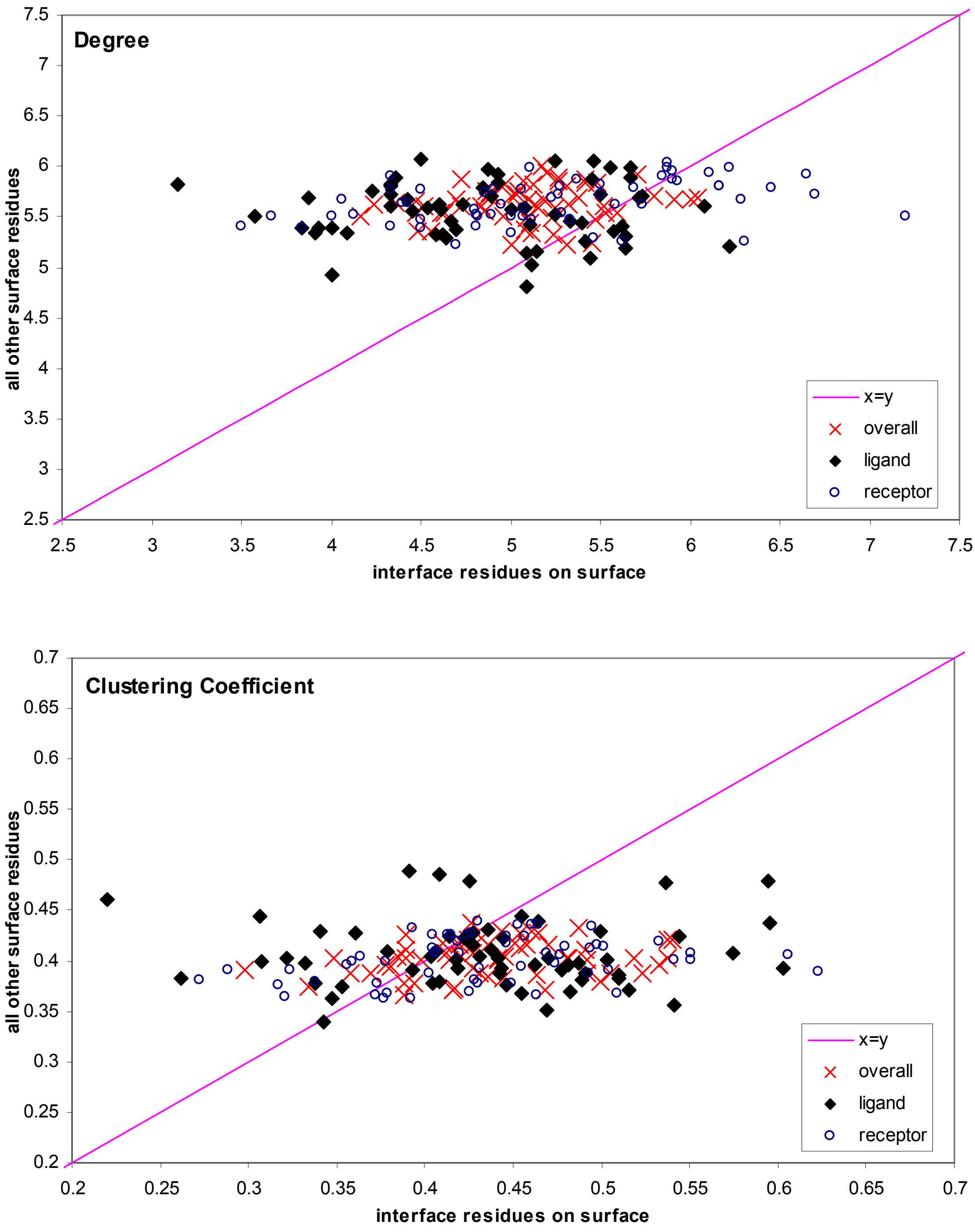


Figure D. 3 Comparison between interface residues and surface residues of (a) degree and (b) clustering coefficient.

In Figure D.3, it is observed that both degree and clustering coefficient does not provide much clue to determine interface residues. The only difference is that, surface residues have a narrow range of both degree and clustering ((≈0,3 – ≈0.5 for *C*, and ≈4.5 – ≈6 for *k*) while interface residues spread more (≈0,2 – ≈0.6 for *C*, and ≈3 – ≈7 for *k*).

When we consider the weighted averages of these two parameters (i.e. strength and weighted clustering coefficient) the results lead slightly different inferences. Figure D.4 demonstrates these differences.

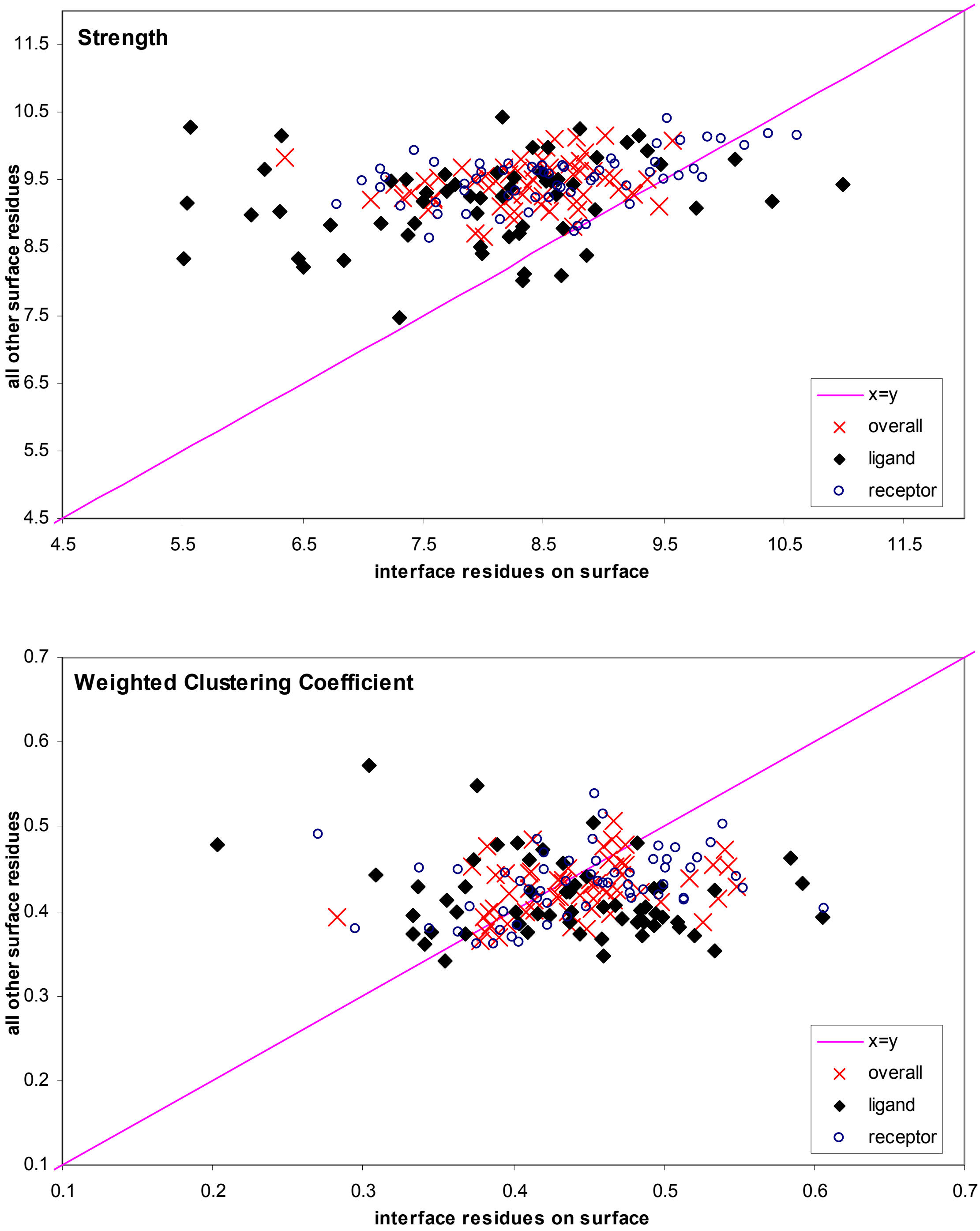


Figure D. 4 Comparison between interface residues and surface residues of (a) strength and (b) weighted clustering coefficient.

In Figure D.4, the same inference made for the clustering in Figure D.3 is valid for the weighted clustering. However, the comparison of average strengths leads better discrimination between interface residues in molten surface and all other surface residues. In all complexes (except only one, considering overall strength), interface residues have lower weighted connectivity than other surface residues.

REFERENCES

1. Abola E.E., F.C. Bernstein, S.H. Bryant, T.F. Koetzle, and J. Weng, *"Protein Data Bank" in Crystallographic Databases - Information Content, Software Systems, Scientific Applications*, eds. F.H. Allen, G. Bergerhoff and R. Sievers, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987. p. 107-132.

2. Bahar, I., B. Erman and A.R. Atilgan, *Direct Evaluation of Thermal Fluctuations in Proteins Using a Single Parameter Harmonic Potential*, Fold. Des., 1997. **2(3)**: p. 173-181.

3. Bahar, I., B. Erman, R.L. Jernigan, A.R. Atilgan and D.G. Covell, *Collective Dynamics of Hiv-1 Reverse Transcriptase: Examination of Flexibility and Enzyme Function*, J. Mol. Biol., 1999. **285**: p. 1023-1037.

4. Barabasi, A.-L, R. Albert and H Jeong, *Internet: Growth Dynamics of the World-Wide Web*, Nature, 1999. **401**: p. 130-131.

5. Barabasi, A.-L. and R. Albert, *Emergence of Scaling in Random Networks*, Science, 1999. **286**: p. 509-512

6. Barabási, A. L., H. Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vicsek, *Evolution of the social network of scientific collaborations*, Physica A, 2002. **311**: p. 590-614

7. Barrat A., M. Barthèlèmy, R. Pastor-Satorras, and A. Vespignani, *The architecture of complex weighted networks*, Proc. Natl. Acad. Sci., 2003. **101(11)**: p. 3747-3752.

8. Barthèlèmy, M. and L.A.N. Amaral, Small-World Networks:Evidence for a Crossover Picture. Phys. Rev. Lett., 1999. 82: p. 3180-3183.

9. Baysal, C. and A.R. Atilgan, *Elucidating the Structural Mechanisms for Biological Activity of the Chemokine Family*, Proteins, 2001. **43**: 150-160.

10. Baysal, C. and A.R. Atilgan, *Relaxation Kinetics and the Glassiness of Proteins: The Case of Bovine Pancreatic Trypsin Inhibitor,* Biophys. J., 2004. **83**: p. 699-705.

11. Baysal, C, P. Akan & A.R. Atilgan, *Small-World Communication of Residues and Significance for Protein Dynamics,* Biophys. J., 2004. **86**: p. 85-91.

12. Bernstein, F.C., T.F. Koetzle , G.J. Williams, E.F. Jr. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *The protein data bank: a computer based archival file for macromolecular structures.* J Mol Biol 1977. **112**: 535–542.

13. Braunstein, L.A, S.V. Buldyrev, R. Cohen, S. Havlin and H.E. Stanley, *Optimal Paths in Disordered Complex Networks*, Phys. Rev. Lett., 2003. **91**, 168701.

14. Braunstein, L.A., S. Sreenivasan, T. Kalisky, S.V. Buldyrev, S. Havlin and H.E. Stanley, *Effect of Disorder Strength on Optimal Paths in Complex Networks*, submitted to Phys. Rev. E, 2004.

15. Chakravarty, S. and Varadarajan, R., Residue Depth: A Novel Parameter for the Analysis of Protein Structure and Stability, Structure, 1999 **7**: p. 723-732.

16. Chen, R., W. Tong, J. Mintseris, L. Li and Z. Weng, *ZDOCK Predictions for the CAPRI Challenge*. Proteins, 2003. **52**: p. 68-73.

17. Goh, K.-I., B. Kahng and D. Kim, *Universal Behavior of Load Distribution in Scale-Free Networks*, Phys. Rev. Lett., 2001. **87**: 278701.

18. Guimerà, R., S. Mossa, A. Turtschi and L.A.N Amaral, Structure and Efficiency of the World-Wide Airport Network, preprint (available online at http://arxiv.org/abs/cond-mat/0312535)

19. McCann, K., A. Hatings and G. R. Huxel, *Weak Trophic Interactions and the Balance of Nature*, Nature, 1998. **395**: p. 794-798.

20. Ming, D., Y. Kong, Y. Wu and J. Ma, *Substructure Synthesis Method for Simulation Large Molecular Complexes*, Proc. Natl. Acad. Sci., 2003. **100**: p. 104-109

21. Miyazawa, S. and R.L. Jernigan, *Residue-Residue Potentials with a Favorable Contact Pair Term and an Favorable High Packing Density Term, for Simulation and Threading*. J. Mol. Biol., 1996. **256**: p. 623-644.

22. Newman, M. E. J., *The Structure of Scientific Collaboration Networks*, Proc. Natl. Acad. Sci., 2001. **98**: p. 404-409

23. Newman, M. E. J., *Scientific collaboration networks. I. Network construction and fundamental results*, Phys. Rev. E, 2001. **64**: 016131.

24. Newman, M. E. J., *Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality*, Phys. Rev. E, 2001. **64**: 016132.

25. Rosen, K. H., *Discrete Mathematics and Its Applications.* 2003, New York, NY, USA: McGrav-Hill Companies, Inc.

26. Strogatz, S.H., Exploring Complex Networks, Nature, 2001. **410**: p. 268-276.

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

28. Voet, D. and J.G. Voet, *Biochemistry (2nd Edition)*. 1995, Somerset, NJ, USA: John Wiley & Sons, Inc.

29. Watts, D.J. and S.H. Strogats, *Collective Dynamics of 'Small World' Networks*. Nature, 1998. **393**: p. 440-442.

30. Watts, D.J., Small *Worlds: the Dynamics of Networks Between Order and Randomness*. 1999, Princeton, NJ, USA: Princeton University Press.

31. Wilson, R. J., and J. J. Watkins, *Graphs: An Introductory Approach*. 1990, New York, USA: Wiley.

32. Woolhouse, M. and A. Donaldson, *Managing Foot and Mouth*, Nature, 2001. **410**: p. 515-517