MONTE CARLO (MC) PATH GENERATION AND SMALL-WORLD NETWORK APPROACH TO IDENTIFY FUNCTIONAL RESIDUES IN PROTEINS

by

Andaç Armutlulu

B.S., Chemical Engineering, Boğaziçi University, 2007

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering

Boğaziçi University

2009

ACKNOWLEDGEMENTS

All the work presented here was conducted in Polymer Research Center of Boğaziçi University.

First of all I would like to thank my thesis supervisor Prof. Türkan Haliloğlu for her unending welcome, guidance, and support throughout my studies.

I would also like to thank my committee members Prof. Kutlu Ülgen and Assist. Prof. Nevra Özer for their time and precious comments.

Many thanks to all members of the PRC family for their friendship and support throughout.

I would also like to mention my dear friend Seren Soner with whom we've shared all the troubles and joys of undergraduate and graduate education.

Very special thanks go to Şölen Ekesan for her help and support throughout my student life in Boğaziçi University.

ABSTRACT

MONTE CARLO (MC) PATH GENERATION AND SMALL-WORLD NETWORK APPROACH TO IDENTIFY FUNCTIONAL RESIDUES IN PROTEINS

Allosteric mechanism of proteins is crucial for their proper functioning. An important aspect underlying this mechanism is the communication pathways connecting the regulatory and active sites of proteins. Monte Carlo (MC) path generation is proposed as a novel methodology to identify the ensemble of these pathways and the residues taking part in the communication. This methodology is applied to determine the allosteric communication pathways in tetramerization domain of the Shaker potassium channel, PSD-95, bovine rhodopsin, and Dictyostelium myosin II. Another aspect derived from MC path generation method is generating a network of residues and finding the functional residues by applying the small-world network approach. The network parameters such as betweenness, closeness, and clustering coefficient are calculated to be used as tools for predicting the functionally important residues. In addition to the mentioned systems, this approach is also applied to the HIV-1 protease system. The suggested pathways and the residues constituting these pathways are mostly in agreement with the previous studies in literature. Also, the majority of the residues proposed as functionally important are consistent with previous studies. Furthermore, the small-world network approach is applied to the HIV-1 protease system; the network parameters are calculated for both mutant and co-evolved structures and compared with the wild-type structure. The highest correlation is observed between the wild-type and the co-evolved structures, which justifies the existence of this mutation for the conservation of network properties. Overall, the proposed computational methodology has the potential to identify the residues mediating the allosteric communication and to detect the functionally important sites in proteins.

ÖZET

PROTEİNLERDEKİ İŞLEVSEL REZİDÜLERİN BELİRLENMESİ İÇİN MONTE CARLO (MC) PATİKA YARATMA VE KÜÇÜK DÜNYA AĞ-YAPI YAKLAŞIMI

Proteinlerin alosterik mekanizması, işlevselliklerini sürdürebilmeleri için çok büyük önem taşımaktadır. Bu mekanizmanın altında yatan önemli bir yön ise proteinlerin düzenleyici ve aktif noktalarını bağlayan iletişim patikalarıdır. Monte Carlo (MC) patika yaratma, bu patika topluluklarını ve yer alan rezidüleri teşhis edebilmek amaçlı ileri sürülmüş yeni bir yöntemdir. Bu yöntem; Shaker potasyum kanalı tetramerizasyon bölgesi, PSD-95, bovine rhodopsin ve Dictyostelium myosin II sistemlerindeki alosterik iletişim patikalarının tanımlanması için uygulanmıştır. MC patika yaratma yönteminden yola çıkılarak elde edilen başka bir yöntem ise rezidü ağ-yapıları oluşturulup küçük-dünya ağ-yapı yaklaşımı uygulanmasıdır. Fonksiyonel olarak önemli rezidülerin tahmin edilmesinde kullanılmak üzere bir araç olarak aradalık, yakınlık ve kümeleme katsayısı gibi ağ-yapı parametreleri hesaplanmıştır. Bahsi geçen sistemlere ek olarak, bu yaklaşım HIV-1 proteaz sistemine de uygulanmıştır. Önerilen patikalar ve bu patikalarda yer alan rezidüler çoğunlukla literatürdeki önceki çalışmalarla uyum göstermektedir. Önerilen fonksiyonel olarak önemli rezidülerin çoğunluğu da önceki çalışmalarla tutarlıdır. İlave olarak, küçük-dünya ağ-yapı yaklaşımı HIV-1 proteaz sistemine uygulanmış, ağ-yapı parametreleri hem mutant hem eşevrim yapıları için hesaplanmış ve yaban tipi yapısıyla karşılaştırılmıştır. Değerler arasındaki en yüksek korelasyon yaban tipi ve eşevrim yapılarıarasında görülmüştür ve bu da bu mutasyonun ağ-yapı parametrelerini koruma amaçlı oluştuğunu göstermektedir. Sonuç olarak, ileri sürülen hesapsal yöntem potansiyel olarak alosterik iletişimi sağlayan ve proteinlerdeki fonksiyonel olarak önemli rezidüleri teşhis edebilmektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

b_i	Betweenness of residues i
C_i	Clustering coefficient of residue i
E	Energy
E_{min}	Minimum energy
E_i	Number of edges between neighbors of i
g_{ij}	Number of different shortest paths between i and j
G	Graph
k_i	Number of neighbors of residue i
kT	Boltzman temperature
l	Step length of shortest paths
M	Total number of atoms in a protein
N	Number of vertices in a network
N	Total number of residues in a protein
N_{ij}	Total number of atomistic contacts between i and j
O_i	Closeness of residue i
P_{ij}	Probability of interaction of i and j
r	Distance
r_{cut}	Cut-off radius
r_{min}	Radius corresponding to the minimum energy
V	Number of vertices in a graph
W_{ij}	Weight of interaction of i and j
ε	Well depth
σ	Collision diameter
Arg	R, Arginine
Asn	N, Asparagine
Asp	D, Aspartic acid
ATD	Anisotropic Thermal Diffusion
BB	B, Backbone

Cys	C, Cysteine
Gln	Q, Glutamine
Glu	E, Glutamic acid
Gly	G, Glycine
His	H, Histidine
HIV	Human Immunodeficiency Virus
Ile	I, Isoleucine
Leu	L, Leucine
LJ	Leonard-Jones
Lys	K, Lysine
MC	Monte Carlo
MD	Molecular Dynamics
Met	M, Methionine
PDB	Protein Data Bank
Phe	F, Phenylalanine
Pro	P, Proline
Ser	S, Serine
Thr	T, Threonine
Trp	W, Tryptophan
Tyr	Y, Tyrosine
Val	V, Valine
WT	Wild-type

1. INTRODUCTION

Dynamic properties of proteins play a crucial role in their activities. Proteins undergo conformational changes which are essential for their functions. An important aspect of the protein dynamics is called "allostery" which is a biological mechanism regulating the activity. Proteins whose binding affinity or catalytic efficiency is modulated by perturbations such as ligand binding or chemical modification at distal sites are defined as "allosteric" (Clarkson *et al.*, 2006). Being an intrinsic property of proteins, allostery is intramolecular signaling between distant sites in a protein adjusting the function and/or flexibility of a protein triggered by the binding of a ligand or another protein, termed an effector, at a site called allosteric site located away from the active site (Goodey and Benkovic, 2008). Effectors are named either allosteric activators or allosteric inhibitors that respectively enhance or inhibit function. All proteins except the fibrous ones with stable conformations are considered as potentially allosteric proteins(Gunasekaran *et al.*, 2004). Since it is a difficult task to uncover the allosteric signaling mechanism experimentally, various computational methods have been developed to elucidate it. Recently, terms such as "allosteric pathways" and "shortest path" have been introduced to explain the signaling mechanism (Lockless and Ranganathan, 1999, Atilgan *et al.*, 2007). Identification of the allosteric pathways and residues mediating allosteric communication is crucial for understanding the functional mechanism as well as the determinants for malfunctioning of a protein.

There have been several studies attempting to identify the allosteric pathways in various protein structures. One of the approaches developed finds the pathway of energetic connectivity on POZ and PDZ domain using the evolutionary data (Lockless and Ranganathan, 1999). That is to suggest a shortest pathway based on the statistical interactions between amino acid positions, i.e. through extracting the evolutionarily-conserved thermodynamic couplings between residue pairs (Süel *et al.*, 2002). Furthermore, a novel anisotropic thermal diffusion method where the heat in the form of kinetic energy propagates from the temperature-coupled residues to uncoupled residues in an anisotropic way was proposed to track the intramolecular signaling pathway in

PSD-95, a member of PDZ domain protein family (Ota and Agard, 2005). In another approach to suggest the allosteric pathways in proteins, weights have been assigned to the residue pairs using a form of knowledge-based potentials and a heterogeneous network of residues has been generated. With the help of Dijktra's algorithm, optimal paths between residues are obtained(Atilgan *et al.*, 2007). In a recent study on myosin, allosteric communication pathways are proposed for three different conformations of myosin connecting the ATP binding site to the lever arm by using a combination of distance constraints between residues and evolutionary data (Tang *et al.*, 2007). In another recent study, intramolecular signaling pathways have been derived by using a systematic computational method based on the Markov process (Chennubhotla and Bahar, 2006).

In addition to the residues mediating the allosteric communication, there are also other functionally important residues such as active sites, binding sites, etc. in a protein structure. In fact, it is highly possible that the latter residues participate in the allosteric interaction networks of residues. Small-world network approach is one of the efficient methods for determining these residues. Network parameters like betweenness, closeness, and clustering coefficient are utilized to achieve this goal.

In previous studies, residues with high betweenness and closeness values are shown to be functionally important ones. In one of them, the residues that play key role in protein folding are found using small-world network approach (Vendruscolo *et al.*, 2002), whereas in other one actives sites are determined (Amitai *et al.*, 2004, Thibert *et al.*, 2005). Also for finding residues regulating allosteric communication (Daily *et al.*, 2008, del Sol and O'Meara, 2005) and key residues in protein-protein interactions (del Sol *et al.*, 2006), this approach has shown to be quite useful.

In the present study, as an alternative to the latter approaches where a large number of homologous proteins, dynamic simulation of proteins, and various graph search algorithms are required, a path generation method based on the Monte Carlo (MC) sampling using a modeled atomistic potential is proposed. This way, also the combinatorial explosion in larger systems caused by trying to generate all possible pathways

can be avoided. The premise here is that instead of a single pathway of interest, an ensemble of pathways could be of significance in the allosteric communication. The pathways generated by MC path generation method are used for the calculation of the previously mentioned network parameters as well.

The proposed method is tested on the systems; POZ and PDZ domain representatives, rhodopsin, and three myosin conformations which have extensively been studied for the allosteric communication pathways. Network parameters are additionally applied to the HIV-1 protease system in order to perform analysis and comparison of its mutant and coevolved structures.

2. MATERIALS AND METHODS

In this section consisting of two main parts, firstly the studied protein structures are introduced briefly and then the methods applied on these structures are explained in detail.

2.1. STUDIED STRUCTURES

Five different protein structures are of interest in the present study. These are the tetramerization domain of the Shaker potassium channel as a POZ domain representative, PSD-95 as a PDZ domain representative, bovine rhodopsin, Dictyostelium myosin II, and HIV-1 protease. Basic information on these structures is available in the following part.

2.1.1. POZ Domain Representative

POZ (Poxvirus and Zinc Finger) domains play key roles in signaling by acting as organizing centers for multiprotein signaling complexes which constitute functional macromolecular units (Ranganathan and Ross, 1997). Because of this fact, these domains are believed to possess allosteric communication mechanisms.

As a representative of the POZ domain, tetramerization domain of the Shaker potassium channel (PDB ID: 1A68) is investigated in the present study whose structure, shown in (Figure 2.1), was determined by X-Ray diffraction method at 1.8 Å resolution (Kreusch *et al.*, 1998). It is a relatively small protein consisting of 87 amino acids. This structure has been studied recently with the aim of finding its allosteric communication pathways. The proposed pathways are between Phe77, located at the interaction surface, and Phe148, which takes part in the binding of other subunits of potassium channels (Lockless and Ranganathan, 1999, Atilgan *et al.*, 2007).

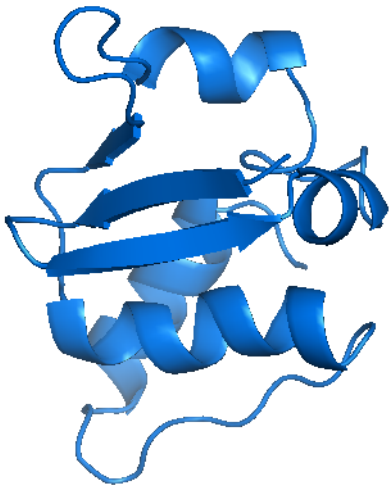

Figure 2.1. Crystal structure of the tetramerization domain of the Shaker potassium channel (1A68)

2.1.2. PDZ Domain Representative

Similar to POZ domains, PDZ (PSD-95 / DLgA / Zo-1) domains are important signaling proteins. They are defined as small globular protein-protein interaction modules recognizing terminal and internal ligands (Harris *et al.*, 2003).

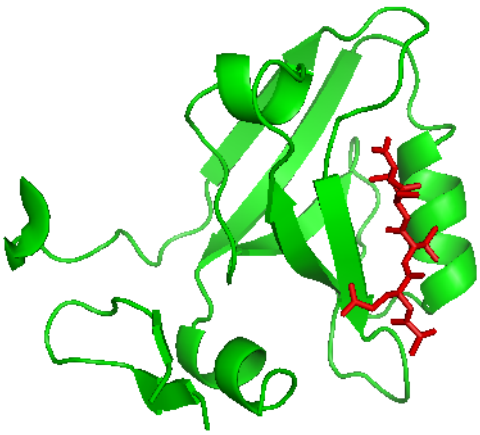

Figure 2.2. Crystal structure of PSD-95 with a C-terminal peptide (1BE9)

The PDZ domain representative of this study is illustrated in (Figure 2.2) which is the third PDZ domain from the synaptic protein PSD-95 in complex with a C-terminal peptide (PDB ID: 1BE9). Its X-ray crystallographic structure was determined at 1.8 Å resolution (Doyle *et al.*, 1996). This structure is also a small one consisting of a single chain and a peptide having 115 and 5 amino acids respectively. In the previous studies on this protein, allosteric communication pathways have been suggested connecting His372 and Leu353 (Lockless and Ranganathan, 1999, Ota and Agard, 2005). His372 has been selected as signal source since it is responsible for ligand specificity (van Ham and Hendriks, 2003) and Leu353 is located just on the opposite face of the ligand-binding pocket.

2.1.3. Myosin

Myosin is a molecular motor protein that transforms the chemical energy obtained from the hydrolysis of ATP into mechanical movement. This function of myosin is mediated by allostery. Three major myosin motor conformations, pre-stroke, rigor, and post-rigor, are identified by X-ray crystallography and biochemical experiments that correspond to discrete stages of the acto-myosin cycle (Houdusse *et al.*, 2000). In the pre-stroke conformation, ATP is hydrolyzed in the catalytic site and the lever arm is in its cocked position. Pre-stroke is followed by the rigor conformation where phosphate and ADP are released and the myosin is strongly bound to actin with its lever arm in the down position. Finally, the conformation where the catalytic stie is occupied by ATP and myosin is dissociated from actin with its lever arm still in the down position is defined as post-rigor (Tang *et al.*, 2007, Cecchini *et al.*, 2008).

In the present study, instead of the whole structure only the motor domain of myosin is investigated with the aim of finding allosteric pathways connecting the ATP binding site to the beginning of the lever arm. This domain has been previously studied and allosteric communication pathways have been proposed for its three different conformations (Tang *et al.*, 2007). The X-ray structures of pre-stroke (PDB ID: 1VOM), post-rigor (PDB IDs: 1MMA, 1MMD, and 1W9I), and rigor (PDB ID: 2AKA) states are available in the literature. In Figure 2.3, pre-stroke conformation of myosin is

Figure 2.3. Crystal structure of Dictyostelium myosin II (1VOM)

illustrated.

2.1.4. Rhodopsin

Rhodopsin is a member of G-protein coupled receptor (GPCR) family that becomes activated when it detects light. Like in the case of other GPCRs, there is an equilibrium between the activated and inactivated forms of rhodopsin. The control of this equilibrium is provided by the isomerization of retinal which is attached to rhodopsin covalently (Kong and Karplus, 2007). In the absence of light, 11-cis-retinal stabilizes rhodopsin in its inactive conformation. Once a photon is absorbed by retinal, it is isomerized from its 11-cis state to the all-trans state and thus rhodopsin becomes activated (Hubbard and Wald, 1952). In Figure 2.4, rhodopsin is illustrated.

Due to the high distance between the perturbation and target sites, an allosteric communication is claimed to be existing. As a result of this, there have been several attempts for finding allosteric pathways connecting these two sites (Kong and Karplus, 2007, Atilgan *et al.*, 2007). In the present study, ensemble of allosteric pathways is available connecting Lys296, the site of retinal isomerization, to Tyr136, a site located in a region undergoing a structural change upon light activation.

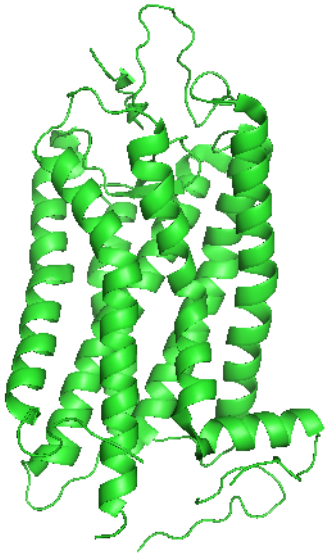

Figure 2.4. Crystal structure of the bovine rhodopsin (1HZX)

2.1.5. HIV-1 Protease

HIV-1 protease is a crucial protein for the life-cycle of HIV. Its function is to cleave the polyproteins at their appropriate sites (Ozer, 2008). It is a homodimer and each of its chains consists of 99 amino acids. An illustration of HIV-1 protease bound with a substrate is given in Figure 2.5. The active site of the protease is made up of residues Asn25, Thr26, and Gly27 which are located at the dimer interface. Among these three residues, Asn25 serves as a catalytic residue in both chains.

HIV-1 protease is capable of recognizing different non-homologous octameric substrate sites. The sequences of the structures with available crystal structures are given in Table 2.1 (Ozer, 2008). Among the given substrates, p1-p6 is investigated in this study. One coevolved and three mutant (D30N, N88D, D30N/N88D) structures are available for HIV-1 protease-substrate complex which are obtained from MD simulations performed in a recent study on substrate recognition and co-evolution in HIV-1 protease (Ozen *et al.*, 2009). The network parameters of these structures are compared with those of the wild type structure.

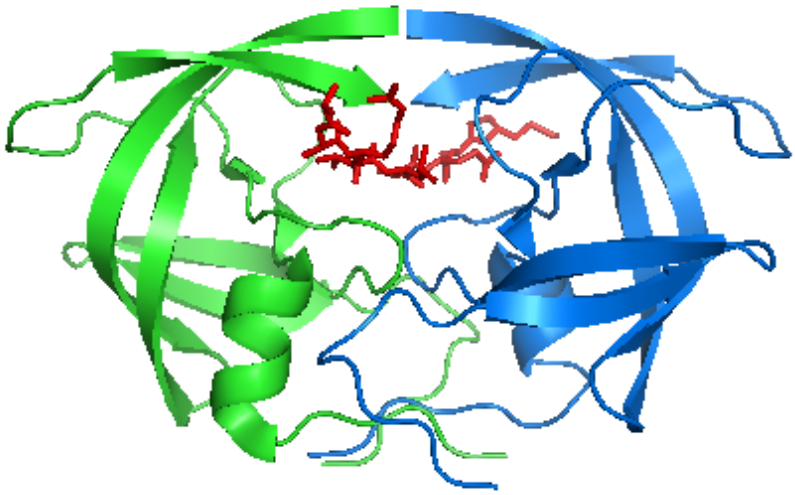

Figure 2.5. Crystal structure of HIV-1 protease complex with chain A shown in blue, chain B in green and substrate (i.e. chain P) in red

Table 2.1. Amino acid sequences of the natural substrate cleavage sites of HIV-1 protease with available crystal structures.

	P4	P3	P2	P1		P1'	P2'	P3'	P4'
capsid-p2	A	R	V	L	*	A	E	A	M
matrix-capsid	S	Q	N	Y	*	P	I	V	Q
nucleocapsid-p1	R	Q	A	N	*	F	L	G	K
p1-p6	P	G	N	F	*	L	Q	S	R
p2-nucleocapsid	A	T	I	M	*	M	Q	R	G
RNase-integrase	R	K	I	L	*	F	L	D	G
reverse transcriptase-RNaseH	A	E	T	F	*	Y	V	D	G

2.2. METHODS

2.2.1. Interresidue Interactions

Throughout the thesis, all of the results and calculations, such as pathway generation and network parameters, are based on the interactions of amino acids which are determined on atomistic scale. The attraction between each atom of two residues except hydrogen atoms is calculated by modified Lennard-Jones 12-6 potential.

$$E(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{2.1}$$

In this formula ϵ, σ and r correspond to the minimum energy between the atoms, collusion diameter of two atoms, and the distance between the two atoms respectively (Ozen, 2008). The first two parameters are obtained experimentally and each type of amino acid has its own ϵ and σ values. Repulsion part of the Lennard-Jones 12-6 potential is not included in the calculations. Therefore, a minimum radius, r_{min}, which corresponds to minimum energy value, is assigned. The value of r_{min} for each type of amino acid is calculated as follows:

$$r_{min} = \sqrt[6]{2}\sigma \tag{2.2}$$

Interaction energies of atom pairs which are in a distance of less than r_{min} to each other are set to E(r_{min}), thus the repulsion effect of the Lennard-Jones potential is disregarded. Similarly, an r_{cut} value of 5.5 Å is introduced to simplify the calculations. Hence, for pairs of atoms in a distance of more than the r_{cut} values, E(r) is set to zero. All inter-atomic interaction energies of a protein with M atoms and N residues are calculated and stored in a square energy matrix with dimensions of MxM as in Table 2.2.

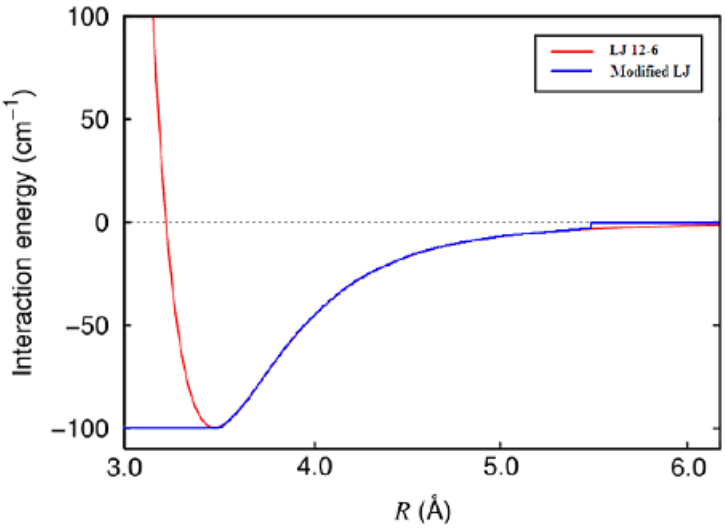

Figure 2.6. Lennard-Jones 12-6 Potential and Modified Lennard-Jones Potential

Once the atomistic interaction matrix is built up, the energy between each atom in every residue couple is added up to obtain a single energy value between those two residues. This way, the atomistic interaction matrix of size MxM is reduced to a smaller symmetric interresidue interaction matrix with dimensions of NxN as in Table 2.3 . An element E(i,j) of this matrix contains the interaction energy value of the i^{th} residue with the j^{th} one.

To sum up, all the information required to calculate inter-residual interaction energies are the x, y, z coordinates of the atoms obtained from PDB codes of proteins (Ranganathan and Ross, 1997) and experimentally obtained Lennard-Jones 12-6 potential parameters, ϵ and σ, specific for each type of amino acid.

Table 2.2. Atomistic interaction energy sample matrix

	GLU 66									**ARG 67**										...
	N	**CA**	**C**	**O**	**CB**	**CG**	**CD**	**OE1**	**OE2**	**N**	**CA**	**C**	**O**	**CB**	**CG**	**CD**	**NE**	**CZ**	**NH1**	**NH2**	**...**
N	0.00	-0.11	-0.17	-0.14	-0.17	-0.16	-0.09	-0.06	0.00	-0.24	-0.10	-0.06	-0.04	-0.06	0.00	0.00	0.00	0.00	0.00	0.00	
CA	-0.11	0.00	-0.08	-0.09	-0.07	-0.07	-0.08	-0.18	-0.18	-0.11	-0.05	-0.07	0.00	-0.07	0.00	0.00	0.00	0.00	0.00	0.00	
C	-0.17	-0.08	0.00	-0.14	-0.12	-0.12	-0.12	-0.15	-0.28	-0.17	-0.08	-0.12	-0.11	-0.12	-0.10	0.00	0.00	0.00	0.00	0.00	
O	-0.14	-0.09	-0.14	0.00	-0.13	-0.13	-0.14	-0.10	-0.32	-0.19	-0.09	-0.14	-0.03	-0.13	-0.06	0.00	0.00	0.00	0.00	0.00	
CB	-0.17	-0.07	-0.12	-0.13	0.00	-0.11	-0.12	-0.27	-0.27	-0.17	-0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
CG	-0.16	-0.07	-0.12	-0.13	-0.11	0.00	-0.12	-0.27	-0.27	-0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
CD	-0.09	-0.08	-0.12	-0.14	-0.12	-0.12	0.00	-0.28	-0.28	-0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
OE1	-0.06	-0.18	-0.15	-0.10	-0.27	-0.27	-0.28	0.00	-0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
OE2	0.00	-0.18	-0.28	-0.32	-0.27	-0.27	-0.28	-0.65	0.00	-0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
N	-0.24	-0.11	-0.17	-0.19	-0.17	-0.15	-0.06	0.00	-0.06	0.00	-0.11	-0.17	-0.19	-0.17	-0.17	-0.07	-0.04	0.00	0.00	0.00	
CA	-0.10	-0.05	-0.08	-0.09	-0.07	0.00	0.00	0.00	0.00	-0.11	0.00	-0.08	-0.09	-0.07	-0.07	-0.07	-0.10	0.00	0.00	0.00	
C	-0.06	-0.07	-0.12	-0.14	0.00	0.00	0.00	0.00	0.00	-0.17	-0.08	0.00	-0.14	-0.12	-0.12	-0.11	0.00	0.00	0.00	0.00	
O	-0.04	0.00	-0.11	-0.03	0.00	0.00	0.00	0.00	0.00	-0.19	-0.09	-0.14	0.00	-0.13	-0.13	0.00	0.00	0.00	0.00	0.00	
CB	-0.06	-0.07	-0.12	-0.13	0.00	0.00	0.00	0.00	0.00	-0.17	-0.07	-0.12	-0.13	0.00	-0.11	-0.11	-0.17	-0.12	-0.08	-0.07	
CG	0.00	0.00	-0.10	-0.06	0.00	0.00	0.00	0.00	0.00	-0.17	-0.07	-0.12	-0.13	-0.11	0.00	-0.11	-0.17	-0.12	-0.13	-0.08	
CD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.07	-0.07	-0.11	0.00	-0.11	-0.11	0.00	-0.17	-0.12	-0.17	-0.17	
NE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.04	-0.10	0.00	0.00	-0.17	-0.17	-0.17	0.00	-0.17	-0.24	-0.24	
CZ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.12	-0.12	-0.12	-0.17	0.00	-0.17	-0.17	
NH1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.08	-0.13	-0.17	-0.24	-0.17	0.00	-0.24	
NH2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.07	-0.08	-0.17	-0.24	-0.17	-0.24	0.00	
⋮																					

Table 2.3. Interresidue interaction energy sample matrix

	GLU 66	**ARG 67**	**...**
GLU 66	0	-2.65	
ARG 67	-2.65	0	
⋮			

2.2.2. Monte Carlo Path Generation

From the interresidue energy matrix, generated using atomistic potential, a statistical weight matrix is formed by taking Boltzmann weights of the energy values.

$$W_{ij} = \exp\left(\frac{-E_{ij}}{kT}\right) \tag{2.3}$$

Each row of this weight matrix is normalized to 1, thus a square probability matrix P is generated with dimensions NxN where N is the residue number. In order to exclude the effect of residue pairs which do not have atomistic contacts, the value of zero is assigned to every minimum element in a row of the probability matrix P. Then the row normalization is conducted again. After this operation an element, P(i,j), of the matrix possesses the interaction probability of the i^{th} residue with the j^{th} one calculated as.

$$P_{ij} = \frac{W_{ij}}{\sum_{j=1}^{N} W_{ij}} \tag{2.4}$$

The probability matrix obtained from atomistic potentials is used to generate pathways by Monte Carlo (MC) path generation method. In this method, firstly, a random number between 0 and 1 is generated. Starting from the i^{th} residue, every non-zero element in the i^{th} row is assigned a range. The range of the first non-zero element j is from 0 to P(i,j). The second non-zero element k has the range of P(i,j) to P(i,j) + P(i,k) whereas the third non-zero element l ranges from P(i,j) + P(i,k) to P(i,j) + P(i,k) + P(i,l). It goes in a similar manner for every non-zero element of the i^{th} row. Depending on the range to which the random number falls, the residue of first step, or in other words, the second residue in the pathway is determined. After each step, the probability matrix is regenerated by nullifying the columns of the visited residues and then normalizing every row to 1. Thus, the pathway is prevented from revisiting the previously visited residues.

The MC path generation algorithm is utilized for three different pathway generation approaches: for finding pathways between an initial and a final residue, for generating pathways of desired length beginning from an initial residue, and for gener-

ating infinitely long pathways. The latter one is different from the first two methods. The reason for defining this method as "infinitely long pathway" is that the generated pathway might be as long as a couple of hundred thousand steps. In order to obtain such a long pathway the limitation to repetition of residues is not set. This long pathway is later used for calculating network parameters whose details are given under the section "small-world network approach".

2.2.3. Analysis of the Pathways

The pathways generated by MC pathway generation algorithm are then subjected to the analysis. The analysis techniques are explained in the following part.

2.2.3.1. Pathway Frequency and Probability. For a proper pathway frequency analysis, the number of pathways to be generated should first be estimated. That is to ensure the sampling of the statistically significant number of generated pathways. Here, several runs of various numbers of pathways will be given. Then, the pathways that are sampled at the most will be considered as the most popular pathways. This is based on the premise that the lowest energy pathway means the highest probable pathway.

In proteins with low residue numbers, this method is expected to work well. In the case of proteins with high residue numbers or in the case of pathway generation attempts between highly distant residues, however, observing popular pathways may not be possible due to the high number of possible pathways. In such cases, pathways are sorted according to their probabilities. Then, the most probable, for example 20, pathways are given as the ensemble of highly probable pathways connecting the given two residues.

2.2.3.2. Residue Frequency. The visiting frequency of the residues could be of interest with functional relevance. This is in how many pathways these residues are encountered. If an MC simulation has generated, for example 50,000 pathways, then one can be sure that the frequency of both initial and final residues is 50,000. Other than

that, the average location of the residues can also be calculated which is obtained by dividing the sum of the step numbers the residue appeared in by the frequency of that residue. Obviously, this value is 1 for the initial residue since it only appears in the initial step. Similar to this information, average pathway length of the residues can be obtained by dividing the length of the pathways they appear in by the frequency of that residue. Initial and final residues have the same values as they appear in every pathway generated. As expected, the average location of the final residue has the same value as its average pathway length.

2.2.4. Small-World Network Approach

In order to calculate the network properties (clustering coefficient, closeness, betweenness), a very long pathway needs to be generated. This generation is performed by MC pathway generation method. But unlike the previous cases, this time the limitation to visiting a previously visited residue is lifted. Thus, it is possible to generate a pathway of as long as 300,000 steps. This very long pathway can be considered as a network G(V,E) of residues where the set of vertices V is the residues themselves and the set of edges E is the visiting frequencies of the residues. Throughout the thesis no distinction is made between graphs and networks, therefore the terms "graph" and "network" are used synonymously.

Due to the fact that the probability of the walk from residue i to residue j is not necessarily equal to the probability of an opposite walk, the visiting frequency f(i,j) will most likely be different from the visiting frequency f(j,i). As a result of this, the generated graph G(V,E) could be considered as a weighted directed graph

Once the weighted directed graph G(V,E) is obtained, firstly, the clustering coefficient of the graph is calculated using the following formula:

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{2.5}$$

For a vertex v_i , which has k_i neighbors, at most k_i(k_i-1)/2 edges can exist between

the neighbors if every neighbor of v_i is connected to every other neighbor of v_i. The clustering coefficient C_i is the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. C, on the other hand, is the average of C_i over all vertices.

$$C = \frac{\sum_{i=1}^{N} C_i}{N} \tag{2.6}$$

For directed graphs like in this case, clustering coefficient is defined as

$$C_i = \frac{E_i}{k_i(k_i - 1)} \tag{2.7}$$

Clustering coefficient is also a key measure of the randomness of a network. The more random a network is, the lower is its clustering coefficient (Watts and Strogatz, 1998). On the contrary, networks whose clustering coefficients are close to 1 are considered to be regular networks. The intermediate region located between the two extremes represents the small-world network, which is suggested by Watts and Strogatz as another regime of connectivity. Small-world networks consist of sub-networks in which almost every vertex is connected to another one. In this type of networks, it is possible to find short pathways between vertices. The following figure helps one to understand the network types.

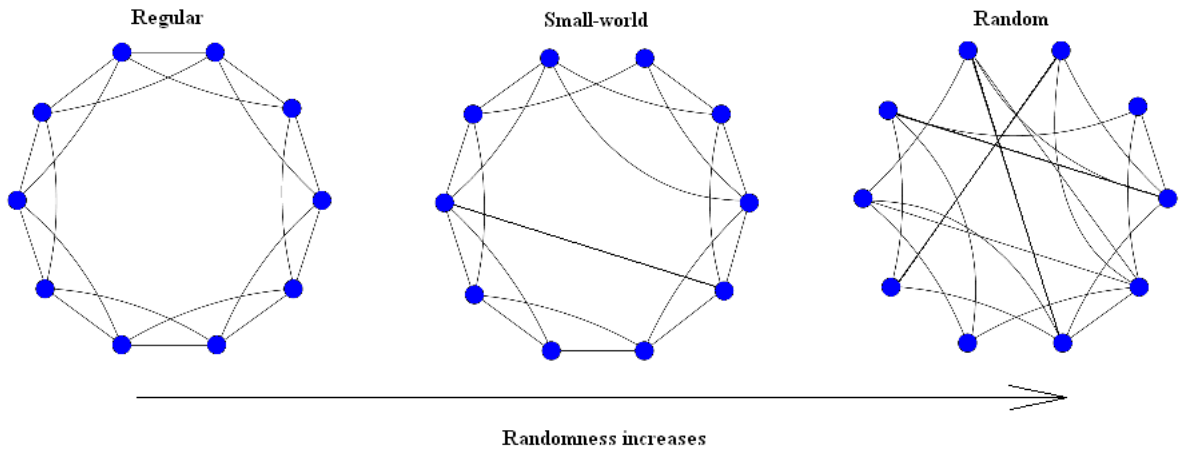

Figure 2.7. Examples of regular, small-world, and random networks

In the Results and Discussion section, it is shown that amino acids form small-

world networks.

Another important measure for networks is the closeness value. It is defined as the inverse of the average shortest pathway length between a vertex v_i and all other vertices of the network and formulated as follows: According to the formula, vertices which tend to have shorter shortest pathways to other vertices within the network are expected to possess higher closeness values.

$$O_i = \frac{N-1}{\sum_{j=i}^{N} l_{ij}} \tag{2.8}$$

The final network measure to be introduced is the betweenness value. The betweenness value for a vertex v_i is the number of shortest pathways pathways between two vertices within the network that pass through the vertex v_i normalized by the total number of pairs. Its formula is given as follows:

$$b_k = \sum_{ij} \frac{g_{ikj}}{g_{ij}} \tag{2.9}$$

Finding the shortest pathway between two vertices of a network is the most important part of calculation of the closeness and betweenness values. Various algorithms are available for this purpose, such as Dijkstra's, Johnson's and Floyd-Warshall algorithms. In the previous studies on networks, these algorithms were been used. Gray et al.(Daily *et al.*, 2008) used the Floyd-Warshall algorithm whereas Del Rio (Thibert *et al.*, 2005) used Dijkstra's Algorithm for tracing the shortest pathways between two vertices.

In this study, a novel algorithm is introduced for finding the shortest pathways. Once a very long pathway was generated by MC pathway generation method, a comprehensive search is conducted to find the between each pair of residues. The pathway which connects a pair with the least number of residues is defined as the shortest pathway. Following example help one to illustrate the procedure more easily.

A - C - B - D - E - B - C - A - D - C - A - E - B - A - C

This is a 15-residue-long pathway consisting of five different residues A, B, C, D, and E. Suppose that it is aimed to find the shortest pathway between C and E. Firstly, it is found in which steps C and E appear. For this given pathway, C appears on the 2^{nd}, 7^{th}, 10^{th}, and 15^{th} positions whereas E is available on the 5^{th} and 12^{th} ones. Two possible pathways exist for traveling from C to E. First one is from the second position to the fifth one and the second pathway is located between 10^{th} and 12^{th} positions. Since the latter one is the shortest among the two pathways, it is defined as the shortest pathway from C to E. The shortest pathway from E to C, however, is not necessarily as long as the opposite pathway. As one can see, there is a two-step pathway from E to C located between 5^{th} and 7^{th} positions.

An important point to be paid attention is that the shortest pathway is not always the most probable one. But for the sake of algorithm's simplicity, it is assumed this way. Another important point is as follows: since MC pathway generation method is based on randomness, it is obvious that in order to obtain better results, a quite long pathway should be generated. However, the longer the pathway is the more time-consuming becomes the algorithm. That's why an optimum pathway length should be decided. To overcome this problem, two separate runs are performed and the results are compared. The length of the pathway to be generated is increased until the difference between the results of two separate runs is insignificant.

3. RESULTS AND DISCUSSION

3.1. PATHWAY ANALYSIS

In this part, pathway analysis results of the POZ and PDZ domain representatives, rhodopsin, and myosin are available.

3.1.1. Pathways of POZ Domain Representative (1A68)

The allosteric mechanism of the tetramerization domain of the Shaker potassium channel, which is a member of POZ domain family, has been examined recently by an evolutionary statistical method (Lockless and Ranganathan, 1999). As a result of this study, an allosteric pathway has been suggested connecting Phe77 with Phe148. These to residues are located at the oligomerization interface and the opposing protein surface respectively. In a later study, two quite similar but one step longer shortest pathways have been proposed between the same residues which have been obtained by the application of graph search algorithms to the network of amino acids generated by using a form of knowledge-based potentials (Atilgan *et al.*, 2007). The pathways proposed by these two studies are shown in Table 3.1

Table 3.1. Suggested pathways in the literature for the POZ domain representative

Paths					**Proposed by**
77	118	149	148		(Lockless and Ranganathan, 1999)
77	118	121	149	148	(Atilgan *et al.*, 2007)
77	122	121	149	148	(Atilgan *et al.*, 2007)

The pathways proposed by the present study are generated by MC path generation method. The ensemble of most popular 10 pathways of two separate runs is illustrated in Table 3.2. As it is seen, the pathway obtained by evolutionary statistical method and one of the pathways suggested by Atilgan et al. are among the ensemble of most popular pathways. However, the former one appears far more frequently. By checking the frequencies of the most popular pathways, it can be concluded that the

Table 3.2. Most popular 10 pathways out of 30,000 and 40,000 generated by MC path generation method (1A68)

30000 runs						
77	118	149	148			76
77	72	149	148			18
77	72	113	148			15
77	115	113	148			14
77	119	118	149	148		12
77	75	72	113	148		11
77	70	111	148			10
77	118	121	149	148		10
77	115	114	113	148		9
77	75	115	113	148		8
40000 runs						
77	118	149	148			187
77	72	149	148			38
77	72	113	148			33
77	118	121	149	148		33
77	115	113	148			28
77	119	118	149	148		23
77	75	115	113	148		22
77	75	115	114	113	148	21
77	76	71	110	111	148	21
77	70	118	149	148		20

pathway consisting of residues Phe77, Phe118, Thr149, and Phe148 is the most dominant one. Statistical analysis of the residues shown in table Table 3.3 support this fact.

The residues available in the ensemble of 30,000 pathways generated by MC path generation method are sorted according to their appearance frequencies in descending order and the most popular ten residues can be seen in Table 3.3. Residues Phe118 and Thr149 are available in the top ten. By looking at the average positions of the most popular residues, located in the third column of the table, it can be said that

Table 3.3. Statistical analysis of the residues in 30,000 pathways of the POZ domain representative

77	30000	1	37.912
148	30000	37.912	37.912
70	19290	14.132	44.778
118	19250	16.469	44.444
76	19057	13.647	44.509
71	18621	16.219	45.155
149	18152	29.132	43.164
72	17776	18.153	44.735
111	17273	22.774	45.29
75	17017	16.459	45.146

Ile70, Arg76, and Phe118 are mostly alternatives to each other since their average position values are nearly the same. Even though Ile70 is more frequently observed in the pathways compared to Phe118, there is only one pathway among the most popular ten which contains Ile70. Hence, it can be decided on that Phe118 is more important in allosteric communication compared to Ile70. In addition to that, Thr149 is similarly important because the only residue having an average position value close to Phe118 is Phe111 which is not only less popular but also not as much available among the most popular pathways.

3.1.2. Pathways of PDZ Domain Representative

As a member of PDZ domain protein family, PSD-95 (PDB ID: 1BE9) has been subject to the studies with respect to its intramolecular signaling pathways. In a recent study by Agard and Ota, an allosteric communication pathway has been observed using a non-equilibrium molecular dynamics simulation method, which is called anisotropic thermal diffusion (ATD) (Ota and Agard, 2005). In another study, an evolutionary statistical method has been used to propose a signaling pathway (Lockless and Ranganathan, 1999). His372 is known as the key residue responsible for ligand specificity in PSD-95 (van Ham and Hendriks, 2003). Hence, in both latter studies, the initial point of the proposed signaling pathway has been considered to be His372. The

destination point of these paths has been decided to be Leu353, which is located on the opposite face of the ligand-binding pocket. Experimental mutagenesis has proved the functional importance of these two residues (Lockless and Ranganathan, 1999).

Table 3.4. Suggested pathways in the literature for the PDZ domain representative

Paths						**Proposed by**
372 - A	7 - P	9 - P	325 - A	347 - A	353 - A	(Lockless and Ranganathan, 1999)
372 - A	327 - A	325 - A	341 - A	353 - A		(Ota and Agard, 2005)

As it is seen, the pathway suggested by ATD method (Ota and Agard, 2005) differs from the one proposed by the evolutionary statistical method (Lockless and Ranganathan, 1999). In the latter one, a more direct route is offered for the connection of His372 and Phe325 passing through Ile327; whereas in the former one, the signal propagation occurs through the peptide.

In the present work, the ensemble of the most popular 10 pathways from two different runs obtained by MC path generation method seems to be somehow a mixture of the paths proposed in both of these previous studies. Nearly half of the most popular pathways pass through the peptide in order to reach Leu353. Other than the residues that appear in the previously suggested pathways, some other residues, Ile338 and Ile359, appear to play key role in the signaling according to the MC path generation method. In another study, where signaling pathway generation is conducted by Markov process (Ozel, 2007), Ile338 has been proposed to be an important residue for the signaling. However, His372 and Phe325 are not identified in this latter study, although the present study identifies these residues in allosteric communication in agreement with the previous studies (Lockless and Ranganathan, 1999, Ota and Agard, 2005).

The statistical analysis of the residues sorted according to their appearance frequencies in descending order is presented in Table 3.6. Here, only the most frequent 20 residues are included.

The residues appearing in the most popular signaling pathways can also be seen among the most frequent residues, which further emphasizes the importance of these

Table 3.5. Most popular 10 pathways out of 30,000 and 40,000 generated by MC path generation method (1BE9)

30000 runs						
372 - A	7 - P	325 - A	353 - A			39
372 - A	327 - A	338 - A	353 - A			37
372 - A	327 - A	326 - A	325 - A	353 - A		27
372 - A	327 - A	325 - A	353 - A			25
372 - A	327 - A	359 - A	353 - A			21
372 - A	5 - P	6 - P	326 - A	325 - A	353 - A	19
372 - A	6 - P	7 - P	325 - A	353 - A		18
372 - A	7 - P	8 - P	9 - P	325 - A	353 - A	16
372 - A	7 - P	326 - A	325 - A	353 - A		16
372 - A	336 - A	359 - A	353 - A			16
40000 runs						
372 - A	327 - A	325 - A	353 - A			39
372 - A	327 - A	338 - A	353 - A			38
372 - A	7 - P	325 - A	353 - A			36
372 - A	336 - A	359 - A	353 - A			34
372 - A	327 - A	326 - A	325 - A	353 - A		30
372 - A	327 - A	359 - A	353 - A			27
372 - A	7 - P	8 - P	9 - P	325 - A	353 - A	25
372 - A	6 - P	339 - A	340 - A	341 - A	353 - A	22
372 - A	7 - P	326 - A	325 - A	353 - A		22
372 - A	327 - A	326 - A	340 - A	341 - A	353 - A	22

residues in signaling. The average positions of the residues with the higher frequencies of appearance indicate which residues are more likely to be visited earlier. For example, the signal propagating from His372 is expected to visit the peptide earlier than the PDZ domain since the average position values of the peptide residues are lower compared to PDZ domain residues. Only the neighboring, chemically bonded, residues of His372 possess lower values than the peptide residues. These neighboring residues are not observed in the popular pathways, hence they are not considered as functionally plausible.

Table 3.6. Statistical analysis of the residues in 30,000 pathways of the PDZ domain representative

353 - A	30000	14.042	14.042
372 - A	30000	1	14.042
327 - A	9414	4.8292	14.078
325 - A	8421	8.5305	13.978
336 - A	7412	5.8654	14.357
7 - B	7299	4.4939	14.209
328 - A	7121	4.8995	14.5
326 - A	6918	6.8848	14.157
338 - A	6882	9.4878	13.676
376 - A	6641	4.5558	15.524
375 - A	6572	4.2511	15.504
337 - A	6491	7.1944	14.46
373 - A	6470	3.2207	15.519
371 - A	6350	3.6268	15.482
6 - B	6333	4.1754	14.121
357 - A	6326	11.405	14.477
359 - A	6011	9.6037	14.344
379 - A	5905	6.8574	15.571
377 - A	5838	4.9709	15.954
5 - B	5658	3.4028	14.36

Another parameter to be taken into consideration is the average pathway lengths of the most frequent residues. These values of Ile327 and Ile338 are lower than the overall average value of 14.042. On the other hand, the peptide residues like Thr7 and Gln6 have values that are above the overall average. As a result of this, it can be expected that the pathways passing through the peptide are generally longer pathways.

3.1.3. Pathways of Rhodopsin

As a member of GPCR family, the signal-transduction mechanism of rhodopsin (PDB ID: 1HZX) was subjected to extensive studies (Kong and Karplus, 2007, Süel *et al.*, 2002, Atilgan *et al.*, 2007). In the study of Atilgan et al., two pathways have been suggested, one of them being weak and the other one strong, which are shown in Table 3.7. These pathways start from Lys296 where isomerization of the retinal occurs and extends to Tyr136 which is located in a region undergoing a structural change as result of light activation.

Table 3.7. Suggested pathways in the literature for rhodopsin

Paths										**Proposed by**
296	298	124	128	132	136					(Atilgan *et al.*, 2007)
296	293	294	297	301	305	136	128	132	136	(Atilgan *et al.*, 2007)

Similar to the case of myosin, rhodopsin has a large number of residues. In addition to that, Lys296 is located 29 A apart from Tyr136. Thus, the ensemble of most probable allosteric pathways are proposed instead of most popular ones for this case. This ensemble can be seen in Table 3.8.

Even though the pathways suggested by Atilgan et. al are not observed exactly, with the exception of Ile305, each residue of the suggested pathways is observed in the ensemble of most probable pathways.

Table 3.8. Most probable 20 pathways out of 30,000 generated by MC path generation method (1HZX)

30000 runs										
296	299	83	79	75	130	131	136			
296	299	298	302	76	75	131	136			
296	297	264	265	261	258	219	222	136		
296	294	264	262	220	224	225	226	136		
296	293	295	265	261	128	131	134	136		
296	297	264	265	261	128	129	130	134	135	136
296	295	264	259	258	128	130	133	136		
296	297	301	302	303	76	71	134	137	136	
296	297	298	265	262	216	217	222	136		
296	297	264	260	258	257	254	132	136		
296	299	83	55	80	79	78	75	131	136	
296	293	43	294	264	262	220	222	136		
296	297	298	301	261	216	217	221	222	136	
296	298	301	302	257	253	254	131	133	136	
296	294	264	262	261	258	219	222	136		
296	295	264	263	262	220	224	225	227	226	136
296	293	294	263	261	125	127	128	132	136	
296	295	265	261	262	220	222	223	225	226	136
296	295	265	262	258	128	131	130	133	136	
296	295	264	263	260	257	258	219	222	136	

3.1.4. Pathways of Myosin

In a recent study on allosteric communication in myosin, the three different conformations of myosin, pre-stroke, post-rigor, and rigor, have been investigated (Tang *et al.*, 2007). A group of pathways were predicted which are believed to couple ATP hydrolysis site to the lever arm recovery stroke. Several different types of myosin, such as Dictyostelium, scallop, and chicken myosin II, were analyzed in the latter study. Here, only the results of Dictyostelium myosin II are shown here since the pathways of all three conformations are available only for this type of myosin.

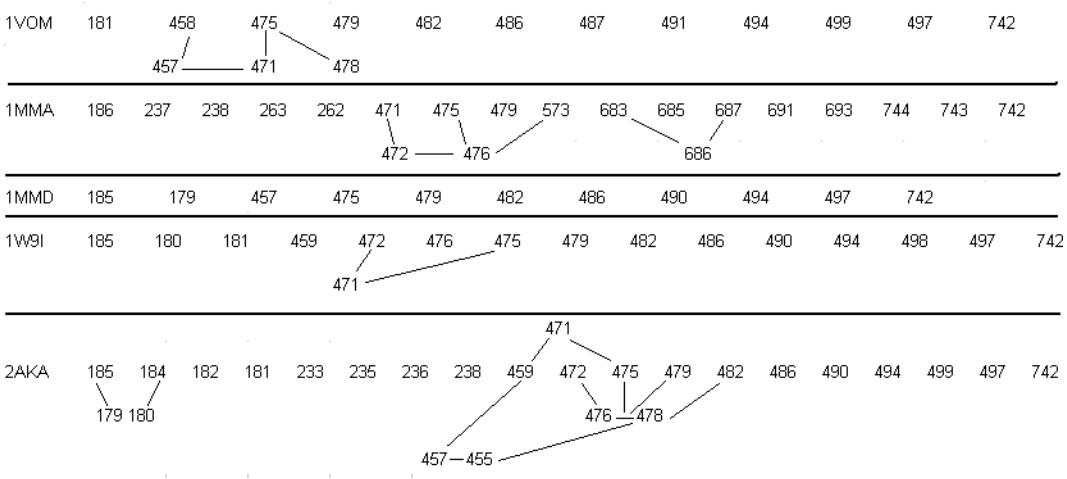

Figure 3.1. Suggested pathways in the literature for myosin

The source of signaling was postulated to be the ATP catalytic site (Tang *et al.*, 2007). Therefore, for the pre-stroke conformation (PDB ID: 1VOM), the initial residue of the pathway was determined to be Ser181 whereas in the case of post-rigor (PDB IDs: 1MMD, 1W9I) and rigor conformations (PDB ID: 2AKA) Lys185 has been chosen. It is assumed that the signal propagates from the ATP catalytic site. But for all the three conformations Thr742 has been selected as the endpoint of the pathways that is located at the beginning of the lever arm.

It is seen that the allosteric communication is maintained by similar pathways in these three different conformations. With the exception of 1MMA, almost the same residues mediate the allosteric communications.

Unlike the POZ and PDZ domain representatives, it is not possible to show an ensemble of most popular pathways in myosin. This is: firstly, the distance between the initial and final points of the pathway is very large and secondly the number of residues in myosin is far too high compared to residue numbers of potassium channel and PSD-95. The number of possible communication pathways increases excessively with the number of residues accessible for the path generation. As a result, instead of the popular pathways, an ensemble of most probable pathways is demonstrated. These pathways can be observed in the Table A.4. Residues which are available in the ensemble of most probable pathways of each protein are shown in bold. Here, only the pathways of pre-stroke conformation (PDB ID: 1VOM) are given. The rest is available in the Appendix.

Although these pathways are not totally identical to the ones in the previous study, a very similar pattern can be seen in these paths too. For instance, just like in the previous study, the pathways of 1MMA are quite different from the pathways of other proteins. Not only are the pathways of the 1MMA longer than the other proteins, but also is the allosteric communication mediated by the residues that do not appear in other proteins' pathways. Excluding 1MMA, it can be concluded that even though the conformation of myosin changes, the residues playing a key role in intramolecular signaling remain mostly the same. This can also be observed by the fact that the residues given in bold make up the great majority of all residues. It can thus be concluded that even though the protein undergoes conformation changes, its allosteric pathways are mostly conserved.

Table 3.9. Most probable 20 pathways out of 50,000 generated by MC path generation method (1VOM)

50000 runs													
181	179	**178**	**652**	**486**	**490**	**695**	**696**	**742**					
181	458	475	478	**481**	**485**	**490**	**695**	**743**	**742**				
181	180	179	**178**	**652**	**482**	**486**	487	**488**	**492**	**497**	**742**		
181	180	675	680	487	**491**	**492**	**496**	**497**	**741**	**742**			
181	180	675	680	487	**488**	**489**	**492**	**496**	**497**	**741**	**742**		
181	180	183	675	680	487	**492**	**497**	**742**					
181	179	**178**	**652**	**482**	**486**	487	**490**	**494**	**497**	**743**	**742**		
181	180	**178**	679	680	487	**490**	**494**	**497**	**742**				
181	183	**655**	**654**	**482**	**486**	487	**491**	**496**	**497**	**741**	**742**		
181	179	**178**	**654**	**482**	**485**	**490**	**494**	498	**497**	**742**			
181	179	**178**	**654**	**652**	**486**	**488**	**492**	**497**	**742**				
181	180	675	677	683	506	505	504	503	502	501	**494**	**497**	**742**
181	457	475	479	483	**488**	**492**	**495**	**496**	**497**	**741**	**742**		
181	180	675	681	686	690	691	692	693	**695**	**696**	**742**		
181	458	573	476	479	**482**	483	**488**	**492**	**497**	**742**			
181	183	**178**	**652**	**486**	**490**	**492**	**497**	**742**					
181	458	573	476	479	483	**484**	**488**	**489**	**492**	**497**	**742**		
181	179	**178**	**654**	**482**	**486**	487	**491**	**495**	498	**497**	**741**	**742**	
181	180	675	681	680	683	506	504	503	502	501	**495**	**497**	**742**
181	182	183	**655**	**654**	**652**	**486**	**490**	**695**	**743**	**742**			

3.2. NETWORK ANALYSIS

3.2.1. Network Analysis for the POZ Domain Representative

Figure 3.2, Figure 3.3, and Figure 3.4 respectively show the betweenness, closeness, and clustering coefficient values of the Shaker potassium channel (PDB ID: 1A68). The residues observed in the ensemble of most popular pathways generated by MC path generation method are also presented in red in these figures.

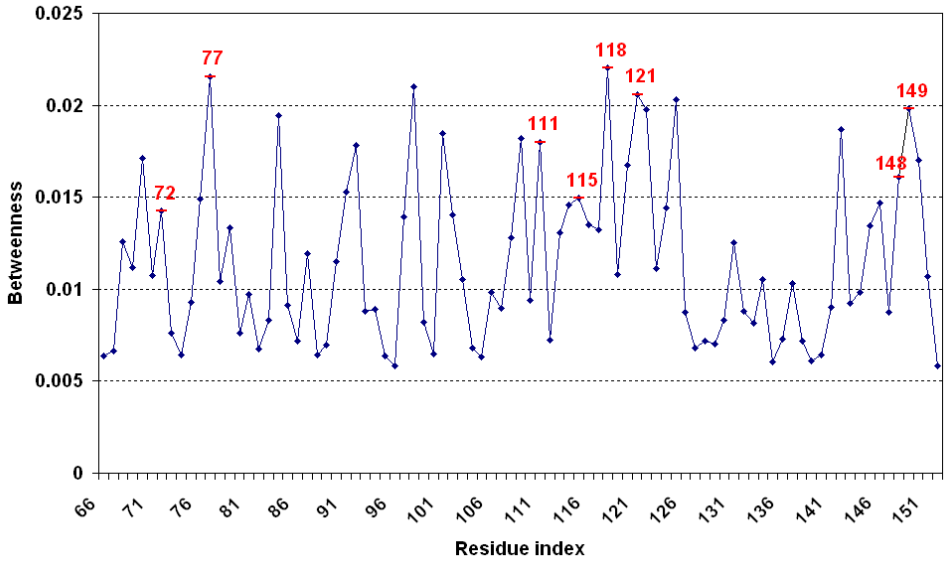

Figure 3.2. Betweenness values for the POZ domain representative

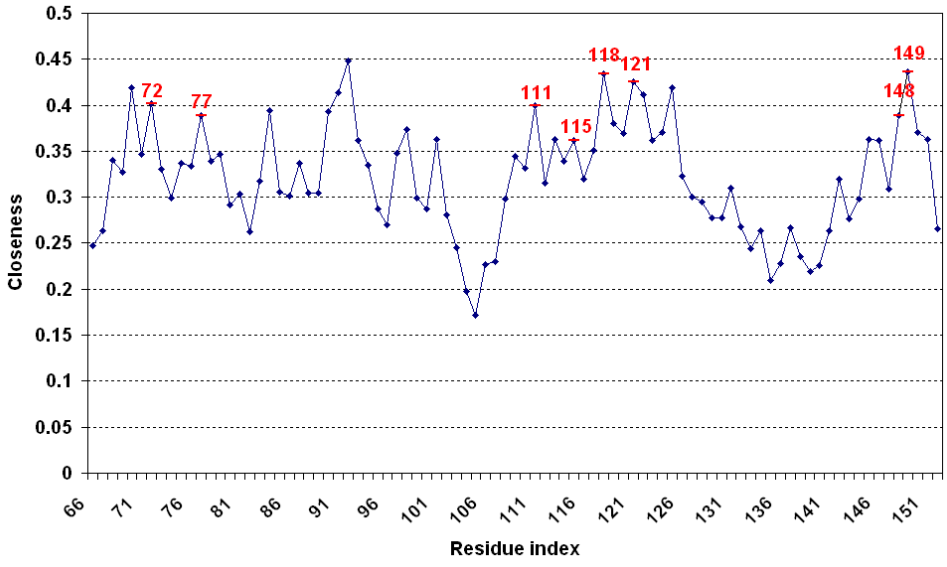

Figure 3.3. Closeness values for the POZ domain representative

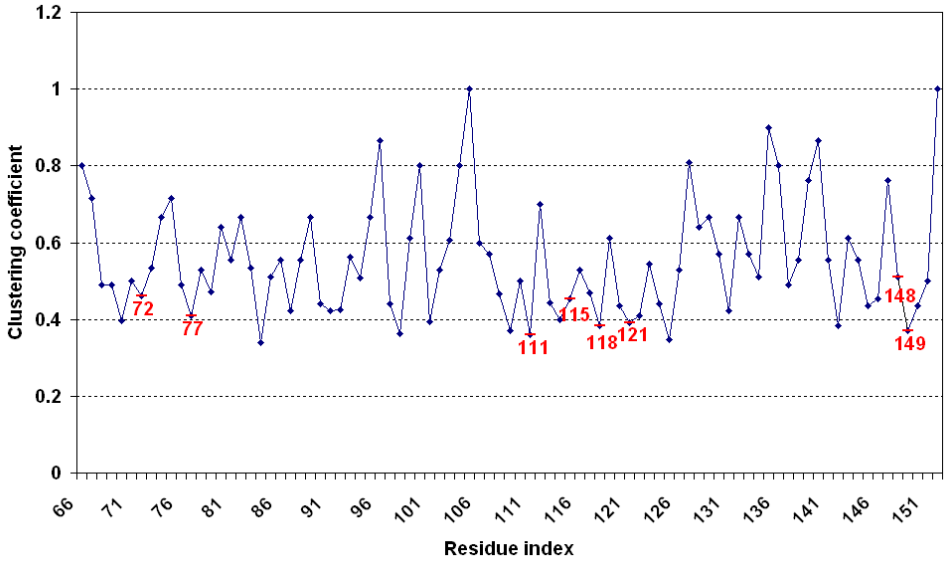

Figure 3.4. Clustering coefficient values for the POZ domain representative

As expected, functionally important residues which are in this case the ones mediating the allosteric communication, appear as local maximum points on the figures of betweenness and closeness values whereas they are observed in the local minima on the figure of clustering coefficient values. The importance of these residues comes from their existence in the ensemble of the most popular 10 allosteric pathways given in the previous section of the results. Especially the residues Phe77, Phe118, Tyr149, and Phe148 which form the most popular pathway in the ensemble have higher betweenness and closeness values than the other labeled residues. Thus, the significance of the most popular pathway is emphasized which has been proposed in a previous study as well (Lockless and Ranganathan, 1999).

Among the unlabeled residues; Ile70, Leu84, Leu92, and Tyr125 might be proposed as functionally important residues since they appear in the extreme points of all figures. Even though Arg98, Tyr109, and Phe142 happen to have very high betweenness values, their closeness values are not similarly extreme. However, these residues might be functionally important as well.

3.2.2. Network Analysis for the PDZ Domain Representative

Network parameters such as betweenness, closeness, and clustering coefficient calculated for PSD-95 can be seen in Figure 3.5, Figure 3.6, and Figure 3.7 respectively, on which functionally important residues are shown in red. These important residues include active sites and sites having contact with a ligand whose information are obtained from PDBsum (Doyle *et al.*, 1996) and those playing key role in intramolecular signaling which are observed both in the ensemble of popular paths generated by MC path generation method and in the recent studies on signaling pathways of PSD-95 (Lockless and Ranganathan, 1999, Ota and Agard, 2005).

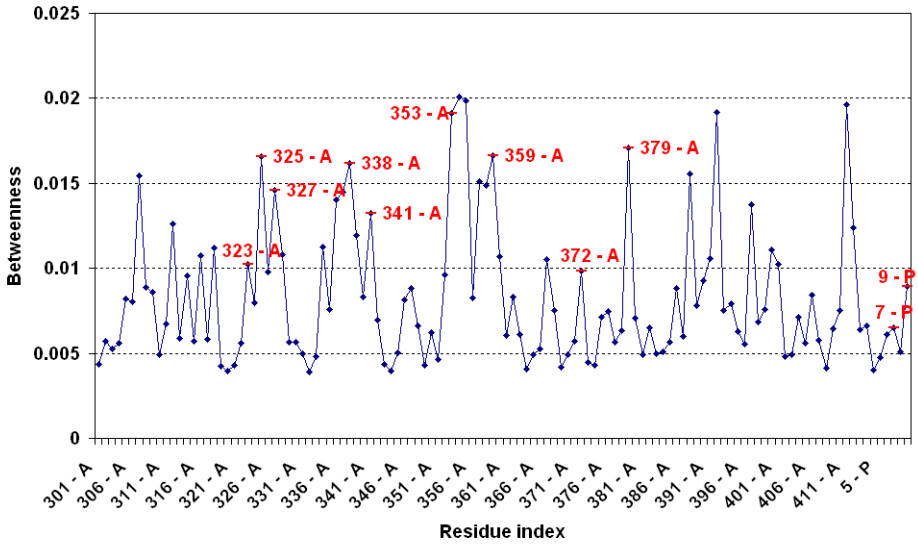

Figure 3.5. Betweenness values for the PDZ domain representative

In the figures of betweenness and closeness values, functionally important residues correspond to maximum points whereas in the clustering coefficient figure they are located in the minima. Compared to PDZ domain residues, betweenness and closeness values of the peptide residues Thr7 and Val9 are hardly above the average. But among all 5 peptide residues, these two residues have the highest values. Hence, they are taken into consideration as well.

Residues like His372, Phe325, and Ile327 are not only allosterically important, but also they are active site residues PDBsum (Doyle *et al.*, 1996). As shown in the previous section of the results, remaining labeled residues are claimed to mediate the

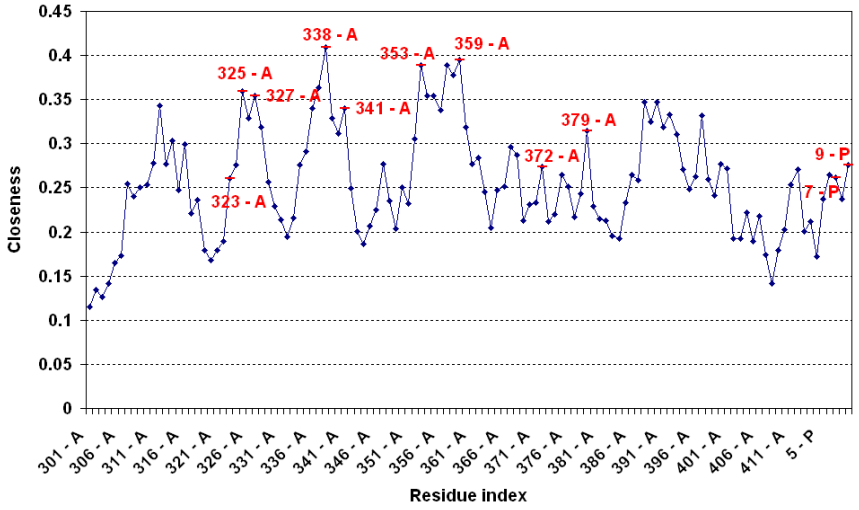

Figure 3.6. Closeness values for the PDZ domain representative

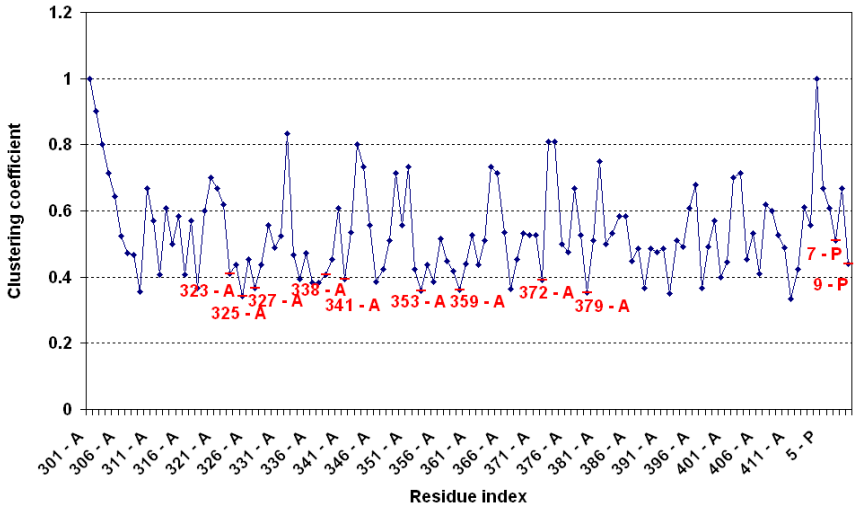

Figure 3.7. Clustering coefficient values for the PDZ domain representative

allosteric communication.

Other than the residues whose functional importance are well known, there are some which might be proposed as possible functionally key residues due to their network parameters' extreme values. These residues include Ile307, Ile388, Tyr392, and Arg411. Their importance might be observed in protein folding or in intermolecular interactions.

3.2.3. Network Analysis for Rhodopsin

Betweenness and closeness values of rhodopsin are shown in Figure 3.8 and Figure 3.9 respectively. Clustering coefficient figure is available in Appendix. Functionally important sites of rhodopsin are labeled red on these figures. Unlike the POZ and PDZ domain representatives, residues appearing in the allosteric pathways generated by MC path generation method are not labeled here because of two reasons. Firstly, the generated pathways are much longer compared to previous cases. As a result of this, too many residues need to be labeled which causes the figure to look more crowded. Secondly, instead of most popular pathways, ensemble of most probable pathways is given for rhodopsin. Hence, all pathways in the ensemble can be similarly important.

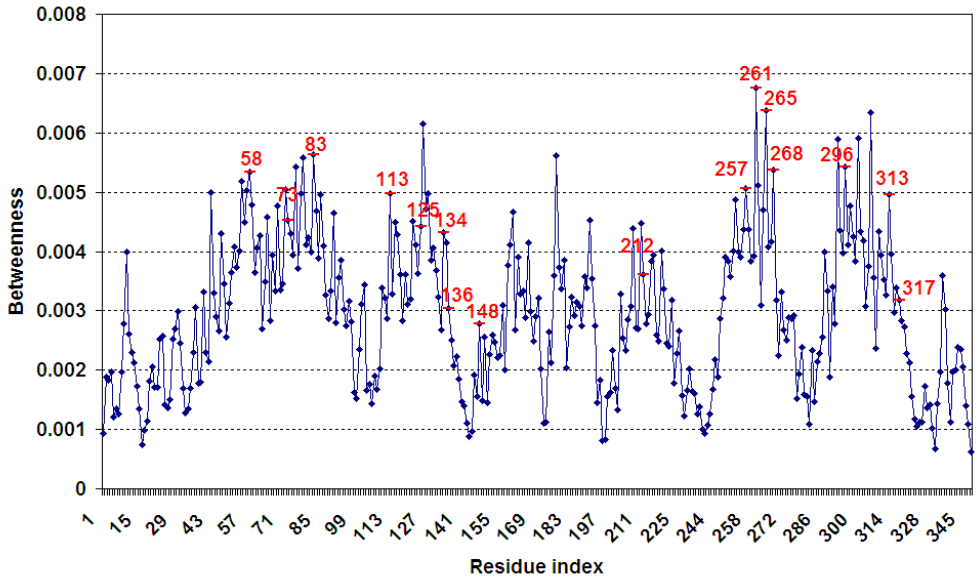

Figure 3.8. Betweenness values for rhodopsin

As it was mentioned previously, the signaling initiates from Lys296 where isomerization of the retinal takes place. Not surprisingly, Lys296 appears as local maximum on both Figure 3.8 and Figure 3.9 . In recent studies, the importance of other labeled residues have been revealed as follows: Asp83 and Glu134 have been suggested to play important role in signal-transduction mechanism (Kong and Karplus, 2007). Glu134 is claimed to be acting as signal amplifier and Asp83 is suggested to introduce a threshold to prevent background noise from activating rhodopsin. It has been found that mutations at Lys296 and Glu113 causes the allosteric control to be lost (Porter *et al.*, 1996).

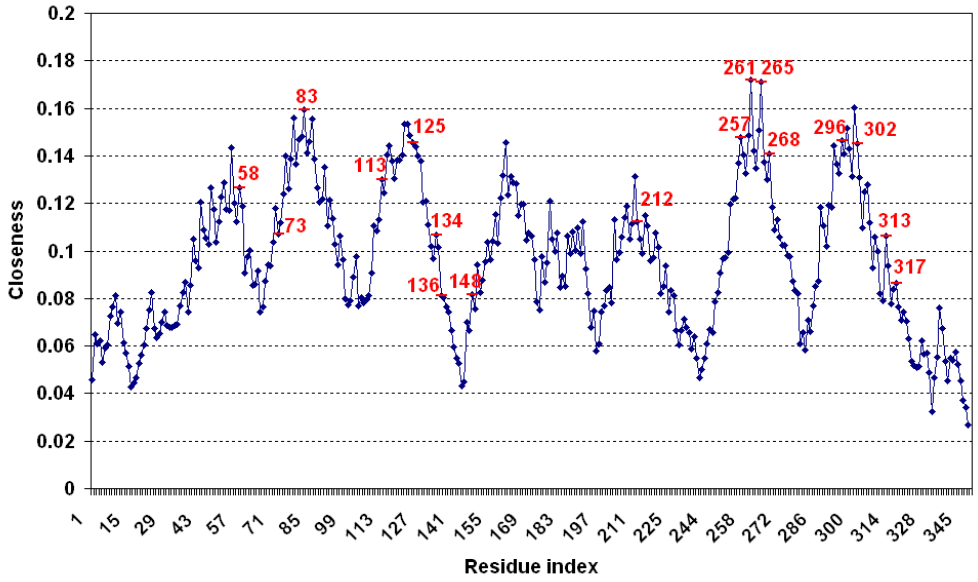

Figure 3.9. Closeness values for rhodopsin

Phe212, Phe261, Trp265, and Tyr268 form the cyclohexenyl ring of retinal which have also been determined as determinants of ligand specificity or receptor activation in several GPCR subfamilies (Gether, 2000, Ballesteros *et al.*, 2001). It is seen that Phe261 is the global maximum of both betweenness and closeness values and Trp265, along with Tyr268, are located close to Phe261. Leu125 has been found similarly important by showing that mutations of Leu125 and Phe261 lead to constitutive receptor activity (Yano *et al.*, 1997, Andres *et al.*, 2001, Garriga *et al.*, 1996). Asn302 has been suggested to be a part of an important motif in GPCRs, which has been referred to stabilizing the inactive conformation (Doyle *et al.*, 1996, Han *et al.*, 1998). Met257 has been determined to participate in constitutive activity of the protein (Han *et al.*, 1998). The remaining labeled residues form two important regions, first one including Glu134, Tyr136, and Phe148, and the second one consisting of Phe313, Met317, Thr58, and Asn73. These two regions have been claimed to containing sites that likely contact the G protein alpha subunit and causes the loss of allosteric control in the case of mutation (Menon *et al.*, 2001, Cai *et al.*, 1999).

Other than these residues, Trp161 and Tyr178 might be proposed as potential functional sites in rhodopsin since they are clear maximum points on the figures of both betweenness and closeness. They also seem to be local minima of the clustering

coefficient values. Residues like Trp126 and Tyr301 might also be proposed but their neighboring residues, Leu125 and Asn302, have already been suggested as functionally important sites.

3.2.4. Network Analysis for Myosin

Figures Figure 3.10, and Figure 3.11 show the betweenness and closeness values of the pre-stroke structure of Dictyostelium myosin II respectively. The figure of the clustering coefficient values can be found in Appendix. Since similar results are obtained, the figures of the other structures are not given separately. On these figures, functionally important sites are labeled red. Due to the same reasons mentioned in the rhodopsin case, the residues appeared in the ensemble of most probable pathways are not labeled on these figures.

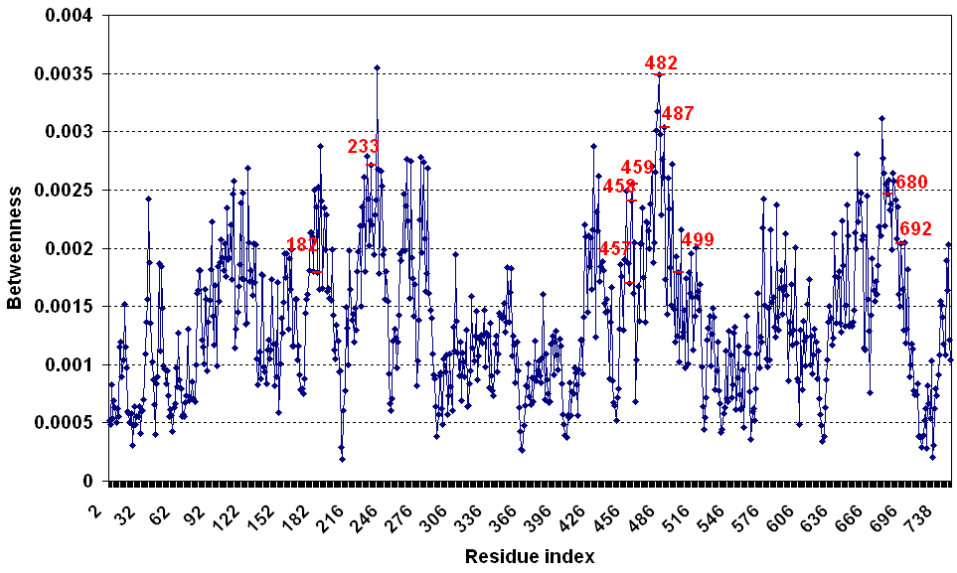

Figure 3.10. Betweenness values for myosin (1VOM)

In addition to the four catalytic residues, Gly182, Asn233, Gly457, and Glu459, there are other labeled residues whose importance is identified by mutational analysis. In one of the studies on mutational analysis, it has been shown that a mutation in Phe458 results in the loss of actin-activated ATPase activity (Sasaki *et al.*, 1998). In another study, the importance of Phe482 has been revealed by finding out that a mutation in this residue causes the hydrolysis of ATP to slow down dramatically (Ito

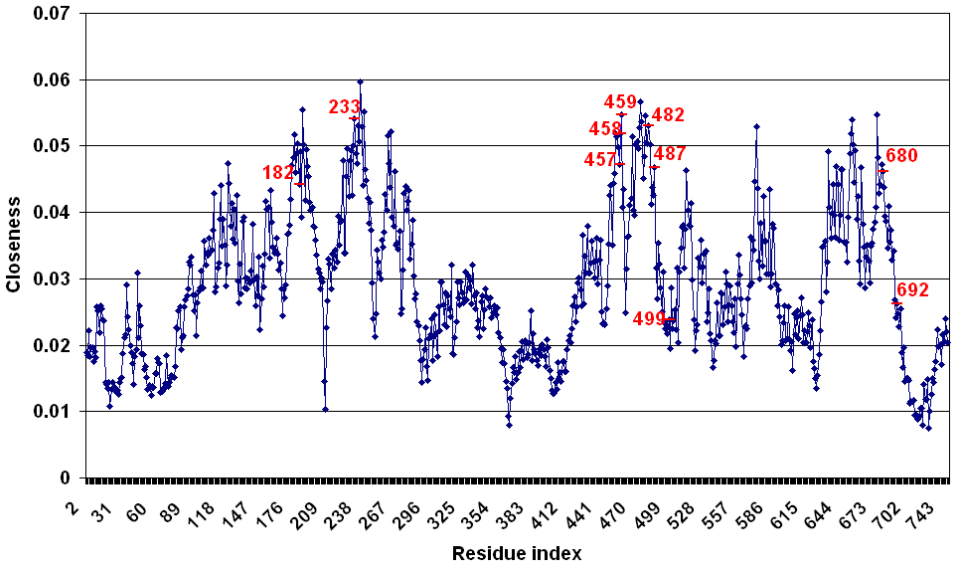

Figure 3.11. Closeness values for myosin (1VOM)

et al, 2003). Also Gly680 has been determined to be important in the same study by showing that a mutation at that site alters the Pi release rate. A mutation in Phe487 has been shown to result in disruption of the normal release from the actin filament in the post-rigor state (Tsiavaliaris *et al.*, 2002). Ile499 has been determined to be critical, since its mutation prevents the lever arm from swinging properly (Sasaki *et al.*, 2003). In the same study, Phe692 has been found to be important as well. A mutation at that site has been claimed to be uncoupling ATPase activity from the lever arm swing.

Other than the residues mentioned in the previous studies, Lys185, Arg238, Arg654, and Gln675 might be speculated to be functionally important since their betweenness and closeness values are among the highest. Therefore, they might be considered as good candidates for studies like mutational analysis. Network properties are used to compare different conformations with each other as well. In this analysis, network parameters are calculated for different conformations of Dictyostelium myosin II and they are compared as shown in figures Figure 3.12 and Figure 3.13. In order to compare different structures properly, the calculated values are normalized to 1. The post-rigor structure with the PDB ID of 1MMA is excluded due to the discrepancies observed in pathway analysis.

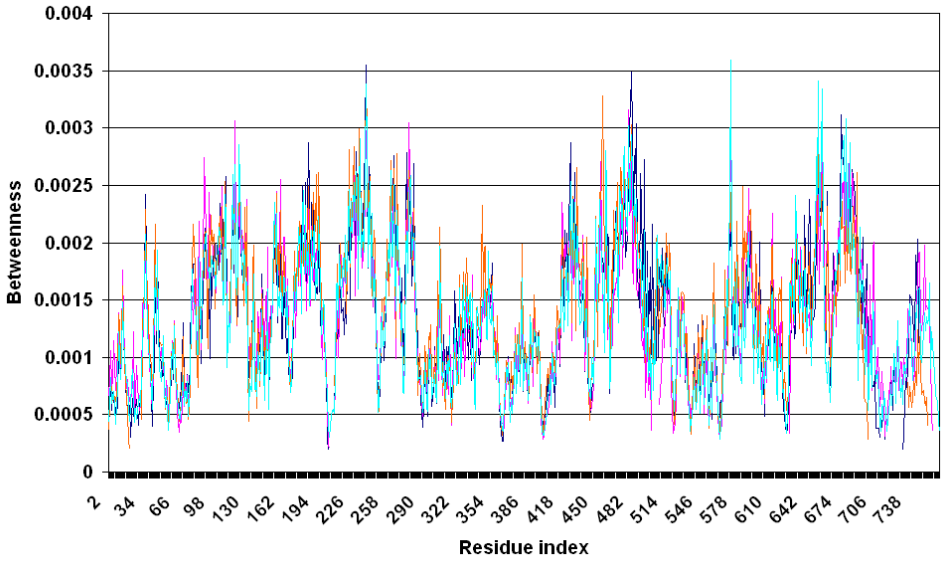

Figure 3.12. Comparison of betweenness values of four different myosin structures. (1VOM, 1MMD, 1W9I, and 2AKA are colored in blue, magenta, orange, and cyan respectively

As shown in Figure 3.12 and Figure 3.13, conformation change does not affect the overall trend very much. Several deviations might have occurred due to missing residues in some of the structures. Other than that, no significant difference is observed. Therefore, it can be proposed that the functionalities of the residues are in this case only slightly dependent on the conformation.

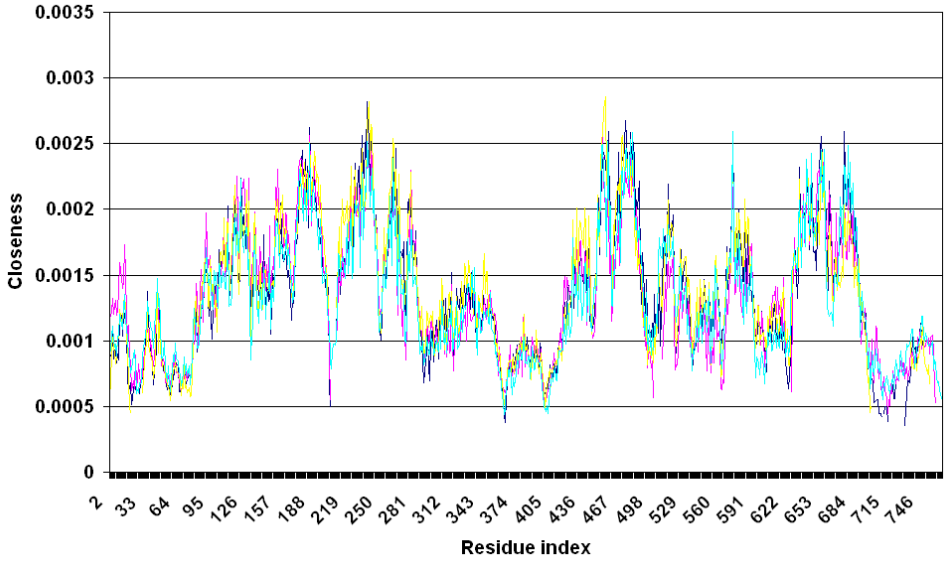

Figure 3.13. Comparison of closeness values of four different myosin structures. (1VOM, 1MMD, 1W9I, and 2AKA are colored in blue, magenta, orange, and cyan respectively

3.2.5. Network Analysis for HIV-1 Protease

HIV-1 protease-substrate complex is the final case on which small-world network approach is applied. The analyzed structures of HIV-1 protease-substrate complex, including wild type, mutant, and coevolved structures, are obtained from MD simulations carried out in a recent study (Ozen *et al.*, 2009). Figure 3.14 and Figure 3.15 show the calculated betweenness and closeness values of the wild type structure respectively. Clustering coefficient results are available in Appendix.

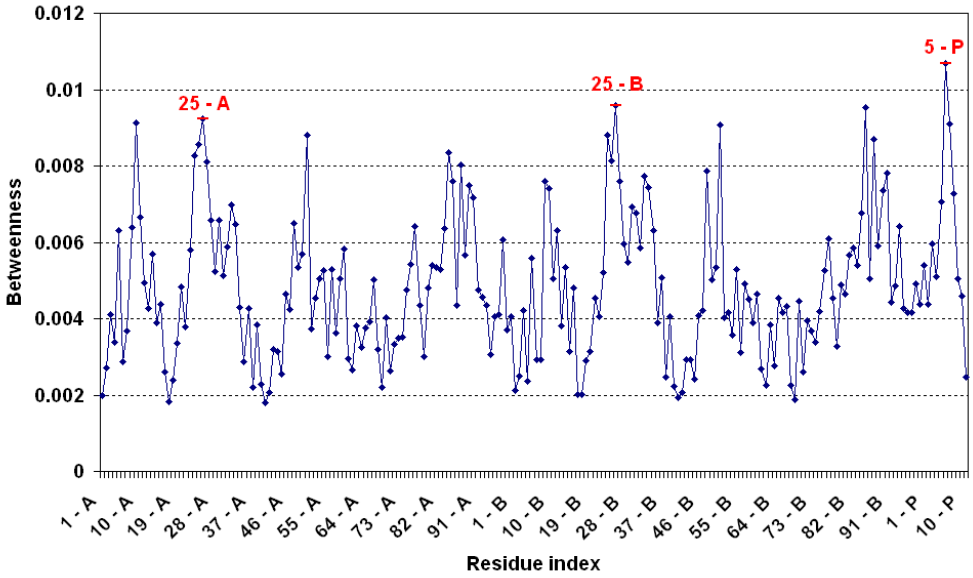

Figure 3.14. Betweennes values of HIV-1 protease

Asn25 is known to serve as the catalytic residue in both chains and it appears as the peaks of both chains in the figure. The peak of the peptide is, on the other hand, the fifth residue which is the cleavage site of the substrate. Other than these residues whose importance is well known, Pro9, Ile50, and Ile84 might be proposed as proper candidates for further studies.

The network properties of HIV-1 protease-substrate complex are also used to compare the mutant and coevolved structures to the wild type structure. The comparisons of betweenness values are available in Figure 3.16, Figure 3.17, Figure 3.18, and Figure 3.19. Similar results are obtained in all parameters, therefore, only the comparison of betweenness values are given.

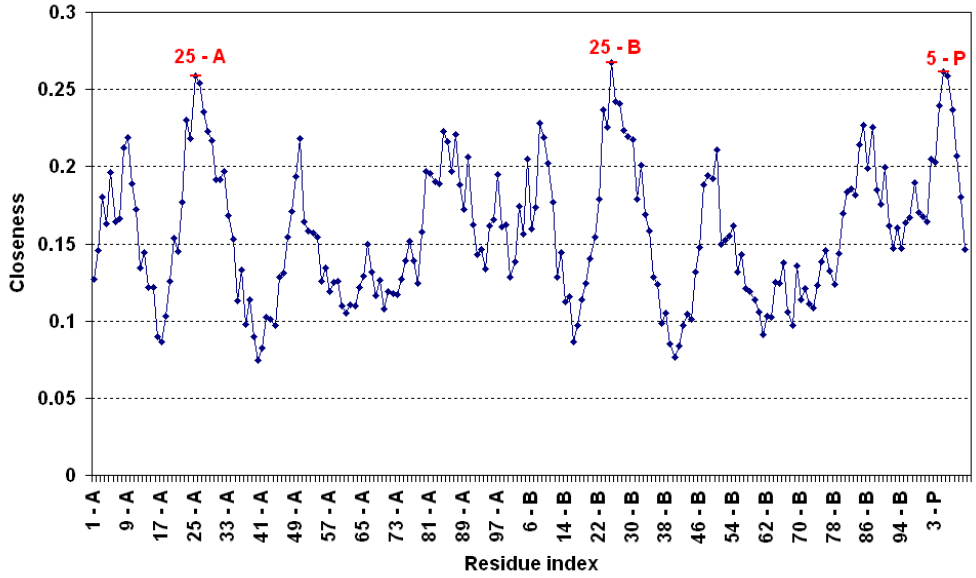

Figure 3.15. Closeness values of HIV-1 protease

In the case of D30N mutation, a significant decrease is observed in the values of the catalytic site of chain B (Asn25) and cleavage site of the peptide (Fifth residue). In the other mutant case, N88D, the difference becomes more significant. Even though the values of catalytic sites are nearly the same, dramatic decreases are observed in the values of Pro9 and Ile50 of chain A, and Ile85 of chain B. Similar deviations are observed in the double mutant case, however, there are other highly deviating sites as well. Especially at Ile54 and Ile85, dramatic increases are observed after double mutation. Finally, wild type structure is compared to the coevolved one and it is seen that the deviations caused by single and double mutations are mostly disappeared. This fact is also supported numerically by calculating the correlation coefficients of all structures' values with respect to wild type structure. The results are shown in Table 3.10. As expected, coevolved structure has the highest correlation coefficient whereas the double mutant has the lowest.

Table 3.10. Correlations between wild type and other structures

	D30N	**N88D**	**D30N/N88D**	**Coevolved**
Wild Type	0.918	0.915	0.896	0.938

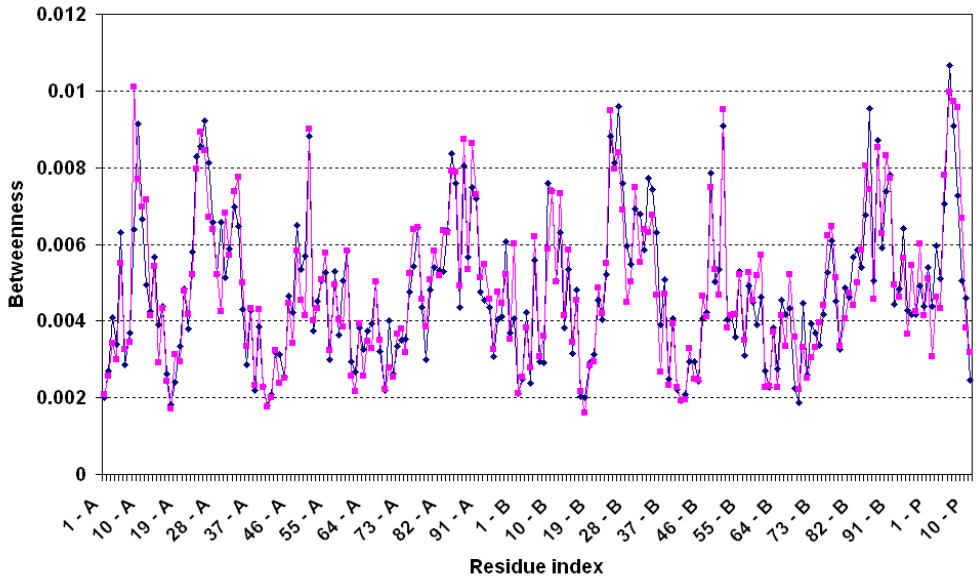

Figure 3.16. Comparison of betweenness values between wild type and D30N mutant (Wild type is shown in blue and D30N in magenta)

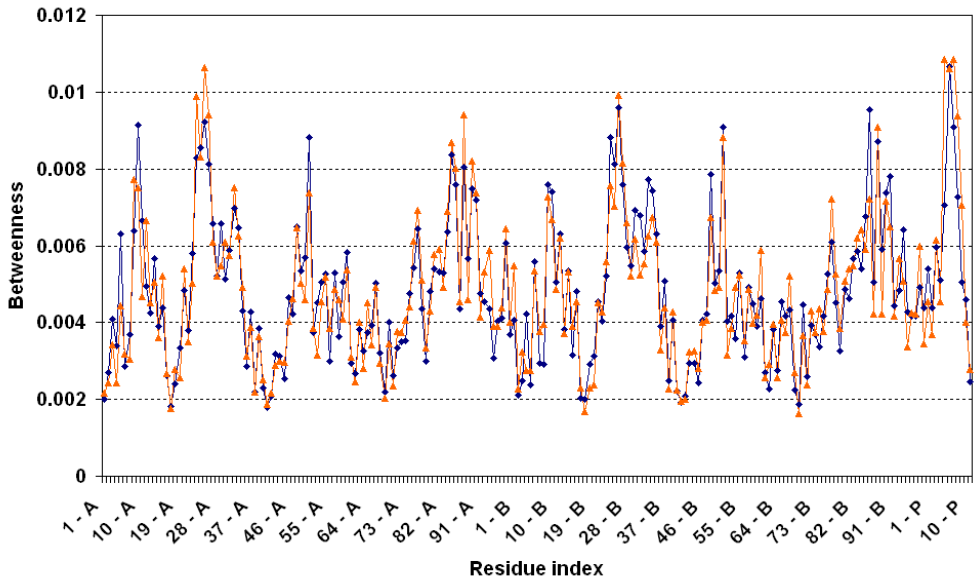

Figure 3.17. Comparison of betweenness values between wild type and N88D mutant (Wild type is shown in blue and N88D in orange)

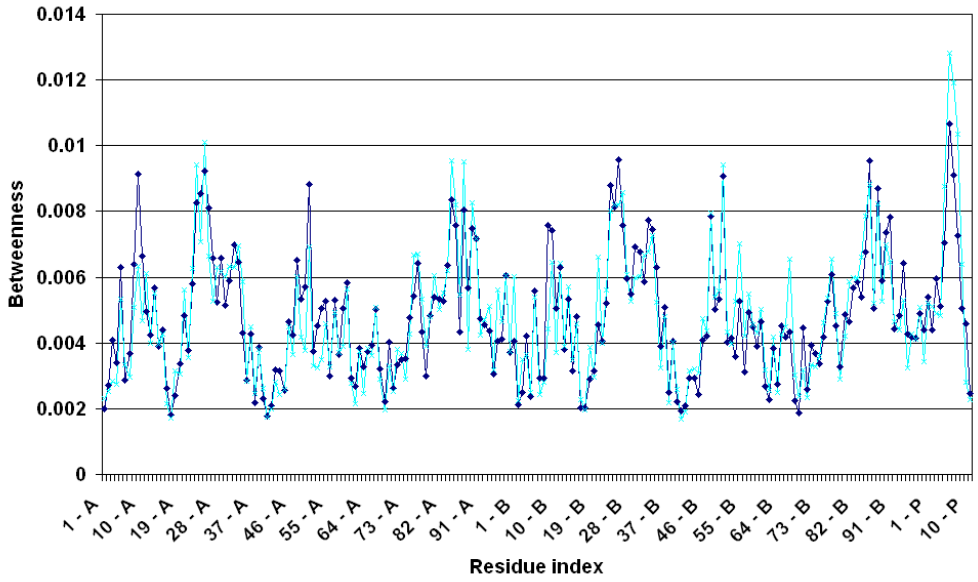

Figure 3.18. Comparison of betweenness values between wild type and double mutant (Wild type is shown in blue and double mutant in cyan)

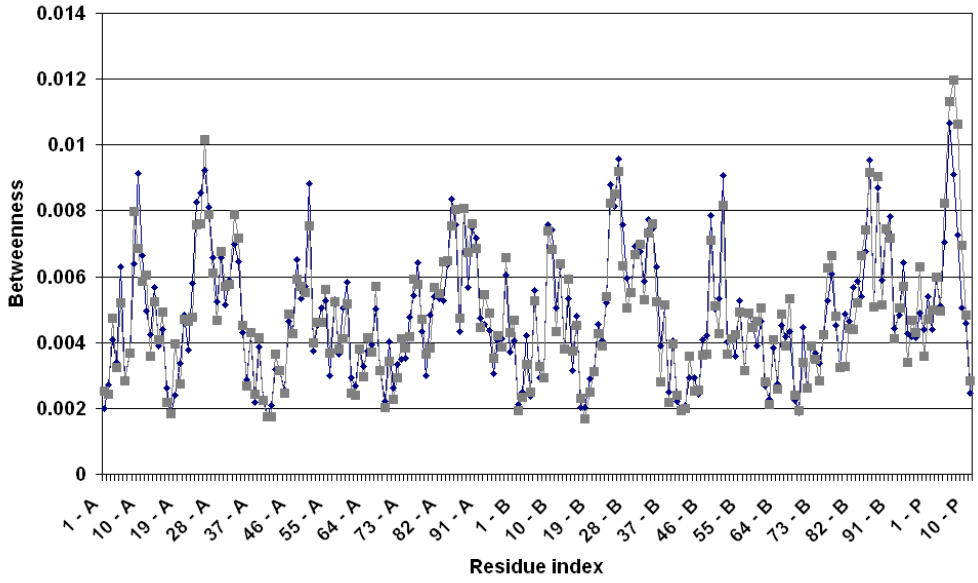

Figure 3.19. Comparison of betweenness values between wild type and coevolved structure (Wild type is shown in blue and coevolved in gray)

4. CONCLUSIONS AND FUTURE WORK

4.1. CONCLUSION

MC path generation method introduced here is shown to be a powerful tool for studying the intramolecular signaling pathways and functional residues underlying the allosteric mechanism as: an ensemble of pathways connecting two known allosteric sites, and identification of important residues by the network parameters of generated pathways.

The pathway analysis suggests that the shortest pathways are mostly defined as the pathways of highest probability, yet it might not always be the case. The higher probability for a specific pathway means a higher frequency for its appearance in particular in the case of small proteins such as POZ and PDZ domain representatives. The pathways with the highest probabilities result in a biologically sound ensemble of pathways, although none of them appears more than once, in the case of large proteins such as rhodopsin and myosin.

The network parameters, such as betweenness and closeness, point to functionally important residues, most of which are verified by the previous studies and some of them are suggested as functionally plausible in the present work. To this end, the small world network approach combined with MC path generation method comes out to be a powerful tool for determining the functional residues.

Since the functionality of residues is related to these network parameters, they are expected to be similar in wild type and coevolved structures of a protein where the protein is known to be functioning properly. This is verified in the HIV-1 protease system by showing the relatively higher correlation between the wild type and coevolved structures as compared to mutant structures. This analysis might be suggested as a novel aspect of the small-world network approach.

4.2. FUTURE WORK

In the calculations of interatomic energies, only L-J 12-6 potential is taken into account which involves the Van der Waal's forces. Another type of force, such as electrostatic forces, might be included as well which might result in more accurate results. The process of finding functional residues and allosteric pathways might become more automated and thus a web server performing these operations can be built.

In addition to the HIV-1 protease, other mutant and coevolved structures can found and subjected to analysis. This way, the robustness of this method can be tested.

APPENDIX A: Pathways of Other Myosin Structures

Table A.1. Most probable 20 pathways out of 50,000 generated by MC path generation method (1MMA)

50000 runs															
186	454	455	478	**482**	**486**	**489**	**492**	**493**	**743**	**742**					
186	187	188	116	117	118	90	92	694	693	745	**743**	**742**			
186	188	116	118	90	93	694	**696**	744	**742**						
186	185	456	455	478	**481**	**484**	**488**	**489**	**492**	**493**	**743**	**742**			
186	188	116	153	118	90	92	**695**	744	**743**	**741**	**742**				
186	241	454	453	176	651	**652**	**485**	**489**	**492**	**493**	**743**	**742**			
186	454	453	478	477	**481**	**484**	**488**	**492**	**493**	**743**	**742**				
186	188	116	153	118	90	92	694	692	744	**742**					
186	185	**178**	**177**	176	**652**	651	649	650	**485**	**489**	**492**	**493**	**743**	**742**	
186	187	185	**178**	**177**	176	**652**	**121**	**489**	**492**	**493**	**743**	**742**			
186	185	**655**	**654**	**122**	**121**	**489**	**492**	**490**	**491**	**493**	**743**	**742**			
186	185	179	**655**	**654**	**122**	**652**	**485**	**489**	**492**	**493**	**743**	**741**	**742**		
186	184	**655**	**654**	**652**	**485**	**489**	**492**	**493**	**743**	740	**742**				
186	454	452	175	651	650	**485**	**489**	**492**	**495**	**496**	**497**	**494**	**493**	**743**	**742**
186	188	116	118	90	92	93	94	**695**	745	**743**	**742**				
186	185	**177**	176	**481**	**486**	**485**	**489**	**492**	**491**	**493**	**743**	**742**			
186	454	455	478	**482**	**485**	**489**	**490**	**491**	**492**	**494**	**493**	**743**	**742**		
186	190	191	135	134	154	118	90	694	693	744	**742**				
186	189	452	453	645	646	**488**	**489**	**490**	**492**	**493**	**743**	**742**			
186	188	116	154	153	118	148	92	**695**	744	**742**					

Table A.2. Most probable 20 pathways out of 50,000 generated by MC path generation method (1MMD)

50000	runs											
185	**655**	**654**	**122**	**121**	**489**	**493**	**497**	**742**				
185	**177**	653	**652**	**121**	**489**	**492**	**496**	**497**	**742**			
185	**177**	176	**481**	**486**	**489**	**492**	**496**	**497**	**742**			
185	**177**	653	123	**121**	**489**	**492**	**496**	**741**	**742**			
185	**177**	176	650	**485**	487	**492**	**496**	**497**	**741**	**742**		
185	455	478	480	**484**	487	**492**	**496**	**497**	**742**			
185	**177**	653	**652**	**485**	**489**	**490**	**695**	697	**742**			
185	**178**	478	**482**	**486**	487	**488**	**492**	**496**	**497**	**742**		
185	**178**	**654**	**122**	**121**	**489**	**488**	**491**	**492**	**496**	**497**	**742**	
185	**177**	176	**652**	**482**	**485**	**489**	**490**	**493**	**497**	**742**		
185	**177**	176	**652**	650	**485**	**489**	**493**	**496**	**497**	**742**		
185	**655**	**654**	**122**	**121**	**489**	**492**	**496**	**741**	**497**	**743**	**742**	
185	**177**	176	651	650	649	**485**	**489**	**492**	**496**	**741**	**742**	
185	455	478	**481**	**484**	**485**	**489**	**492**	**496**	**497**	**742**		
185	455	478	**482**	487	**491**	**495**	**496**	**741**	**742**			
185	**655**	**654**	**122**	**121**	**489**	**488**	487	**491**	**492**	**496**	**741**	**742**
185	**177**	653	**652**	651	650	**485**	**488**	**491**	**493**	**497**	**742**	
185	**177**	653	**652**	**485**	**486**	487	**488**	**493**	**497**	**743**	**742**	
185	**655**	**654**	**482**	483	487	**488**	**490**	**695**	**743**	**742**		
185	**655**	**654**	**122**	**482**	**485**	**488**	**491**	**495**	**497**	**742**		

Table A.3. Most probable 20 pathways out of 50,000 generated by MC path generation method (1W9I)

50000	runs											
185	**177**	176	**652**	**485**	**489**	**492**	**493**	**497**	**742**			
185	**177**	653	**652**	**121**	**489**	**490**	**493**	**496**	**497**	**742**		
185	**177**	176	651	650	**485**	**488**	**492**	**493**	**497**	**742**		
185	179	**178**	653	**652**	**121**	**489**	**493**	**496**	**497**	**742**		
185	**177**	176	651	**652**	**121**	**489**	**493**	**496**	**497**	**742**		
185	179	**178**	**177**	176	650	**485**	**489**	**493**	**497**	**742**		
185	455	478	**481**	**484**	**488**	**491**	**495**	**497**	**742**			
185	**177**	176	**652**	**485**	**489**	**492**	**491**	**493**	**496**	**497**	**742**	
185	179	**178**	**177**	176	**652**	**485**	**489**	**493**	**497**	**741**	**742**	
185	**178**	**654**	**122**	**482**	**486**	**485**	**489**	**492**	**496**	**497**	**742**	
185	454	453	478	645	646	**488**	**492**	**496**	**497**	**742**		
185	**655**	**177**	176	650	**485**	**489**	**492**	**496**	**497**	**742**		
185	**178**	**654**	**122**	**121**	**489**	**492**	**493**	**496**	498	**497**	**742**	
185	**177**	176	650	651	**652**	**482**	**485**	**489**	**493**	**497**	**742**	
185	**177**	176	651	**652**	**482**	**485**	**489**	**490**	**494**	**497**	**742**	
185	457	475	480	**484**	487	**491**	**496**	**497**	**741**	**742**		
185	184	**655**	**654**	**122**	119	120	**121**	**489**	**493**	**496**	**497**	**742**
185	**178**	653	**652**	**485**	**489**	**486**	**490**	**493**	**497**	**742**		
185	179	**655**	**654**	**122**	123	**121**	**489**	**492**	**496**	**497**	**742**	
185	**177**	176	**481**	**482**	**486**	**490**	**492**	**495**	**497**	**742**		

Table A.4. Most probable 20 pathways out of 50,000 generated by MC path generation method (2AKA)

50000	runs													
185	**177**	653	**652**	**482**	487	**491**	**495**	**497**	**742**	0	0	0	0	0
185	**177**	176	650	**485**	**489**	**490**	**491**	**492**	**497**	**742**	0	0	0	0
185	**177**	176	175	651	650	**485**	**489**	**490**	**695**	**743**	**742**	0	0	0
185	**177**	176	**652**	651	650	**485**	**484**	**488**	**492**	**493**	**496**	**497**	**742**	0
185	**178**	**654**	123	**121**	**485**	**489**	**490**	**494**	**497**	**742**	0	0	0	0
185	179	**178**	**654**	**122**	**121**	**485**	**489**	**491**	**492**	**496**	**497**	**742**	0	0
185	**178**	**654**	**122**	119	120	**121**	**489**	**490**	**494**	**497**	**741**	**742**	0	0
185	**177**	653	**652**	123	**122**	**121**	**489**	**493**	**496**	**497**	**742**	0	0	0
185	**177**	**178**	**654**	653	**652**	**482**	487	**488**	**491**	**492**	**493**	**497**	**742**	0
185	**655**	**178**	**654**	**122**	**121**	**489**	**488**	**490**	**695**	**743**	**742**	0	0	0
185	**177**	653	**654**	479	483	**486**	487	**491**	**492**	**497**	**742**	0	0	0
185	**177**	175	176	651	650	**485**	**486**	487	**491**	**495**	498	**497**	**742**	0
185	**177**	653	**652**	**485**	**490**	**495**	**494**	499	740	**742**	0	0	0	0
185	**177**	653	**654**	**652**	**482**	**121**	**485**	**489**	**490**	**493**	**497**	**742**	0	0
185	187	189	191	159	158	**121**	**489**	**493**	**497**	**742**	0	0	0	0
185	454	453	175	174	650	**485**	**486**	**490**	**494**	**495**	**496**	**497**	**742**	0
185	**177**	653	**654**	**122**	**121**	**486**	**484**	**488**	**491**	**493**	**497**	**742**	0	0
185	**177**	176	**652**	**485**	**488**	**491**	**490**	**495**	**492**	**493**	**497**	**742**	0	0
185	**178**	**654**	**122**	**121**	**489**	**488**	**492**	**494**	498	**741**	**497**	**742**	0	0
185	**177**	176	**652**	**482**	**486**	**489**	**490**	503	502	501	499	498	**497**	**742**

APPENDIX B: Clustering Coefficient Figures

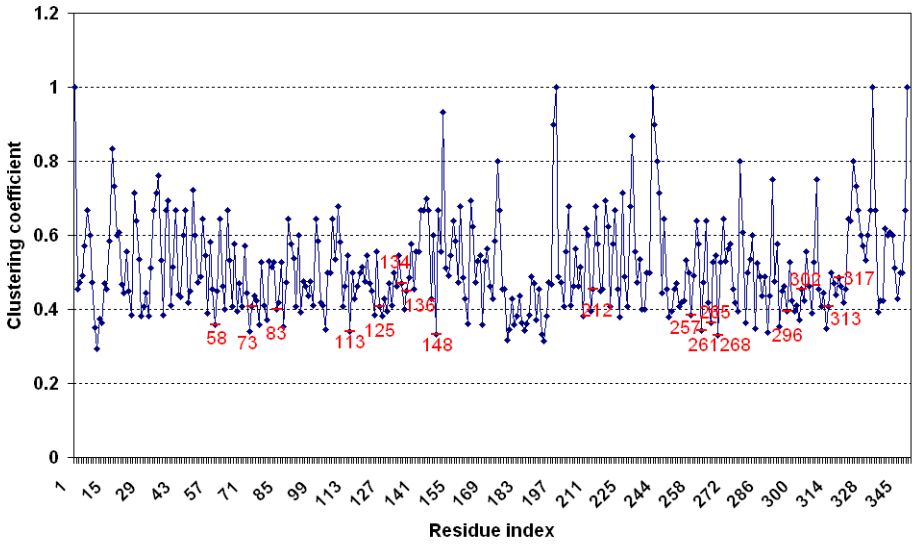

Figure B.1. Clustering coefficient values for rhodopsin

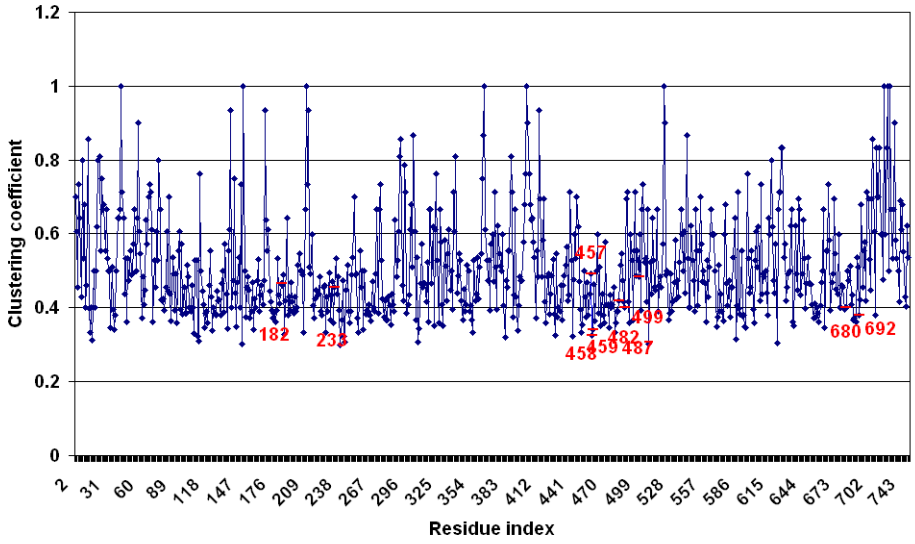

Figure B.2. Clustering coefficient values for myosin (1VOM)

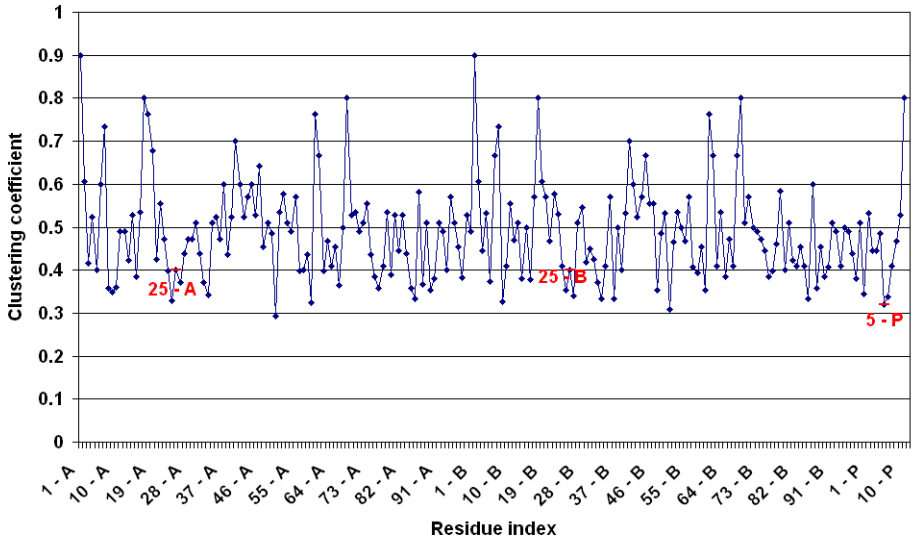

Figure B.3. Clustering coefficient values for HIV-1 protease

REFERENCES

Amitai, G., A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger and S. Pietrokovski, 2004, " Network analysis of protein structures identifies functional residues", *J Mol Biol*, Vol. 344, pp. 1135-1146.

Andres, A., P. Garriga and J. Manyosa, 2001, " Mutations at position 125 in transmembrane helix III of rhodopsin affect the structure and signalling of the receptor.", *Eur. J. Biochem.*, Vol. 268, pp. 5696-5704.

Atilgan, A. R., D. Turgut and C. Atilgan, 2007, "Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication", *Biopyhsical J.*, Vol. 92, pp. 3052-3062.

Ballesteros, J. A., L. Shi and J. A. Javitch, 2001, " Structural mimicry in G protein-coupled receptors: implications of the high-resolution structure of rhodopsin for structurefunction analysis of rhodopsin-like receptors", *Mol. Pharmacol*, Vol. 60, pp. 1-19.

Bernstein, E. E., T. F. Koetzle, G. J. B. Williams, J. E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. J. Tasumi, 1977, "The Protein Data Bank: a computer-based archival file for macromolecular structures", *Journal of Molecular Biology*, Vol. 117, pp. 535-542.

Cai, K. et al., 1999, " Single-cysteine substitution mutants at amino acid positions 306-321 in rhodopsin, the sequence between the cytoplasmic end of helix VII and the palmitoylation sites: sulfhydryl reactivity and transducin activation reveal a tertiary structure.", *Biochemistry*, Vol. 38, pp. 7925-7930

Cecchini, M., A. Houdusse and M. Karplus, 2008, " Allosteric communication in myosin V: from small conformational changes to large directed movements", *PLOS Computational Biology*, Vol. 4, Issue 8.

Chennubhotla, C. and I. Bahar, 2006, "Markov Propagation of allosteric effects in biomolecular systems: application to GroEL-GroES", *Molecular Systems Biology*, Article no: 36.

Clarkson, M. W., S. A. Gilmore, M. H. Edgell and A. L. Lee, 2006, "Dynamic coupling and allosteric behavior in a non-allosteric protein", *Biochemistry*, Vol. 45, pp. 7693-7699

Daily, M. D., T. J. Upadhyaya and J. J. Gray, 2008, " Contact rearrangements form coupled networks from local motions in allosteric proteins", *Proteins Structure Function Bioinformatics*, Vol. 71, pp. 455-466.

del Sol, A., H. Fujihashi, D. Amoros and R. Nussinov, 2006, " Residues crucial for maintaining short paths in network communication mediate signaling in proteins", emphMol Syst Biol, Vol. 2.

del Sol, A. and P. O'Meara, 2005, " Small-world network approach to identify key residues in protein-protein interaction", *Proteins*, Vol. 58, pp. 672-682.

Doyle, D. A., A. Lee, J. Lewis, E. Kim, M. Sheng and R. MacKinnon, 1996, "Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ", *Cell Press*, Vol. 85, pp. 1067-1076.

Garriga, P., X. Liu and H. G. Khorana, 1996, " Structure and function in rhodopsin: correct folding and misfolding in point mutants at and in proximity to the site of the retinitis pigmentosa mutation Leu-125?Arg in the transmembrane helix C", *Proc. Natl. Acad. Sci. USA*, Vol. 93, pp. 4560-4564.

Gether, U., 2000, " Uncovering molecular mechanisms involved in activation of G proteincoupled receptors", *Endocr. Rev.*, Vol. 21, pp. 90-113.

Goodey, N. M. and S. J. Benkovic, 2008, "Allosteric regulation and catalysis emerge via a common route", *Nature Chemical Biology*, Vol. 4, pp. 474-482.

Gunasekaran, K., B. Ma and R. Nussinov, 2004, "Is allostery an intrinsic propoerty of all dynamic proteins?", *Proteins:Stucture Function and Bioinformatics*, Vol.57, pp. 433-443.

Han, M., S. O. Smith, and T. P. Sakmar, 1998, " Constitutive activation of opsin by mutation of methionine 257 on transmembrane helix 6", *Biochemistry*, Vol. 37, pp. 8253-8261.

Han, M., S. O. Smith and T. P. Sakmar, 1998, " Constitutive activation of opsin by mutation of methionine 257 on transmembrane helix 6", *Biochemistry*, Vol. 37, pp. 8253-8261.

Harris, B. Z., F. W. Lau, N. Fujii, R. K. Guy and W. A. Lim, 2003, " Role of electrostatic interactions in PDZ domain ligand recognition", *Biochemistry*, Vol. 41, pp. 2797-2805.

Houdusse, A., A. G. Szent-Gyorgyi and C. Cohen, 2000, " Three conformational states of scallop myosin", *Proc. Natl Acad. Sci. USA*, Vol. 97, pp. 11238-11243.

Houdusse, A., A. G. Szent-Gyorgyi and C. Cohen, 2000, " Three conformational states of scallop myosin S1", *Proc. Natl Acad. Sci. USA*, Vol. 97, pp. 11238-11243

Hubbard, R. and G. Wald, 1952, " Cis-trans isomers of vitamin A and retinene in the rhodopsin system", *J. Gen. Physiol*, Vol. 36, pp. 269-315.

Ito, K., T. Q. Uyeda, Y. Suzuki, K. Sutoh and K. Yamamoto, 2003, " Requirement of domain-domain interaction for conformational change and functional ATP hydrolysis in myosin", *J. Biol. Chem*, Vol. 278, pp. 31049-31057.

Kong, Y. and M. Karplus, 2007, " The signaling pathway of rhodopsin", *Structure*, Vol. 15, 611-623.

Kreusch A., P. J. Pfaffinger, C. F. Stevens and S. Choe, 1998, "Crystal structure of the tetramerization domain of the Shaker potassiumchannel", *Nature*, Vol. 392, pp.

945-948.

Lockless, S. W. and R. Ranganathan, 1999, "Evolutionarily conserved pathways of energetic connectivity in protein families", *Science*, Vol. 286, pp. 295-299.

Menon, S. T., M. Han and T. P. Sakmar, 2001, " Rhodopsin: structural basis of molecular physiology", *Physiol. Rev.*, Vol. 81, pp. 1659-1688.

Ota, N. and A. D. Agard, 2005, "Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion", *J. Mol. Biol.*, pp. 1-10 .

Ozel, S., 2007, *A computational approach for analyzing the communication in proteins*, M.S. Thesis, Boğaziçi University.

Ozen, A., T. Haliloglu and C. Schiffer, 2009, " Using molecular dynamics to investigate substrate recognition and co-evolution in HIV-1 protease", *Biophysical Journal*, Vol. 96, Issue 3.

Ozen, A., 2008, *Molecular dynamics of substrate recognition and co-evolution in HIV-1 protease*, M.S. Thesis, Boğaziçi University.

Ozer, N., 2008, *Recognition and binding processes in HIV-1 protease*, PhD Thesis, Boğaziçi University

Porter, J. E., J Hwa and D. M. Perez, 1996, " Activation of the α1b-adrenergic receptor is initiated by disruption of an interhelical salt bridge constraint", *J. Biol. Chem.*, Vol. 271, pp. 28318-28323

Ranganathan, R. and E. M. Ross, 1997, "PDZ Domain proteins: scaffolds for signaling complexes", *Curr. Biol.*,Vol. 7, pp. R770-R773.

Süel, G. M., S. W. Lockless, M. A. Wall and R. Ranganathan, 2002, "Evolutionarily conserved networks of residues mediate allosteric communication in proteins", *Nature Structural Biology*, Vol. 10, pp. 59-68.

Sasaki, N., T. Shimada and K. Sutoh, 1998, " Mutational analysis of the switch II loop of Dictyostelium myosin II", *J. Biol. Chem*, Vol. 273, 20334-20340.

Sasaki, N., R. Ohkura and K. Sutoh, 2003, " Dictyostelium myosin II mutations that uncouple the converter swing and ATP hydrolysis cycle", *Biochemistry*, Vol. 42, pp. 90-95.

Tang, S., J. C. Liao, A. R. Dunn, R. B. Altman, J. A. Spudich and J. P. Schmidt, 2007, "Predicting allosteric communication in myosin via a pathway of conserved residues", *J. Mol. Bio*, Vol. 373, pp. 1361-1373.

Thibert, B., D. E. Bredesen and G. del Rio, 2005, " Improved prediction of critical residues for protein function based on network and phylogenetic analyses", *BMC Bioinformatics*, Vol. 6, pp. 213.

Tsiavaliaris, G., S. Fujita-Becker, R. Batra, D. I. Levitsky, F. J. Kull, M. A. Geeves and D. J. Manstein, 2002, " Mutations in the relay loop region result in dominantnegative inhibition of myosin II function in Dictyostelium.", *EMBO Rep.*, Vol. 3, 1099-1105.

Van Ham, M. and W. Hendriks, 2003, " PDZ domainsglue and guide", *Mol. Biol. Rep.*, Vol. 30, pp. 69-82.

Vendruscolo, M., N. V. Dokholyan, E. Paci and M. Karplus, 2002, " Small-world view of the amino acids that play a key role in protein folding", *Phys Rev E Stat Nonlin Soft Matter Phys*, Vol. 65.

Watts, D. J. and S. H. Strogatz, 1998 " Collective dynamics of 'small-world' networks", *Nature*, Vol. 393, pp. 440-442.

Yano, K., L. D. Kohn, M. Saji, A. Okuno and G. B. Cutler, 1997, " Phe576 plays an important role in the secondary structure and intracellular signaling of the human luteinizing hormone/chorionic gonadotropin receptor.", *J. Clin. Endocrinol. Metab.*, Vol. 82, pp. 2586-2591.