PREDICTION OF ALLOSTERIC KEY RESIDUES AND THEIR ROLE IN PROTEIN FOLDING

by

Şölen Ekesan

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 and part of it was funded by TUBITAK scientific research projects funds.

First of all I would like to thank my thesis supervisor Prof. Türkan Haliloğlu for her unending welcome, guidance, understanding and encouragement through 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; especially to Burcu Aykaç Fas and Nevra Özer.

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

Finally I would like to express my gratitude to Andaç Armutlulu who deserves most of the credit for helping me intellectually, mentally and physically to complete these studies, and for being there for me all the time

ABSTRACT

PREDICTION OF ALLOSTERIC KEY RESIDUES AND THEIR ROLE IN PROTEIN FOLDING

Allostery, an aspect of protein dynamics, is crucial in regulation of protein activity. It is believed that allostery is maintained through communication pathways of key residues residing within distant allosteric sites. Prediction of these key residues therefore, would be a milestone in understanding protein allostery. In this study, prediction of functional residues is carried out by a newly proposed Monte Carlo (MC) path generation method, where the protein structure is considered as a network of amino acid residues and inter-residue interactions are described by a potential function. Study of the effect of the type of the potential function used, is carried out with four different potential functions, among which atomistic potential function is found to be the best to describe the interactions. Three different approaches of MC path generation are studied; 1) generating paths between two residues (BTR), 2) generating paths with specific number of steps (PSNS) and analyzing residue frequencies, and 3) generating infinite step paths (ISP) and calculating network parameters such as closeness, betweenness and clustering coefficient. In studying Shaker potassium channel and HIV-1 protease systems using these approaches, paths are generated in ensembles rather than obtaining a single shortest path. MC path generation is also applied to study the communication within and between different monomers of a structure and different structures of a protein. Combined information from these approaches reveals a list of functionally important residues, such as catalytic, binding and allosterically important sites. The role of these proposed residues in protein folding is studied through trajectories of protein folding simulations, algorithm of which is based on robotic motion planning. Through the folding trajectories, residue contacts are analyzed and residues that form initial contacts and conduct folding are identified. Interestingly these residues are noticed to be among those that display high closeness and betweenness values in pathway anal-

ysis carried out for the native state of the proteins. Overall, this study suggests that communication pathways are evolutionarily conserved and MC path generation is an effective method for prediction of residues that are important in both allostery and protein folding.

ÖZET

ALOSTERİK ANAHTAR REZİDULARIN TAHMİN EDİLMESİ VE PROTEİN KATLANMASINDAKİ ROLÜ

Protein dinamiğinin bir parçası olarak alosteri, protein aktivitesinin denetiminde önemli bir rol teşkil eder. Alosterinin, birbirinden uzak iki alosterik nokta arasında bulunan anahtar rezidular üzerinden geçen bilgi patikaları aracılığıyla sağlandığına inanılır. Bu anahtar rezidulların tahmin edilmesi, protein alosterisini anlamak açısından bir kilometre taşı oluşturacaktır. Bu çalışmada, işlevsel rezidulların tahmini yeni ileri sürülen Monte Carlo (MC) patika yaratma yöntemi ile gerçekleştirilmiştir. Bu yöntemde protein yapısı bir amino asit ağ yapısı olarak ele alınır ve rezidular-arası etkileşmeler bir potansiyel fonksiyon ile tanımlanır. Kullanılan potansiyel fonksiyon türünün etkileri dört farklı potansiyel fonksiyon denenerek çalışılmış ve etkileşmeleri en iyi tanımlayan potansiyel fonksiyon olarak atomistik potansiyel fonksiyon seçilmiştir. MC patika yaratma, üç farklı yaklaşım ile incelenmiştir; 1) iki rezidu arası patika yatarımı (BTR), 2) belirli sayıda adımdan oluşan patika yaratımı (PSNS), ve 3) sınırsız sayıda adımdan oluşan patika yaratımı ve yakınlık, aradalık ve kümeleme katsayısı gibi ağ yapı parametrelerinin hesaplanması (ISP). Bu yaklaşımlar kullanılarak; Shaker potasyum kanalı ve HIV-1 proteaz gibi sistemlerin çalışılmasında, bir tek en kısa patika elde etmekten ziyade patika toplulukları yaratılmıştır. Ek olarak, MC patika yaratma, monomer içi ve monomerler arası ve de aynı proteinin faklı yapılarındaki iletişim çalışmalarına da uygulanmıştır. Elde edilen bu bilgilerin birleştirilmesi sonucu katalitik, bağlanma ve alosterik fonksiyonlar açısından önemli rezidulların bir listesi açığa çıkmıştır. Bu ileri sürülen rezidulların protein katlanmasındaki rolleri, robotik hareket planlama kavramlarına dayanan bir algoritma ile yapılan protein katlanma simulasyonları üzerinden çalışılmıştır. Katlanma rotaları boyunca rezidulların yaptığı kontaklar incelenmiş ve katlanmaya yön veren ilk kontakları oluşturan rezidular teşhis edilmiştir. İlginçtir ki bulunan bu rezidular patika analizlerinde yüksek yakındalık ve aradalık değerleri

gösteren rezidular arasında yer almaktadır. Sonuç olarak bu çalışma, iletişim patikalarının evrimsel olarak korunduğunu ve MC patika yaratma yönteminin hem alosteri hem protein katlanmasındaki önemli rezidüları tahmin etmekte başarılı bir yöntem olduğunu önermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

a_{ij}	Affinity between i and j
b_k	Betweenness
C_α	Alpha carbon
C_v	Clustering coefficient
d_j	Local interaction density
d_v	Number of neighbors of v
E	Energy
E_{min}	Minimum energy
E_v	Number of edges between neighbors of v
g_{ij}	Number of different shortest paths
kT	Boltzman temperature
l	Fixed distance between robot bodies, bond length
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
N_i	Total number of heavy atoms
O_i	Closeness
p	Robot body
P_{ij}	Probability of interaction of i and j
PP	Path probability
Q	Normalized probability matrix
r	Distance
R	Probability range matrix
r_{cut}	Cut-off radius
r_{min}	Radius corresponding to the minimum energy
S	Number of steps of a path
s	Torsion angles of robot bodies

u	Eigenvector
W_{ij}	Weight of interaction of i and j
w_0	Initial conformation
w_g	Target conformation
α	Description of α
γ	Hookean force constant
Γ	Kirchoff matrix
ε	Well depth
θ	Virtual bond angle
λ	Eigenvalue
ρ	Radius of a robot body
σ	Collision diameter
σ	Squashing function
σ_d	Sharpening function
ϕ	Pseudo dihedral angle
φ	Artificial potential function
$\check{\varphi}$	Intrinsic artificial potential function

3-D	Three-dimensional
Arg	R, Arginine
Asn	N, Asparagine
Asp	D, Aspartic acid
ATD	Anisotropic Thermal Diffusion
BB	B, Backbone
BJ	Bahar Jernigan
Cys	C, Cysteine
DNA	Deoxyribonucleic acid
Gln	Q, Glutamine
Glu	E, Glutamic acid
Gly	G, Glycine
GNM	Gaussian Network Model

HFF	High Frequency Fluctuations
His	H, Histidine
HIV	Human Immunodeficiency Virus
Ile	I, Isoleucine
Leu	L, Leucine
LJ	Leonard-Jones
LR	Long range
Lys	K, Lysine
MC	Monte Carlo
MD	Molecular Dynamics
Met	M, Methionine
MSF	Mean squared fluctuations
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PDB	Protein Data Bank
Phe	F, Phenylalanine
PI	Protease Inhibitor
PR	Protease
Pro	P, Proline
RMSD	Root mean squared deviation
RNA	Ribonucleic acid
SS	S, Side-chain
Ser	S, Serine
SR	Short range
TD	Thomas Dill
Thr	T, Threonine
Trp	W, Tryptophan
Tyr	Y, Tyrosine
Val	V, Valine
WT	Wild-type

1. INTRODUCTION

Proteins are dynamic entities undergoing different amplitude and time scales of motions, which are essential in their functioning. Scales range from femtoseconds of atomistic fluctuations to hours of protein folding, with the in between rigid body motions of binding, hinge bending and allosteric transitions. Protein dynamics covering all these motions is, therefore, crucial in protein studies.

Allostery, an aspect of protein dynamics, is a regulation mechanism for protein activity. It is defined as the conformational change at one site (i.e due to binding), causing a functional change at a distant site. The proteins, in which binding affinity or catalytic efficiency is modulated by binding or chemical perturbations at distal sites are defined as *allosteric* (Clarkson *et al.*, 2006). All proteins are argued to be potentially allosteric (Gunasekaran *et al.*, 2004), suggesting allostery based regulation for all proteins and revealing the importance of allostery. The connection between the allosteric sites defines the allosteric mechanism. An approach for how this connection is maintained is through communication pathways of key amino acid residues residing in between the allosteric sites. Prediction of these key residues would reveal crucial information regarding the allosteric mechanism, i.e. activity regulation.

Communication pathways have been the subject of a number of studies (Lockless and Ranganathan, 1999, Ota and Agard, 2005, Süel *et al.*, 2002, Atilgan *et al.*, 2007, Tang *et al.*, 2007) on systems such as POZ and PDZ domains, GPCR family and myosin. Novel terms such as *allosteric pathways* and *shortest path* are introduced (Lockless and Ranganathan, 1999, Atilgan *et al.*, 2007). Different approaches have been developed, one of which studies thermodynamic residual couplings along the evolution using statistical interactions between amino acid positions (Lockless and Ranganathan, 1999). One other approach studies the communication in terms of thermal diffusion by anisotropic thermal diffusion (ATD) simulations, where the initial residue is heated and the propagation of heat in the form of kinetic energy reveals a pathway of communication (Ota and Agard, 2005). In another approach (Atilgan

et al., 2007) the *shortest path* is sought using Dijkstra's algorithm, where the weights between the residues are defined by Thomas-Dill knowledge-based potential (Thomas and Dill, 1996).

All of these methods have some limitations. First of all any two methods, upon the study of the same system, do not obtain the same paths. This suggests that either the methods are inefficient, or and rather that the communication does not occur through a single *shortest path*, but through ensemble of probable paths. The approach presented by Atilgan *et al.* (2007) would be useless for systems with any unknown allosteric site. Since allostery is an intrinsic property of the proteins, communication information starting from a single site should be obtainable without the necessity of another site. Lockless and Ranganathan's (1999) method is also very limited since it requires evolutionary data which is not available for all structures, and also due to their time consuming in depth analysis. The approach by Ota and Agard (2005) is limited to small proteins due to the computer expense of simulations. A novel approach that; gives ensemble of probable paths, can study communication from a single residue, does not need evolutionary data, and has no structure limitations could converge all these methods into one.

Another aspect of protein dynamics, protein folding, is the conformational changes the protein goes through in reaching a unique native structure according to the evolutionarily designed interactions among amino acid residues (Nagao *et al.*, 2005). It remains to be the most major unsolved problem in protein dynamics. Finding the intermediate structures along the folding trajectory is crucial in understanding the folding mechanism. Experimental and computational studies are being carried out to be able to determine these intermediate conformations. Experimental studies through protein engineering and NMR techniques (Kmiecik and Kolinski, 2007) can only identify few intermediate conformations, whereas all is needed to fully map the folding process. On the other hand computational studies in atomistic detail cannot be carried out due to computational expenses and the sufficiency of interatomic interactions on a such time scale are in question. Coarse-grained simulations are proposed to provide basic information on the folding (Ulutas *et al.*, 2009).

In this study, prediction of residues residing in the communication pathways is carried out by a newly proposed MC path generation method, which provides an ensemble of probable paths. This method is used in three different approaches; 1) generating paths between two residues (BTR), 2) generating paths with specific number of steps (PSNS) and analyzing residue frequencies, and 3) generating infinite step paths (ISP) and calculating network parameters such as clustering coefficient, closeness and betweenness. POZ domain representative Shaker potassium channel and HIV-1 protease complex structure are the studied systems. Apart from these three approaches applied to both systems, different substrate structures of HIV-1 protease and connection between two monomers of ca-p2 substrate-bound HIV-1 protease complex structure are studied. On the other hand protein folding is simulated for both of the systems introduced using an algorithm based on robotic motion planning to search for the role of functional key residues suggested by the native-state network parameters during the folding.

2. MATERIALS AND METHODS

2.1. Studied Protein Structures

2.1.1. POZ and PDZ domain representatives

POZ (poxvirus and zinc finger) and PDZ (Post-Synaptic Density-95/DLG/ZO-1) domains are specific protein-protein interaction (protein-binding) domains with conserved sequence elements (Bardwell and Treisman, 1994, Fanning and Anderson, 1999). These domains are known to play important roles in protein signaling (Fanning and Anderson, 1999, Ranganathan and Ross, 1997, Tsunoda *et al.*, 1997).

Representative structures for both of the domains are selected to carry out the calculations. Shaker potassium channel (PDB code:1A68) (Kreusch *et al.*, 1998), is selected as a representative of POZ domain and the third PDZ domain (PDZ-3) of PSD-95 (PDB code:1BE9) (Doyle *et al.*, 1996) is selected as a representative of PDZ domain, as was done by Atilgan *et al.* (2007) and Ota and Agard (2005) respectively.

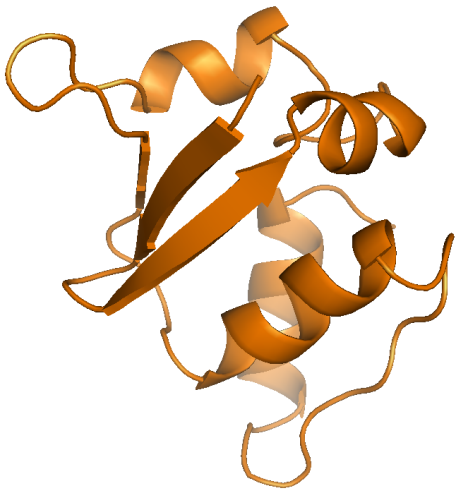

Figure 2.1. Tetramerization domain of Shaker potassium channel structure (1A68) in cartoon representation

Tetramerization domain of Shaker potassium channel (Figure 2.1) is a small protein, consisting of a single chain with 87 residues. Paths are generated between residues Phe77 and Phe148 (Lockless and Ranganathan, 1999, Atilgan *et al.*, 2007). Phe77 is the residue at the interaction surface and Phe148 is the residue acting in binding of other subunits of K^+ channel (Lockless and Ranganathan, 1999).

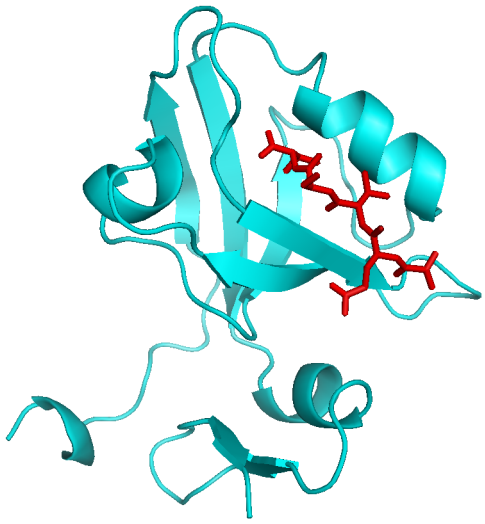

Figure 2.2. PSD-95 structure cartoon representation with chain A shown in cyan and chain P (i.e. substrate) in red.

PDZ-3 of PSD-95 (Figure 2.2) is also a small protein consisting of single chain of 115 residues and a five residue c-terminal of a peptide. In this study residue numbers from the PDB are used to identify the amino acid positions. Paths are generated between residues His372 and Leu353. His372 is the position critical in ligand specificity and Leu353 is a position on the opposite side from the ligand pocket (Lockless and Ranganathan, 1999).

2.1.2. HIV-1 Protease

HIV (Human Immunodeficiency Virus) is the virus found to be responsible for AIDS and HIV-1 is a member of this retrovirus family. HIV protease is an essential protein in HIV maturation due to its important role of processing the vital polyproteins (Chou, 1996).

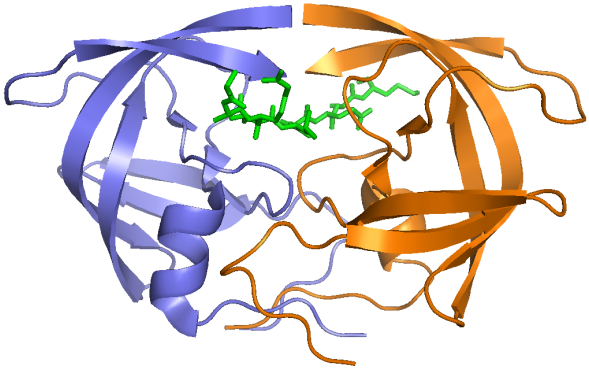

Figure 2.3. HIV-1 protease structure cartoon representation with chain A shown in purple, chain B in orange and substrate (i.e. chain P) in green.

HIV-1 protease is active as a homodimer, with each monomer consisting of a single chain of 99 amino acid residues (Figure 2.3). Even though the protein itself is symmetric, it recognizes different substrate sites some of which are summarized in Table 2.1 and the corresponding crystal structures are given in Table 2.2.

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

The single active site (residues Asp25-Thr26-Gly27) of the protease is formed upon dimerization and the closure of the flaps, and residues Asp25 of each monomer act as catalytic site (Wlodawer and Erickson 1993). The hinge residues determined for the protease are 56&78 69&73 and loops through residues 36-44 and 12-22 (Ozer, 2008). From these the tips, corresponding to residues 56, 40, 17 and 69 are studied.

Table 2.2. Substrate bound structures of HIV-1 protease used in this study

	PDB code	Reference
Substrates		
capsid-p2 (ca-p2)	1F7A	Prabu-Jeyabalan et al., 2000
matrix-capsid (ma-ca)	1KJ4	Prabu-Jeyabalan et al., 2002
nucleocapsid-p1(nc-p1)	1TSU	Prabu-Jeyabalan et al., 2004
p1-p6	1KJF	Prabu-Jeyabalan et al., 2002
p2-nucleocapsid (p2-nc)	1KJ7	Prabu-Jeyabalan et al., 2002
RNase-integrase (rh-in)	1KJH	Prabu-Jeyabalan et al., 2002
rev.trans.-RNaseH (rt-rh)	1KJG	Prabu-Jeyabalan et al., 2002

2.2. Interaction Potentials

Interaction potential function is the key to calculate the energetic interaction between residue pairs, which is the basis of the pathways of communication. Statistical (knowledge) and physical based potentials are tried; values of all of which are calculated differently as explained below.

2.2.1. Atomistic Potential

Atomistic potential (Ozen, 2008) is an empirical physical based potential in atomistic level considering the attraction between each atom like Lennard-Jones 12-6 potential.

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

where ϵ and σ correspond to the minimum energy and the collision diameter between the two atoms, respectively. r corresponds to the distance between the two atoms.

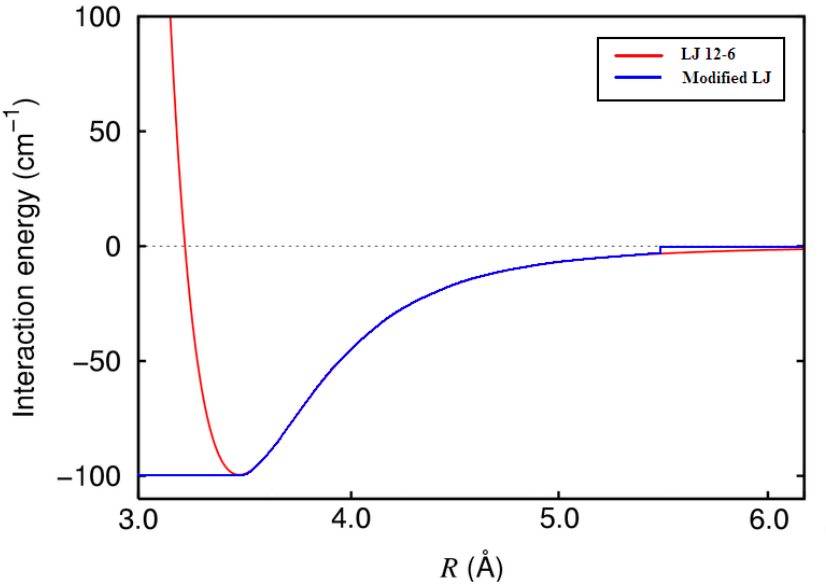

Figure 2.4. Lennard-Jones 12-6 Potential and Modified Lennard-Jones Potential

Modifications are introduced (see Figure 2.4) considering the over punishment of the accidental get-togethers in crystal structures (Ozen, 2008). New parameters r_{min} (radius corresponding to the minimum energy) and E_{min} (minimum energy) are introduced such that;

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

$$E_{min} = E(r_{min}) \tag{2.3}$$

and the energy corresponding to distances below r_{min} is set to an energy value of E_{min}. Similarly an r_{cut} of value 5.5 Å is introduced to simplify the calculations so that;

$$\text{If} \quad r < r_{cut} \quad \text{then} \quad E = 0 \tag{2.4}$$

The Van der Waals parameters are experimentally obtained and listed for each atom in each amino acid residue (Ozen, 2008). The necessary x, y, z coordinates data are available from the PDB file of the protein.

For a protein with M atoms and N residues, all atom-atom interaction energies are calculated and stored in a square energy matrix with MxM dimensions.

Table 2.3. Atomistic interaction energy sample matrix

	GLU 66 (kT)									**ARG 67 (kT)**										...
	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	
⋮																					

Then the energy between each atom in every residue couple is summed to obtain a single energy value between those two residues. The atomistic energy matrix thus reduces to the residue energy matrix with dimensions NxN.

Table 2.4. Residual interaction energy sample matrix

	GLU 66 (kT)	**ARG 67 (kT)**	**...**
GLU 66	0	-2.65	
ARG 67	-2.65	0	
⋮			

The probability of occurrence of an interaction between residue pairs i and j, are considered to be proportional to the Boltzmann weight and normalized considering all

interactions of i^{th} residue as

$$W_{ij} = \exp\left(\frac{-E_{ij}}{kT}\right) \tag{2.5}$$

$$P_{ij} = \frac{W_{ij}}{\sum_{i=1}^{N} W_{ij}} \tag{2.6}$$

The probability matrix of dimensions NxN thus comprises the normalized probabilities for the interaction of N residues with the remaining N-1 residues. The diagonal elements of the matrix present zero values. Here, the premise is that the inter-residue interaction energy describes the energy flow in the structure.

2.2.2. Bahar-Jernigan (BJ) Potential

Bahar-Jernigan (BJ) potential function (Bahar and Jernigan, 1997, Bahar *et al.*, 1997) is a low resolution statistical (knowledge-based) potential function designed for conformational energy calculation of proteins. This low resolution potential models the amino acid residues as two center points; center of side-chain atoms (S) and center of backbone atoms (B). Residue specific energy values are defined from data of radial distribution for 302 protein structures. The potential consists of long-range (LR) and short-range (SR) interaction energies in terms of discrete values specified for intervals of distances and angles respectively.

The long-range (LR) (Bahar and Jernigan, 1997) and short-range (SR) (Bahar *et al.*, 1997) interactions are considered separately and combined to obtain an overall energy value as,

$$E(\Phi) = E_{SR}(\Phi) + E_{LR}(\Phi) \tag{2.7}$$

where Φ is a given conformation. The long-range interactions depend on the distances between side-chain-side-chain (SS), side-chain-backbone (SB) and backbone-backbone (BB) residue centers (Bahar and Jernigan, 1997). Overall conformational LR interaction energy is calculated considering all of these interactions through every residue

pair.

$$E_{LR}(\Phi) = \sum_{i=1}^{N}\sum_{j=1}^{N} E_{SS}(r_{ij}) + \sum_{i=1}^{N}\sum_{j=1}^{N} E_{SB}(r_{ij}) + \sum_{i=1}^{N}\sum_{j=1}^{N} E_{BB}(r_{ij}) \tag{2.8}$$

where N is the number of residues and r_{ij} is the distance between i^{th} and j^{th} residues.

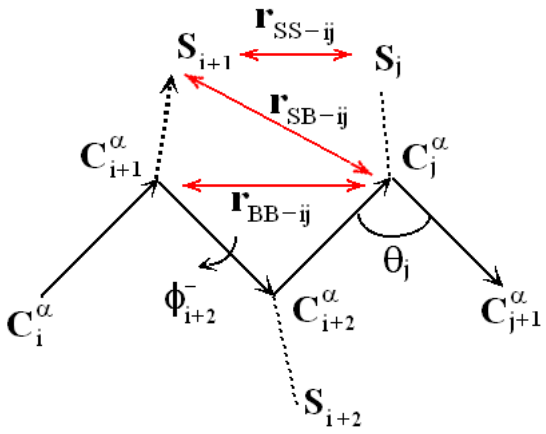

Figure 2.5. Representation of the parameters of BJ long range interactions

Short-range interactions, on the other hand, are function of bond and rotational angles of the virtual bonds between the backbone centers (Bahar *et al.*, 1997). Short-range interactions are defined as independent interactions and coupling interactions.

$$E_{SR}(\Phi) = E_{SR-independent}(\Phi) + E_{SR-coupling}(\Phi) \tag{2.9}$$

The coupling interactions refer to the coupled effects of bond (θ) and rotational (ϕ) angles on the structure where as the independent interactions, as the name implies, are independent of these couplings.

$$E_{SR-independent}(\Phi) = \sum_{i=2}^{N-1} E_i(\theta) + \sum_{i=3}^{N-1}\left[\frac{E_i(\phi^-) + E_{i-1}(\phi^+)}{2}\right] \tag{2.10}$$

$$E_{SR-coupling}(\Phi) = \sum_{i=3}^{N-1} \Delta E_i(\theta, \phi^-) + \Delta E_{i-1}(\theta, \phi^+) + \sum_{i=3}^{N-2} \Delta E_i(\phi^+, \phi^-) \tag{2.11}$$

In order to calculate energies of residue pair interactions rather than the overall conformational energy, the above LR and SR interaction energy calculations are modified. LR interactions of each residue pair is calculated as;

$$E_{LR-ij} = E_{SS-ij}(r_{SS-ij}) + E_{SB-ij}(r_{SB-ij}) + E_{SB-ji}(r_{SB-ji}) + E_{BB-ij}(r_{BB-ij}) \quad (2.12)$$

SR interactions of five consecutive residues (i.e. n=i+4) are considered to be significant. Therefore conformational SR energy for a protein of five residues (n=5) is divided up such that the sum of all residue pair interaction energies would equal to the overall conformational energy.

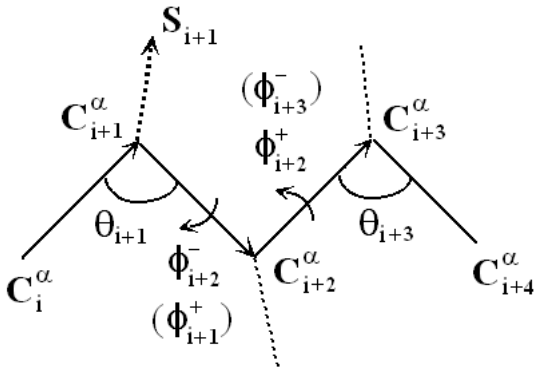

Figure 2.6. Corresponding numbering of virtual parameters for BJ energy calculation

So E_{SR-ij} are calculated according to contributions of up to five consecutive residues as follows;

1. If $j = i + 2$

$$E_{SR-ij} = E_{i+1}(\theta) \quad (2.13)$$

2. If $j = i + 3$

$$E_{SR-ij} = \frac{E_{i+2}(\phi^-) + E_{i+1}(\phi^+)}{2} + \Delta E_{i+2}(\theta, \phi^-) + \Delta E_{i+1}(\theta, \phi^+) \quad (2.14)$$

3. If $j = i + 4$

$$E_{SR-ij} = \Delta E_{i+2}(\phi^+, \phi^-) \tag{2.15}$$

4. Else

$$E_{SR-ij} = 0 \tag{2.16}$$

2.2.3. Thomas-Dill (TD) Potential

Thomas Dill potential function (Thomas and Dill, 1996) is a low resolution knowledge based potential function derived by generating pair-wise additive amino acid energy scores from known protein structures. The low resolution model defines the residues as the carbon-beta atoms, and the energy values are calculated according to the distances between these centers and the types of the residues (Thomas and Dill, 1996). A single set of energy values and two distance limits (upper denoted as U and lower denoted as L) are available and the energy is calculated according to three conditions;

1. If $0 < r_{ij} < L$ and $r_{ij} \neq 0$

$$E_{ij} = (E_{ij}, \text{set}) \tag{2.17}$$

2. If $L < r_{ij} < U$

$$E_{ij} = (E_{ij}, \text{set}) \left[\frac{(r_{ij} - U)^2 (2r_{ij} - 3L + U)}{(U - L)^3} \right] \tag{2.18}$$

3. If $U < r_{ij}$ and $r_{ij} = 0$

$$E_{ij} = 0 \tag{2.19}$$

The defined value for the upper limit U is 9 Å and for the lower limit L is 6 Å.

2.2.4. Markov Affinity Potential

Markov affinity potential function is a model potential function used in the Markov process of network communications (Chennubhotla and Bahar, 2006), which defines the affinity between two residues in terms of the number of atomistic contacts between the two residues and the total number of heavy atoms present in both residues.

Two atoms are considered to be in contact if they are closer in distance than a defined cutoff radius r_{cut}. If N_{ij} is the total number of atomistic contacts and N_i is the total number of heavy atoms in i^{th} residue than the affinity is;

$$a_{ij} = \frac{N_{ij}}{\sqrt{N_i N_j}} \tag{2.20}$$

Another definition is the local interaction density d_j which corresponds to the total number of contacts a residue has.

$$d_j = \sum_{i=1}^{N} a_{ij} \tag{2.21}$$

Interaction probability is then obtained as;

$$P_{ij} = \frac{a_{ij}}{d_j} \tag{2.22}$$

2.3. Path Generation & Analysis

Any network is defined by its vertices V and edges E as G(V,E). Proteins can be considered as a network of amino acid residues, such that each amino acid residue defines a vertex (node) in the network. Definition of edges depends on the use of the network. In this study the edges are defined as the probabilities P_{ij} of interaction between the amino acid residues (see section 2.3.1). The overall network represents all possible interactions within pairs of residues throughout the protein (i.e. all possible

paths). Therefore the term *path generation* in this study is used for the sampling of more probable connections between two nodes of the network.

There are many different path generation methods (Lockless and Ranganathan, 1999, Ota and Agard, 2005, Atilgan *et al.*, 2007). In the present work, a novel methodology is introduced by applying a Monte Carlo (MC) approach to generate paths of interactions between residues. Analysis of the generated paths is as important as the path generation itself. The generated paths are analyzed in various ways to bring up different aspects of the generated paths.

2.3.1. Monte Carlo Path Generation

Monte Carlo (MC) path generation is based on the concepts of the MC Method (Metropolis and Ulam, 1949), where random numbers are used to decide upon a number of possible solutions to sample a subset of results that would represent the whole solution set. In case of path generation the whole solution set is all the possible paths in the protein. The random numbers decide upon which residue will be next in the path according to the probabilities of interaction between the amino acid residues.

The probabilities P_{ij} are obtained from the energies E_{ij} (see section 2.2) calculated from the interaction potentials (except for Markov Affinity) according to the Boltzmann weight;

$$P_{ij} = \exp \frac{-E_{ij}}{kT} \tag{2.23}$$

The probability matrix P is formed so that P_{ij} corresponds to the probability of residue j (column) to follow residue i (row) in the path. The probability matrix P is row normalized so that the total probability of the path from i to all the other residues (columns) is 1, forming the normalized probability matrix Q. In order to obtain the matrix of unique ranges of probabilities R, the values are summed up for each row, so that the range of the first column starts with 0 and the range of the last column ends with 1. At this point generating a random number between 0 and 1 defines the next

step in the path, by choosing the residue with the range covering the random number.

Three different path generation methods are used in this study; paths between two end residues (BTR), paths with specific number of steps (PSNS), and infinite step paths (ISP). BTR: Paths between two end residues are generated starting from one end residue (starting node) and generating successive steps, edges that connect the residues, towards the other end residue (target node) SSN: The generation of paths with specific step number ends regardless of the final residue when the specified step is reached. In both of these methods selection of already visited residues are prohibited to avoid reverse stepping by setting their probabilities to zero after the visit. The probability matrix P is re-normalized (i.e. dynamic probability matrix) and the range matrix Q is re-created. For both methods significant numbers of paths are generated to ensure a statistical significant sampling.

Unlike the previous two methods, ISP are generated allowing loops in the path (i.e. visited residues are not eliminated). Also, instead of an ensemble of paths, a single infinitely long path (e.g. 200,000) is generated. This way the starting and end residues are not significant in a path of infinite steps.

2.3.2. Path Analyses

Generated paths are used to obtain three parameters that comply the path information; path frequency, residue frequency and path probability. These are mostly applied to the first two path generation methods.

2.3.2.1. Path Frequency. Path frequency is the frequency of occurrence of each distinct path within the ensemble of generated paths. Paths are sorted according to this value in order to reveal the top most probable paths in the ensemble.

2.3.2.2. Residue Statistics. Statistics of the data of all the amino acid residues in the paths are calculated. The calculated parameters include; the frequency of the

appearance of residues, average positioning of the residues within the paths and average step size of the paths the residue appears in.

2.3.2.3. Path Probability. Probability of each path is calculated. The probabilities are calculated as the paths were generated, through the dynamic probability matrix, P, by simply multiplying the probabilities of each step.

$$PP(n) = \prod_{i=1}^{S-1} P_{i-i+1} \tag{2.24}$$

n denotes the specified path, S denotes the number of steps, and P_{i-i+1} denotes the probability between residues of steps i and $i+1$.

2.4. Network Parameters

Any network has the properties as: clustering coefficient, closeness and betweenness. Since proteins are defined as network of amino acid residues, the mentioned parameters can be calculated for these protein networks. The characteristics of networks are that they are both weighted (i.e. edges are not equal) and directed (i.e. probability forward does not equal backward).

Some of these network properties require the shortest paths between all of the pairs of vertices (residues). The shortest paths are extracted from the generated infinite step path (ISP) (see section 2.3.1). In a network of N vertices, using the information of shortest paths between any residue i and j; the step length of the shortest paths l_{ij}, the number of different shortest paths g_{ij}, the number of different shortest paths visiting residue k g_{ikj} are calculated for all residues, as elaborated below.

2.4.1. Clustering coefficient, C_v

Clustering coefficient is a measure of how clustered a vertex is in the network. It is defined as the ratio of number of edges between neighbors of residue v and the

number of total possible contacts between the neighbors of v. Unlike the other two parameters, clustering coefficient does not require any path generation and can easily be calculated from the connectivity of the network.

$$C_v = \frac{E_v}{d_v(d_v - 1)} \tag{2.25}$$

where d_v is the number of neighbors of v and E_v is the number of edges between these neighbors.

2.4.2. Closeness, O_i

Closeness is a measure of how close the vertex i is to all the other vertices in the network. It is defined as the reciprocal of the average lengths of shortest paths between vertex i and all the vertices.

$$O_i = \frac{N - 1}{\sum_{j=i}^{N} l_{ij}} \tag{2.26}$$

2.4.3. Betweenness, b_k

Betweenness is the need for the vertex k in connection of the vertices in the network. It is defined as the sum of the ratio of number of paths residue k is present in all the different shortest paths between all residue pairs.

$$b_k = \sum_{ij} \frac{g_{ikj}}{g_{ij}} \tag{2.27}$$

2.5. Folding Simulations

This is a computational method that makes use of robotic motion planning approach (Ulutas *et al.*, 2009) to predict the mechanism of protein folding dynamics.

Order constrained navigation applies very well to protein folding due to total correspondence of the parameters.

Order constrained navigation consists of a sequence of independent robot bodies p with defined radii ρ linked to each other with a fixed distance l, starting from an initial conformation w_0 with initial torsion angles s_0 moving towards a target conformation w_g with target torsion angles s_g in 3-D space (Ulutas *et al.*, 2009).

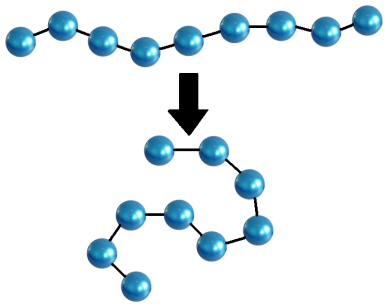

Figure 2.7. Order constrained navigation initial conformation w_0 and target conformation w_g representations.

A simplified energy model is used for the whole system with a novel approach of using multiplicative formulation instead of the classical additive formulation. Overall a single artificial potential function φ takes care of both collision free navigation and intrinsic geometry of the goal conformation. This function consists of three functions; sharpening function σ_d squashing function σ and intrinsic artificial potential function $\check{\varphi}(w)$ (Ulutas *et al.*, 2009).

$$\varphi(w) = \sigma_d \circ \sigma \circ \check{\varphi}(w) \tag{2.28}$$

These functions are defined as;

$$\sigma = \frac{x}{x+1} \tag{2.29}$$

$$\sigma_d = x^{1/k} \tag{2.30}$$

$$\check{\varphi}(w) = \frac{1000{\gamma_T}^k(w)}{\beta(w)} \tag{2.31}$$

The intrinsic artificial potential function $\breve{\varphi}(w)$ is the ratio of distance of torsion angles from the target conformation ${\gamma_T}^k(w)$ and the distance from free space boundary $\beta(w)$ in Cartesian coordinates.

$$ {\gamma_T}^k(w) = ||w - w_g||^2 \tag{2.32} $$

$$ \beta(w) = \prod \beta_{mn}(w) \tag{2.33} $$

The dynamics of the simulations adapts a first order model (Ulutas *et al.*, 2009). The initial condition is the initial conformation w_0 and the negative gradient of the potential function is used as the force acting on the system. The dynamics considers the contribution of each robot's potential function in driving the overall conformation geometry.

This method is applied to protein folding by defining the independent robot bodies as C_α atoms of amino acid residues in the protein sequence with virtual radii and by connecting them trough a link of virtual bonds and torsion angles (Ulutas *et al.*, 2009). The target conformation in this case would be the native state of the protein and the initial conformation would be the linear sequence of amino acid residues.

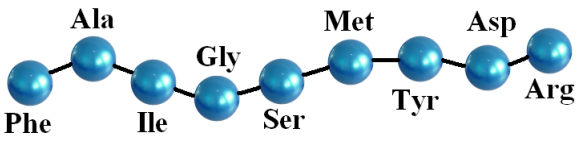

Figure 2.8. Representation of application to protein folding.

The user input parameters of the simulation are the target conformation of the protein (i.e. native state w_g) in PDB format, and the constant link length l and robot body (amino acid) radii ρ. Some restrictions apply to the latter two parameters such as $(\rho_i + \rho_j) < l$ and $2(\rho_i + \rho_j) > l$ so that collision of two amino acids i and j (first condition) and allowing a third amino acid in between the two amino acids (second condition) are avoided. Through simulations l is used as 3.8 Å and ρ is used as 1.0 Å.

2.5.1. Cluster Analysis

Results of the folding simulations give hundreds of conformations, and it is not possible nor necessary to analyze all these conformations in depth. By clustering the number of conformations to be analyzed can be reduced to reasonable amounts (i.e. 10 to 30 conformations). Cluster analysis is the grouping of objects according to a similarity measure, in this case the root mean square deviations (RMSD) between the conformations. The number of clusters is indirectly proportional to the the user-assigned cut-off RMSD value.

In this work, MMTSB Toolset's (Feig *et al.*, 2004) *kclust* utility is used to perform conformational clustering. It uses a high-performance clustering algorithm, k-means. The clustering proceeds by first randomly selecting a collection of frames and assigning them to their own clusters. Centroids for these clusters are calculated. The iteration continues over all other frames, until all frames are over, so that each frame is assigned to a cluster with the closest centroid, and at each step new centroids for each cluster are calculated.

3. RESULTS AND DISCUSSION

3.1. Paths With Different Potentials

The proposed path generation method in the present work is tested on a set of protein structures, which have previously been studied (Lockless and Ranganathan, 1999, Ota and Agard, 2005, Atilgan *et al.*, 2007). These studies report paths of residue communications on POZ and PDZ domains. PDB codes of the representative structures taken are 1A68 (Kreusch *et al.*, 1998) and 1BE9 (Doyle *et al.*, 1996). Paths between residues that are defined in the former studies are generated using each four of the potential functions. Instead of suggesting a single path as the *shortest path*, an ensemble of paths are analyzed to suggest a list of *probable paths* and when possible, identify the *most probable path*. For each ensemble top 20 paths are presented, according to their occurrence frequency in the ensemble of paths.

3.1.1. Paths of POZ domain representative (1A68)

This first case is taken to be more deterministic in potential selection, since the paths in literature for this case specifically give the connected sequence of the residues within the path, rather than a list of probable residues. These previously suggested paths are summarized in Table 3.1. The ensembles of paths generated here are compared to former in terms of residues that lie along the paths.

Table 3.1. Suggested paths in the literature for POZ domain

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

First, the ensemble of paths are generated with the atomistic potential (Section 2.2.1). From the results in Table 3.2, it can be seen that the longest path step-wise consists of six residues (five steps). Even though no limitations for the path length were introduced, the top probable paths only consist of short paths. Another important point is that it is very clearly seen that the shortest path is not necessarily the most probable path. The second path in the list is longer than the third and fourth, but is distinctly more frequently observed than they are. These two points together justify the use of the most probable paths instead of shortest paths.

Table 3.2. 1A68 Paths using atomistic potential (30,000)

Atomistic Paths						**Path Freq**
77 - A	118 - A	149 - A	148 - A			160
77 - A	118 - A	121 - A	149 - A	148 - A		43
77 - A	72 - A	113 - A	148 - A			30
77 - A	72 - A	149 - A	148 - A			22
77 - A	115 - A	114 - A	113 - A	148 - A		20
77 - A	118 - A	114 - A	113 - A	148 - A		20
77 - A	119 - A	118 - A	149 - A	148 - A		20
77 - A	75 - A	115 - A	113 - A	148 - A		19
77 - A	70 - A	118 - A	149 - A	148 - A		17
77 - A	71 - A	110 - A	111 - A	148 - A		17
77 - A	70 - A	111 - A	148 - A			16
77 - A	115 - A	113 - A	148 - A			16
77 - A	76 - A	71 - A	110 - A	111 - A	148 - A	14
77 - A	75 - A	115 - A	114 - A	113 - A	148 - A	13
77 - A	75 - A	72 - A	113 - A	148 - A		12
77 - A	71 - A	110 - A	111 - A	149 - A	148 - A	11
77 - A	76 - A	75 - A	72 - A	149 - A	148 - A	11
77 - A	75 - A	72 - A	149 - A	148 - A		10
77 - A	122 - A	121 - A	149 - A	148 - A		10
77 - A	76 - A	71 - A	72 - A	113 - A	148 - A	9

In these paths, there is a clear break between the most probable path and the second probable path, since it is observed four times more than the second path in the sample. This most probable path is the exact path proposed by Lockless and Ranganathan (1999) by an evolutionary method. Atilgan *et al.* (2007) report that their method was not able to regenerate this path and propose two other paths. The second probable path in the atomistic paths is exactly the same as one of these paths, and the top 20 also includes their second path. The probability rankings of the paths suggested by Atilgan *et al.* (2007) differ from the rankings obtained in these paths by the atomistic potentials. Overall, it is seen that the paths generated by the atomistic potential sample all the paths previously suggested in the previous studies (Lockless and Ranganathan, 1999, Atilgan *et al.*, 2007).

Table 3.3 presents the top 20 probable paths generated using BJ potential. The paths are quite different from those generated using the atomistic potential; such that some residues are never seen and some new residues are introduced into the paths. The paths are not exactly same as the paths in the literature, but only close to them. Some steps of the paths are either totally missing or include some other residues. This proves that the paths generated by different potential functions may differ, which implies the importance of the choice of the potential function used in describing the inter-residue interaction in path generation.

TD paths on the other hand are more similar to atomistic paths in terms of residues and path lengths. Again both of the paths suggested by Atilgan *et al.* (2007) are seen within the probable paths but this is the case, since TD is the potential that Atilgan *et al.* (2007) used for path generation. The path suggested by Lockless and Ranganathan (1999) is not seen among the probable paths. So in terms of regenerating the paths available in literature paths by atomistic potential covers the paths by TD potential and is superior.

Table 3.3. 1A68 Paths using BJ potential (30,000)

BJ Paths				**Path Freq**
77 - A	111 - A	148 - A		105
77 - A	73 - A	148 - A		53
77 - A	118 - A	148 - A		53
77 - A	117 - A	148 - A		44
77 - A	72 - A	148 - A		41
77 - A	114 - A	148 - A		41
77 - A	121 - A	148 - A		39
77 - A	92 - A	148 - A		28
77 - A	109 - A	148 - A		25
77 - A	110 - A	148 - A		25
77 - A	74 - A	111 - A	148 - A	15
77 - A	69 - A	101 - A	148 - A	11
77 - A	74 - A	112 - A	148 - A	11
77 - A	72 - A	111 - A	148 - A	9
77 - A	75 - A	112 - A	148 - A	9
77 - A	108 - A	111 - A	148 - A	9
77 - A	117 - A	111 - A	148 - A	9
77 - A	72 - A	101 - A	148 - A	8
77 - A	118 - A	111 - A	148 - A	7
77 - A	84 - A	92 - A	148 - A	7

The overall picture of paths generated using Markov Affinity show tendency to be longer, which is a draw back because even though we suggest that shortest paths are not necessarily the most probable paths, the negative effect of path length in the path probability is a fact. The path suggested by Lockless and Ranganathan (1999) is by far the most probable path here. It was the same case with the atomistic potential.

Considering the overall results obtained by the different potential functions, the atomistic potential eliminates the others in path generation due to; its ability to re-generate the paths available in the literature, sensibility in terms of path length and residues in the path. Therefore the selected potential for path generation is the atomistic potential.

Table 3.4. 1A68 Paths using TD potential (30,000)

TD Paths					**Path Freq**
77 - A	72 - A	111 - A	148 - A		49
77 - A	76 - A	148 - A			33
77 - A	118 - A	111 - A	148 - A		30
77 - A	122 - A	121 - A	149 - A	148 - A	29
77 - A	70 - A	111 - A	148 - A		20
77 - A	79 - A	148 - A			18
77 - A	75 - A	72 - A	111 - A	148 - A	15
77 - A	70 - A	92 - A	91 - A	148 - A	14
77 - A	75 - A	71 - A	148 - A		13
77 - A	69 - A	70 - A	111 - A	148 - A	13
77 - A	122 - A	118 - A	111 - A	148 - A	13
77 - A	122 - A	124 - A	149 - A	148 - A	11
77 - A	70 - A	92 - A	149 - A	148 - A	10
77 - A	118 - A	121 - A	149 - A	148 - A	10
77 - A	69 - A	76 - A	148 - A		9
77 - A	72 - A	113 - A	148 - A		8
77 - A	75 - A	76 - A	148 - A		8
77 - A	72 - A	118 - A	111 - A	148 - A	8
77 - A	118 - A	72 - A	111 - A	148 - A	8
77 - A	72 - A	110 - A	111 - A	148 - A	8

Table 3.5. 1A68 Paths using Markov Affinity (30,000)

Markov Paths								**Path Freq**
77 - A	118 - A	149 - A	148 - A					110
77 - A	75 - A	74 - A	73 - A	113 - A	148 - A			31
77 - A	118 - A	117 - A	114 - A	113 - A	148 - A			30
77 - A	118 - A	117 - A	116 - A	115 - A	114 - A	113 - A	148 - A	27
77 - A	76 - A	75 - A	74 - A	73 - A	113 - A	148 - A		22
77 - A	118 - A	117 - A	116 - A	114 - A	113 - A	148 - A		21
77 - A	76 - A	75 - A	74 - A	73 - A	112 - A	113 - A	148 - A	21
77 - A	75 - A	115 - A	114 - A	113 - A	148 - A			20
77 - A	75 - A	74 - A	73 - A	112 - A	113 - A	148 - A		20
77 - A	76 - A	75 - A	115 - A	114 - A	113 - A	148 - A		20
77 - A	76 - A	75 - A	72 - A	73 - A	113 - A	148 - A		17
77 - A	76 - A	71 - A	110 - A	111 - A	148 - A			16
77 - A	76 - A	74 - A	73 - A	113 - A	148 - A			16
77 - A	70 - A	71 - A	110 - A	111 - A	112 - A	113 - A	148 - A	16
77 - A	76 - A	71 - A	110 - A	111 - A	112 - A	113 - A	148 - A	16
77 - A	75 - A	72 - A	73 - A	112 - A	113 - A	148 - A		13
77 - A	70 - A	71 - A	110 - A	111 - A	148 - A			12
77 - A	75 - A	72 - A	73 - A	113 - A	148 - A			12
77 - A	118 - A	149 - A	150 - A	148 - A				11
77 - A	76 - A	71 - A	111 - A	112 - A	113 - A	148 - A		11

3.1.2. Paths of PDZ domain representative PSD-95 (1BE9)

The path generation for the PDZ domain representative is not as straight forward as the previous case, since the paths reported in the literature for the PDZ domain are not as consistent in terms of path residues. Also Lockless and Ranganathan (1999) do not suggest a connected path but rather give a set of residues that are most likely to be present in the paths. The first two paths in Table 3.6 are the same except insertion of two peptide residues. Lockless and Ranganathan (1999) do not actually propose two different paths, but report that the peptide residues also have a possibility to be in the path. On the other hand Ota and Agard (2005) propose a direct connected path.

Table 3.6. Suggested paths in the literature for PDZ domain

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

The set of paths generated using the atomistic potential (Table 3.7) do not contain the exact same paths as those listed in Table 3.6 which also show differences among them but there are very close similarities. The residues suggested by both Lockless and Ranganathan (1999) and Ota and Agard (2005) are seen through out the paths except for Ala347. This residue is present in the ensemble of paths (data not shown) but not in the top 20 probable paths. This should not mean that the atomistic potential at the present application fails to generate the right paths. On the contrary, there are paths that are very close to those proposed, with extra steps of residues in the vicinity such as Asn326, Ile328, and Phe340 or with exclusion of some steps. This shows that paths generated using atomistic potential are consistent with both methods used by Lockless and Ranganathan (1999) and Ota and Agard (2005), even though these methods are based on two different concepts evolution and anisotropic thermal diffusion.

Table 3.7. 1BE9 Paths using atomistic potential (40,000)

Atomistic Paths							**Path Freq**
372 - A	7 - P	325 - A	353 - A				14
372 - A	327 - A	338 - A	353 - A				11
372 - A	336 - A	337 - A	338 - A	353 - A			10
372 - A	327 - A	325 - A	353 - A				9
372 - A	336 - A	357 - A	354 - A	353 - A			8
372 - A	336 - A	357 - A	353 - A				7
372 - A	6 - P	326 - A	325 - A	353 - A			7
372 - A	6 - P	339 - A	338 - A	353 - A			7
372 - A	327 - A	359 - A	353 - A				6
372 - A	336 - A	359 - A	353 - A				6
372 - A	336 - A	338 - A	353 - A				5
372 - A	327 - A	326 - A	325 - A	353 - A			5
372 - A	6 - P	339 - A	340 - A	341 - A	353 - A		5
372 - A	327 - A	326 - A	340 - A	341 - A	353 - A		5
372 - A	328 - A	327 - A	326 - A	325 - A	353 - A		5
372 - A	327 - A	326 - A	338 - A	353 - A			4
372 - A	5 - P	6 - P	326 - A	338 - A	353 - A		4
372 - A	6 - P	326 - A	340 - A	341 - A	353 - A		4
372 - A	7 - P	8 - P	9 - P	325 - A	353 - A		4
372 - A	5 - P	6 - P	339 - A	340 - A	341 - A	353 - A	4

Occurrence of the top paths in 3.7 are very close to each other, therefore no path is distinguished as the most probable path. Some paths are totally made up of residues referred in the literature (e.g. first and fourth), and some paths are just modified versions of the proposed paths, such as neighboring residues replace each other. Some totally new paths along with new residues are introduced as well.

The paths generated using BJ potential are again very short (i.e. three residues two steps). Also the residues making up the paths are not close to those suggested by Lockless and Ranganathan (1999) or Ota and Agard (2005). Some of the key residues are seen rarely with some neighboring residues. So path generation using BJ potential is not close to being able to regenerate the evolutionary (Lockless and

Ranganathan, 1999) or simulation (Ota and Agard, 2005) path results.

As it was in the case of POZ domain, again the paths generated using TD potential are very close to those generated using the atomistic potential. Most paths include the residues of the paths suggested. Still these paths have the deficiency of short path length. A path made of all the residues suggested by Ota and Agard (2005) is not present among the top 20 possible paths, as it were in the atomistic paths, so again TD potential is not preferable over the atomistic potential in this case study too.

Table 3.8. 1BE9 Paths using BJ potential (40,000)

BJ Paths				**Path Freq**
372 - A	359 - A	353 - A		79
372 - A	325 - A	353 - A		69
372 - A	326 - A	353 - A		57
372 - A	338 - A	353 - A		54
372 - A	339 - A	353 - A		52
372 - A	337 - A	353 - A		40
372 - A	327 - A	353 - A		15
372 - A	379 - A	359 - A	353 - A	11
372 - A	379 - A	388 - A	353 - A	9
372 - A	9 - P	345 - A	353 - A	8
372 - A	326 - A	347 - A	353 - A	8
372 - A	326 - A	359 - A	353 - A	8
372 - A	8 - P	339 - A	353 - A	7
372 - A	8 - P	342 - A	353 - A	7
372 - A	325 - A	388 - A	353 - A	7
372 - A	378 - A	359 - A	353 - A	7
372 - A	378 - A	388 - A	353 - A	7
372 - A	8 - P	346 - A	353 - A	6
372 - A	325 - A	316 - A	353 - A	6
372 - A	338 - A	392 - A	353 - A	6

Table 3.9. 1BE9 Paths using TD potential (40,000)

TD					**Path**
Paths					**Freq**
372 - A	371 - A	353 - A			55
372 - A	327 - A	325 - A	353 - A		51
372 - A	373 - A	353 - A			48
372 - A	327 - A	338 - A	353 - A		41
372 - A	331 - A	353 - A			40
372 - A	328 - A	338 - A	353 - A		33
372 - A	7 - P	353 - A			30
372 - A	336 - A	338 - A	353 - A		22
372 - A	327 - A	9 - P	325 - A	353 - A	16
372 - A	373 - A	341 - A	353 - A		10
372 - A	376 - A	327 - A	325 - A	353 - A	10
372 - A	7 - P	341 - A	353 - A		8
372 - A	331 - A	314 - A	353 - A		8
372 - A	331 - A	325 - A	353 - A		8
372 - A	331 - A	341 - A	353 - A		8
372 - A	371 - A	314 - A	353 - A		8
372 - A	327 - A	338 - A	341 - A	353 - A	8
372 - A	376 - A	9 - P	325 - A	353 - A	8
372 - A	371 - A	338 - A	353 - A		7
372 - A	373 - A	347 - A	353 - A		7

Finally the paths generated using Markov Affinity, do not show any better qualities, in terms of path length and path residues, as was the case in POZ domain representative. The paths are too long (i.e. up to ten residues) and this length is caused by the repetitive connections made between neighbor residues. The paths are made up of both the residues suggested in the literature and some new additional residues. Even though this would make this potential function a good candidate, the presence of the peptide residues in almost each of the paths is not reasonable. Lockless and Ranganathan (1999) suggests that peptide residues may be in the path but it is not a must, and Ota and Agard (2005) do not see the peptide residues within their simulations.

Table 3.10. 1BE9 Paths using Markov Affinity (40,000)

Markov Paths										**Path Freq**
372 - A	5 - P	6 - P	339 - A	340 - A	341 - A	353 - A				33
372 - A	5 - P	6 - P	339 - A	355 - A	354 - A	353 - A				16
372 - A	5 - P	6 - P	339 - A	338 - A	353 - A					14
372 - A	5 - P	6 - P	339 - A	338 - A	355 - A	354 - A	353 - A			14
372 - A	7 - P	6 - P	339 - A	340 - A	341 - A	353 - A				11
372 - A	6 - P	339 - A	340 - A	341 - A	353 - A					9
372 - A	6 - P	339 - A	355 - A	354 - A	353 - A					9
372 - A	5 - P	6 - P	7 - P	8 - P	9 - P	325 - A	340 - A	341 - A	353 - A	8
372 - A	6 - P	339 - A	338 - A	353 - A						7
372 - A	5 - P	6 - P	7 - P	327 - A	326 - A	340 - A	341 - A	353 - A		6
372 - A	6 - P	339 - A	340 - A	341 - A	354 - A	353 - A				5
372 - A	329 - A	328 - A	339 - A	340 - A	341 - A	353 - A				5
372 - A	7 - P	8 - P	9 - P	325 - A	340 - A	341 - A	353 - A			5
372 - A	330 - A	329 - A	328 - A	327 - A	326 - A	338 - A	353 - A			5
372 - A	5 - P	6 - P	7 - P	8 - P	9 - P	325 - A	324 - A	341 - A	353 - A	5
372 - A	5 - P	6 - P	339 - A	338 - A	354 - A	353 - A				4
372 - A	5 - P	6 - P	328 - A	339 - A	340 - A	341 - A	353 - A			4
372 - A	6 - P	327 - A	326 - A	339 - A	340 - A	341 - A	353 - A			4
372 - A	7 - P	8 - P	9 - P	325 - A	326 - A	340 - A	341 - A	353 - A		4

When all the results are combined, as it was in the case of POZ domain, the atomistic potential function produces paths that are in good agreement with the results of previous studies at the most. This was expected because among all, the atomistic potential function is the most realistic potential, a physical based potential. It is based on van der Waals interactions, which are the main component of interactions within pathways (Ota and Agard, 2005). From this point on all the paths are generated by atomistic potential.

3.2. Paths Between Two Residues

Generation of paths between two residues (BTR) is the method with the most straight forward results. As seen in previous results (Section 3.1) paths of different

steps can be produced, presenting a variety of probable paths. A case study is done with HIV-1 protease structure (PDB:1F7A) between the catalytic residue Asp25 and the least fluctuating (hinge) residues 56 and 69, and most fluctuating residues 17 and 40 (Section 2.1.2). Since Asp25 is both binding and catalytic site, it is clear that if a signal would be carried through the protein, the origin would be this site. Hinge residues are chosen with the assumption that communication between them and active site is most likely to occur, since hinges are functional in protein dynamics (Ozer, 2008) and most fluctuating residues are chosen to check for presence of allosteric sites.

Table 3.11. Paths of ca-p2 (1F7A) BTR: 25-A and 56-A

BTR: 25-A and 56-A (1F7A)							**Freq**
Paths							**Path**
25 - A	84 - A	32 - A	56 - A				6
25 - A	4 - P	47 - A	46 - A	56 - A			3
25 - A	4 - P	47 - A	56 - A				2
25 - A	5 - P	4 - P	32 - A	56 - A			2
25 - A	5 - P	48 - A	47 - A	56 - A			2
25 - A	23 - A	82 - A	80 - A	79 - A	56 - A		2
25 - A	24 - A	23 - A	82 - A	80 - A	79 - A	56 - A	2
25 - A	28 - A	4 - P	32 - A	56 - A			2
25 - A	28 - A	29 - A	30 - A	2 - P	47 - A	56 - A	2
25 - A	28 - A	30 - A	45 - A	56 - A			2
25 - A	84 - A	32 - A	47 - A	56 - A			2
25 - A	84 - A	32 - A	76 - A	56 - A			2
25 - A	84 - A	33 - A	32 - A	56 - A			2
25 - A	84 - A	80 - A	79 - A	56 - A			2
25 - A	85 - A	84 - A	33 - A	77 - A	56 - A		2
25 - A	85 - A	84 - A	80 - A	79 - A	56 - A		2
25 - A	86 - A	31 - A	32 - A	56 - A			2

Val56 is one of the proposed (Ozer, 2008) hinge residues. Paths between the active site residue 25 and residue 56 (Table 3.11 often go through the peptide residues 4 - P and 5 - P (corresponding to P2 and P1 respectively). Another clear pattern is the connection of residues around 84 and 32. Also the consecutive steps of the paths

mostly consist of long distance contacts rather than neighbor residues. Overall the paths are 4-5 step paths, considered as short paths.

Table 3.12. Paths of ca-p2 (1F7A) BTR: 25-A and 69-A

BTR: 25-A and 69-A (1F7A)						**Freq**
Paths						**Path**
25 - A	24 - A	66 - A	69 - A			11
25 - A	24 - A	99 - B	69 - A			10
25 - A	90 - A	93 - A	69 - A			9
25 - A	24 - A	66 - A	67 - A	69 - A		8
25 - A	85 - A	66 - A	69 - A			6
25 - A	85 - A	66 - A	67 - A	69 - A		5
25 - A	85 - A	66 - A	68 - A	69 - A		4
25 - A	24 - A	11 - A	12 - A	67 - A	69 - A	3
25 - A	24 - A	11 - A	67 - A	69 - A		3
25 - A	24 - A	66 - A	68 - A	69 - A		3
25 - A	24 - A	85 - A	66 - A	69 - A		3
25 - A	24 - A	99 - B	1 - A	69 - A		3
25 - A	26 - A	24 - A	66 - A	69 - A		3
25 - A	23 - A	24 - A	99 - B	69 - A		2
25 - A	24 - A	11 - A	66 - A	69 - A		2
25 - A	24 - A	66 - A	70 - A	69 - A		2
25 - A	24 - A	97 - B	99 - B	69 - A		2
25 - A	90 - A	94 - A	93 - A	69 - A		2
25 - A	24 - A	66 - A	67 - A	68 - A	69 - A	2
25 - A	90 - A	91 - A	94 - A	93 - A	69 - A	2

The other hinge residue selected as a final residue for the paths is His69. Paths generated BTR 25 and 69 (Table 3.12) also show patterns of specific residues. The most obvious pattern is the triangular interactions between residues 24-66-85. Even though its rare, residues of the second monomer are present in the paths (i.e. 99 - B and 97 - B). Again paths consist of long distance residue connections and path lengths can be considered as short.

Table 3.13. Paths of ca-p2 (1F7A) BTR: 25-A and 17-A

BTR: 25-A and 17-A (1F7A)								**Freq**
Paths								**Path**
25 - A	24 - A	11 - A	12 - A	19 - A	17 - A			5
25 - A	24 - A	66 - A	65 - A	14 - A	17 - A			5
25 - A	85 - A	66 - A	65 - A	14 - A	17 - A			5
25 - A	23 - A	21 - A	19 - A	17 - A				4
25 - A	23 - A	21 - A	20 - A	19 - A	18 - A	17 - A		4
25 - A	24 - A	66 - A	14 - A	17 - A				4
25 - A	84 - A	85 - A	13 - A	18 - A	17 - A			4
25 - A	23 - A	22 - A	20 - A	18 - A	17 - A			3
25 - A	24 - A	11 - A	10 - A	21 - A	20 - A	19 - A	17 - A	3
25 - A	24 - A	11 - A	13 - A	18 - A	17 - A			3
25 - A	24 - A	22 - A	21 - A	19 - A	17 - A			3
25 - A	84 - A	83 - A	20 - A	18 - A	17 - A			3
25 - A	85 - A	13 - A	14 - A	16 - A	17 - A			3
25 - A	85 - A	13 - A	15 - A	16 - A	17 - A			3
25 - A	85 - A	13 - A	15 - A	17 - A				3
25 - A	85 - A	13 - A	19 - A	17 - A				3
25 - A	85 - A	13 - A	20 - A	19 - A	18 - A	17 - A		3
25 - A	85 - A	13 - A	14 - A	17 - A				2
25 - A	85 - A	13 - A	18 - A	17 - A				2
25 - A	23 - A	11 - A	12 - A	19 - A	17 - A			2

Overall paths ending with the hinge residues showed similar characteristics in terms of path length and the residues within. Paths ending with one of the most fluctuating residues Gly17 do not show these characteristics. First of all the paths are a lot longer and mostly consist of the pairs of the sequential neighbors located on the same loop with Gly17. Interestingly the same pattern of triangular 24-66-85 interactions is also present in this set of paths.

Finally the paths ending with the another highly fluctuating residue Gly40 show similar characteristics (Table3.14) with the former such as the extra long path lengths and step by step communication through sequentially close residues. The pattern of communication around residues 84 and 32 are also present here (as was present in paths

between 25 and 56 listed in Table 3.11).

Table 3.14. Paths of ca-p2 (1F7A) BTR: 25-A and 40-A

BTR: 25-A and 40-A (1F7A)									**Freq**
Paths									**Path**
25 - A	28 - A	30 - A	45 - A	44 - A	43 - A	41 - A	40 - A		4
25 - A	28 - A	30 - A	45 - A	44 - A	43 - A	42 - A	41 - A	40 - A	2
25 - A	28 - A	30 - A	45 - A	76 - A	59 - A	40 - A			2
25 - A	84 - A	32 - A	76 - A	75 - A	59 - A	38 - A	39 - A	40 - A	2
25 - A	84 - A	33 - A	34 - A	36 - A	37 - A	39 - A	40 - A		2
25 - A	85 - A	31 - A	75 - A	77 - A	36 - A	37 - A	38 - A	40 - A	2
25 - A	85 - A	84 - A	33 - A	34 - A	36 - A	38 - A	39 - A	40 - A	2
25 - A	85 - A	84 - A	33 - A	36 - A	38 - A	39 - A	40 - A		2
25 - A	90 - A	93 - A	89 - A	73 - A	59 - A	38 - A	40 - A		2

Overall, two specific patterns of residue interactions are observed in the paths generated between the active site and least/most fluctuating residues. Here it is suggested that if these end residues are important in communication within the protease in terms of dynamics, then these residues are key residues setting the backbone of communication.

3.3. Paths with Specific Number of Steps (PSNS)

Information on the allosteric site of proteins is not always available (as in the case of HIV-1 protease), preventing the analysis of path generation BTR. At this point, paths with specific number of steps (PSNS) are generated, with the suggestion that important sites involved in allostery can be determined by generating open ended paths. In order to see the effectiveness of the method, first a testing structure with known sites (POZ domain representative, Shaker potassium channel) is studied.

Ten residue (nine step) paths are generated with PSNS method starting from Phe77, the site at the interaction surface (Lockless and Ranganathan, 1999). Studying the actual paths did not reveal any specific results (data not shown) therefore residue

frequencies for each step (shown in different colors) of the path are plotted on a single graph (Figure 3.1). The most obviously crowded region is around the starting residue as expected. The second crowded region is through 116-126, a region covering residues 118,121 and 122, which were the first and second steps of the paths suggested in literature (Section 3.1.1). The final steps of those paths, residues 149 and 148 are also in an amplified region within residues 146-151. Frequencies of this region are not as high as the former, but this is due to the increase in the step number.

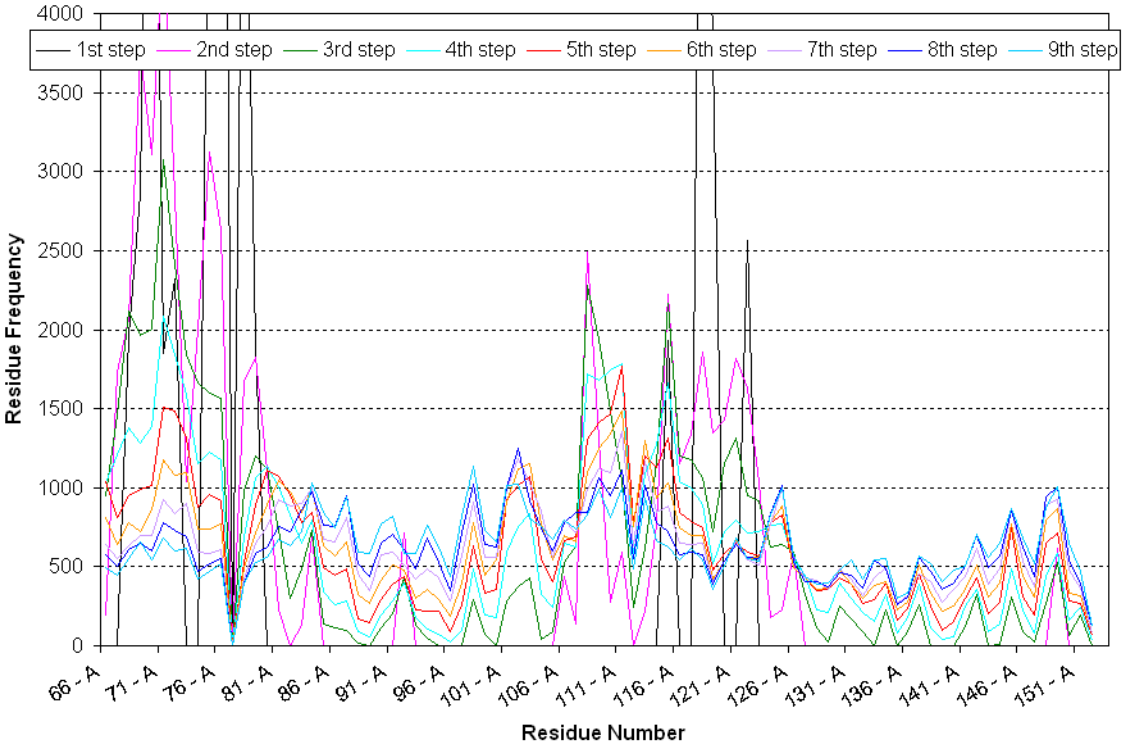

Figure 3.1. Residue frequencies of 1A68 10-residue (9-step) paths starting from 77-A. Residue frequencies of each step are shown in different color line summarized in the legend.

Other regions around the height of 146-151 are also present in the path. One of which is through residues 106-116. This region covers the residues present in paths suggested by this study (Section 3.1.1). One final region that attracts attention is through residues 96-106. These residues are neither present in the paths suggested in literature nor this study. This may suggest that there is another site in communication with Phe77.

These results show that important sites in communication with a known residue (i.e. active site, binding site) may be observed through residue frequency analysis on PSNS. HIV-1 protease here is a case study for that. Again ten residue (nine step) paths are generated, using the structure of ca-p2 (1F7A) starting from the catalytic residue Asp25 and the residue frequency plot is prepared (Figure 3.2).

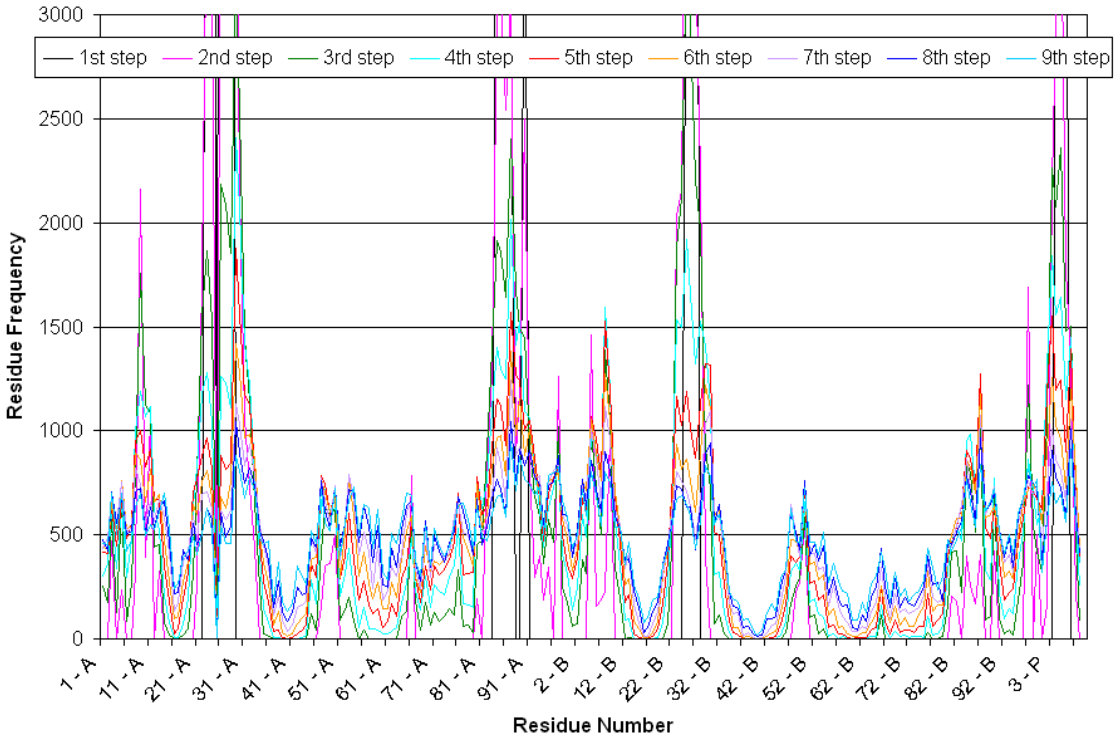

Figure 3.2. Residue frequencies of ca-p2 (1F7A) 10-residue (9-step) paths starting from 25-A. Residue frequencies of each step are shown in different colored line as shown in the legend.

Four main regions are observable on the plot through residues 23-31 A, 84-91 A, 23-28 B and 3-7 P. These represent the first and second steps of any path from Asp25. Residues in the sequential neighborhood of 25 (regions 23-31 A and 23-28 B) were expected. The other two regions are important in construction of any path from Asp25. Most of the residues suggested to be important in previous paths (Section 3.2), such as 24 - A, 84 - A, 85 - A , 4 - P and 5 - P, are present in these regions.

Rest of the observable peaks in the graph are single residues rather than regions. The second set, which probably constructs third and fourth steps, includes residues 9 A, 97 A, 5 B, 9 B, 87 B, 97 B. Residues 97 A and B were observed in the previous paths BTR, but the others are newly introduced. Presence of these residues as peaks might suggest the existence of another site in communication with Asp25.

The third set of peaks, that are relatively lower and probably make up the last few steps of the paths consist of residues 47 A, 50 A, 53 A, 54 A and 66 A. The presence of residue 66 completes the important patterns suggested. Again the other residues might suggest the existence of another site, but overall it is seen that the paths generated BTR ending with residues 56, 69, 17 and 40 were significant, because the same patterns are observed without any force determining the direction of the paths. The information from residue Asp25 naturally flows through these patterns.

3.4. Infinite Step Paths (ISP) and Network Parameters

Communication paths without specific end residues studied in the previous section brings to attention the case with one more missing parameter, the starting residue. In order to study that, infinite step paths (ISP) are generated. The algorithm requires a starting residue but when a path of infinite steps is generated, it does not matter what the starting residue was. The ISP generated is analyzed by calculating the network parameters.

The first case again is Shaker potassium channel (1A68) to see the effectiveness of the method. An ISP of 200,000 steps is generated and the parameters (i.e. clustering coefficient, closeness and betweenness) are calculated for each residue. In order to see the complete protein, the parameters for each residue are plotted. Clustering coefficient (Figure 3.3), closeness (Figure 3.4) and betweenness (Figure 3.5) are shown below.

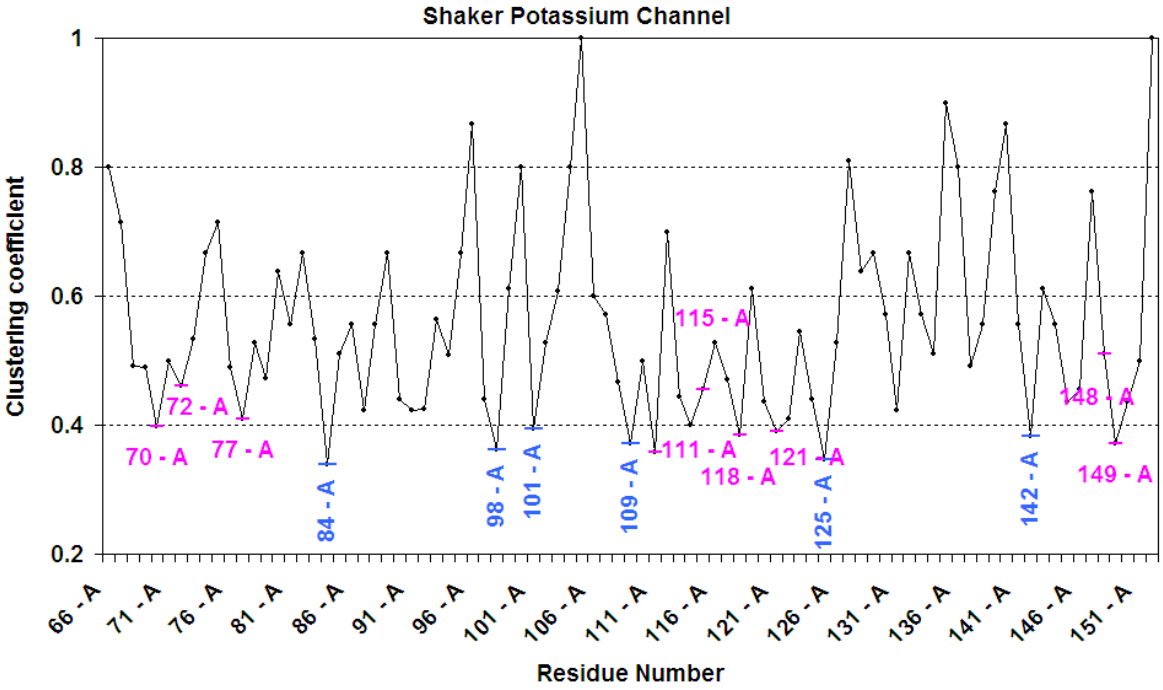

Figure 3.3. Plot of clustering coefficient values of Shaker potassium channel (1A68). Residues labeled in pink show the resdiues present in paths BTR and PSNS and labeled in blue show other residues corresponding to minimum points.

It is obvious that the minimums of the clustering coefficient graph (Figure 3.3) present all the residues suggested to construct a path (shown in pink). So it is suggested here that the minimums of the clustering coefficient correspond to important residues. Apart from those residues present in the paths, there are other minimums. It is proposed that the latter residues; 84, 98, 101, 109, 125 and 142, may have unknown functions regarding protein dynamics such as folding or binding.

In the case of closeness and betweenness parameters, proposed path residues correspond to the peaks of the graphs (Figure 3.4 and Figure 3.5 respectively). The peaks other than the path residues, again are proposed as important residues. So all three network parameters calculated, could clearly predict the path residues including the start and end residues for the Shaker potassium channel.

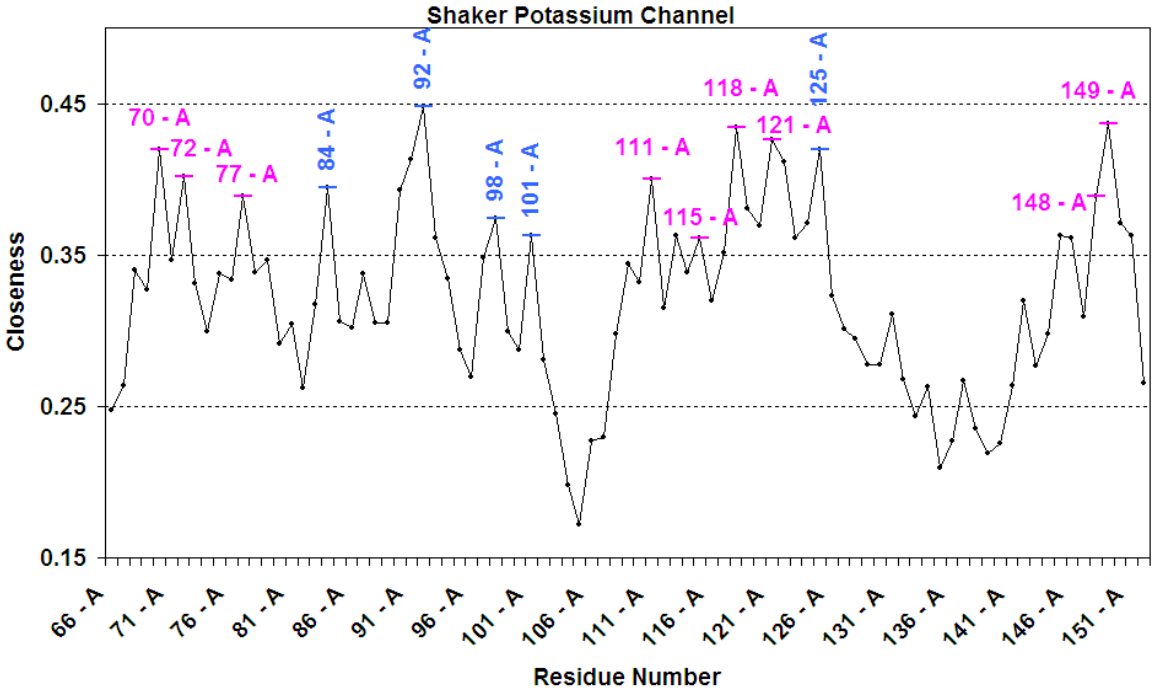

Figure 3.4. Closeness values of Shaker potassium channel (1A68). Residues labeled in pink show the resdiues present in paths BTR and PSNS and labeled in blue show other residues corresponding to minimum points.

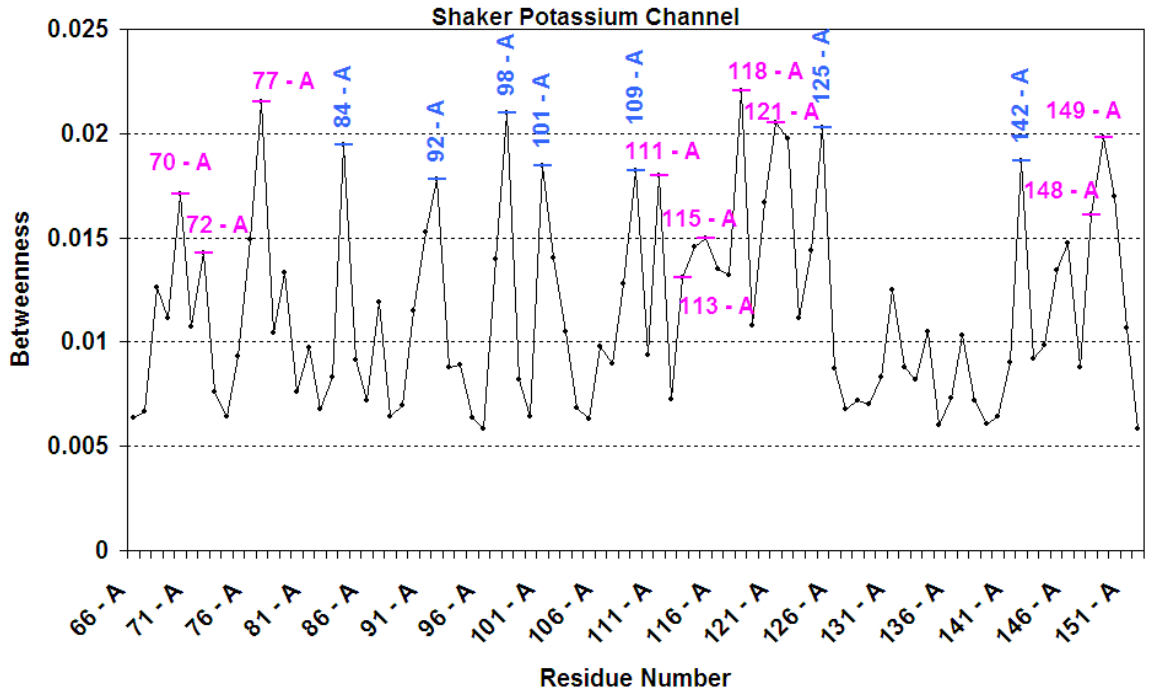

Figure 3.5. Betweenness values of Shaker potassium channel (1A68). Residues labeled in pink show the resdiues present in paths BTR and PSNS and labeled in blue show other residues corresponding to minimum points.

Next the same parameters are calculated for HIV-1 protease structure (1F7A). The minimums of the clustering coefficient and the peaks of closeness and betweenness parameters are studied to identify the functional residues of HIV-1 protease. Clustering coefficient results (Figure 3.6), emphasize residues 50, 66, 87 and 97 of both monomers and residues 5 and 6 of the substrate (i.e. cleavage) all shown in pink. The catalytic sites 25 A and B correspond to a local maximum. The other residues mentioned in previous sections are also labeled, but do not correspond to any specific site on the graphs.

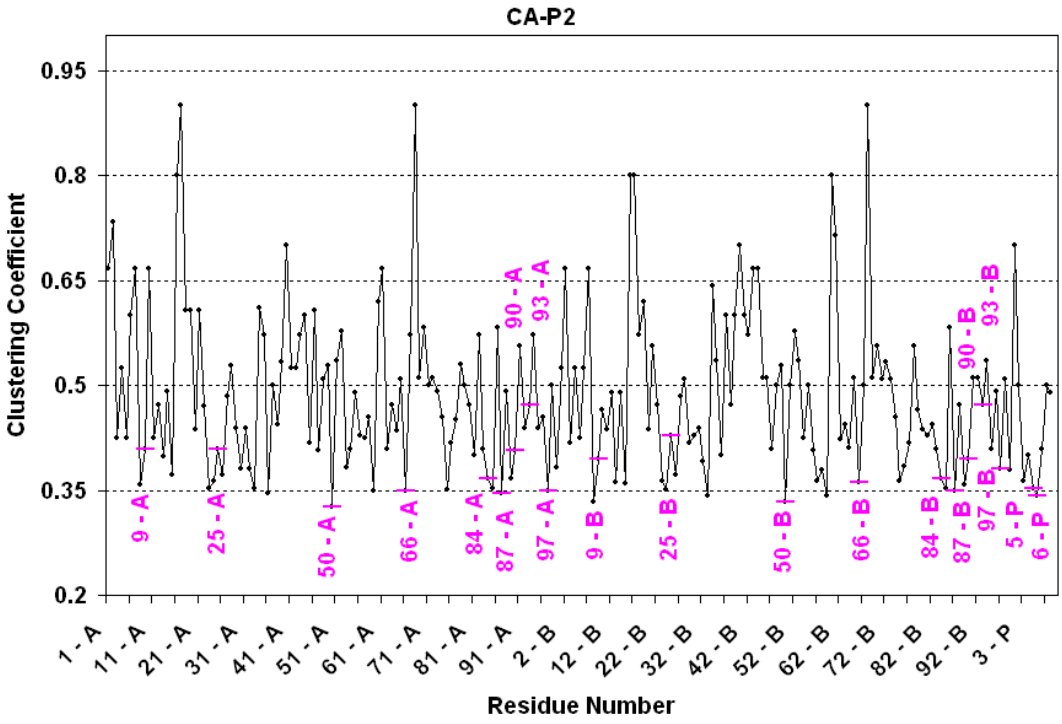

Figure 3.6. Clustering coefficient values of ca-p2 (1F7A). Residues labeled in pink show the resdiues present in paths and other minimums.

Results of closeness and betweenness parameters (Figures 3.7 and 3.8) on the other hand clearly emphasize residues 25A, 25B and 5 and 6P, which are in act functionally most important residues; the catalytic site and substrate cleavage. Residues corresponding to other relatively lower minimums (shown in pink) are proposed as important residues, most of which correspond to functional residues proposed by PDBsum database (Laskowski, 2000). The residues seen in the paths suggested in previous sections are also shown and they mostly correspond to local minimums.

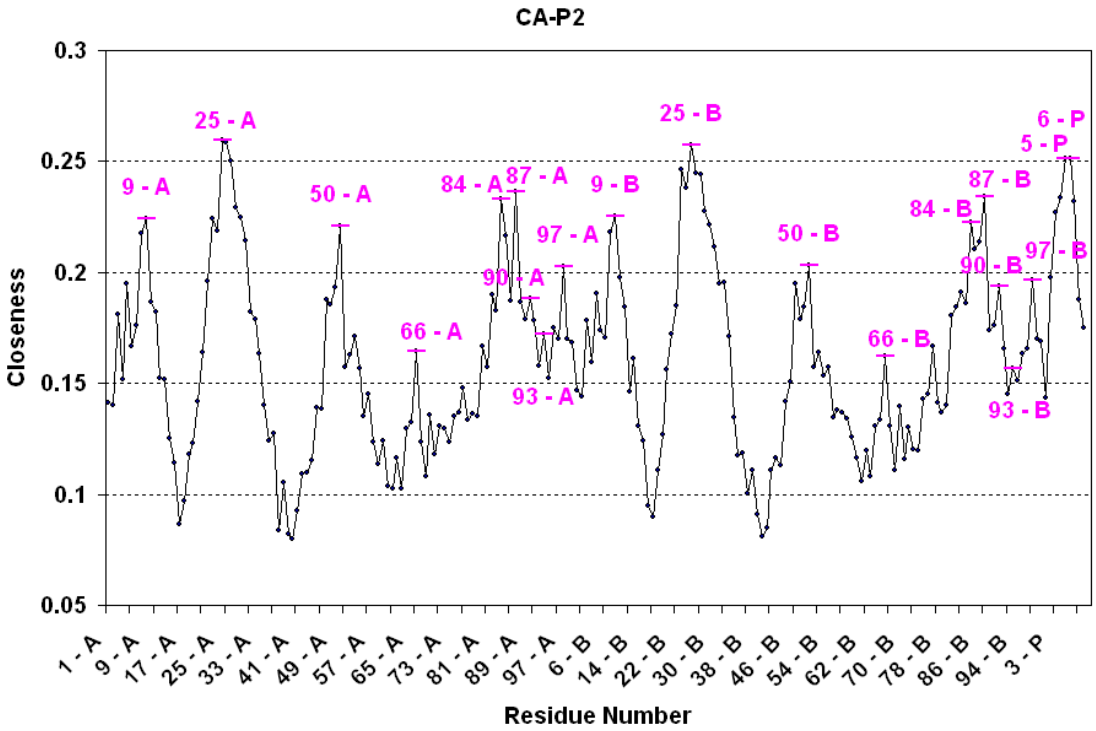

Figure 3.7. Plot of closeness values of ca-p2 (1F7A). Residues labeled in pink show the resdiues present in paths BTR and PSNS and other peaks.

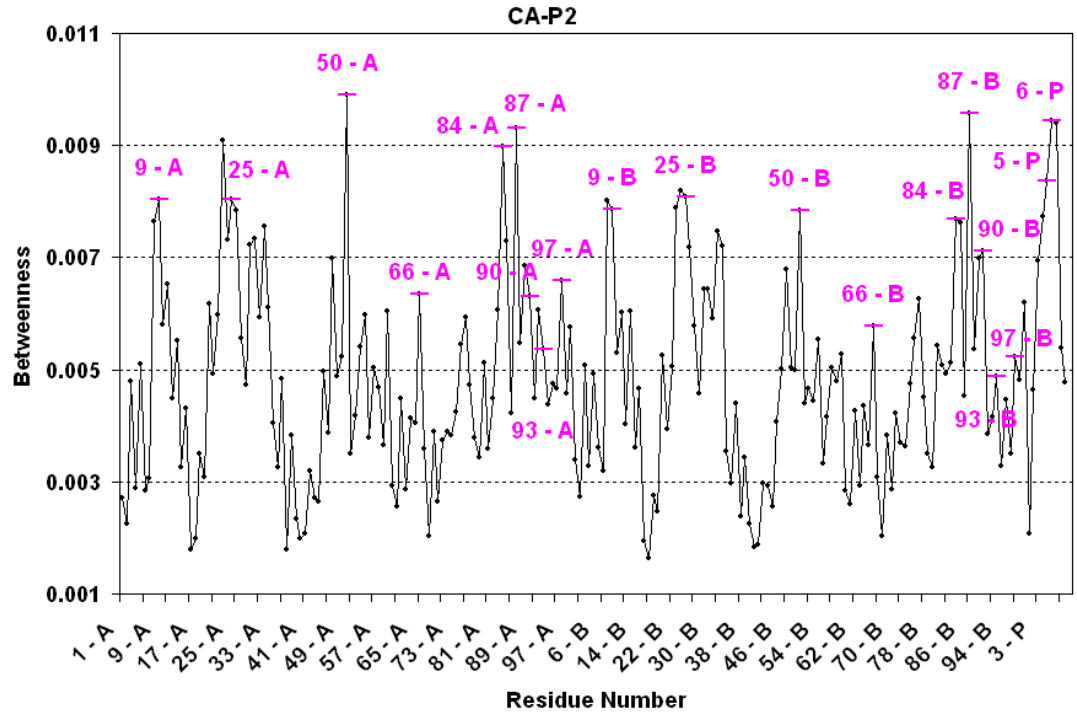

Figure 3.8. Plot of betweenness values of ca-p2 (1F7A). Residues labeled in pink show the resdiues present in paths BTR and PSNS and other peaks.

So these results overall suggest that the network parameters can predict functionally important residues in proteins. Some of these residues were present in the paths presented in the previous sections, suggesting their efficiency.

3.5. Conservation of Paths through all HIV-1 protease complex structures

In this study so far communication pathways of proteins are generated on a single structure, but there are questions regarding how the paths would be affected when conformational changes occur in the structure. To study this, here paths are generated BTR 25 and 69 for all the listed complex structures of HIV-1 protease (Table 2.2). Ending residue is chosen as 69 because, it has been proposed to be an allosteric site for HIV-1 protease complex structures (Lindgren, 2004).

Paths of ca-p2 (1F7A) (Table 3.15) are presented again for easy comparison. The paths of all structures are generated under same conditions and are shown below (Table 3.15-3.21). The first observation is that the most probable path in all structures are exactly the same, with different frequencies. Three other three step paths are also present in all the structures in top six, but their rankings differ from structure to structure. Ranking of all these paths are the same only in two structures 1KJF and 1KJH. There is an extra three step path in 1TSU that is not present in any other structure.

The closeness values of all structures align nearly perfectly, except a few residues. This was expected after the results obtained from the paths BTR. The deviations in the lines present the 3-D deviations of the structures.

Both of the results (paths BTR and network parameters) clearly show that communication within the protein is conserved through different substrate complexes. If the communication pathways are necessary for the protein to function as proposed, since all the structures are known to function without problems the expectation was conservation of the paths. Lockless and Ranganathan (1999) had also suggested that paths occurred in the early stages of evolution and were conserved throughout.

Table 3.15. Paths BTR 25-69 on ca-p2 (1F7A) structure.

BTR: 25-A and 69-A (1F7A)						**Path**
Paths						**Freq**
25 - A	24 - A	66 - A	69 - A			11
25 - A	24 - A	99 - B	69 - A			10
25 - A	90 - A	93 - A	69 - A			9
25 - A	24 - A	66 - A	67 - A	69 - A		8
25 - A	85 - A	66 - A	69 - A			6
25 - A	85 - A	66 - A	67 - A	69 - A		5
25 - A	85 - A	66 - A	68 - A	69 - A		4
25 - A	24 - A	11 - A	12 - A	67 - A	69 - A	3
25 - A	24 - A	11 - A	67 - A	69 - A		3
25 - A	24 - A	66 - A	68 - A	69 - A		3
25 - A	24 - A	85 - A	66 - A	69 - A		3
25 - A	24 - A	99 - B	1 - A	69 - A		3
25 - A	26 - A	24 - A	66 - A	69 - A		3

Table 3.16. Paths BTR 25-69 on ma-ca (1KJ4) structure.

BTR: 25-A and 69-A (1KJ4)						**Path**
Paths						**Freq**
25 - A	24 - A	66 - A	69 - A			22
25 - A	85 - A	66 - A	69 - A			15
25 - A	24 - A	99 - B	69 - A			8
25 - A	24 - A	66 - A	68 - A	69 - A		8
25 - A	24 - A	66 - A	67 - A	69 - A		7
25 - A	90 - A	93 - A	69 - A			6
25 - A	85 - A	66 - A	67 - A	69 - A		6
25 - A	24 - A	66 - A	65 - A	69 - A		5
25 - A	24 - A	11 - A	66 - A	69 - A		4
25 - A	24 - A	11 - A	67 - A	69 - A		4
25 - A	24 - A	66 - A	67 - A	68 - A	69 - A	4
25 - A	84 - A	85 - A	66 - A	69 - A		4
25 - A	23 - A	24 - A	66 - A	67 - A	69 - A	3
25 - A	24 - A	85 - A	66 - A	67 - A	69 - A	3

Table 3.17. Paths BTR 25-69 on nc-p1 (1TSU) structure.

BTR: 25-A and 69-A (1TSU) Paths					**Path Freq**
25 - A	24 - A	66 - A	69 - A		20
25 - A	85 - A	66 - A	69 - A		13
25 - A	24 - A	99 - B	69 - A		12
25 - A	24 - A	93 - A	69 - A		10
25 - A	90 - A	93 - A	69 - A		5
25 - A	90 - A	91 - A	93 - A	69 - A	4
25 - A	85 - A	24 - A	93 - A	69 - A	3
25 - A	86 - A	85 - A	66 - A	69 - A	3
25 - A	90 - A	89 - A	71 - A	69 - A	3
25 - A	90 - A	92 - A	70 - A	69 - A	3
25 - A	90 - A	93 - A	66 - A	69 - A	3

Table 3.18. Paths BTR 25-69 on p1-p6 (1KJF) structure.

BTR: 25-A and 69-A (1KJF) Paths					**Path Freq**
25 - A	24 - A	66 - A	69 - A		15
25 - A	85 - A	66 - A	69 - A		9
25 - A	24 - A	99 - B	69 - A		7
25 - A	90 - A	93 - A	69 - A		5
25 - A	23 - A	24 - A	66 - A	69 - A	5
25 - A	24 - A	66 - A	67 - A	69 - A	5
25 - A	85 - A	66 - A	68 - A	69 - A	5
25 - A	23 - A	24 - A	99 - B	69 - A	3
25 - A	24 - A	66 - A	65 - A	69 - A	3
25 - A	24 - A	99 - B	1 - A	69 - A	3
25 - A	26 - A	85 - A	66 - A	69 - A	3
25 - A	84 - A	85 - A	66 - A	69 - A	3
25 - A	85 - A	66 - A	67 - A	69 - A	3

Table 3.19. Paths BTR 25-69 on p2-nc (1KJ7) structure.

BTR: 25-A and 69-A (1KJ7)						**Path**
Paths						**Freq**
25 - A	24 - A	66 - A	69 - A			22
25 - A	85 - A	66 - A	69 - A			12
25 - A	90 - A	93 - A	69 - A			7
25 - A	24 - A	99 - B	69 - A			6
25 - A	23 - A	24 - A	66 - A	69 - A		4
25 - A	24 - A	66 - A	67 - A	69 - A		4
25 - A	90 - A	93 - A	66 - A	69 - A		4
25 - A	24 - A	66 - A	65 - A	68 - A	69 - A	3
25 - A	24 - A	90 - A	93 - A	69 - A		3
25 - A	24 - A	97 - B	98 - B	99 - B	69 - A	3
25 - A	85 - A	66 - A	65 - A	69 - A		3
25 - A	85 - A	66 - A	67 - A	68 - A	69 - A	3
25 - A	90 - A	95 - A	94 - A	93 - A	69 - A	3

Table 3.20. Paths BTR 25-69 on rh-in (1KJH) structure.

BTR: 25-A and 69-A (1KJH)						**Path**
Paths						**Freq**
25 - A	24 - A	66 - A	69 - A			18
25 - A	85 - A	66 - A	69 - A			10
25 - A	24 - A	99 - B	69 - A			8
25 - A	90 - A	93 - A	69 - A			5
25 - A	24 - A	66 - A	67 - A	68 - A	69 - A	5
25 - A	23 - A	24 - A	66 - A	69 - A		4
25 - A	24 - A	66 - A	68 - A	69 - A		4
25 - A	26 - A	24 - A	66 - A	69 - A		4
25 - A	24 - A	85 - A	66 - A	69 - A		3
25 - A	85 - A	66 - A	68 - A	69 - A		3
25 - A	90 - A	92 - A	93 - A	69 - A		3
25 - A	90 - A	94 - A	93 - A	69 - A		3

Table 3.21. Paths BTR 25-69 on rt-rh (1KJG) structure.

BTR: 25-A and 69-A (1KJG) Paths							**Path Freq**
25 - A	24 - A	66 - A	69 - A				18
25 - A	24 - A	99 - B	69 - A				13
25 - A	85 - A	66 - A	69 - A				9
25 - A	90 - A	93 - A	69 - A				7
25 - A	85 - A	64 - A	69 - A				5
25 - A	24 - A	99 - B	1 - A	69 - A			4
25 - A	26 - A	24 - A	66 - A	69 - A			4
25 - A	23 - A	22 - A	11 - A	67 - A	69 - A		3
25 - A	24 - A	66 - A	67 - A	68 - A	69 - A		3
25 - A	24 - A	66 - A	68 - A	69 - A			3
25 - A	24 - A	97 - B	2 - A	1 - A	99 - B	69 - A	3
25 - A	26 - A	90 - A	93 - A	69 - A			3
25 - A	84 - A	85 - A	66 - A	67 - A	69 - A		3
25 - A	84 - A	85 - A	90 - A	93 - A	69 - A		3
25 - A	90 - A	91 - A	92 - A	93 - A	69 - A		3

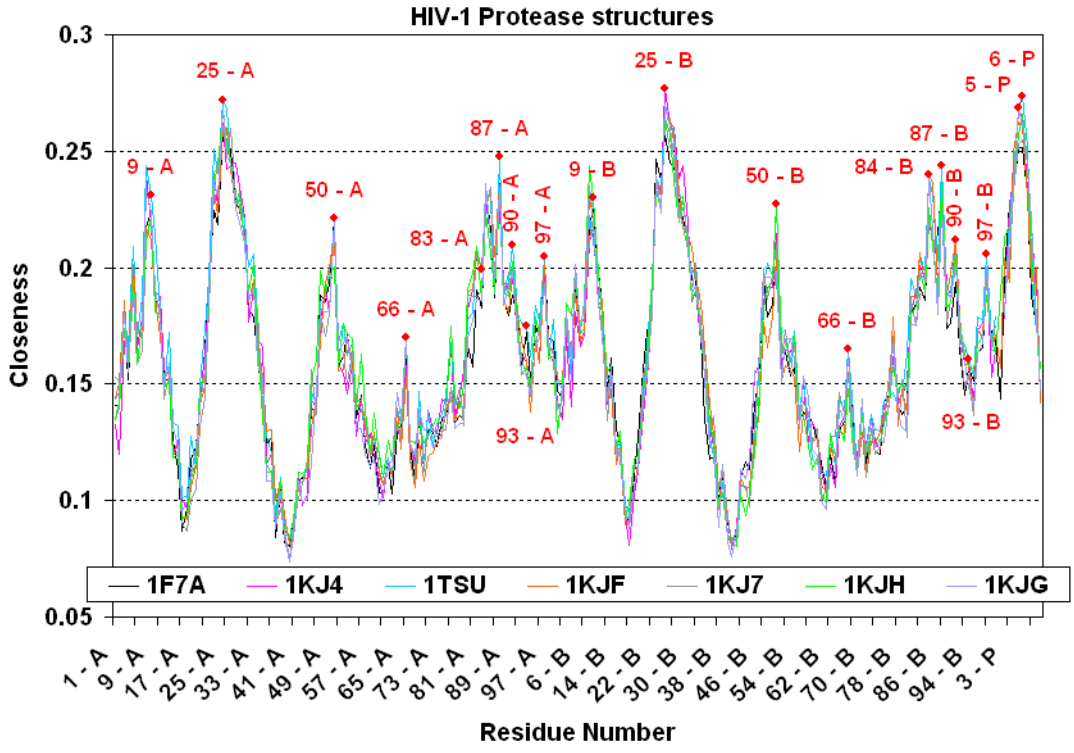

Figure 3.9. Closeness parameters of all HIV-1 protease complex structures plotted in a single graph. Each colored line denotes a different structure.

3.6. Paths Between Monomers of Structure

Another question regarding the nature of paths is the communication between and within different monomers of a protein. Again here the studies are carried on HIV-1 protease complex structure ca-p2 (1F7A), and paths BTW 25 and 69 are generated for all monomer pairs A-A, B-B, A-B and B-A (Tables 3.22-3.25).

Table 3.22. CA-P2 paths BTR 25 A - 69 A

BTR: 25-A and 69-A (1F7A)						**Path**
Paths						**Freq**
25 - A	24 - A	66 - A	69 - A			11
25 - A	24 - A	99 - B	69 - A			10
25 - A	90 - A	93 - A	69 - A			9
25 - A	24 - A	66 - A	67 - A	69 - A		8
25 - A	85 - A	66 - A	69 - A			6
25 - A	85 - A	66 - A	67 - A	69 - A		5
25 - A	85 - A	66 - A	68 - A	69 - A		4
25 - A	24 - A	11 - A	12 - A	67 - A	69 - A	3
25 - A	24 - A	11 - A	67 - A	69 - A		3
25 - A	24 - A	66 - A	68 - A	69 - A		3
25 - A	24 - A	85 - A	66 - A	69 - A		3
25 - A	24 - A	99 - B	1 - A	69 - A		3
25 - A	26 - A	24 - A	66 - A	69 - A		3
25 - A	23 - A	24 - A	99 - B	69 - A		2
25 - A	24 - A	11 - A	66 - A	69 - A		2
25 - A	24 - A	66 - A	70 - A	69 - A		2
25 - A	24 - A	97 - B	99 - B	69 - A		2
25 - A	90 - A	94 - A	93 - A	69 - A		2
25 - A	24 - A	66 - A	67 - A	68 - A	69 - A	2

First the effect of asymmetry is studied by comparing the paths within monomer A and B (Table 3.22 and 3.23). Both paths consist of same residues that were proposed in the previous sections. There are a few exact paths but mostly the paths are different by single residues. So the structural asymmetry has not affected paths BTR 25-69.

Table 3.23. CA-P2 paths BTR 25 B - 69 B

BTR: 25-B and 69-B (1F7A)						**Path**
Paths						**Freq**
25 - B	85 - B	66 - B	69 - B			13
25 - B	24 - B	66 - B	69 - B			12
25 - B	90 - B	93 - B	69 - B			10
25 - B	24 - B	99 - A	69 - B			9
25 - B	24 - B	99 - A	1 - B	69 - B		6
25 - B	24 - B	90 - B	93 - B	69 - B		5
25 - B	23 - B	24 - B	99 - A	1 - B	69 - B	4
25 - B	24 - B	11 - B	67 - B	68 - B	69 - B	4
25 - B	24 - B	66 - B	68 - B	69 - B		3
25 - B	85 - B	66 - B	68 - B	69 - B		3
25 - B	90 - B	91 - B	92 - B	93 - B	69 - B	3
25 - B	24 - B	3 - B	99 - A	69 - B		2
25 - B	24 - B	66 - B	65 - B	69 - B		2
25 - B	26 - B	24 - B	99 - A	69 - B		2
25 - B	26 - B	90 - B	93 - B	69 - B		2
25 - B	26 - B	97 - A	1 - B	69 - B		2
25 - B	84 - B	85 - B	66 - B	69 - B		2
25 - B	85 - B	66 - B	67 - B	69 - B		2
25 - B	86 - B	90 - B	93 - B	69 - B		2
25 - B	90 - B	93 - B	70 - B	69 - B		2

A communication between the monomers is expected since the protein is active as a dimer. Paths between the monomers (Tables 3.24 and 3.25) by definition cannot be exact. Here rather than similarities in the paths, introduction of new residues are sought. These new residues would be important in connecting the two monomers. Analysis of the paths resulted in identification of new important residues 1, 3, 97 and 99. Residues 3 and 97 correspond to local peaks of closeness and betweenness parameters. Residue 97 was reported in PSNS analysis to suggest presence of other communication sites within the protein. So the suspected site could be the same site in the other monomer. Communication pathways generated between the monomers may be as important as any communication pathway.

Table 3.24. CA-P2 paths BTR 25 A - 69 B

BTR: 25-B and 69-A (1F7A)							**Path**
Paths							**Freq**
25 - B	25 - A	90 - A	93 - A	69 - A			3
25 - B	26 - B	24 - A	66 - A	69 - A			3
25 - B	27 - B	24 - A	66 - A	69 - A			3
25 - B	25 - A	24 - A	66 - A	69 - A			2
25 - B	25 - A	86 - A	85 - A	66 - A	69 - A		2
25 - B	26 - A	25 - A	24 - A	99 - B	69 - A		2
25 - B	26 - B	24 - A	99 - B	1 - A	69 - A		2
25 - B	26 - B	97 - B	1 - A	69 - A			2
25 - B	26 - B	97 - B	2 - A	1 - A	69 - A		2
25 - B	26 - B	97 - B	98 - B	99 - B	1 - A	69 - A	2
25 - B	26 - B	97 - B	98 - B	99 - B	93 - A	69 - A	2
25 - B	27 - A	25 - A	24 - A	99 - B	69 - A		2
25 - B	27 - B	23 - A	22 - A	11 - A	66 - A	69 - A	2
25 - B	27 - B	24 - A	99 - B	69 - A			2

Table 3.25. CA-P2 paths BTR 25 B - 69 A

BTR: 25-A and 69-B (1F7A)								**Path**
Paths								**Freq**
25 - A	26 - A	97 - A	99 - A	69 - B				5
25 - A	25 - B	24 - B	99 - A	69 - B				4
25 - A	26 - A	24 - B	66 - B	67 - B	69 - B			3
25 - A	26 - B	97 - A	98 - A	99 - A	69 - B			3
25 - A	25 - B	24 - B	66 - B	69 - B				2
25 - A	26 - A	24 - B	66 - B	69 - B				2
25 - A	26 - A	26 - B	97 - A	3 - B	2 - B	1 - B	69 - B	2
25 - A	26 - A	97 - A	1 - B	69 - B				2
25 - A	26 - A	97 - A	3 - B	2 - B	1 - B	99 - A	69 - B	2
25 - A	26 - A	97 - A	99 - A	1 - B	69 - B			2
25 - A	26 - B	97 - A	1 - B	69 - B				2
25 - A	27 - A	24 - B	66 - B	69 - B				2

3.7. Folding Simulations

Protein folding of the structures studied in path generation (e.g. Shaker potassium channel and HIV-1 protease) are carried out using a previously developed program (Ulutas *et al.*, 2009), that is based on robotic motion planning (Section 2.5). Folding trajectories are obtained in terms of snapshots of 3-D conformations. Here these snapshots are analyzed to see how the proteins are folded, and how the key residues proposed in the previous sections by pathway analysis behave through the trajectory.

3.7.1. Folding Trajectories

The fold of Shaker potassium channel, 1A68 (87 residues) succeeded in 367 snapshots. The snapshots are clustered with an RMSD cut-off radius of 4 Å and 33 clusters are obtained. Best members of each cluster are obtained and structures of those snapshots are analyzed. A sample of representative snapshots are summarized in Figure 3.10. It was seen that the protein folds very slowly and helices just start to form around snapshot number 245 (Figure 3.10.a). The first interactions form locally between sequential neighbor residues, and the helices are constructed (Figure 3.10.a-b). Then the formation of tertiary contacts derives the chain collapse. This takes a lot shorter time compared to forming of helices (i.e. through snapshots 320-367, Figure 3.10.c-f).

Similarly folding trajectories are obtained for a single monomer of the HIV-1 protease structure (1F7A chain A). Folding of the 99 residues succeeded in 954 snapshots. Because of this an RMSD cut-off radius of 5 Å is chosen to get a more reasonable number of clusters and 22 clusters are obtained. An increase in the number of snapshots was expected since the protein has more residues, but this is not the reason for a difference this huge. The folding simulation was stuck on a single conformation, that is thought to be a local minimum, through snapshots 516-751 and also it was seen that the local interactions took place a lot later in the simulation (snapshot #397) than it was in the previous case. Representative snapshots of ca-p2 (1F7A) are summarized in Figure 3.11. Again it is seen that first the secondary (local) then the tertiary (distant) interactions form along the trajectory.

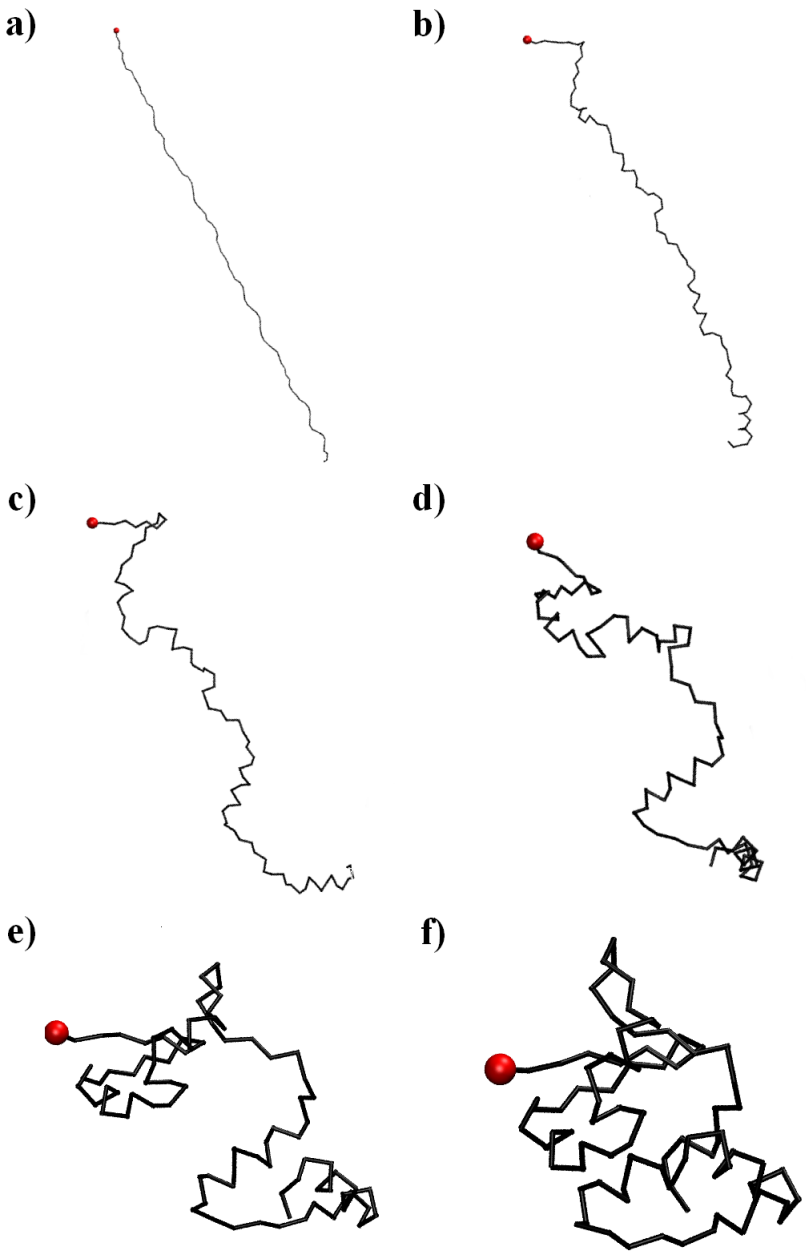

Figure 3.10. Sample snapshots of cluster best member conformations from the folding trajectory of 1A68. a) Cluster #3 Snapshot #245, b) Cluster #16 Snapshot #320; c) Cluster #20 Snapshot #345, d) Cluster #23 Snapshot #359, e) Cluster #25 Snapshot #364, f) Cluster #26 Snapshot #367.

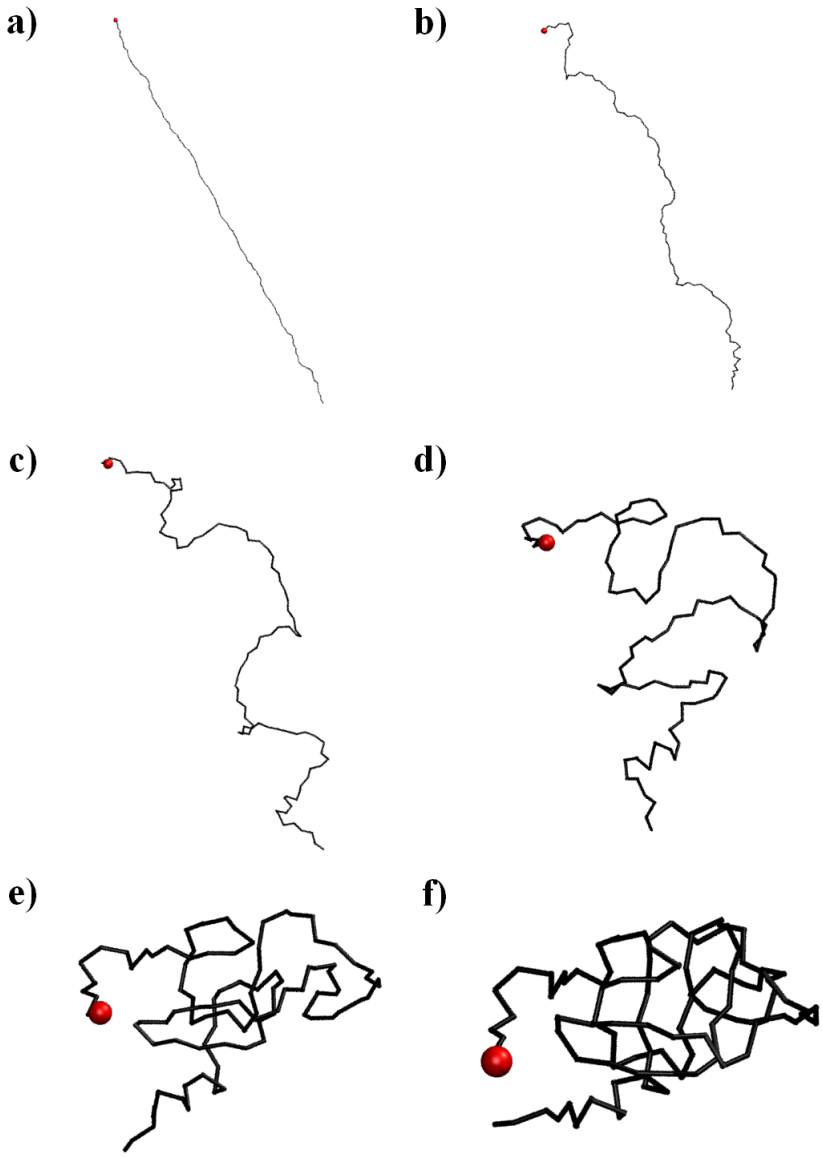

Figure 3.11. Sample snapshots of cluster best member conformations from the folding trajectory of 1F7A. a) Cluster #22 Snapshot #397, b) Cluster #11 Snapshot #480, c) Cluster #14 Snapshot #501, d) Cluster #17 Snapshot #512, e) Cluster #19 Snapshot #751, f) Cluster #1 Snapshot #954.

3.7.2. Contact Map Analysis of Fold Trajectories

Protein folding is basically how amino acid residues make contacts with each other until the protein reaches its folded state, i.e. the conformational visits of the chain led by the formation of these contacts. Here contact map analyses are made for all best member snapshots. It was seen in the trajectory that first local then tertiary contacts are formed. Here, the roles of the key residues proposed in communication path results in forming these interactions are sought.

Among all the analyzed contact maps of Shaker potassium channel a sample representing the whole trajectory is summarized in Figure 3.12. The first contacts (Figure 3.12.a) are formed within two sets of residues 70-78 and 146-152, which cover the allosterically most important residues 77 and 148 as well as two other residues proposed by path analysis 70, 72 and 149. The next contacts in the fold are around residues 92-98, and in fact residues 92 and 98 both were proposed to be important residues in previous sections. These regions thicken in the following contact maps and some patterns are visible around the diagonal, but until snapshot #364 no distinctively tertiary contacts are observed. This most distinctive contact is between the residues 72 and 111 both of which are among the proposed key residues proposed in this work. As it can be seen, most of the residues proposed in the previous sections (path elements and other important resdiues) start to form contacts in early stages of the folding.

Summary of the contact maps of HIV-1 protease trajectory is shown in Figure 3.13. The first observable contacts are around residues 9-11, 18-20 and 90-93 (Figure 3.13.a). Residues 9, 90 and 93 are among the functionally important residues proposed in the previous sections. Next contacts around residues 64-68, 84-87 and 96-98 become observable (Figure 3.13.b). All these ranges cover residues previously proposed by path analysis; 66, 84, 87 and 97. In Figure 3.13.c-d contacts around residues 50 and 25 complete the set of proposed residues. In this case also the tertiary contacts are observed towards the end of the simulation (i.e. snapshot #512).

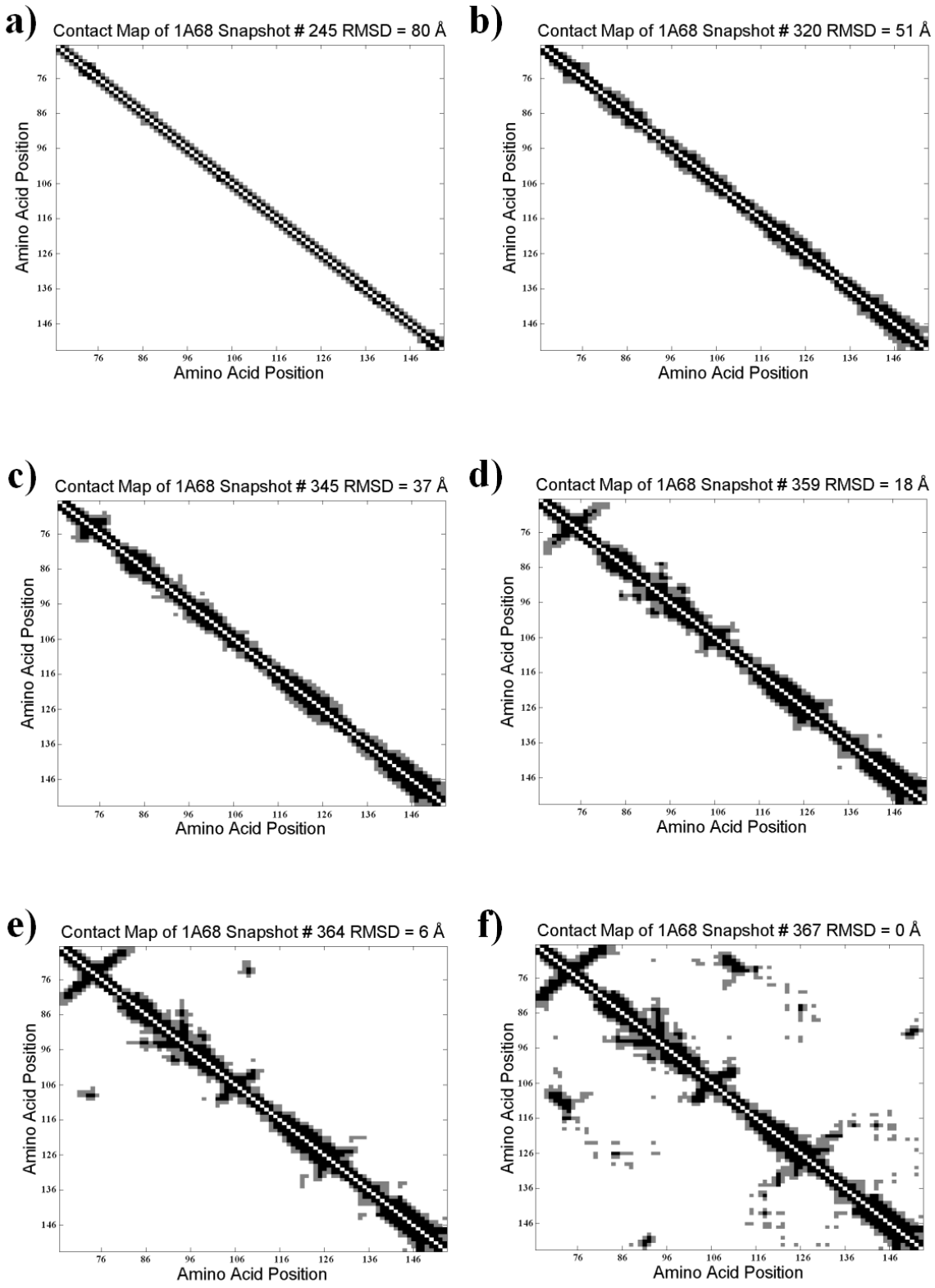

Figure 3.12. Sample contact maps of cluster best member conformations from the folding trajectory of 1A68. RMSD values correspond to the RMSD between that specific snapshot and the target conformation.

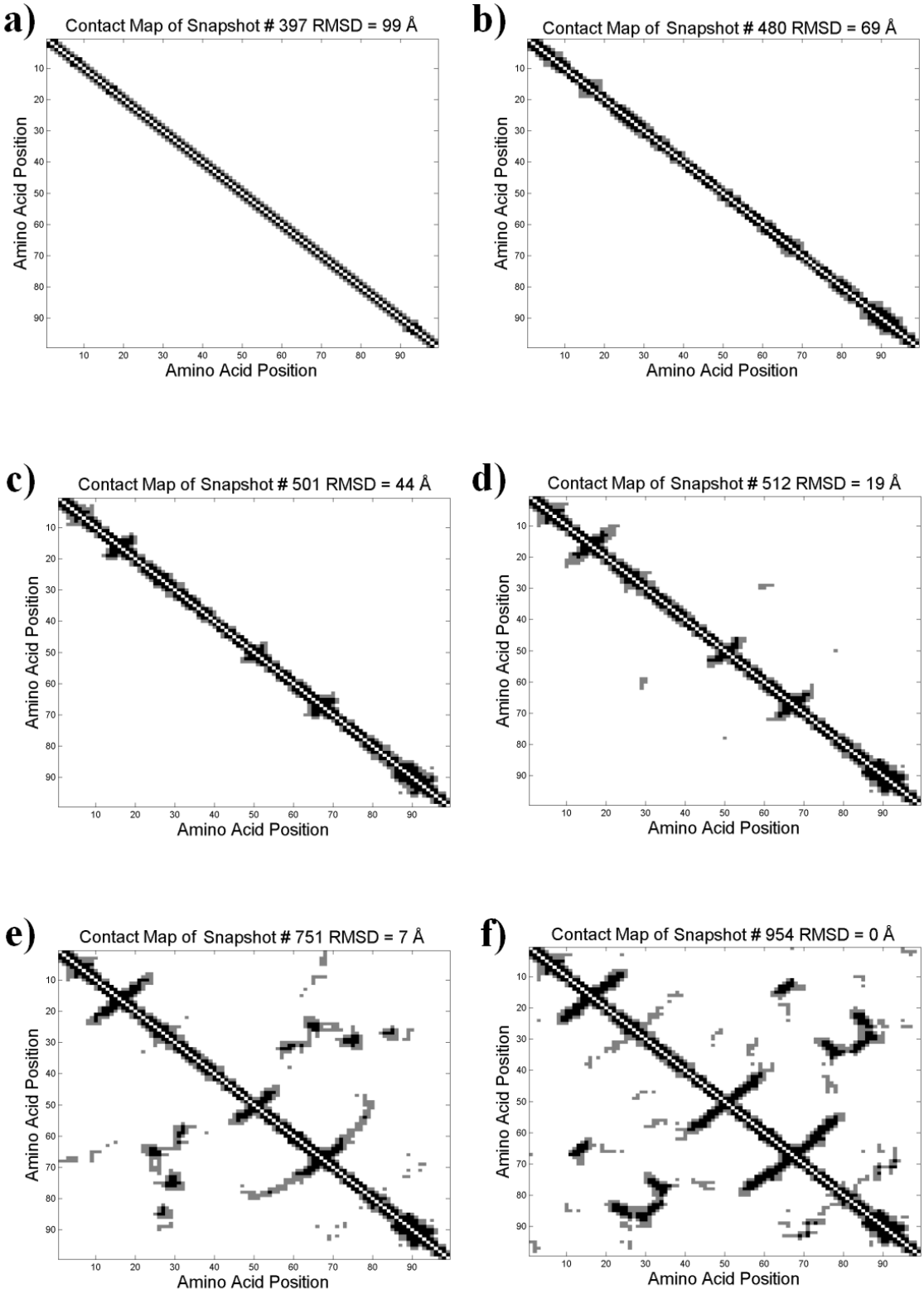

Figure 3.13. Sample contact maps of cluster best member conformations from the folding trajectory of 1F7A.

Combining the results of both simulations, it is obvious that the residues suggested by path generation are important key residues in protein dynamics with functions in either allostery or protein folding. Overall, it is seen that some residues may even be functionally important for both dynamic behaviors. These folding results are only intended to provide a basic understanding of the folding mechanism rather than giving the actual intermediate states, but even with this basic understanding the role of key residues can be observed clearly.

4. CONCLUSIONS

Information on functionally important residues in proteins is crucial for understanding any type of protein dynamics. In this study, functionally important residues such as catalytic and binding sites are shown to be predicted through generation of inter-residue communication pathways by a newly proposed MC method.

MC path generation produced ensemble of paths which were more informative than a single shortest path in predicting functionally important residues. The importance of the type of the potential function used in path generation is observed through studies made on four different potential functions, among which atomistic potential function described the interaction between protein residues the best.

Three different ways of MC path generation, BTR, PSNS, ISP separately made contributions to functionally important residue prediction. BTR paths revealed the residues present in the repeated patterns of interactions to be functionally important. PSNS revealed residues communicating naturally without a forced direction to be functionally important. ISP in the form of network parameters, closeness, betweenness and clustering coefficient, revealed that peaks of closeness and betweenness and minimums of clustering coefficient plots correspond to functionally important residues.

Finally protein folding trajectories revealed that the folding mechanisms in general consist of two stages; forming of local interactions and tertiary interactions. While local interactions construct the secondary structures, tertiary interactions are the key in driving the folding. The residues that form the initial tertiary contacts during the folding of the two proteins Shaker potassium channel and HIV-1 protease are interestingly among the residues that display high closeness and betweenness values with the inter-residue pathway analysis in the native state of these proteins.

Further, the studies on different structures of a protein revealed that changes in protein structure that do not affect the function also do not affect the communi-

cation pathways or the characteristics of the residue network, so the idea that communication pathways are evolutionarily conserved is intensified. Then the studies on intra-molecular and inter-molecular communication in monomers of a homodimer revealed that communication within polymeric proteins may occur both within and between monomers. Structural asymmetry did not affect paths within functional sites of the monomers. On the other hand paths between the monomers revealed important residues making the connection.

REFERENCES

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.

Bahar, I., A. Atilgan and B. Erman, 1997. "Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential". *Fold. Des.*, Vol. 2, pp. 173-181.

Bahar, I., M. Kaplan and R. L. Jernigan, 1997, "Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches", *Proteins: Structure, Function, and Genetics*, 29:292-308.

Bahar, I., A. R. Atilgan, M. C. Demirel and B. Erman, 1998, "Vibrational dynamics of proteins: Significance of slow and fast modes in relation to function and stability", *Phys. Rev. Lett.*, Vol. 80, pp. 2733-2736.

Bahar, I. and R. L. Jernigan, 1997, "Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation", *Journal of Molecular Biology*, 195-214

Bardwell, V. J. and R. Treisman, 1994, "The POZ domain: A conserved protein-protein interaction motif", *Genes Dev.*, Vol. 8, pp. 1664-1677.

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.

Brenman, J. E., J. R. Topinka, E. C. Cooper, A. W. McGee, J. Rosen, T. Milroy, H. J. Ralston and D. S. Bredt, 1998, "Localization of Postsynaptic Density-93 to Dendritic Microtubules and Interaction with Microtubule-Associated Protein 1A", *J. Neurosci.*, Vol. 18, pp. 8805-8813.

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

Chou, K.C., 1996, "Prediction of human immunodeficiency virus protease cleavage sites in proteins", *Analytical Biochemistry*, Vol. 233, pp. 1-14.

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. and J. Gray, 2009, "Allosteric communication occurs via networks of tertiary and quaternary motions in proteins", *PLOS Computational Biology*, Vol. 5, Issue 2.

Demirel, M. C., A. R. Atilgan, R. L. Jernigan, B. Erman and I. Bahar, 1998. "Identification of kinetically hot residues in proteins". *Protein Sci.*. Vol. 7, pp. 2522-2532.

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.

Ertekin, A., R. Nussinov and T. Haliloglu, 2006, "Association of putative concave protein-binding sites with the fluctuation behavior of residues", *Protein Science*, 15:2265-2277.

Fanning, A. S. and J. M. Anderson, 1999, "PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane", *J. Clin. Invest.*, Vol. 103, pp. 767-772.

Feig, M., J. Karanicolas and C.L. Brooks, 2004 "MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology", *J. Mol. Graph. Model.*, Vol. 22(5), pp. 377-395, 2004.

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.

Haliloglu, T., I. Bahar and B. Erman, 1997, "Gaussian dynamics of folded proteins", *Phys. Rev. Lett.*, Vol. 79, pp. 3090-3093.

Haliloglu, T., O. Keskin, B. Ma and R. Nussinov, 2005, "How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hotspots and conserved residues", *Biophys. J.*, Vol. 88, pp. 1552-1559.

Haliloglu, T., E. Seyrek and B. Erman, 2008, "Prediction of binding sites in receptor-ligand complexes with the Gaussian Network Model",*Physical Review Letters*, Vol. 100, pp. 228102.

Haliloglu, T. and B. Erman, 2009, "Analysis of Correlations between Energy and Residue Fluctuations in Native Proteins and Determination of Specific Sites for Binding", *Physical Review Letters*, Vol. 102, pp. 088103.

Harpaz, Y, N. Elmasry, A. R. Fersht and K. Henrick, 1994, "Direct observation of better hydration at the N terminus of an α-helix with glycine rather than alanine as the N-cap residue", *Proc. Natl. Acad. Sci. USA*, Vol. 91, pp. 311-315.

Kmiecik, S. and A. Kolinski, 2007, "Characterization of protein-folding pathways by reduced-space modeling", *PNAS*, Vol. 104, pp. 12330-12335.

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.

Laskowski, R. A., 2000, "PDBsum: summaries and analyses of PDB structures", *Nucleic Acid Research*, Vol. 29, pp. 221-222.

Lindgren, M. T., *Exploring inhibitors of HIV-1 protease: Interaction studies with applications for drug discovery*, PhD Thesis, Uppsala University.

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

Metropolis, M. and S. Ulam, "The Monte Carlo Method", *Journal of the American Statistical Association*, Vol. 44, No. 247, pp. 335-341.

Nagao, C., T. P. Terada, T. Yomo and M. Sasai, 2005, "Correlation between evolutionary structural development and protein folding", *PNAS*, Vol. 102, pp. 18950-18955.

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

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

Ozkan, S. B., G. S. Dalgin and T. Haliloglu, 2003, "Unfolding events of Chymotrypsin Inhibitor 2 (CI2) revealed by Monte Carlo (MC) simulations and their consistency from structure-based analysis of conformations", *Polymer*, Vol. 45, pp. 581-595.

Prabu-Jeyabalan, M., E. Nalivaika and C. A. Schiffer, 2000, "How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease", *Journal of Molecular Biology*, Vol. 301, pp. 1207-1220.

Prabu-Jeyabalan, M., E. Nalivaika and C. A. Schiffer, 2002, "Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes", *Structure*, Vol. 10, pp. 369-381.

Prabu-Jeyabalan, M., E. A. Nalivaika, N. M. King and C. A. Schiffer, 2004, "Structural basis for co-evolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease", *Journal of Virology*, Vol. 78, pp. 12446-12454.

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.

Sewing, S., J. Roeper and O. Pongs, 1996, "Kv beta 1 subunit binding specific for shaker-related potassium channel alpha subunits", *Neuron*, Vol. 16, pp. 455-463.

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. Biol.*, Vol. 373, pp. 1361-1373.

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

Tsai, C. J., A. del Sol and R. Nussinov, 2009, "Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms", *Mol. BioSysts.*, Vol. 5, pp. 207-216.

Tsunoda, S., J. Sierralta, Y. Sun, R. Bodner, E. Suzuki, A. Beckner, M. Socolich and C. S. Zucker, 1997,“ A multivalent PDZ-domain protein assembles signalling complexes in a G-protein-coupled cascade”,*Nature*, Vol. 388, pp. 243-249.

Ulutas, B., T. Haliloglu and I. Bozma, 2009, “Folding Pathways Explored with Artificial Potential Functions”. *Phys. Biol.*, submitted.

Wlodawer, A. and J. Erickson, 1993, “Structure-based inhibitors of HIV-1 protease”, *Annual Review of Biochemistry*, Vol. 62, pp. 543-585.