EFFECT OF LIGAND BINDING ON PROTEIN DYNAMICS. A TIME SERIES ANALYSIS

by

Sena Başkan

B.S., Mathematics Engineering, Yıldız Teknik University, 2004

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 Computational Science and Engineering

Boğaziçi University

2008

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor Prof. Pemra Doruker Turgut for all her support, guidance, and especially for her tolerances. I am grateful to my thesis co-supervisor Asst. Prof. Burak Alakent for his patience and time he spent for me. I thank Bülent Balta and Sertan Cansu for helping and providing me with the necessary data by performing simulations. I shouldn't forget to thank my friend Özlem Türe for providing me with articles and answering my questions.

I acknowledge the financial support of TUBITAK (Project No: 104M247).

Any thank is insufficient to all of my elementary and high school teachers, undergraduate and graduate instructors. I want to express my special thanks to my family, especially my mother, for their love and support.

ABSTRACT

EFFECT OF LIGAND BINDING ON PROTEIN DYNAMICS. A TIME SERIES ANALYSIS

Internal dynamics of proteins can be analyzed by the help of simulation techniques, one of which is molecular dynamics (MD) simulation. Principal component analysis is commonly applied on large-sized MD simulation data to extract the functionally relevant essential modes (principal components). Recently, Alakent *et al.* analyzed the principal components from MD simulation data by time series models and interpreted the obtained parameters in terms of protein fluctuations.

In this thesis, MD trajectories of two enzymes, dihydrofolate reductase (DHFR) and triosephosphate isomerase (TIM), are investigated by performing time series analysis on the principal components. Two independent MD trajectories of 3.2 ns duration are used for the free (apo) and ligand-bound (liganded) states of each enzyme. Model parameters extracted from two independent runs are similar for the same state of each enzyme, which indicates the reliability of the analysis, and shows the extent of the effect of different conformational substates on the protein vibrational frequency density. DHFR has been analyzed with two different ligands which differ from each other by only hydride ion. The contribution of NADPH to the collective motions of DHFR is higher compared to those of $NADP^+$. It is also seen that the collective motions in NADPH bound DHFR are similar to the unliganded form, while the collective character of the motions in the $NADP^+$ bound DHFR is lost. It is found that unliganded forms of both DHFR and TIM are more flexible than liganded form. For both TIM and DHFR, low frequencies shift to higher frequencies after the ligands bind. Vibrational frequencies of $NADP^+$ bound DHFR are lower than those of NADPH bound DHFR. Briefly, it can be concluded that ligand binding affects the collectivity of fluctuations, vibrational low frequency density and anharmonic motions of proteins, and particulary vibrational frequencies of the proteins may have importance on the catalytic cycle. All these effects should be taken into consideration for examing binding affinities of proteins.

ÖZET

LİGAND BAĞLANMASININ PROTEİN DİNAMİKLERİNE ETKİSİ. BİR ZAMAN SERİLERİ ANALİZİ

Proteinlerin iç dinamikleri simulasyon tekniklerinin yardımıyla, örneğin moleküler dinamikler (MD) simulasyonuyla analiz edilebilirler. Büyük boyutlu olan MD verisine sıklıkla uygulanan ana bileşenler analizi sonucunda proteinin işleviyle ilişkili oldukları bilinen esas modlar (ana bileşenler) elde edilir. Yakın zamandaki bir çalışmada, Alakent ve çalışma arkadaşları MD verisinden elde edilen esas modları zaman serileri modelleriyle incelemiş ve elde edilen parametreleri protein dalgalanmaları açısından yorumlamışlardır.

Bu tezde, dihidrofolat redüktaz ve triozfosfat izomeraz enzimleri, MD verilerinden elde edilen ana bileşenlerine zaman serileri analizi uygulanarak incelenmiştir. Enzimlerin serbest (apo) ve ligand-bağlı halleri üzerinde gerçekleştirilen MD simulasyonlarının 3.2 ns uzunluğundaki ikişer bağımsız kısımları kullanılmıştır. İkişer bağımsız örnekten çıkarılan model parametrelerinin aynı hal için birbirlerine benzer çıkmaları analizin güvenilirliğini ve farklı konformasyonel altdurumların proteinin vibrasyonel frekans yoğunluğu üzerindeki etkisinin derecesini göstermiştir. DHFR, birbirlerinden bir hidrid iyonu farkı olan iki ayrı ligand ile incelenmiştir. NADPH ligandının DHFR proteinin beraber hareketlerine katılımı $NADP^+$ ligandından yüksektir. NADPH bağlı DHFR proteinindeki beraber hareketlerin serbest haldekine benzediği, $NADP^+$ bağlı haldeki beraber hareketlerin ise kaybolduğu gözlemlenmiştir. Hem DHFR hem de TIM proteinlerinin serbest hallerinin ligand bağlı hallerinden daha esnek oldukları görülmüştür. Her iki proteinde de ligand bağlanması düşük frekansları daha yüksek frekanslara doğru kaydırmıştır. $NADP^+$ bağlı DHFR proteinin vibrasyonel frekanslarının NADPH bağlı halden daha düşük olduğu görülmüştür. Özetle, ligand bağlanmasının proteinlerin salınımlarının beraberliğini, vibrasyonel düşük frekans yoğunluğunu ve harmonik olmayan hareketlerini etkilediği; özellikle vibrasyonel frekansların katalitik döngü üzerinde önemi olabileceği sonuçlarına varılmıştır. Proteinlerin birbirleri ile bağlanma eğilimi incelenirken bu etkenlerin hepsi göz önüne alınmalıdır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

C^{α}	Alpha-carbon
Å	Angstrom
Kcal	Kilocalory
M	Mean
#	number
ns	nano second
ps	pico second
ΔG_c	Free energy of catalyzed reaction
ΔG_u	Free energy of uncatalyzed reaction
ACF	Autocorrelation function
CDF	Cumulative distribution function
DHAP	Dihydroxyacetone phosphate
DHFR	Dihydrofolate reductase
GAP	Glyceraldehyde 3-phosphate
MD	Molecular dynamics
MSF	Mean square fluctuation
$NADP^+$	Nicotinamide adenine dinucleotide phosphate
NADPH	Reduced form of nicotinamide adenine dinucleotide phosphate
NMA	Normal mode analysis
PC	Principal component
PCA	Principal component analysis
PDB	Protein data bank
PDF	Probability distribution function
RMSD	Root mean square deviation
TIM	Triosephosphate isomerase
vs	versus

1. INTRODUCTION

Proteins are macromolecules that are responsible for important biological functions such as enzymatic catalysis and signal transduction. They are polymers, made up of amino acids (residues) linked together to form a specific linear sequence. Every protein with a distinct sequence functions about a specific three-dimensional, well-defined structure, which is called the native state. Diversity of functions performed by proteins is caused by the diversity of those native states. However, functions of proteins do not depend only on their structures, but also on their internal dynamics.

As the proteins fluctuate about their native state, they make transitions in between a large number of energy valleys, named conformational substates[1]. Parts of the protein molecule exhibit collective motion, whereas other parts move locally. Theoretical information about protein dynamics has been acquired by normal mode analysis (NMA) and molecular dynamics (MD) simulation studies[2-5]. Those studies have shown that collective motions are directly related with the protein's function such as displacements in the active site of lysozyme[6], or hinge bending motion in thermolysin[5]. Because of this, one must analyze the collective motions and accompanying conformational changes in a protein in order to understand the relationship between its dynamics and relevant functions.

MD simulations, which are becoming feasible due to the shortening of computational times as computer technologies evolve, generate trajectories depending on the classical equations of motions. The trajectories comprise the Cartesian coordinates of the atoms in a protein sampled during the simulation. Data obtained from MD simulations have high dimensionality due to sampling and system size (number of atoms). Principal Component Analysis (PCA), which is a statistical method, has been applied to overcome this high-degree of freedom problem[4,6,7-13]. PCA provides principal components (PCs), which can represent the original data without loss of information. Alakent *et al.*[10-13], applied time series analysis on the projection of the MD simulation data on the PCs and showed PCA to be a reliable method in extracting important dynamics information from the protein fluctuations. Alakent *et al.*[10-13] used linear stochastic time series model parameters to

approximate the protein motions as interminima and intraminimum fluctuations on the multidimensional energy surface.

In this thesis, the dynamics of enzymes, dihydrofolate reductase (DHFR) and triosephosphate isomerase (TIM) are studied using PCA and time series analysis. DHFR has special biological functions which makes it of pharmacological interest. TIM is an important enzyme in glycolysis[14]. Enzymes perform their functions by molecular interactions with ligands, which bind to specific parts of the protein and alter the behavior of the protein. This alteration on the dynamics is essential for the completion of protein's function. In the current study, analysis of free and ligand bound DHFR and TIM are carried out. MD simulations are performed to mimic proteins' motions in free and ligand bound states. Every execution, that is every run is sampled twice, from two different parts (at different places in the potential energy surface). PCA is applied on the data obtained from each sample to obtain the PCs. Collective coordinates are acquired by projecting the data on the PCs. Time series analysis is applied on the collective coordinates and linear stochastic time series models are derived at the end of the analysis. The purpose is to see how the dynamic behavior changes with respect to different parts of the energy surface and different runs some of which are of free states and some of ligand-bound states of the proteins. MD simulations are carried out by the commercial software package AMBER 8.0[15]. System Identification Toolbox on MATLAB 7.0 (student version) is used for determining appropriate time series models. PYMOL (version 0.99rc6) is used for protein illustrations.

This thesis is structured as follows. In Section 2, protein interactions with ligands for the specific cases of DHFR and TIM are explained briefly. In Section 3, specific details of the MD simulation and its protocols are given. In Section 4, PCA and time series analysis with their application procedures on the MD data are shown. Results and Discussion, Section 5, involves the analysis of free and ligand bound DHFR and TIM. Their internal motions in time are modeled by time series analysis and those model parameters are interpreted in terms of protein fluctuations. Section 6 is Conclusion, where the summary of the study with recommendations are given.

2. PROTEIN DYNAMICS AND INTERACTIONS WITH LIGANDS

Every living cell contains proteins in its structure, some of which function as enzymes. Enzymes are biological catalysts that accelerate the rate of chemical reactions taking place in living cells, like signaling and motor activities of cells. Deficiency of enzymes causes failure in some biological functions. For a chemical reaction to occur, activation-energy barrier, that is the higher-energy region between two consecutive chemical species in a reaction, should be overcomed. The higher this barrier, the slower the reaction and the more difficult the chemical step is to achieve[16]. Catalysts accelarate the reactions by decreasing this barrier. Figure 2.1 illustrates energetics of catalysis of a hypothetical reaction from substrate S to product P. For the uncatalyzed reaction (blue curve) a single barrier determines the rate at which product is formed. In the presence of a catalyst, intermediate barriers are overcomed. Free energy of uncatalyzed reaction (ΔG u) is considerably higher than that of catalyzed reaction (ΔG c).

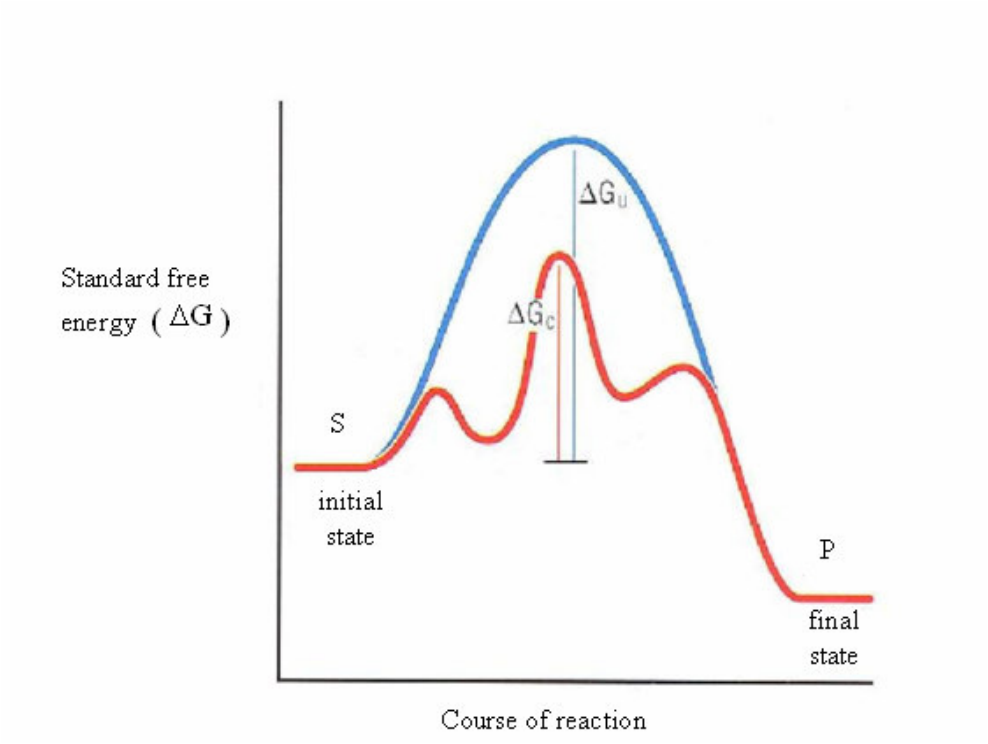


Figure 2.1. Energetics of catalysis [16]

During the catalytic reactions, ligands are bound to enzymes. Interaction between the enzyme and its ligand, during which conformational changes occur, is essential for

enzyme's functionality. Ligands usually contact with only a small number of residues, however, their effect often propagates to other portions of the protein[17]. Most significant changes occur around their some specific residue groups, named as loops. Below is the thermodynamic cycle for the binding of ligand A to an enzyme E[18], which is accompanied by a free energy change:

$$A + E \rightleftharpoons AE \qquad (2.1)$$

DHFR and TIM enzymes, which will be examined in the thesis in the context of ligand binding, are presented briefly below.

2.1. Dihydrofolate Reductase (DHFR)

DHFR is found in many organisms from Escherichia coli to human being. It performs important functions; for instance, it is required for normal folate metabolism in prokaryotes and eukaryotes[19]. The importance of its clinical role has made DHFR a target of enzymological studies[6]. It catalyzes the reduction of 7, 8-dihydrofolate (DHF) or folate to 5, 6, 7, 8-tetrahydrofolate (THF). It is principally responsible for maintaining intra-cellular pools of THF. THF is the cofactor used in synthesis of several important metabolites, one of which is thymidylate, a building block of DNA. Because of this, DHFR has clinical importance and been recognized as a drug target for inhibiting DNA synthesis in rapidly proliferating cells such as cancer cells or bacterial or malarial infections. Since 1950s, researchers have been working on its kinetics and structure[20].

DHFR is bound with the ligands $NADP^+$ (Nicotinamide adenine dinucleotide phosphate) and reduced form of it, NADPH for the maintenance of THF. This is a catalytic cycle which has the following intermediate steps, each of which is a complex: DHFR/DHF/NADPH (DH), DHFR/THF/$NADP^+$ (TP) and DHFR/THF/NADPH (TH). It is summarized as a reaction below[21]:

$$DHF + NADPH + H^+ \xrightarrow[DHFR]{} THF + NADP^+ \qquad (2.2)$$

The catalytic cycle is summarized in Figure 2.2.

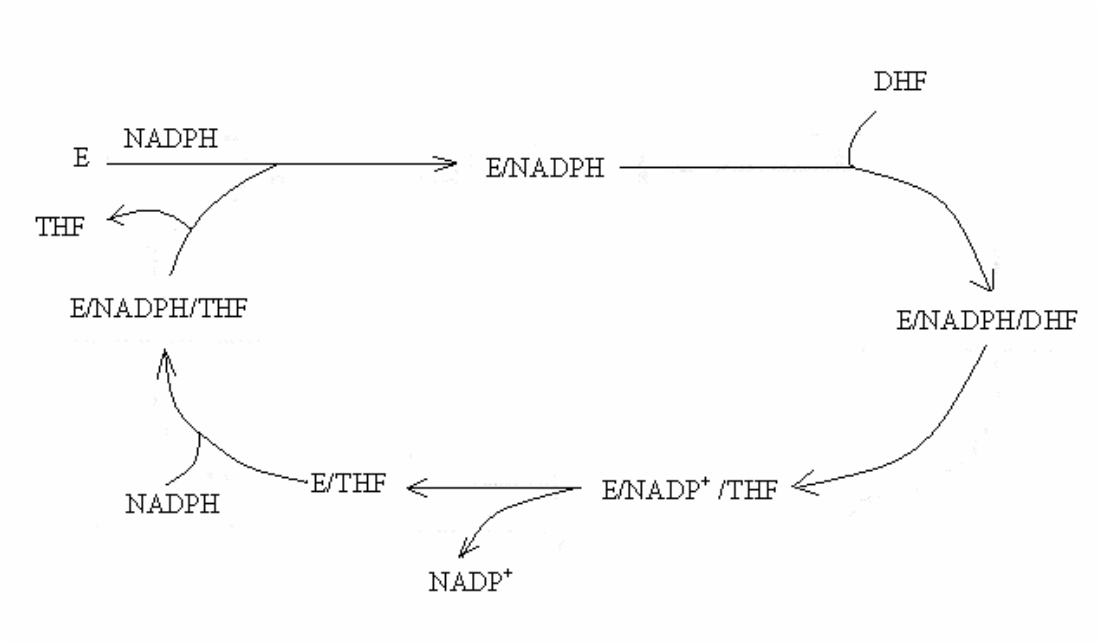


Figure 2.2. Catalytic cycle of maintenance of THF[22]

DHFR comprises 159 residues. The main regions of motion are in the M20 loop (residues 9-24), the neighboring FG loop (residues 117-131), the distant CD loop (residues 64-71), GH loop (residues 142-149) and the hinge region (the area connecting the two subdomains of DHFR) (Figure 2.3).

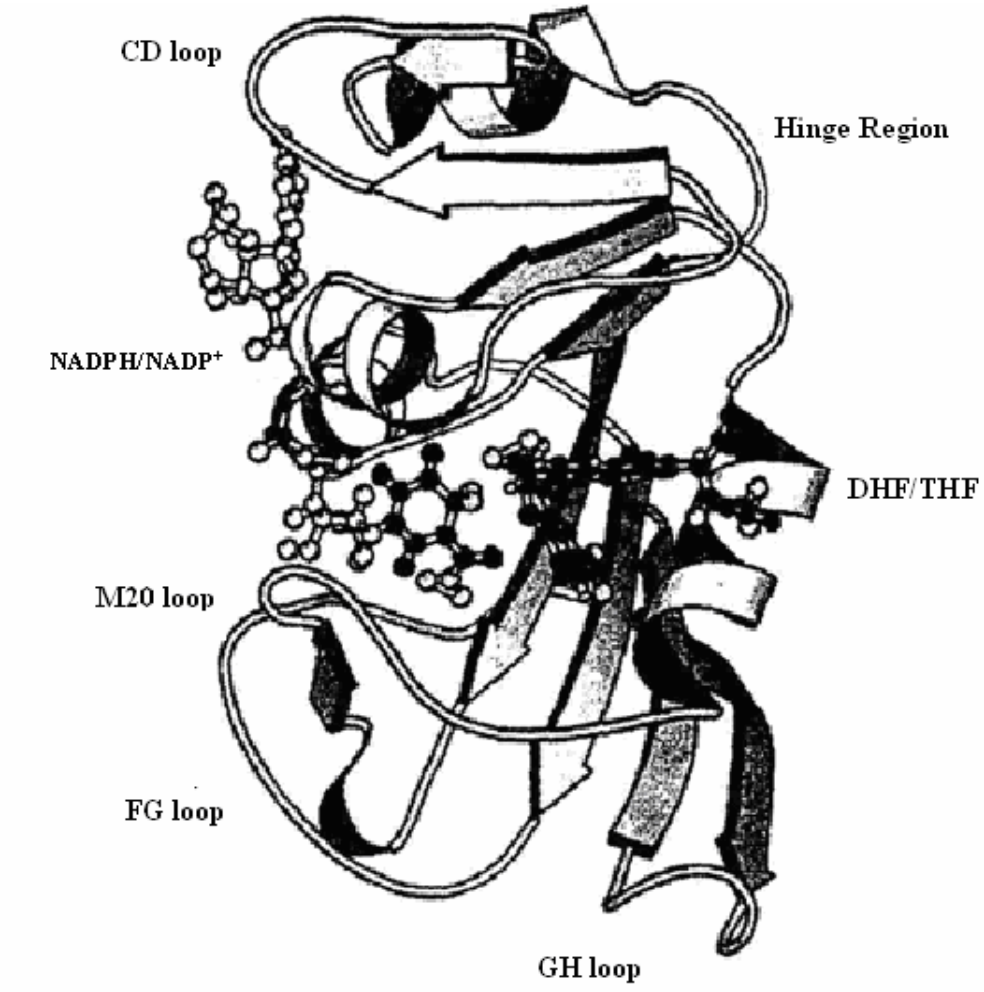


Figure 2.3. DHFR illustration bound with NADPH/NADP+ and DHF/THF[22]

It was found that the loop dynamics and conformations are sensitive to which ligands are bound to the protein[22]. In this thesis, the effects of the ligands $NADP^+$ and NADPH on the dynamics of DHFR will be investigated.

2.2. Triosephosphate Isomerase (TIM)

TIM, for which numerous experimental and computational studies exist, is found in nearly every organism including animals such as mammals and insects as well as in fungi, plants and bacteria[23]. The catalytic reaction of TIM was one of the first clear demonstrations of the role of protein dynamics in promoting and controlling chemical reactivity[24]. TIM plays an important role in fourth step of glycolysis and is essential for efficient energy production. It catalyses the interconversion between DHAP

(Dihydroxyacetone phosphate) and GAP (D-glyceraldehyde 3-phosphate)[14]. The reaction is illustrated below:

$$\text{DHAP} \underset{\text{TIM}}{\rightleftharpoons} \text{GAP} \qquad (2.3)$$

TIM comprises 494 residues. Native TIM is active as a dimer. Loop 6 (residues 166-176 in each subunit) of TIM protects the ligand from solvent exposure during catalysis by closing over the active site. Same loop opens and closes in the free (apo) state, too. Figure 2.4 illustrates free (light green) and ligand-bound (grey) dimer TIM conformations. DHAP, the ligand, is illustrated with black and active site residues are in pink. On the right monomer, Loop 6 is observed open (green) in the free state, and closed (dark blue) in the bound state.

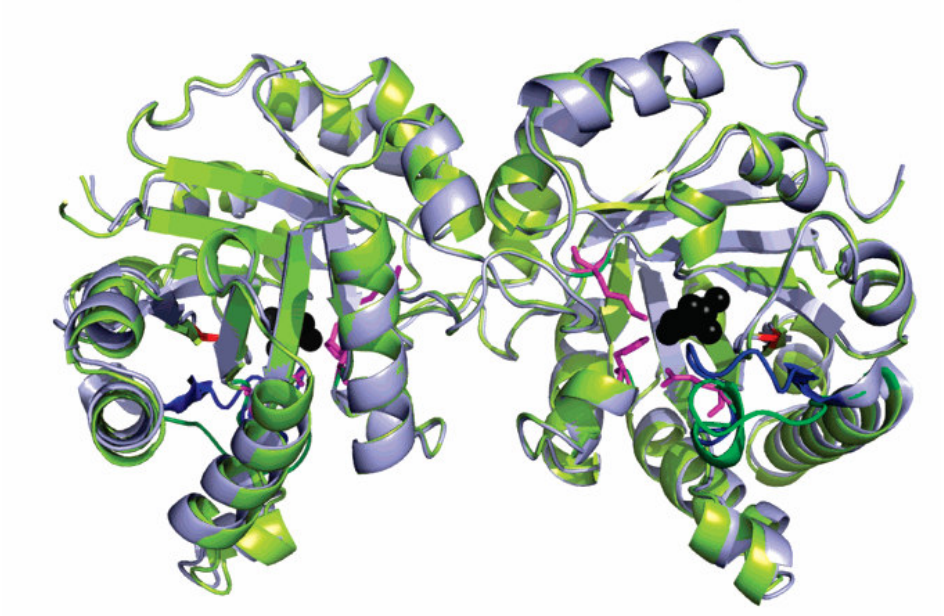


Figure 2.4. Dimer TIM illustration[14] (PDB code:8TIM)

In this thesis, dynamics of free TIM when loop 6 is open, will be compared with its closed form. Additionally, the effect of ligand DHAP on the dynamics of TIM will be investigated.

3. MOLECULAR DYNAMICS SIMULATION OF PROTEINS

3.1. Molecular Dynamics (MD) Simulation of Biomolecules

MD is a deterministic computer simulation technique, which is commonly used to investigate the dynamics of biomolecular systems. It is possible to describe a molecular system as a function of variables[1]. Based on the Cartesian coordinates of the system the classical equations of motion can be solved to find the atomic trajectories in phase space:

$$\boldsymbol{f}_i = m_i \frac{d^2 \boldsymbol{x}_i}{dt^2} = m_i \boldsymbol{a}_i \tag{3.1}$$

Here t represents time, m_i is the mass of i^{th} particle, $\boldsymbol{a}_i$ is the acceleration vector of the molecule i, while $\boldsymbol{x}_i$ and $\boldsymbol{f}_i$ are the three dimensional vectors of Cartesian coordinates, and force term respectively. Force term $\boldsymbol{f}_i$ is the gradient of the potential energy of i^{th} particle. Experimental X-ray conformations are usually used as the initial protein structures in MD simulations and obtained from the Protein Data Bank (PDB)[25].

The potential energy (V) surface used in MD simulations for the calculation of forces is called the "forcefield". The quality of any empirical forcefield determines the validity of the results obtained from an MD simulation. A forcefield consists of a functional form, which usually employs a combination of internal coordinates of the molecule and appropriate force constants. The forcefield employed in this thesis is named AMBER ff03 forcefield[26]. It is an "all-atom" forcefield. It includes bond-stretching, angle-bending, dihedral, van der Waals and electrostatic terms.

A well-established forcefield is a necessary but not sufficient condition to guarantee the success of a simulation. One should also consider the limitations imposed by the computational power that depend on the number of particles, and increase the efficiency of the MD simulations with additional algorithms. These algorithms are based on realistic and reasonable assumptions, such as the application of nonbond cutoffs and periodic boundary conditions. When it is necessary to simulate a protein in solvent environments to mimic the

real situations, it is not possible to surround the molecule with infinitely many solvent molecules. One possible solution to this problem is to implement periodic boundary conditions (PBC); the cubic box is replicated in all three dimensions, forming a lattice of identical cubes. The surrounding imaged atoms exert forces on the real atoms in the interior cube and these force terms are used to calculate the energies of the real atoms. Motions of imaged atoms are calculated by symmetry operations, thus energy calculations need to be performed and stored only for the real atoms.

MD simulation of a protein mainly consists of three stages: minimization, equilibration and data collection. The starting three-dimensional structure of a protein is generally taken as the native folded structure, which has been determined experimentally by X-ray crystallography. The aim of energy minimization is to bring the molecule to a conformation where its potential energy is at a minimum, based on the forcefield being utilized. However, finding a global minimum is a difficult problem due to the nonlinear nature of the protein interactions with many degrees of freedom. Thus, rather than the global minimum, a local minimum close to the starting structure is found.

The aim of equilibration is to bring the system to a favorable conformation at the target temperature and pressure. Initial velocities of the particles should be given to start the integration of equation of motion which are generated from a Maxwell-Boltzman distribution at the specified temperature. The velocity of particles, which is a microscopic quantity, is related to a macroscopic quantity, temperature. Target temperature is maintained by adjusting the velocities of the atoms. Once the equilibrium is reached, data is collected for at least 10-20 ns.

MD simulation runs produce M (sample number) snapshots. A snapshot is a data file in which there exists three-dimensional (3-D) Cartesian coordinates of the atoms of the simulated protein at successive sampling times. Those coordinates belong to representative conformations generated by the simulation. The first data file contains the initial coordinates, while the last one contains the last sampled conformation's coordinates.

3.2. MD Simulation Protocols for the DHFR and TIM

The X-ray crystallography coordinates, used as the starting conformation for MD simulations of both proteins, are taken from PDB. PDB codes are 1RA1 and 1RX1 for the free and ligand bound DHFR, respectively, while PDB codes are 8TIM and 1TPH for the free and ligand bound TIM, respectively. Simulations are performed in explicit water, at 300K for durations of about 20-60 ns. Data (Cartesian Coordinates History) are collected for sampling window sizes (simulation length) of 3.2 ns, with a sampling interval of 0.8 ps in all cases. Simulation condition is constant NPT. The Ewald summation technique with the particle-mesh method[27] is used to calculate long-range interactions with a cutoff distance of 9 Å. A periodic cubic box of 58 Å dimensions is used after solvation of the protein in explicit water molecules (TIP3P)[28]. Constant pressure periodic boundary conditions are used with isotropic position scaling. Energy minimization is performed using 50 cycles of steepest descent algorithm, followed by conjugate gradient until the root mean square (RMS) gradient per atom has reached 0.01 kcal/mol/Å. Each trajectory is started with velocity assignments according to the Boltzmann distribution at either 10 or 20 K, and the temperature is gradually raised to 300 K and the temperature is kept at 300 K using the weak coupling algorithm[29]. Integration of Newton's equations of motion is carried out with the Verlet algorithm[30]. A time step of 2 fs is used by the implementation of SHAKE algorithm for the bonds involving hydrogens[14].

4. STATISTICAL METHODS FOR ANALYZING THE MD TRAJECTORIES

4.1. Principal Components Analysis (PCA)

It is difficult to analyze the simulation data in the form of a large matrix. One must firstly reduce the dimension of the data, without losing important information. PCA is a useful method for this purpose. A data matrix **X**, having a size of M x q can be reduced to smaller matrices. PCA does this by first using the Covariance matrix, **C**, which is a nonsingular matrix. **C** can be decomposed into a "loadings" matrix **P,** consisting of orthogonal eigenvectors, and a diagonal matrix of eigenvalues **Λ,** whose nonzero values correspond to the variation along each principal component. **Λ** has the optional rxr size, where r is the number of principal components, and **P** has qxr size. Equation 4.1 gives the q×q **C** matrix and its decomposition.

$$\mathbf{C} = \frac{(\mathbf{X}-\overline{\mathbf{X}})^{T}(\mathbf{X}-\overline{\mathbf{X}})}{M-1} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}^{T} \tag{4.1}$$

Projection of the data matrix onto the principal components (axes) can be found as

$$\mathbf{T} = (\mathbf{X}-\overline{\mathbf{X}})\,\mathbf{P} \tag{4.2}$$

where **T** is called the "scores" matrix. Each column in **T** matrix, t_i, has Mx1 size, which is called the "scores" vector. $\mathbf{T} = [t_1\ t_2 \ldots t_i \ldots t_{n-1}\ t_r]$ has Mxr size.

4.1.1. Application Procedure of PCA on the Data Obtained From MD Trajectories

In the current study, all of the MD simulation ensembles are of 4000 sample size taken at equal intervals of 0.8 ps. Each sample (snapshot) consists of the Cartesian coordinates of the protein atoms. Amadei *et al.* showed Cartesian coordinates' history of the backbone C^{α} atoms to be adequate in capturing the essential variation in a protein's

motion and they applied PCA on the C^{α} coordinates obtained from MD simulations[6]. Figure 4.1 illustrates a small portion of a MD trajectory file in PDB format.

					x	y	z
ATOM	1	N	MET	1	-1.781	-15.684	-1.890
ATOM	2	H1	MET	1	-2.151	-16.373	-2.528
ATOM	3	H2	MET	1	-2.390	-15.273	-1.198
ATOM	4	H3	MET	1	-1.082	-16.113	-1.301
ATOM	5	CA	MET	1	-1.250	-14.550	-2.700
ATOM	6	HA	MET	1	-0.310	-14.926	-3.104
ATOM	7	CB	MET	1	-2.038	-14.263	-3.921
ATOM	8	HB2	MET	1	-3.087	-14.511	-3.758
ATOM	9	HB3	MET	1	-2.041	-13.204	-4.178
ATOM	10	CG	MET	1	-1.569	-15.025	-5.164
ATOM	11	HG2	MET	1	-0.584	-14.697	-5.496
ATOM	12	HG3	MET	1	-1.439	-16.081	-4.929
ATOM	13	SD	MET	1	-2.581	-14.823	-6.550
ATOM	14	CE	MET	1	-1.543	-15.886	-7.544
ATOM	15	HE1	MET	1	-1.787	-16.896	-7.214
ATOM	16	HE2	MET	1	-1.875	-15.741	-8.571
ATOM	17	HE3	MET	1	-0.491	-15.627	-7.415
ATOM	18	C	MET	1	-0.970	-13.303	-1.883
ATOM	19	O	MET	1	-1.394	-13.166	-0.805
ATOM	20	N	ILE	2	-0.108	-12.402	-2.416
ATOM	21	H	ILE	2	0.383	-12.703	-3.245
ATOM	22	CA	ILE	2	0.353	-11.142	-1.726
ATOM	23	HA	ILE	2	-0.326	-10.880	-0.915
ATOM	24	CB	ILE	2	1.873	-11.237	-1.247
ATOM	25	HB	ILE	2	2.385	-11.437	-2.188
ATOM	26	CG2	ILE	2	2.319	-9.896	-0.620
ATOM	27	1HG2	ILE	2	3.408	-9.878	-0.597
ATOM	28	2HG2	ILE	2	1.889	-9.015	-1.096
ATOM	29	3HG2	ILE	2	1.826	-9.824	0.350
ATOM	30	CG1	ILE	2	2.042	-12.381	-0.277
ATOM	31	2HG1	ILE	2	1.680	-12.153	0.726
ATOM	32	3HG1	ILE	2	1.500	-13.272	-0.594
ATOM	33	CD1	ILE	2	3.553	-12.750	-0.091
ATOM	34	1HD1	ILE	2	3.617	-13.705	0.430
ATOM	35	2HD1	ILE	2	4.075	-12.752	-1.048
ATOM	36	3HD1	ILE	2	3.946	-11.937	0.520
ATOM	37	C	ILE	2	0.195	-10.040	-2.775

Figure 4.1. A sample of a MD trajectory file belonging to one of the snapshots

Cartesian coordinates' history of the backbone C^{α} atoms are placed in a (Mx3N) matrix **X**, where N (3N corresponds to q above) is the number of residues and M denotes the number of snapshots along the trajectory (Figure 4.2). The i^{th} row of matrix X includes Cartesian coordinate values of C^{α} atoms of i^{th} sample. x_{ij} stands for the x coordinate value of the j^{th} residue, so do y_{ij} and z_{ij}.

$$\begin{bmatrix}
x_{11} & y_{11} & z_{11} & x_{12} & y_{12} & z_{12} & \cdots & x_{1j} & y_{1j} & z_{1j} & \cdots & x_{1N} & y_{1N} & z_{1N} \\
x_{21} & y_{21} & z_{21} & x_{22} & y_{22} & z_{22} & \cdots & x_{2j} & y_{2j} & z_{2j} & \cdots & x_{2N} & y_{2N} & z_{2N} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\
x_{i1} & y_{i1} & z_{i1} & x_{i2} & y_{i2} & z_{i2} & \cdots & x_{ij} & y_{ij} & z_{ij} & \cdots & x_{iN} & y_{iN} & z_{iN} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\
x_{M1} & y_{M1} & z_{M1} & x_{M2} & y_{M2} & z_{M2} & \cdots & x_{Mj} & y_{Mj} & z_{Mj} & \cdots & x_{MN} & y_{MN} & z_{MN}
\end{bmatrix}$$

Figure 4.2. Cartesian coordinates history matrix **X** with size (Mx3N)

M x 3N matrix is computationally difficult to handle. PCA is applied to remove this problem. Eigenvalue decomposition is applied on covariance matrix of **X** with size 3Nx3N to find the loadings, and original **X** matrix is projected on these loadings to find the t_i vectors, which are the collective coordinates showing the time evolution of the protein along the i^{th} principal axis. By this way, t_i vectors to be analyzed by time series analysis are acquired.

4.2. Time Series Analysis

A time series is a set of observations generated sequentially in time. If the set is continuous, time series is said to be continuous, otherwise it is discrete. A statistical time series can be regarded as a realization of a *stochastic process* due to the fact that it can be defined as a dynamical system dependent on probabilistic laws.

It is assumed that there is some regularity in any stochastic process generating the time series. A frequently valuable way to view such regularity is through the concept of *stationarity*[31]. A stochastic process is strictly stationary if its probability distribution

does not change with respect to time. If the *mean* (μ) and *variance* (σ^2) of the process is the same at all points in time and the *covariance* between any two values of the series depends only on their distance apart in time, not on their absolute location in time, it is weakly stationary.

Box and Jenkins (1970) proposed the use of a class of models called *autoregressive integrated moving average (ARIMA) models,* and developed a methodology for fitting to data an appropriate member of this class. The ARIMA model-building approach is based on the following two restrictions:

1. The forecasts are linear functions of the sample observations.
2. The aim is to find efficiently parameterized models – that is, models that provide an adequate description of the characteristics of an observed time series with as few parameters as possible (the principal of parsimony).

4.2.1. Autoregressive Models

$z_1, z_2, \ldots, z_M$ denote a series of M discrete observations, the first order autoregressive (AR) model for these data is

$$z_t = C + \emptyset_1 z_{t-1} + a_t \tag{4.3}$$

In equation (4.3), $\emptyset_1$ is the autoregressive parameter, and a_t is a random variable, assumed to be independently identical distributed with zero mean and constant variance, σ_a^2 at all time periods t (white noise). It is assumed that there is no correlation between a_t and the error in any other time period. The parameter C is included to allow for the fact that the time series z_t can have non-zero mean. For generality, for any positive integer p, requiring that forecasts of future values be a linear function of the p most recent observations,

$$z_t = C + \emptyset_1 z_{t-1} + \emptyset_2 z_{t-2} + \ldots + \emptyset_p z_{t-p} + a_t \tag{4.4}$$

is the autoregressive model of order p, denoted AR(p).

μ denoting the mean of the series, we can substitute,

$$\check{z}_i = z_i - \mu \tag{4.5}$$

for all i = 1, 2, 3, . . ., M.

4.2.2. Moving Average Models

In moving average (MA) model, the value of a time series is expressed as a linear function of current and past values of a white noise process.

$$\check{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \tag{4.6}$$

The above process is the moving average model of order q, denoted MA(q), where θ_i (i = 1, 2, . . ., q) are moving average parameters.

4.2.3. Autoregressive (Integrated) Moving Average Models

Sometimes adequate representation of a series with an autoregression or moving average model can only be achieved with quite high-order models, while a model involving both autoregressive and moving average terms requires a relatively small total numbers of parameters[32]. The principal of parsimony suggests that the mixed autoregressive moving average formulation is better, in that case.

The stationary time series $\check{z}_t$ is said to be generated by an autoregressive moving average model of order (p,q), denoted ARMA(p,q), if

$$\check{z}_t = \emptyset_1 \check{z}_{t-1} + \emptyset_2 \check{z}_{t-2} + \ldots + \emptyset_p \check{z}_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \tag{4.7}$$

Using the backward shift operator B as

$$B^n \check{z}_t = \check{z}_{t-n} \tag{4.8}$$

ARMA(p,q) model can be represented as the following.

$$Ø(B)ž_t = \theta(B)a_t \tag{4.9}$$

Ø(B) and θ(B) are respectively autoregressive and moving average operators, where

$Ø(B) = 1 - Ø_1B - \ldots - Ø_pB^p$ and $\theta(B) = 1 - \theta_1B - \ldots - \theta_qB^q$

Autoregressive moving average models are successfully used to represent the behavior of stationary time series. For the ones which are nonstationary, differencing is necessary to obtain a stationary series. Difference operator is defined as follows:

$$w_t = ž_t - ž_{t-1} = (1-B)ž_t = \nabla ž_t \tag{4.10}$$

If we consider a nonstationary time series $ž_t$, and d denoting the number of differencing to make the original series stationary,

$$Ø(B)(1-B)^d ž_t = \theta(B)a_t \tag{4.11}$$

is called as an autoregressive integrated moving average model of order (p, d, q), denoted ARIMA(p, d, q). Usually one differencing is enough to make the resulting series stationary. In this case, an ARMA(p,q) model can be written as the following.

$$Ø(B)w_t = \theta(B)a_t \tag{4.12}$$

4.2.4. Autocorrelation Functions of AR, MA and ARMA Processes

The covariance between z_t and z_{t+k} is called the *autocovariance* at lag *k*, which is constant for a stationary process, defined by

$$\gamma_k = \text{cov}(z_t, z_{t+k}) = E[(z_t - \mu)(z_{t+k} - \mu)] \tag{4.13}$$

Autocovariances are difficult to interpret because their magnitudes depend on the units of measurement of the data. It is invariably preferable to work with correlations, which provide a scale-free measure of the strength of linear association. The correlations between values of a time series separated by *k* time periods are called the autocorrelations of the process, and denoted ρ_k, so that

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma_z^2} \tag{4.14}$$

Autocorrelation function (ACF) for an AR process satisfies the same form of the difference equation satisfied by the process itself.

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \ldots + \phi_p \rho_{t-k} \ , \ k>0 \tag{4.15}$$

which can be written as

$$\phi(B)\rho_k = 0 \tag{4.16}$$

where *B* operates on *k*. The general solution of Equation 4.16 is

$$\rho_k = A_1 G_1^k + A_p G_2^k + \ldots + A_p G_p^k \tag{4.17}$$

where G_1^{-1}, G_2^{-1}, ..., G_p^{-1} are the roots of the characteristic equation $\phi(B) = 0$ and A_1, A_2, . . . ,A_p are coefficients. If the roots are real, each term $A_i G_i^k$ geometrically decays to zero as *k* increases, which is referred to as a *damped exponential* (Figure 4.3a). If the pair of roots are complex, a damped sine wave (pseudo periodic behavior) term is encountered (Figure 4.3b) in the autocorrelation function as $A\mathrm{d}^k \sin(2\pi f_0 + F)$, where d is the damping factor, f_0 is the frequency in cycles and *F* is the phase. Frequency and damping factors can be found for any AR(2) or ARMA(2, *i*) process (*i* can be any positive integer) with complex roots by using the following relations:

$$\mathrm{d} = \sqrt{-\phi_2} \tag{4.18}$$

$$f_0 = \frac{1}{2\pi}\cos^{-1}\left(\frac{\phi_1}{2\sqrt{-\phi_2}}\right) \tag{4.19}$$

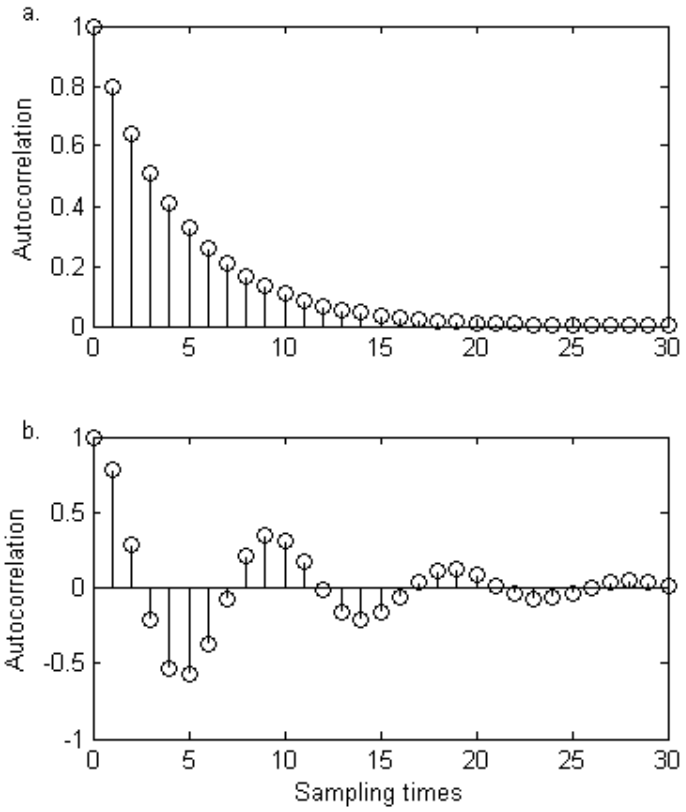


Figure 4.3. ACF of processes with (a) real , (b) complex G_1^{-1}, G_2^{-1}, ..., G_p^{-1} roots

The autocorrelation function of MA(q) process, unlike a AR process, is zero beyond the order q as given in Equation 4.20.

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \ldots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \ldots + \theta_q^2}, & k \le q \\ 0, & k > q \end{cases} \tag{4.20}$$

The autocorrelation function of a MA(1) process is represented in Figure 4.4a, where the autocorrelation function is nonzero only at lags zero and one.

For an ARMA(p,q) process, the autocorrelations at the first q lags are dependent both on the AR and MA parameters. The autocorrelation function of an ARMA(1,1) process is represented as an example in Figure 4.4b.

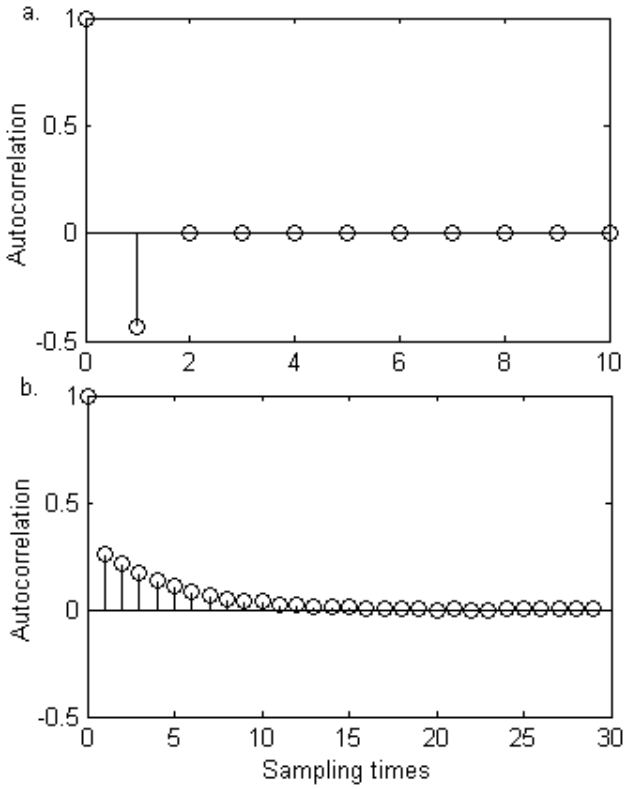


Figure 4.4. ACFs described by (a) an MA(1) process (b) an ARMA(1,1) process

4.2.5. Application Procedure of Time Series on the Scores Obtained From PCA

We need to understand the mechanism which generates t_i scores, so we apply time series analysis to find the best representative ARIMA(p,d,q) model for each of the scores, which are related to protein's internal motion.

Model identification is done according to the statistical properties of residuals, i.e. estimates of random shocks (a_t). Many appropriate models of different orders can be constructed to the available data. In choosing the most appropriate time series model for the scores trajectories, the following points are taken into consideration:

- Model order should be as low as possible for the principle of parsimony.
- Residuals of the representative parsimonious model should have autocorrelation function which is between confidence limits and they should be Gaussian distributed. If those are not satisfied, order of the model is increased by one until the conditions are satisfied.
- If there are very close factors to each other in AR and MA polynomial equations, these are canceled out, consequently lower ordered models are tested.
- To make the comparison of the model parameters possible, models ordered as similar as possible with each other are selected for each modes.
- It is checked if the acquired frequency is realistic with respect to the range of vibrational frequencies of proteins, if unrealistically slow frequency is encountered, a different model is tested.

The most encountered models are ARIMA(3,1,2) and ARMA(3,1). ARIMA(3,1,2) model is in the form of equation 4.21, where z_t stands for the t scores in question:

$$(1 - Ø_3B)(1 - Ø_1B - Ø_2B^2)\ \nabla z_t = (1 - \theta_1B)\ (1 - \theta_2B)a_t \qquad (4.21)$$

As explained in section 4.2.4, if the roots of the characteristic equation $\phi(B) = 0$ are real, this implies an overdamped system, whereas an underdamped behavior is observed if at least two of the roots are complex. Underdamped behavior yields frequencies of the system.

Underdamped behavior can be examined by a graph of $Ø_1$ and $Ø_2$ coefficients, shown in Figure 4.5. The gray section, representing the inequality "$Ø_1^2 + 4\ Ø_2 < 0$" is where underdamped behavior is encountered[31]. Triangle represents the stationary region.

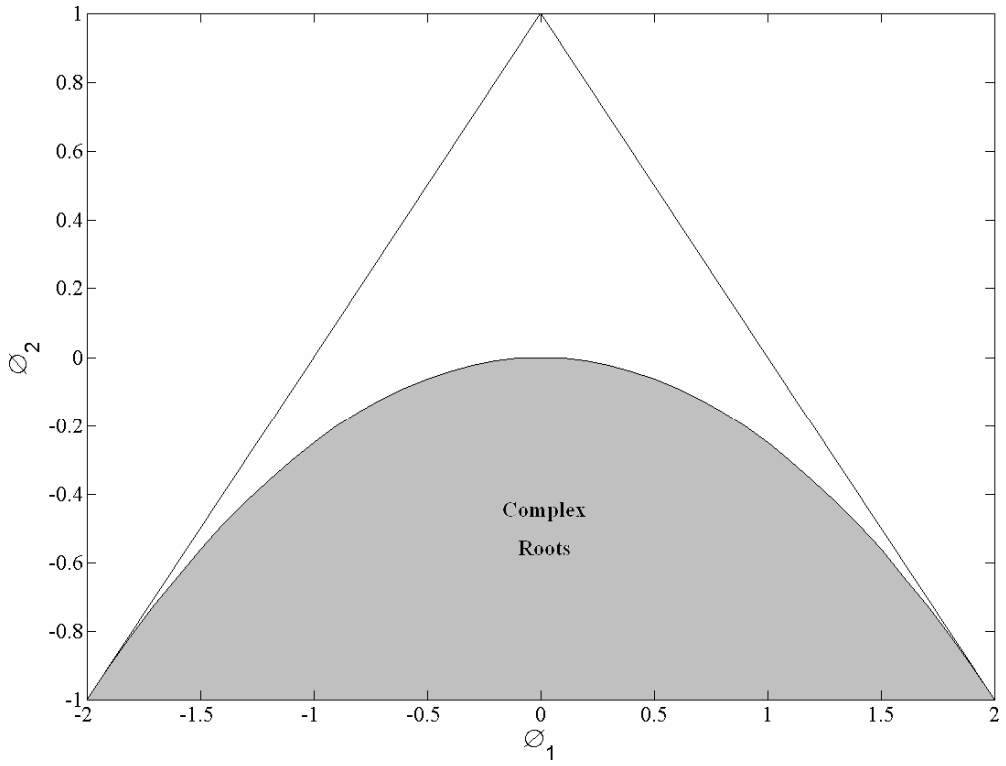


Figure 4.5. Boundaries for stationarity and underdamped behavior

5. RESULTS AND DISCUSSION

Analyses of DHFR and TIM dynamics are carried out in this section. Average conformations of the runs are compared with each other. Mean square fluctuations (MSF) of residues during the runs are calculated and PCs comparisons are carried out. For both DHFR and TIM, 60 PCs are analyzed using time series models. The differences between the free and ligand-bound states of DHFR and TIM are investigated in terms of internal dynamics. Additionally for TIM, the differences between the open and closed forms of the functionally important loop 6 in the free state are investigated.

5.1. Investigation of the Dynamics of DHFR

Three different MD runs are performed on DHFR, namely the free (apo) state (Run A1), $NADP^+$ bound state (Run A2) and NADPH bound state (Run A3). Each of the three runs is sampled twice from different 3.2 ns long parts, which possibly represent different regions on the potential energy surface. In Table 5.1, the notations that will be used in this thesis are summarized. A more detailed summary about the runs is given in Appendix.

Table 5.1. Notations for DHFR runs

	1^{st} Sample	2^{nd} Sample
Free State	Run A1-1	Run A1-2
$NADP^+$ Bound State	Run A2-1	Run A2-2
NADPH Bound State	Run A3-1	Run A3-2

Two different PCA calculations are made for the ligand bound states. In Case 1, only C^α atom coordinates of the DHFR are taken into consideration without the ligand. In Case 2, both C^α atoms of the DHFR and seven atoms (nodes) (see Figure 5.1 for the selection) from the ligand are considered. Taking seven nodes in $NADP^+$ roughly leads to the same mass-per node ratio in the ligand and the protein.

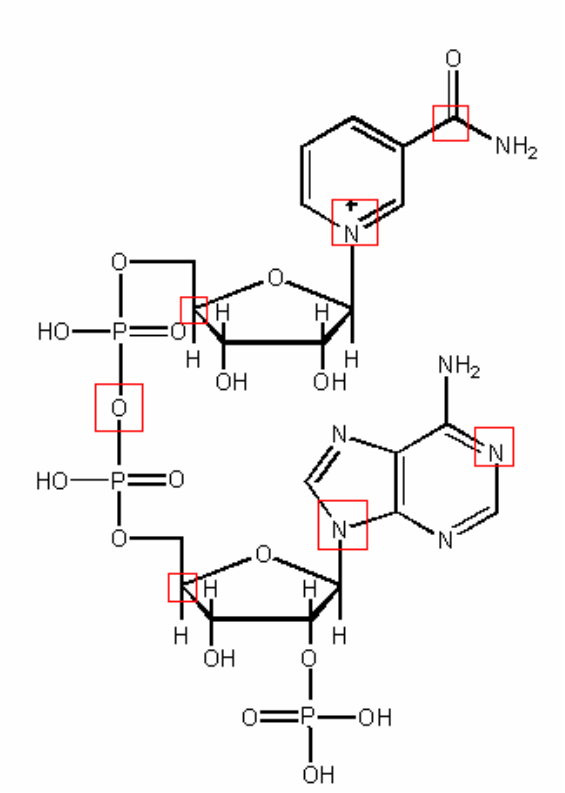


Figure 5.1. $NADP^+$ formula (the atoms in red rectangles are selected in Case 2)

5.1.1. Average Conformation and MSF of Free and Bound States for Case 1

The root-mean-square distance (RMSD) between the average conformations of Run A1-1 and Run A1-2 (inherent variability) is 0.842 Å. RMSD between Run A1-1 and Run A2-1 is 1.271 Å. RMSD between Run A1-1 and Run A3-1 is 1.090 Å. Thus, the average conformation of the NADPH bound state is closer to the free state than $NADP^+$ bound state is. Comparisons of the average conformations of different runs are illustrated in Figure 5.2 (Runs A1, A2, A3 are respectively in black, blue and red). In the NADP+ bound state, M20 loop is closer to FG loop (Figure 5.2b) compared to the other forms. GH loop is seen in two different conformations, which seems to be independent of the liganded form of the protein. CD loop, on the other hand, adopts different conformations for the free and liganded forms.

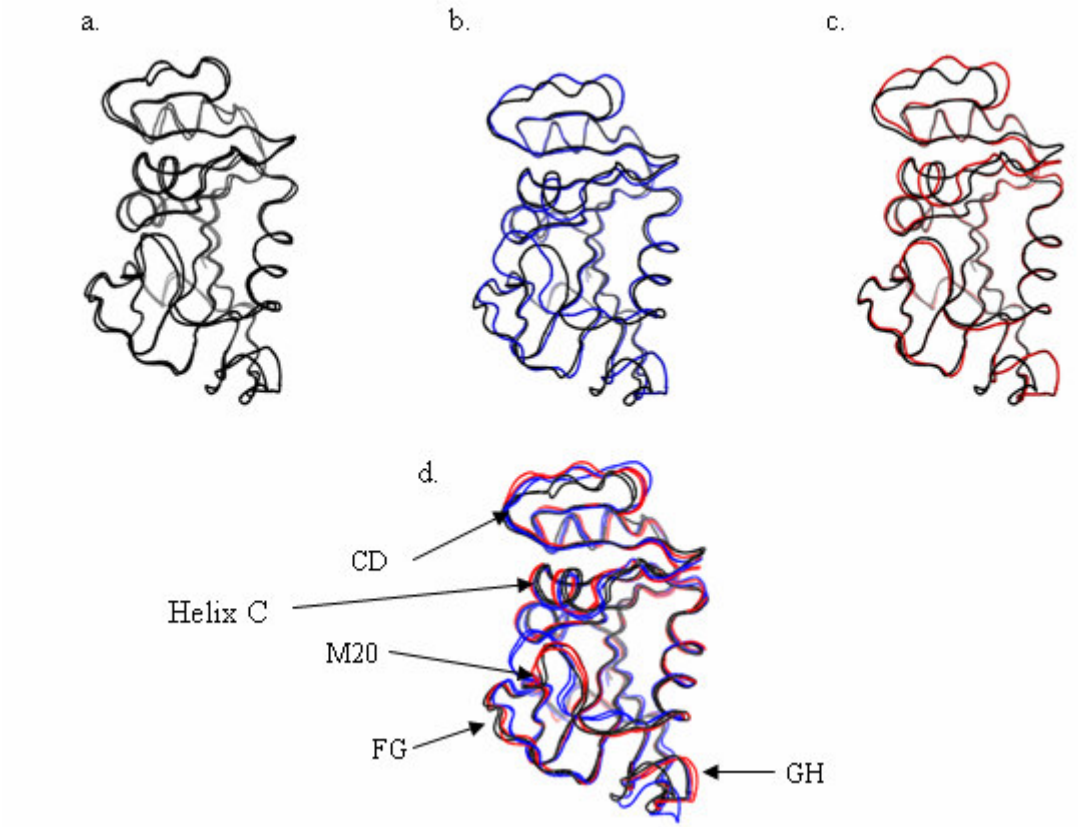


Figure 5.2. Average conformations of (a) Run A1-1 and A1-2, (b) Run A1-1 and A2-1, (c) Run A1-1 and A3-1, (d) Run A1-1, A1-2, A2-1, A2-2, A3-1, A3-2

Figure 5.3 shows the MSF of residues (Run Ai-1 and Ai-2 are averaged, i=1,2,3). Average MSF per C^{α} atoms in Run A1-1 and Run A1-2 is respectively 0.71 $Å^2$ and 0.55 $Å^2$ It is seen that M20 loop (residues 9-24) motion is restricted in the $NADP^+$ bound state (Run A2), exhibiting a significant MSF decrease compared to inherent variability. Along residues 40-55, which corresponds to the helix C (residues 44-50, helix region just above the M20 loop) and surrounding residues, free DHFR is more mobile than the bound states. The case that FG loop (residues 117-131), which is in neighbourhood of M20 loop, is most mobile is the NADPH bound one. The effect of NADPH on M20 loop is to increase the fluctuations of residue 18, while the rest of the loop behaves similar to the unliganded (free) case. The case that GH loop (residues 142-149) is most mobile is the $NADP^+$ bound state. In summary, binding of ligand affects not only the fluctuations of the binding region (M20 loop) but also other distant regions of the protein, as it can be seen in Figure 5.3.

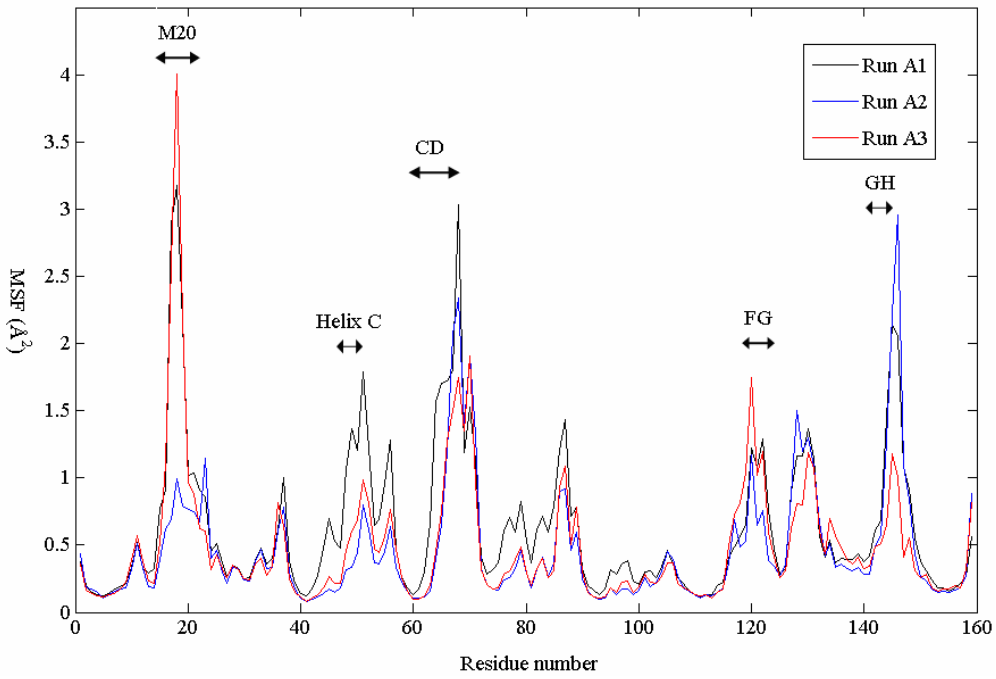


Figure 5.3. MSF of residues along the DHFR runs

5.1.2. PCA of Free and Bound States for Case 1

Table 5.2 gives the sums of eigenvalues of all runs, that is, the total MSF of all residues. It is apparent that most flexible state is free state, NADPH bound state is a bit more flexible than $NADP^+$ bound one.

Table 5.2. Sums of eigenvalues of DHFR run samples

Free State	Run A1-1	113.5 Å^2
	Run A1-2	88.1 Å^2
$NADP^+$ Bound State	Run A2-1	71.9 Å^2
	Run A2-2	75.9 Å^2
NADPH Bound State	Run A3-1	81.2 Å^2
	Run A3-2	77 Å^2

The first 5, 10 and 60 PCs of Run A1-1 explain 53%, 65% and 90% of C^α fluctuations respectively. Eigenvalues, which are the total MSF of residues along the PCs, and the percentage variance explanation of the modes of Run A1-1 and A1-2 can be seen

in Figure 5.4a-b. It can be seen that values are similar except the first modes, which shows that the first mode in run A1-1 has a highly collective character.

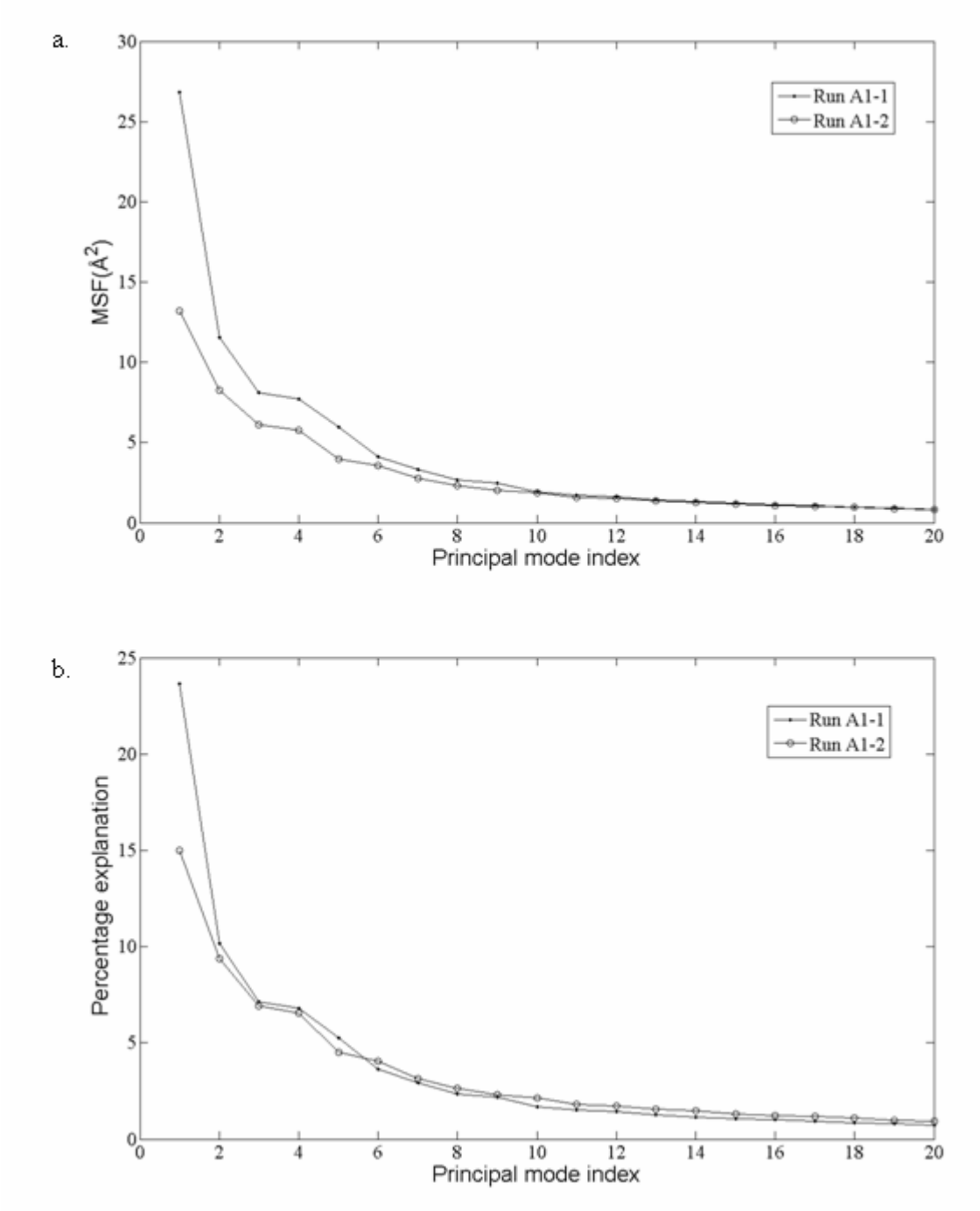


Figure 5.4. (a) Eigenvalues, (b) their percentage explanation of the total variance for the first 60 PCs (Run A1-1,2)

Figure 5.5 shows that the different samples from Runs A2 and A3 are similar in terms of the explanation power of their eigenvalues, indicating that the collectivity of PCs is similar within the same states.

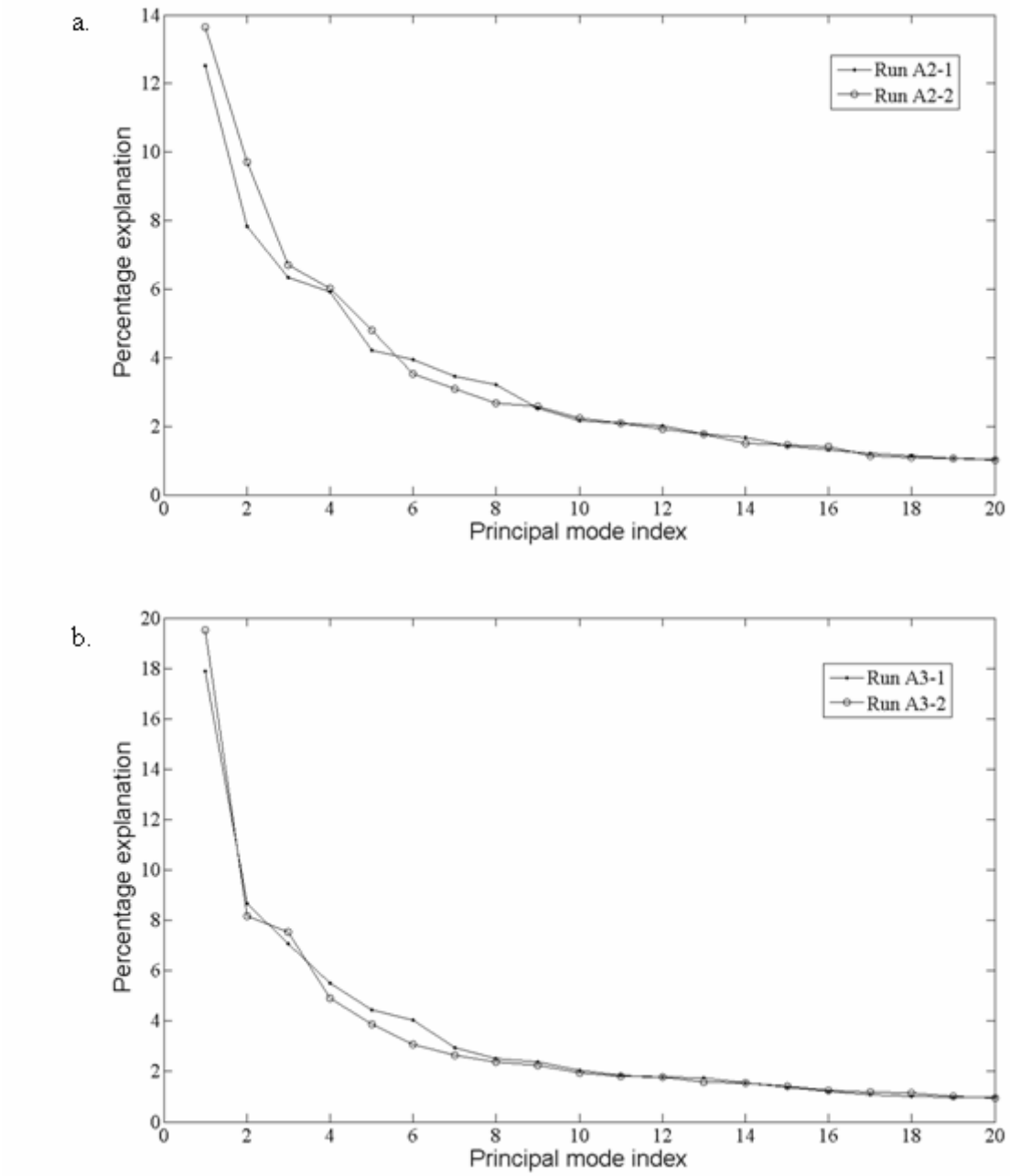


Figure 5.5. Eigenvalues' percentage explanation of the total variance for the first 60 PCs for (a) Run A2-1,2, (b) Run A3-1,2

Due to the fact that Run Ai-1 and Run Ai-2 (i = 1,2,3) exhibit similar behavior in terms of eigenvalues, two sample runs for each DHFR state are averaged for comparison (Figure 5.6a). It is observed that free DHFR fluctuates more than bound states along low indexed (collective) modes. To understand if these fluctuations are caused by dominance of collective motions of free DHFR or not, percentage explanation of eigenvalues should

be observed (Figure 5.6b). Percentage explanation reveals that ligand bound states, especially $NADP^+$ bound one, are a bit less collective than free state, especially for the first few eigenvectors.

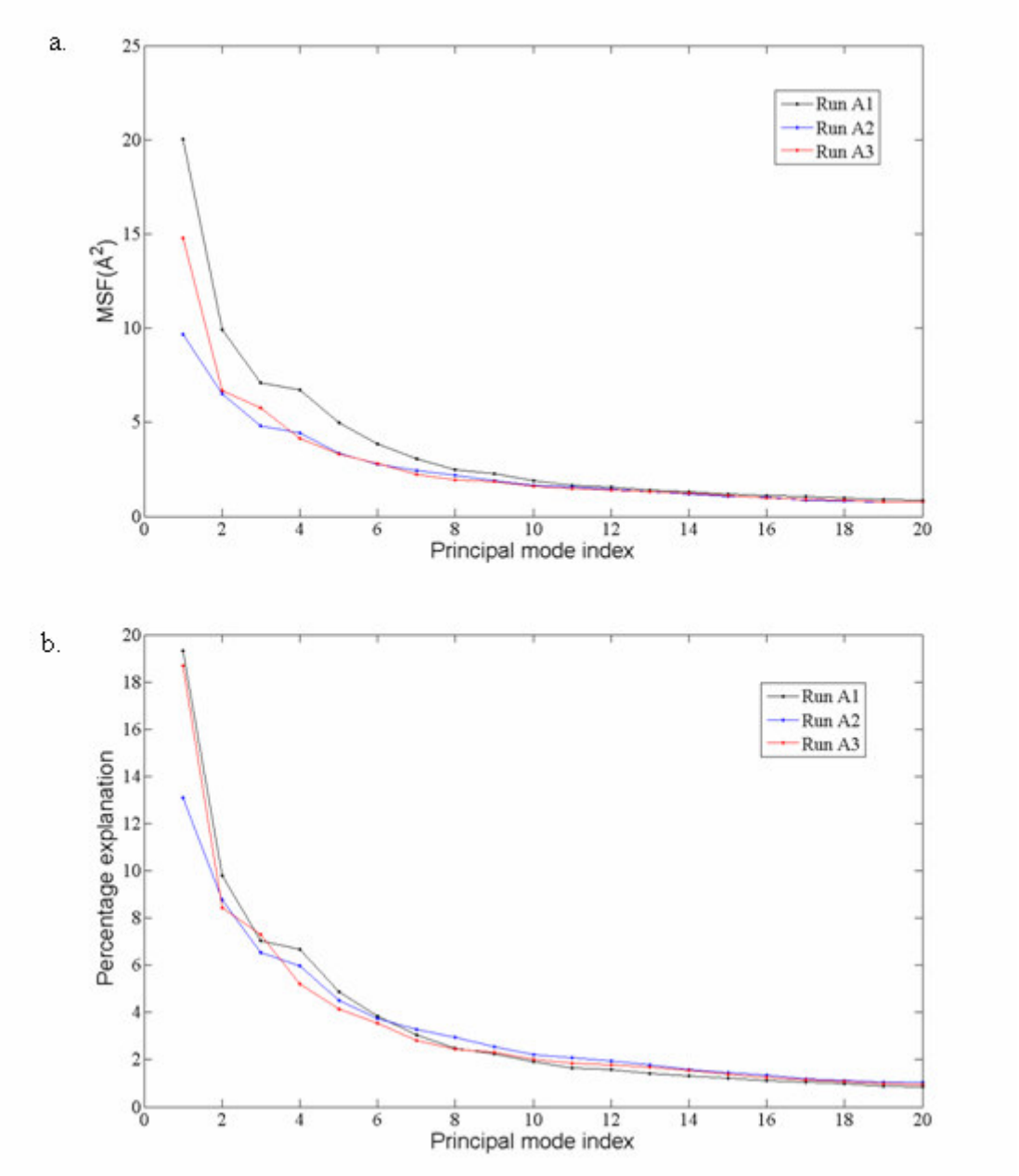


Figure 5.6. (a) Eigenvalues, (b) their percentage explanation of the total variance for the first 60 PCs (Runs A1, A2, A3)

It is observed that DHFR makes a twisting type of motion, as the upper part of M20 loop moves (see Figure 5.7) opposite to the lower parts of M20. To investigate this

twisting motion of the subdomains of DHFR, illustrations of DHFR prepared by PYMOL are shown from two different perspectives (side and top views) in Figure 5.7.

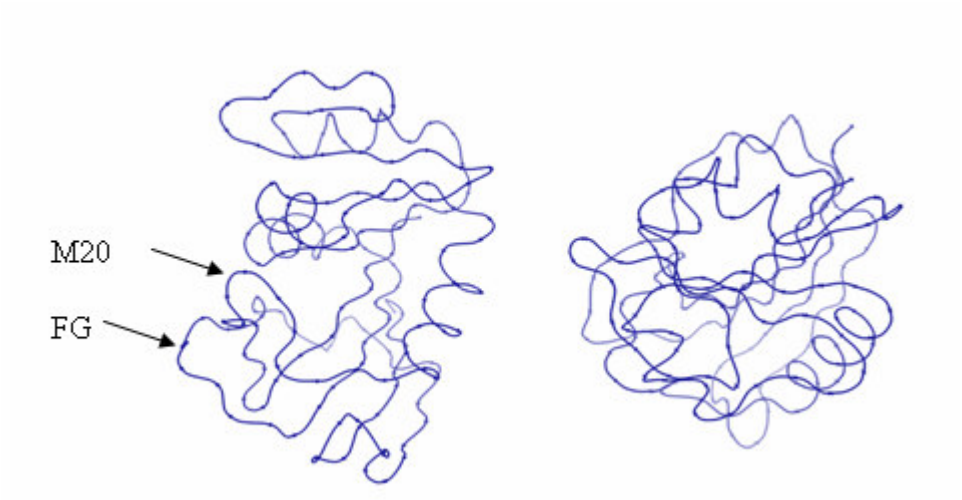


Figure 5.7. Native state C^{α} illustrations of DHFR (left: side view, right: top view)

Displacement vector illustrations (with the trace of DHFR omitted) of PC 1 from the same perspectives help one to understand collective motions better. Figure 5.8 includes displacement vector illustrations of Run A1-1 (free), Run A2-1 ($NADP^+$ bound) and Run A3-1 (NADPH bound) samples. Red dots represent the C^{α} atoms, and blue lines stand for direction of motion of C^{α}atom along PC 1. In first perspective of free DHFR (Figure 5.8a), one can see the closing motion of the M20 loop, while helix C moves in the opposite direction. The twisting motion can be seen better in the second perspective, where all the parts above M20 loop rotates in the opposite direction to the regions below of M20 loop. Therefore, both the opening/closing of the M20 loop and a global twisting motion is clearly seen in the unliganded case. In the first perspective of $NADP^+$ bound DHFR (Figure 5.8b) GH loop seems to have a local motion; twisting motion is not seen at all. In the first perspective of NADPH bound DHFR (Figure 5.8c), it is seen that the opening/closing of M20 loop is even better pronounced than the unliganded DHFR. However, the top view shows that the collective twisting motion is lessened in the presence of NADPH.

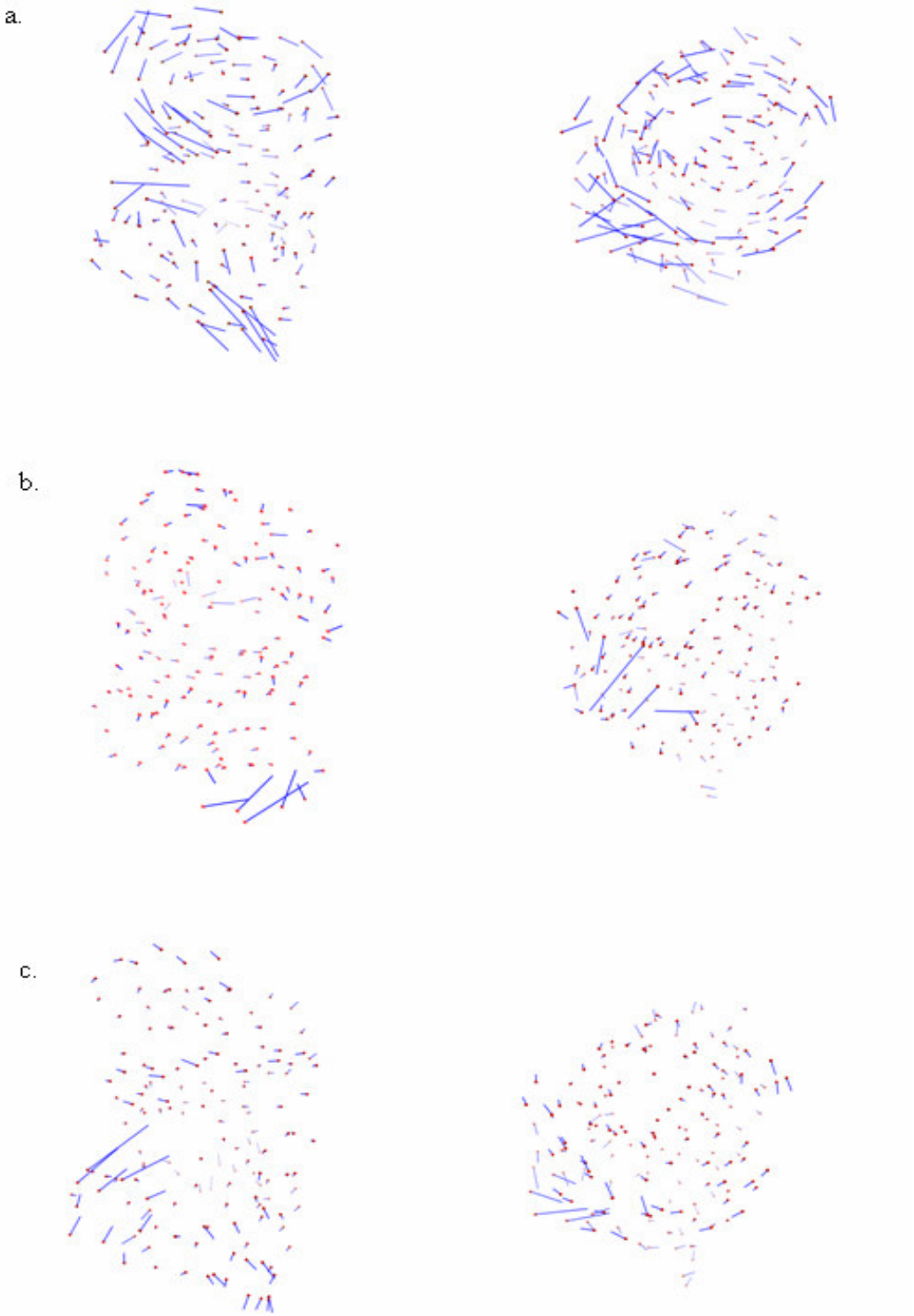


Figure 5.8. Vector field illustrations along PC 1 of Run (a) A1-1, (b) A2-1, (c) A3-1

Collectivity of a protein's motion can be described as cooperative (at the same time) movement of different parts of the protein. To have a clearer idea about the collective motions in DHFR, a number of snapshots are produced by moving the protein along PC 1

and PC 2, within ranges comparable to the correspondings MD runs for each of the six different samples (Run Ai-1, Ai-2, i=1,2,3). Figure 5.9 illustrates these superimposed snapshots of the two Run A1 samples. Along PC 1 of Run A1-1, opening/closing of the M20 loop onto the DHF binding site is clearly seen, with the collective network of residues. GH loop moves to the M20 loop. Helix C and CD loop move opposite to M20 loop. This type of anticorrelated motion of helix C and CD loop with the M20 loop was also observed in another study[22]. Along PC 2 of Run A1-1, lateral motion of the M20 loop and its opening/closing onto NADPH binding site is clearly seen, with the collective network of residues. GH loop moves parellel to the M20 loop. Along PC 1 of Run A1-2, the collective character is evident including the M20 loop motions, however the M20 loop motion is not exactly of opening/closing type. It should be recalled that the the percentage variance explained by the first eigenvalue in Run A1-1 has been the highest, nearly 25%, compared to 15% of the first eigenvalue in Run A1-2. Helix C and CD loop move opposite to M20 loop. Along PC 2 of Run A1-2, lateral motion of the M20 loop is seen, with the collective network of residues. GH loop moves parellel to the M20 loop. Helix C and CD loop move opposite to M20 loop. One can summarize that though there may be differencies, both unliganded runs show similar collective behavior of residues.

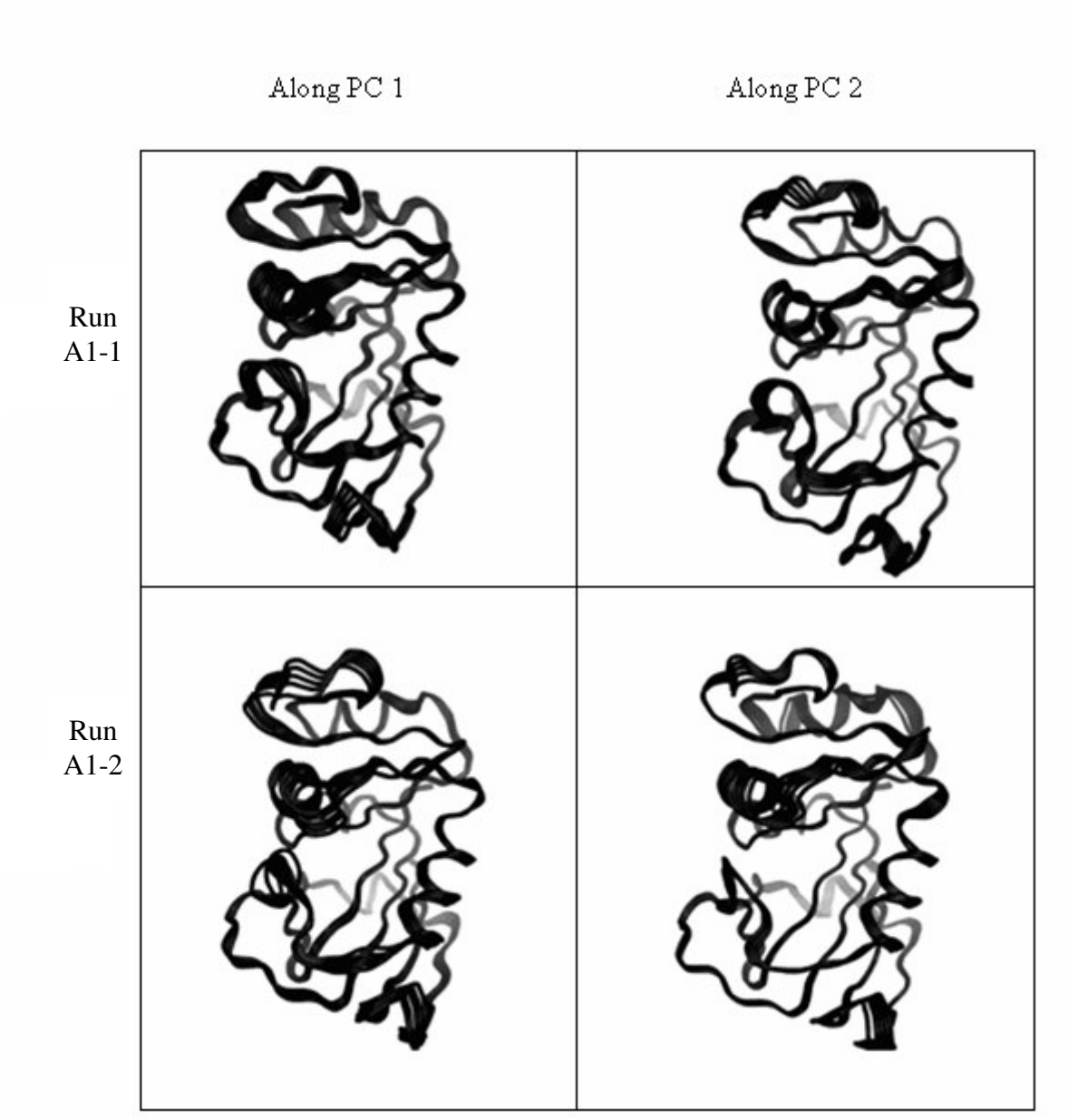


Figure 5.9. Projections of the free DHFR conformations onto PCs 1 and 2

Figure 5.10 illustrates the snapshots of two Run A2 samples along PC 1 and PC 2. Along PC 1 of Run A2-1, CD loop dominates the motion. M20 loop does not participate the motion along this mode. Along PC 2 of Run A2-1, CD loop and M20 loop moves together, however M20 loop makes a lateral motion, not a opening/closing type of motion. Along PC 1 of Run A2-2, CD and GH loops dominate the motion. M20 loop does not participate the motion along this mode. Along PC 2 of Run A2-2, GH loop dominates the motion, however some collective character is also present. A slight tendency of the opening/closing motion of the M20 loop can be seen. It is possible to say that the

collectivity of motions seen in the unliganded DHFR is drastically reduced in the presence of NADP+, and the closing of the M20 loop cannot be seen in the low indexed PCs. The motios with high MSF are restricted to the local regions of the protein.

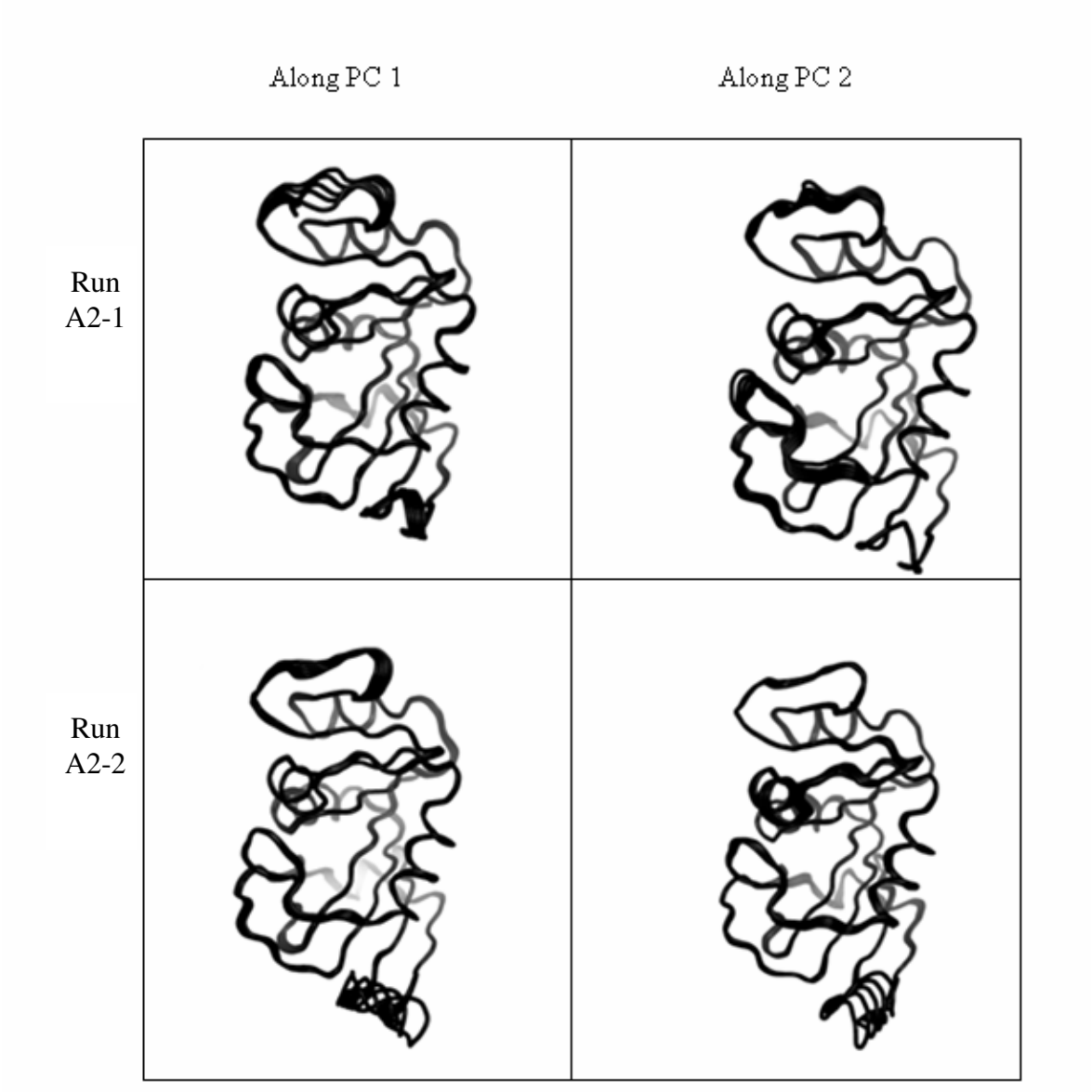


Figure 5.10. Projections of the NADP+ bound DHFR conformations onto PCs 1 and 2

Figure 5.11 illustrates the snapshots of two Run A3 samples along PC 1 and PC 2. Both along PC 1 and PC 2, opening/closing motion of M20 loop is more apparent than free state. Though the collectivity of the protein motions is similar to the unliganded form, the contributions of CD, GH loops and helix C to the collective motions are lessened.

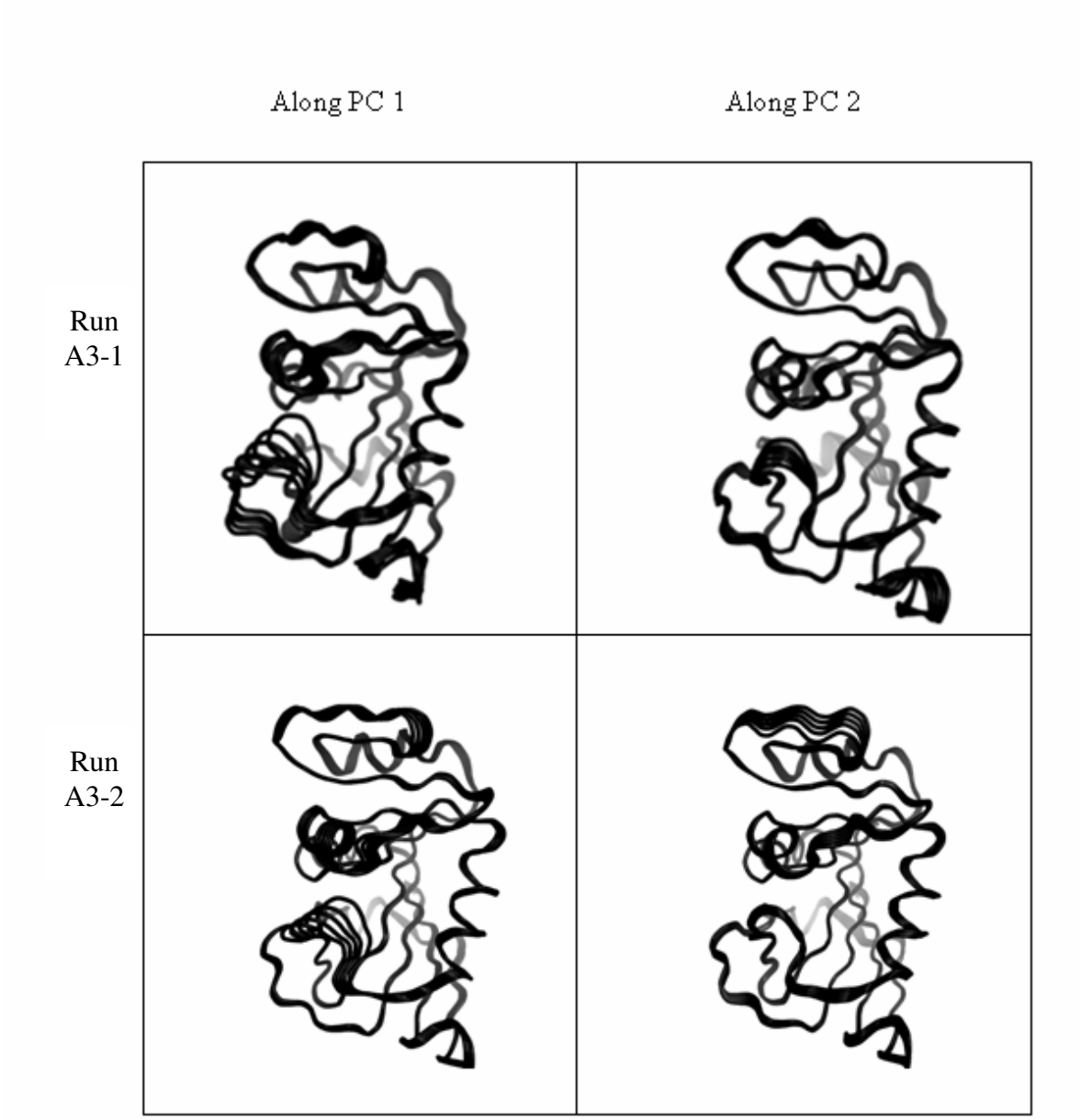


Figure 5.11. Projections of the NADPH bound DHFR conformations onto PCs 1 and 2

5.1.3. Derivation of Time Series Models of Run A1-1 for Case 1

The t_1 scores, which are the trajectory projections along the above shown PC 1 of Run A1-1, have a variance of 26.8 $Å^2$ and explain 23.6 % of the total variation, as shown in Figure 5.4b. This time series is apparently nonstationary as it is seen from Figure 5.12a that the levels of the scores are not constant during the run. In addition, probability density function (PDF) of the scores reveals the non-Gaussian distribution (Figure 5.12b).

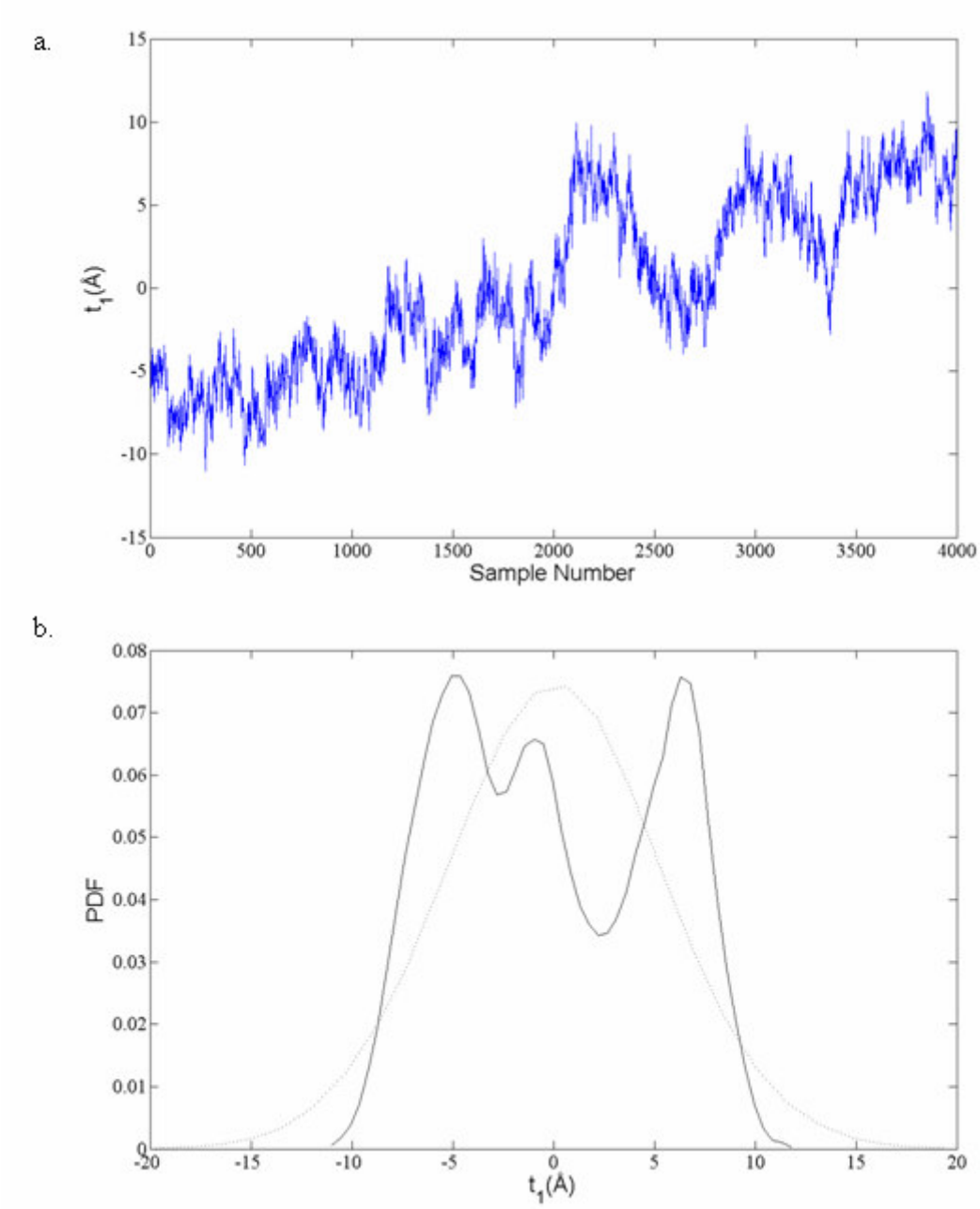


Figure 5.12. (a) Time trajectory, (b) PDF of the t_1 scores belonging to Run A1-1 (dashed lines show the Gaussian distribution with the same mean and variance)

Another clue that makes one think that a time series is nonstationary is the autocorrelation function of the series. Autocorrelation function of t_1 scores fails to die out to zero (Figure 5.13).

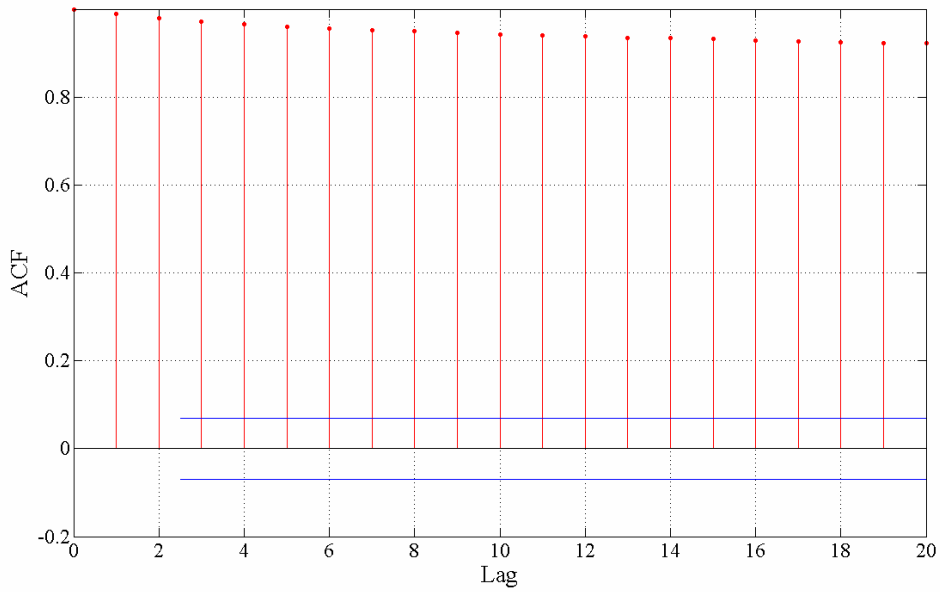


Figure 5.13. Autocorrelation function of t_1 scores belonging to Run A1-1

It is expected that application of the difference operator once ($d = 1$) on t_1 scores will lead to a stationary time series by removing the trend. Taking $w_t = \nabla z_t$ (z_t will represent the score being mentioned, i.e. t_1 for this case), w trajectory and its PDF are shown in Figure 5.14a-b. The resulting process (w_t) has a constant mean and variance and its distribution is Gaussian.

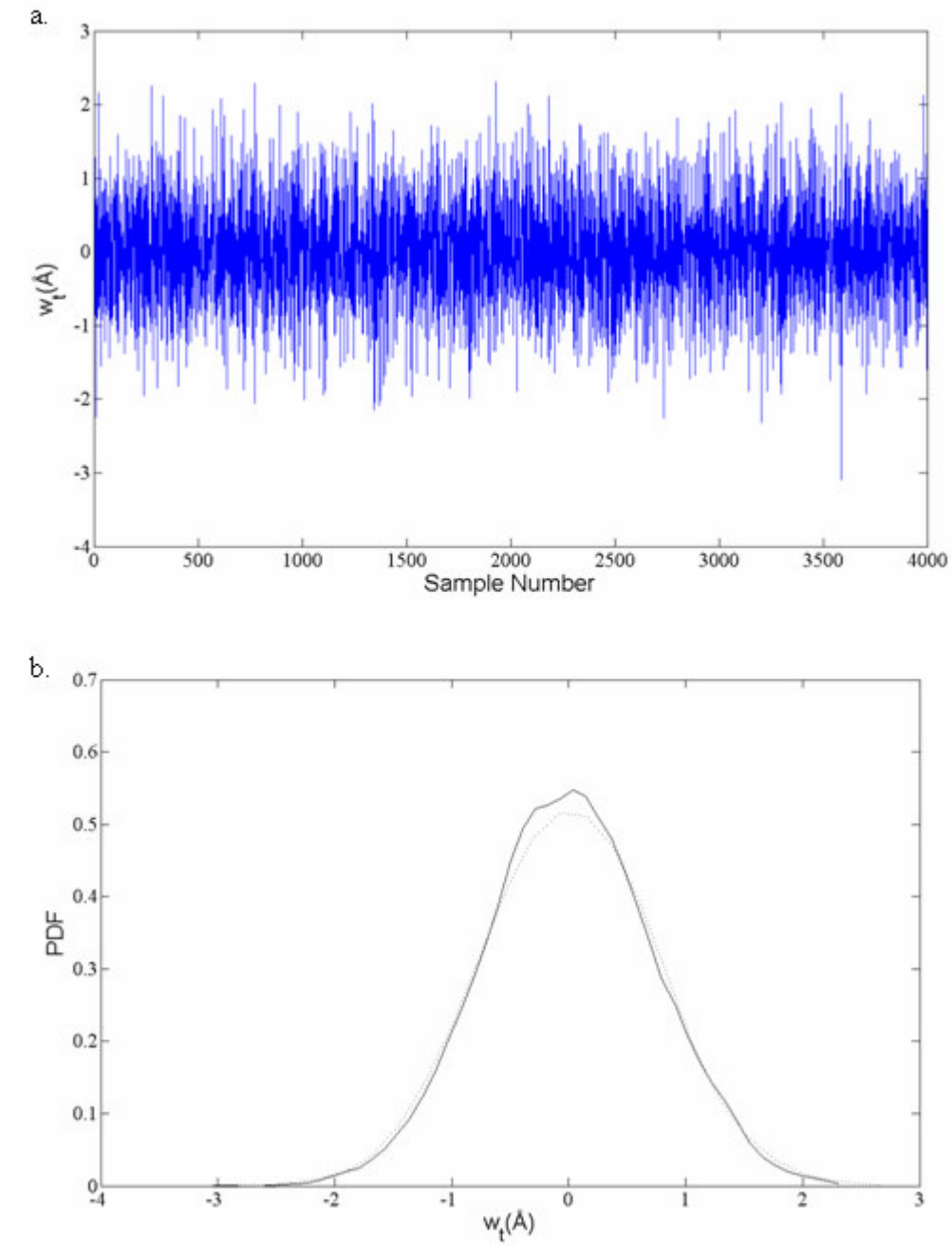


Figure 5.14. (a) w_t trajectory obtained by differencing the t_1 scores (b) PDF of w_t

In a previous study[1], motion of a small protein in water has been explained by ARIMA(2,1,1) or ARIMA(2,1,2) models. AR(2) part corresponds to pseudo-periodic intraminimum oscillations, while I(1) corresponds to interminimum motions. MA part represents the autocorrelation between the random forces acting on the protein, which correspond to intramolecular forces in the protein and intermolecular forces between the protein and the water molecules. It is seen that stochastic time series models derived in this

study, though an additional AR terms is added (see below), still conform the protein fluctuations model described above.

Autocorrelation of w_t (Figure 5.15) makes one suspect the existence of an pseudo-periodic motion, requiring p parameter to be at least two for the ARMA(p, q) process. In the same figure, dashed lines show the 95% confidence limits for zero autocorrelation.

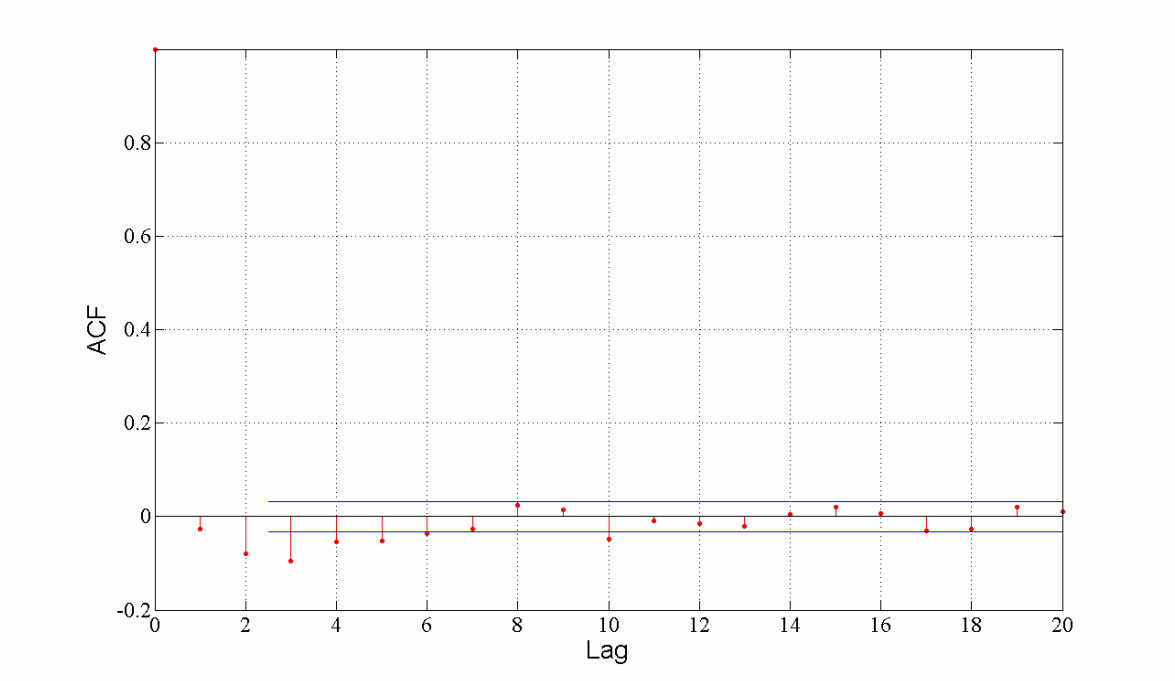


Figure 5.15. Autocorrelation function of w_t

MATLAB System Identification Toolbox is used to find the least-squares estimates of ARIMA(2,1,1), ARIMA(2,1,2), ARIMA(3,1,1) and ARIMA(3,1,2) processes. Considering the residuals (see the next paragraph) and the time series models for other modes, ARIMA(3,1,1) is found to be the most appropriate one. The obtained parameters and σ_a^2, variance of the residuals, are given below:

$$(1 - 0.7445B)(1 - 0.1005B - 0.0247B^2)\,\nabla z_t = (1 - 0.9137B)\,a_t\,,\ \sigma_a^2 = 0.5008\ \text{Å}^2 \qquad (5.1)$$

It is seen in Figure 5.16a that the residuals (a_t) have a Gaussian distribution. The autocorrelation function of the residuals does not significantly exceed the 95% limits (dotted lines) for the first 50 lags, as seen in Figure 5.16b.

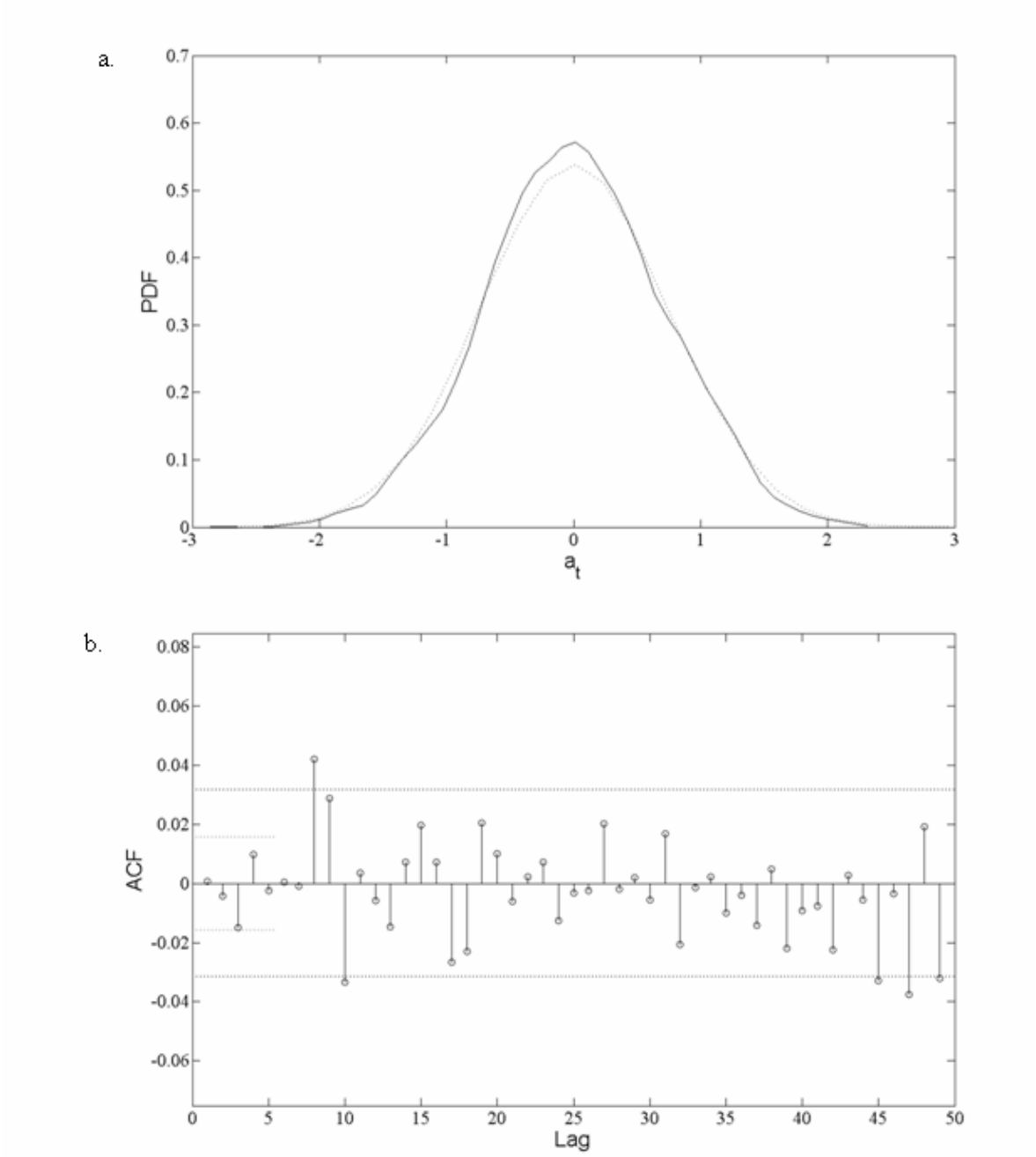


Figure 5.16. (a) PDF (b) autocorrelation function of the residuals

AR characteristic equation has the (1 - 0.1005B - 0.0247B^2) polynomial term, which is 2nd order. Roots of this equation are real, which show the existence of an overdamped behavior. Had the roots been complex, pseudo-periodic oscillatory (underdamped) behavior would have been detected. The overdamped behavior of the low indexed modes is attributed to the dampening effect of water [13].

The difference operator shows the nonstationary behavior along the mode. The two other terms are the (1-0.7445B) term in the AR polynomial and (1-0.9137B) term in the MA polymonial, which do not cancel each other. Had they cancelled, then we would have a simple AR(2) overdamped or underdamped process. The existence of this additional AR term will be discussed below.

The rest of time series analysis of Run A1-1 modes (scores) is done. First 17 scores are nonstationary. However, there appear a few stationary modes among these 17 modes, as well as a few nonstationary modes among the rest of the stationary 43 modes. Among nonstationary models, the one with highest number of parameters is ARIMA(3,1,2), which corresponds ARMA(3,2) for stationary models. For generality, nonstationary models are in the form:

$$(1 - \emptyset_3 B)\,(1 - \emptyset_1 B - \emptyset_2 B^2)\,\nabla z_t = (1 - \theta_1 B)(1 - \theta_2 B)a_t \qquad (5.2)$$

Stationary models are in the form:

$$(1 - \emptyset_3 B)(1 - \emptyset_1 B - \emptyset_2 B^2)\, z_t = (1 - \theta_1 B)(1 - \theta_2 B)a_t \qquad (5.3)$$

As representative examples, time series models extracted for t_2, t_{20}, t_{40} and t_{60} scores are given below:

t_2: $(1–0.945B)(1–0.572B+0.082B^2)\nabla z_t = (1–0.983B)(1\text{-}0.566B)\, a_t,\ \sigma_a^2 = 0.5685$ (5.4)

t_{20}: $(1–0.962B)(1–0.258B\text{-}0.013B^2)\nabla z_t = (1–0.735B)a_t,\ \sigma_a^2 = 0.3316$ (5.5)

t_{40}: $(1 - 1.015B + 0.093B^2)\, z_t = (1 - 0.666B)\, a_t,\ \sigma_a^2 = 0.2333$ (5.6)

t_{60}: $(1 - 0.884B)\, z_t = (1 - 0.597B)(1 - 0.077B)\, a_t,\ \sigma_a^2 = 0.1455$ (5.7)

Mode 2 (t_2) is a typical nonstationary mode expressed by ARIMA(3,1,2) model. $(1–0.572B+0.082B^2)$ term in the characteristic equation denotes the intraminimum motion. t_2 exhibits underdamped behavior, with a frequency of 1.308 cm^{-1}. Difference operator shows

the interminimum motion. The (1–0.945B) term in the AR polynomial equation and the (1–0.983B) term in the MA polynomial equation, though they are close, do not cancel each other. These additional AR and MA terms have not been encountered in the previous studies [10-13]. However, proteins of current study are larger than the one in the mentioned studies. Presumably, more terms are required to approximate the nonlinear behavior of larger proteins, which seems to be the case here. Therefore, it is assumed that these parameters are not significant in describing protein fluctuations, but only used to derive the time series models in a statistically acceptable way. To confirm this hypothesis, $Ø_3$ and θ_1 values are compared in Figure 5.17. For the nonstationary modes, the θ_1 term is always slightly higher than $Ø_3$ term, and they do not cancel other. For the stationary modes, the θ_1 term is always lower than $Ø_3$ term, due to the absence of an additional difference operator. For both types of modes, $Ø_3$ and θ_1 terms are well correlated. Comparison of $Ø_3$ versus θ_1 values of the other runs does not show any significant difference between runs (figures not shown), thus it is concluded that these additional terms only serve the purpose to make a better linear approximation.

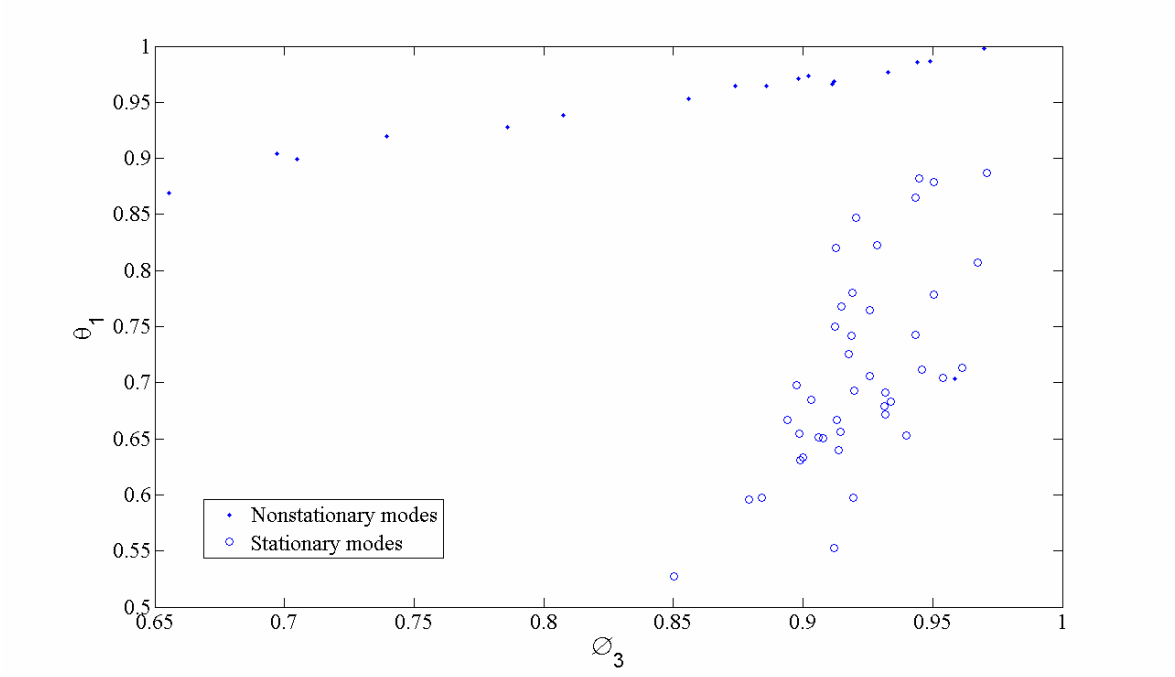


Figure 5.17. $Ø_3$ versus θ_1 graphic for Run A1-1

5.1.4. Comparison of Time Series Models of Free and Bound States for Case 1

Run A1-2, Run A2-1,2 and Run A3-1,2 data are analyzed in the same way. Table 5.3 shows all the derived models. Most of the nonstationary models encountered are ARIMA(3,1,2). Most of the stationary models are ARMA(3,1). These models and the corresponding modes are shown in italic in the table.

Table 5.3. Number of principal modes with respect to their model orders (DHFR)

Type and order of the model	# of modes (A1-1)	# of modes (A1-2)	# of modes (A2-1)	# of modes (A2-2)	# of modes (A3-1)	# of modes (A3-2)
ARIMA(1,1,1)	2	0	1	0	0	0
ARIMA(1,1,2)	0	0	0	0	2	0
ARIMA(2,1,1)	0	0	0	0	1	1
ARIMA(2,1,2)	0	0	0	0	1	0
ARIMA(3,1,1)	1	2	6	4	4	6
ARIMA(3,1,2)	*13*	*7*	*4*	*10*	*3*	*4*
ARMA(1,1)	1	1	0	0	1	0
ARMA(1,2)	3	6	3	0	10	3
ARMA(2,1)	10	16	5	5	5	1
ARMA(2,2)	8	4	5	4	13	4
ARMA(3,1)	*18*	*22*	*26*	*18*	*9*	*24*
ARMA(3,2)	4	2	10	19	11	17

Parameters of the time series models obtained for different runs are compared to see possible differences between different unliganded and liganded forms. Neglecting the additional AR and MA terms, the terms to be compared are the variance of random shocks (σ_a^2) term, the ($1 - \theta_2 B$) term in the MA polynomical equation, and ($1 - Ø_1 B - Ø_2 B^2$) term in the AR characteristic equation. Comparisons of each of these terms are presented in the following paragraphs.

In Figure 5.18, variance of random shocks driving the system (σ_a^2) belonging to the two sample averages of free and ligand bound states are compared. It is seen that σ_a^2 terms of the free state are generally higher than those of bound states. Difference is more significant for the first three modes, the low indexed modes. It is also seen that NADP+ bound state has the lowest σ_a^2 for most of the low indexed modes. This can be explained by the fact that lowest indexed modes stand for the most highly anharmonic motion; directions along which interminimum motions are highly affective. It is seen that interminimum motions are more pronounced in the unliganded form, whereas NADP+ restricts these motions. It should also be pointed out that these low indexed modes correspond to slow and concerted (collective) type of motions, that is, functional motions of protein. When a protein is bound with a ligand, it is plausible that the collective fluctuations along these lower indexed modes will be restrained.

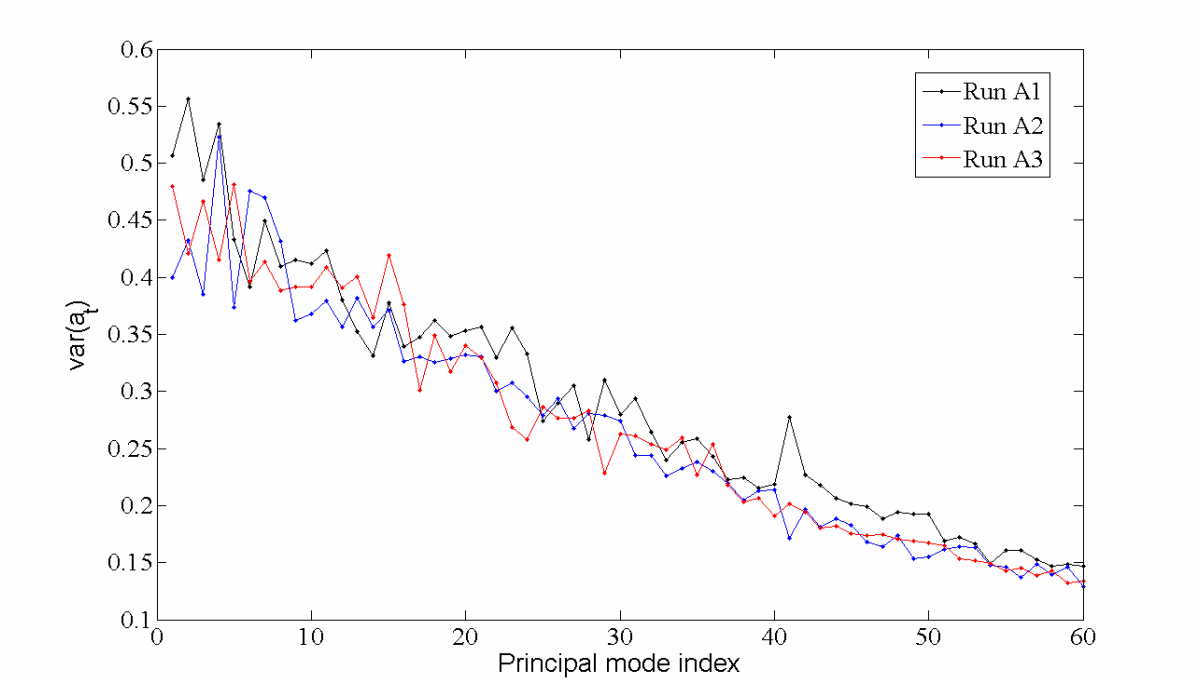


Figure 5.18. Comparison of residual variances with respect to modes for DHFR

In a previous study[1], MA parameters have been used to reveal differences between vacuum-water environment: it was seen that the existence of water drastically changed the autocorrelation of random force terms acting on the protein In the current study, boxplot analysis of θ_2 is done to see possible differences between free ve bound states. In boxplot

diagrams, samples, after being filtered of outliers, are divided into four quartiles. The line at the center shows the median of the samples. A boxplot can be imagined as a bird's eye view of the PDF of the samples. Boxplots of θ_2 in different run samples (Figure 5.19a-c), show that there is a considerable sample-to-sample variation in terms of this parameter: the range of this parameter can be in between -1 to 1, and most of this range is covered in some of the boxplots. These comparisons in between the simulation samples of the protein in the same form give insight about how much variability can be tolerated when comparing the free DHFR behavior with the ligand bound ones.

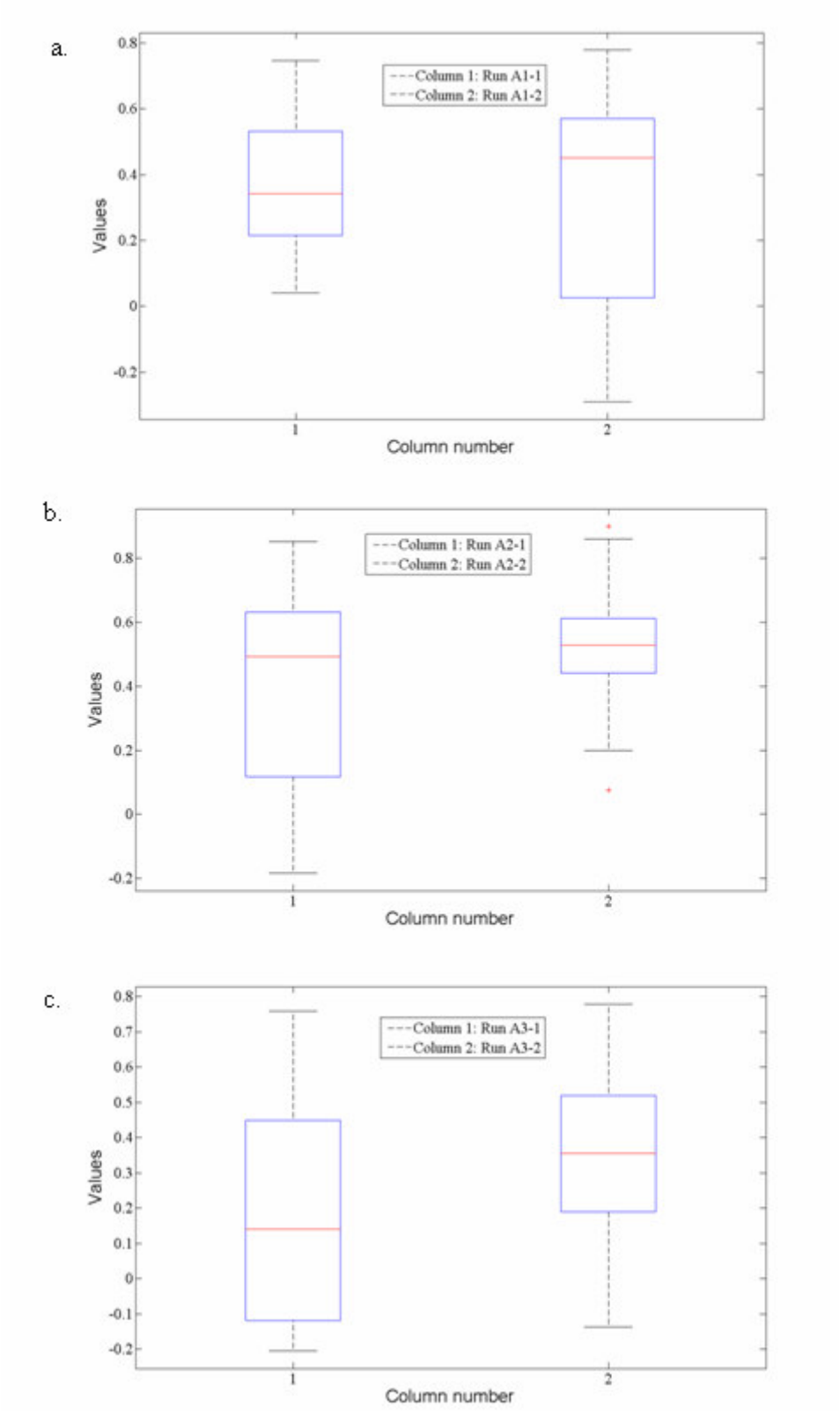


Figure 5.19. Boxplot of θ_2 roots of (a) Run A1 (b) Run A2 (c) Run A3

Run Ai-1 and Ai-2 (i = 1,2,3) are appended, and the boxplot of all the of θ_2 values in Figure 5.20 do not give any conclusive result about the differences of the autocorrelation of the random force term in between the free and bound forms. The autocorrelation of the random forces which the NADP+ bound DHFR is experiencing may be different from the other two forms, but more simulations are required to make this point clear.

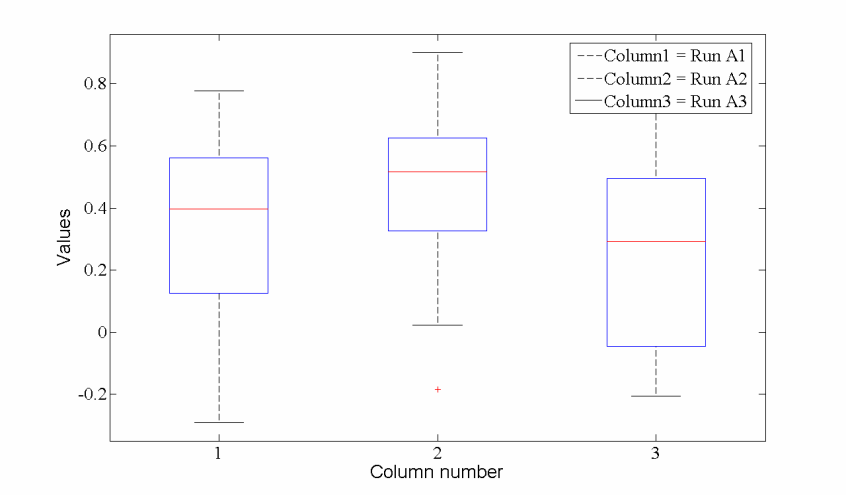


Figure 5.20. Boxplot of θ_2 roots of DHFR runs

It is possible to obtain the frequencies of the pseudo-periodic intraminimum motion by using the $(1 - Ø_1B - Ø_2B^2)$ term in the AR characteristic equation. If the roots are complex, then that mode is underdamped. It is known that water has dampening effect on modes, so it is expected that especially low indexed modes are overdamped, e.g. PC 1. In Table 5.4, numbers of underdamped modes among the first 60 modes are shown.

Table 5.4. The number of underdamped modes for Case 1

	CASE 1
RUN A1-1	24
RUN A1-2	13
RUN A2-1	20
RUN A2-2	22
RUN A3-1	10
RUN A3-2	22

We can examine frequencies by using histograms. Figure 5.21 is the comparison of frequencies of all runs. It is seen that the lowest frequencies are in the unliganded DHFR, whereas the number of lowest frequencies is slightly reduced in NADP+ bound DHFR, and the number of lowest frequencies is remarkably reduced in NADPH bound DHFR.

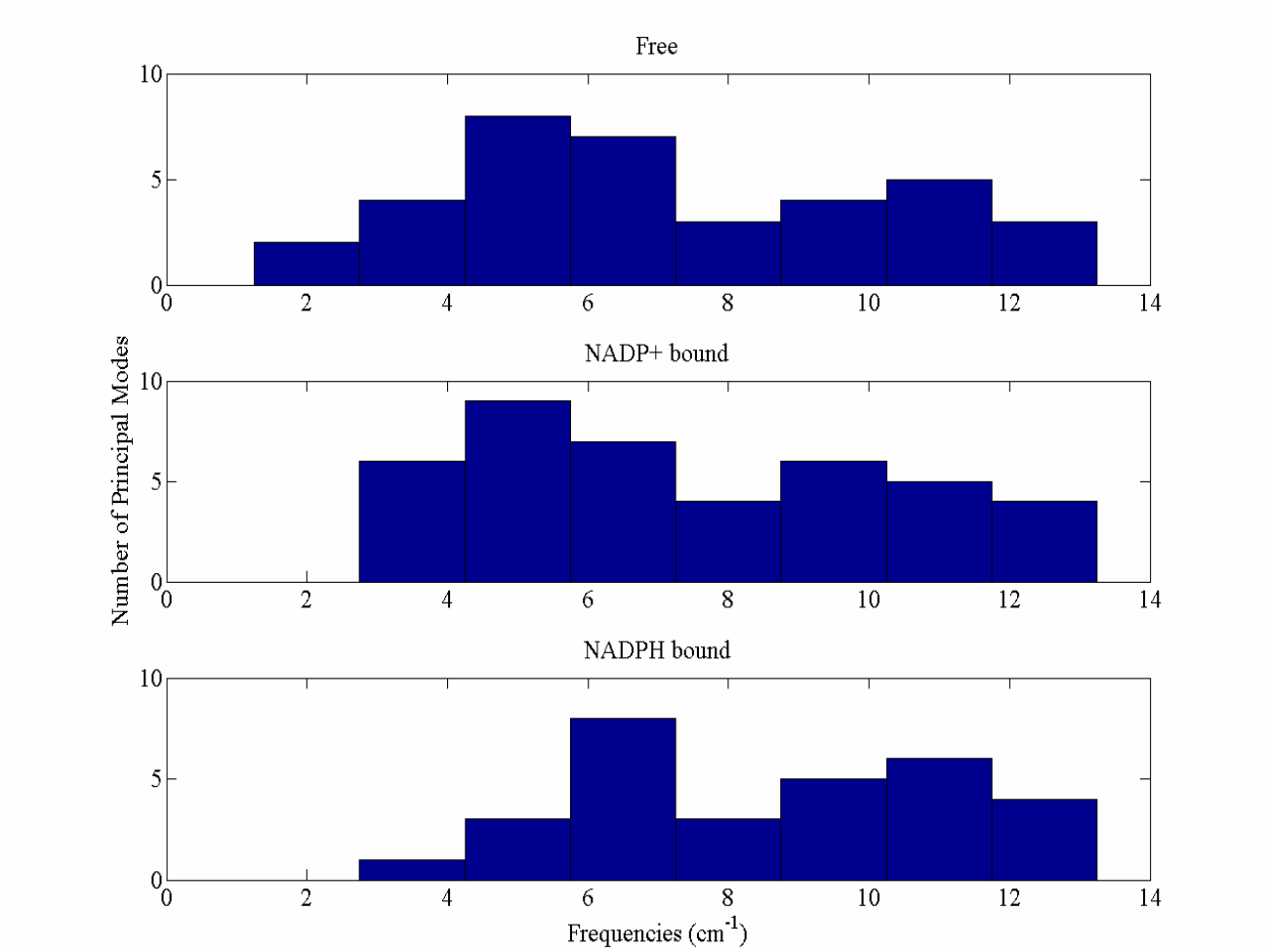


Figure 5.21. Histogram graphics of DHFR frequencies (Case 1)

Figure 5.22a-c shows the plot of cumulative probability density functions (CDF) of frequencies of the principal modes in Runs A1-1, A1-2, A2-1, A2-2, A3-1, A3-2. It is also observed that there is not significant difference between two samples of each run, especially for the free and NADP+ bound cases. It is observed that frequencies of NADPH bound case samples are remarkably different, this may be caused by the fact that Run A3-1 sample yields low number of underdamped modes (see Table 5.4).

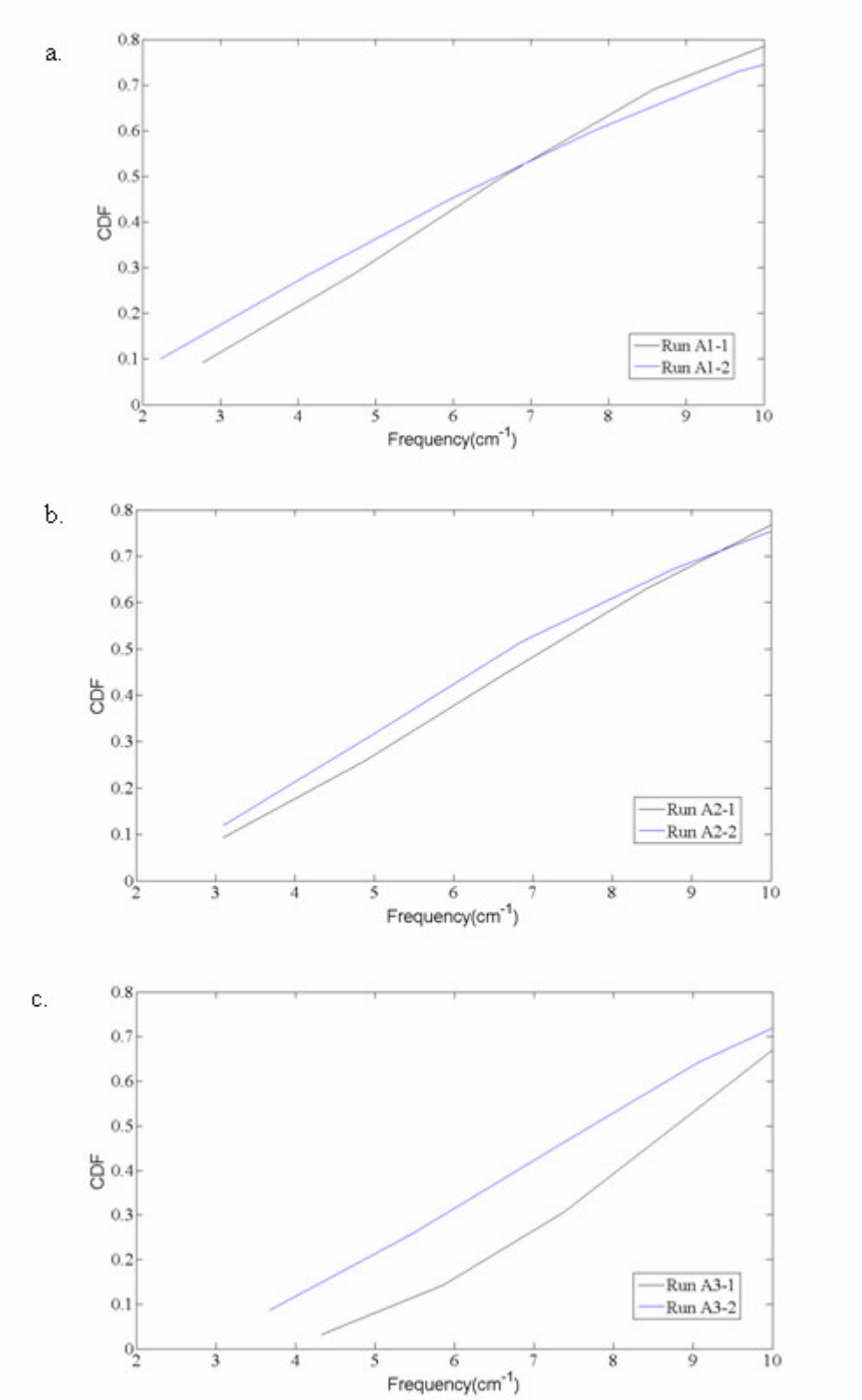


Figure 5.22. CDF of frequencies of Runs (a) A1 (b) A2 (c) A3

CDF of the frequencies acquired from all free and bound state samples is shown in Figure 5.23. Here, it is seen more clearly that free and NADP+ bound states have lower frequencies than NADPH bound state. In the literature, it was shown experimentally and by using NMA [33-35] that ligand binding shifted the lowest frequencies to higher values experimentally. The results presented in this thesis support those findings.

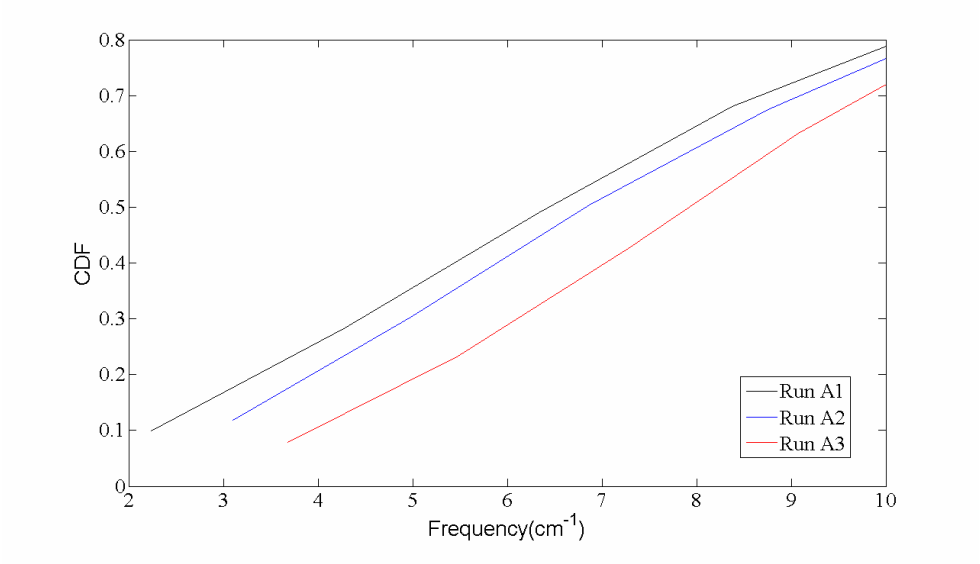


Figure 5.23. CDFs of frequencies (cm^{-1}) (two samples of each run appended)

Damping factors are proportional to the friction experienced by the protein. Damping factors of acquired models exhibit declining behavior with respect to modes as expected (figures not shown). Figure 5.24 is the boxplot of damping factors of three different states. There is not a remarkable difference between the unliganded and different liganded forms.

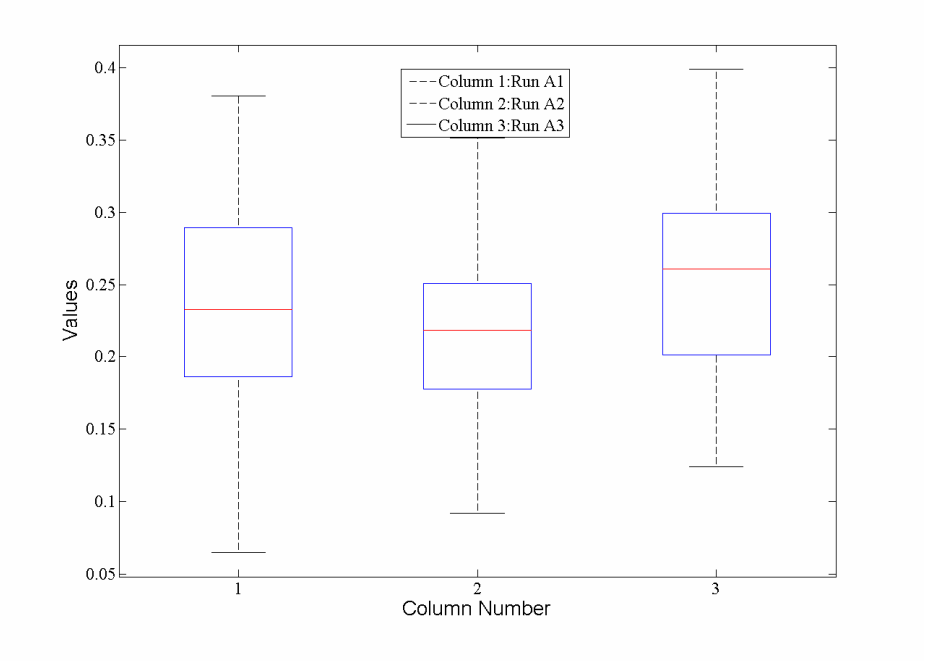


Figure 5.24. Boxplot of damping factors (DHFR)

In the previous paragraphs, the frequencies, thus the underdamped modes have been compared. On the other hand, fluctuations along more than half of the 60 PCs in each run are found to be overdamped. Therefore, these overdamped modes should also be compared. Figure 5.25 shows the $Ø_1$ and $Ø_2$ parameters of the AR characteristic equation, which are responsible from the intraminimum motions. The AR parameters of free and $NADP^+$ bound states, are shown on Figure 5.25a, while unliganded state is compared with NADPH bound one in Figure 5.25b. Although it is difficult to reach a conclusion, it is seen that $Ø_1$ parameter of the DHFR in the NADP+ bound state are higher compared to those in the other states. This means that the relaxation times of the modes in the NADP+ bound state DHFR are higher.

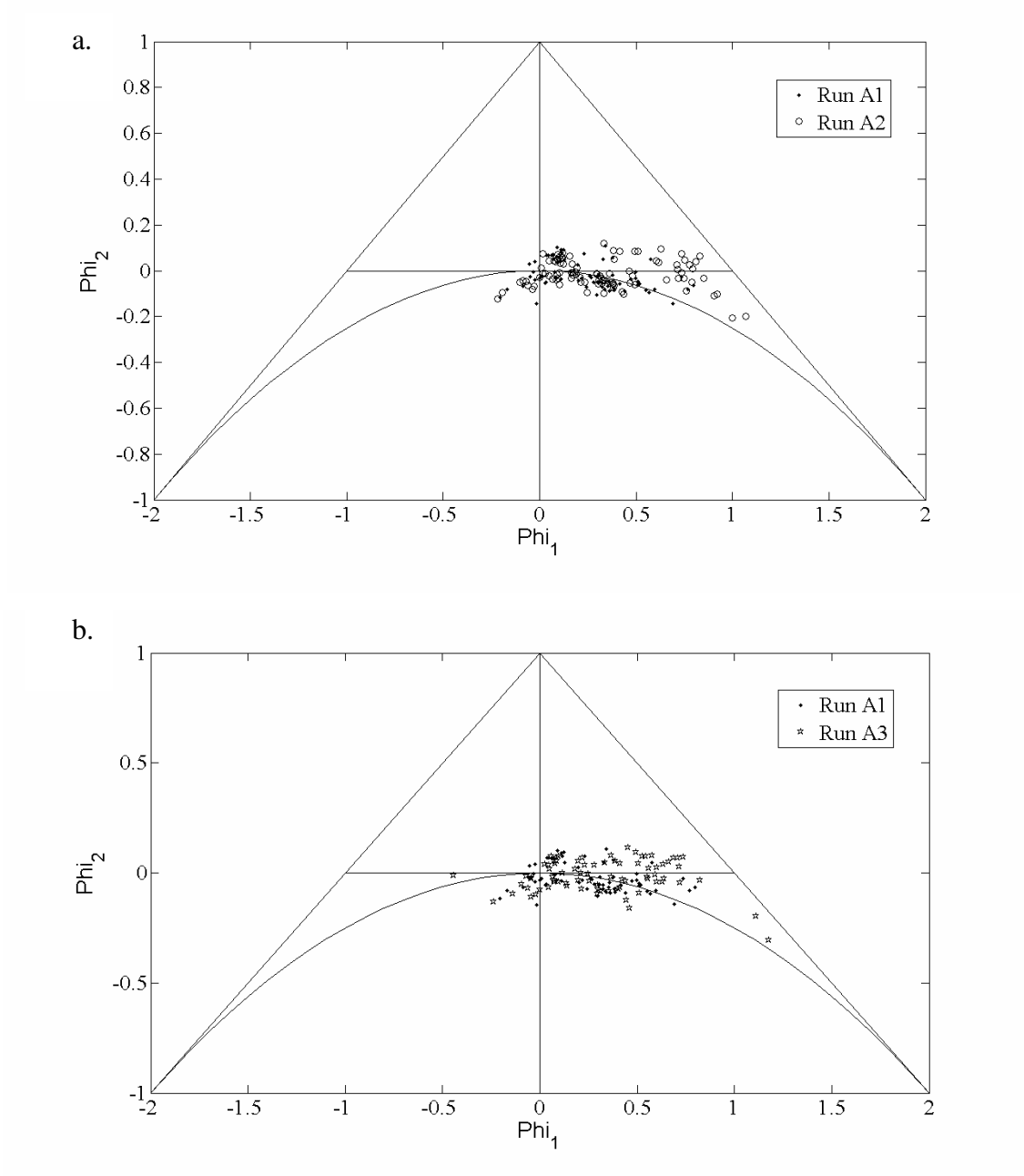


Figure 5.25. Comparison of $Ø_1$ and $Ø_2$ parameters of (a) Run A1 and A2, (b) Run A1 and A3

5.1.5. Analysis of Case2

For this part, extra seven nodes belonging to the ligand, are taken into consideration along with the C^{α} atom coordinates in PCA. The purpose of this analysis is twofold. First, it is aimed to find the contribution of the ligand nodes to the collective motions of the

protein-ligand complex. Second, it is desired to find the lowest vibrational frequencies of the binary complex, so that a better comparison with the literature can be done.

Figure 5.26 shows the percentage variability of the ligands' nodes explained by the first 5 PCs. That is, this is the graph showing how much percentage of the ligand's total motion can be explained by its motion along first 5 PCs. The larger the percentage explained is, the more ligand moves in a coordinated way with the protein. It is seen that NADPH contributes to the motion of the protein much more than $NADP^+$ ligand, which may be attributed to NADPH binding to DHFR 100 times stronger than $NADP^+$ binds to DHFR [18].

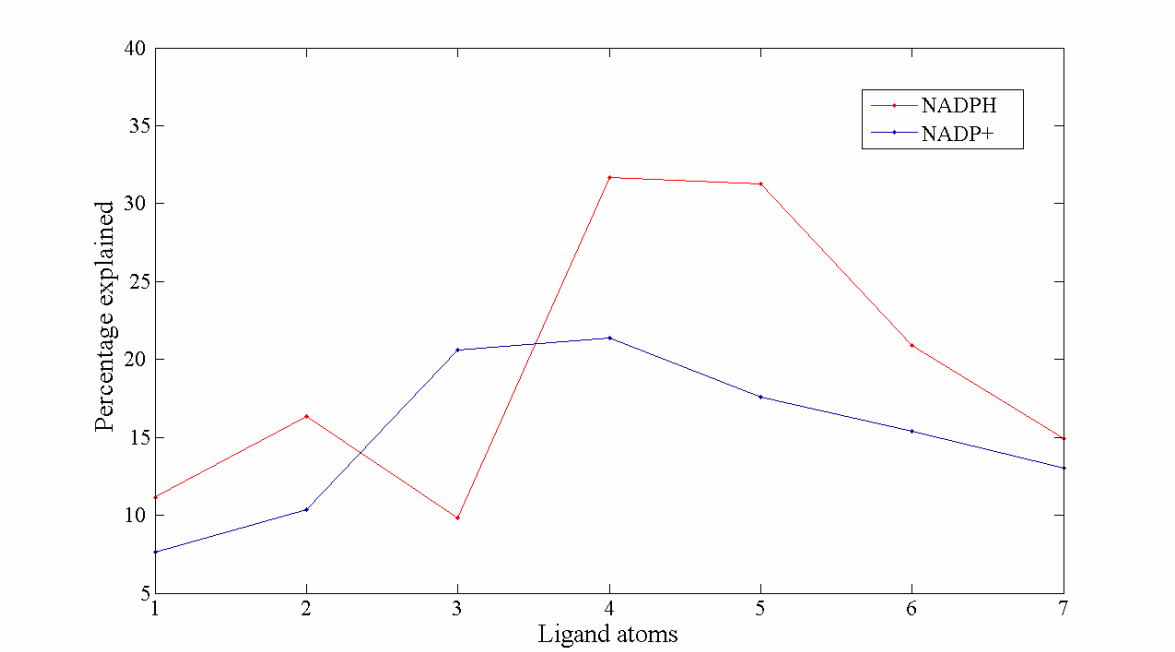


Figure 5.26. Percentage variability of the nodes taken from NADPH and $NADP^+$ explained by the first 5 PCs

ARIMA model types of Case 2 are similar to the ones of Case 1, either. Table 5.5 reveals the number of underdamped modes captured among the 60 modes for Case 1 and Case 2.

Table 5.5. The number of underdamped modes for Case1 and Case 2

	CASE 1	CASE 2
RUN A1-1	24	24
RUN A1-2	13	13
RUN A2-1	20	18
RUN A2-2	22	22
RUN A3-1	10	17
RUN A3-2	22	21

Figure 5.27 is the frequency histogram obtained for Case 2. It can be observed that frequencies of ligand bound states are in approximately (3 cm^{-1} – 13 cm^{-1}) interval in Case 1 (see Figure 5.21), whereas they are in (1 cm^{-1} – 15 cm^{-1}) interval in Case 2, which means that vibrational frequencies of the binary complex are lower compared to the protein in liganded form. It is also seen that, protein-ligand binary complex possesses a number of lower frequency vibrational modes compared to those of the unliganded protein. In a previous experimental study[36], similar results have been obtained. CDF of all frequencies of three states is shown in Figure 5.28, where it is seen more clearly that NADPH-DHFR complex possesses only a few lowest frequency modes, while the density of rest of the low frequencies in the unliganded DHFR is higher. Considering also the results obtained in case 1, that is the frequencies obtained by only taking the protein atoms, it is seen that lowering of the frequencies in the binary complexes is not due to the increase of the flexibility of the protein, but higher mass of the complex[37].

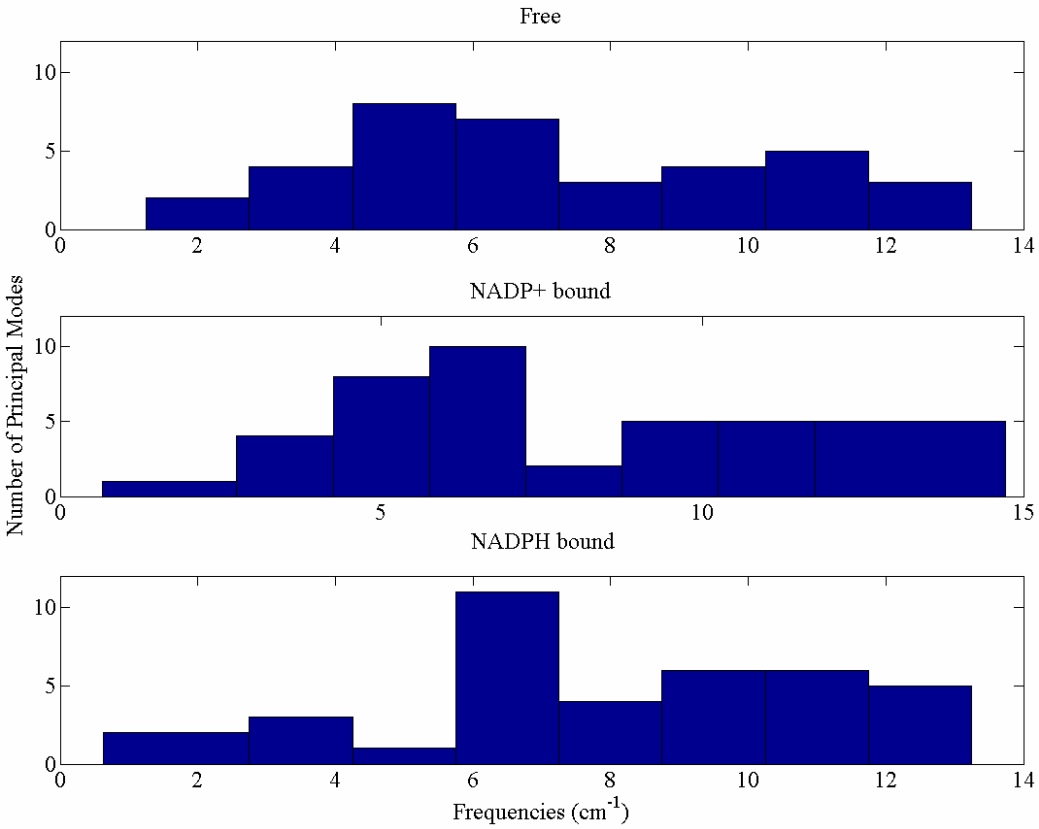


Figure 5.27. Histograms of DHFR frequencies (Case 2)

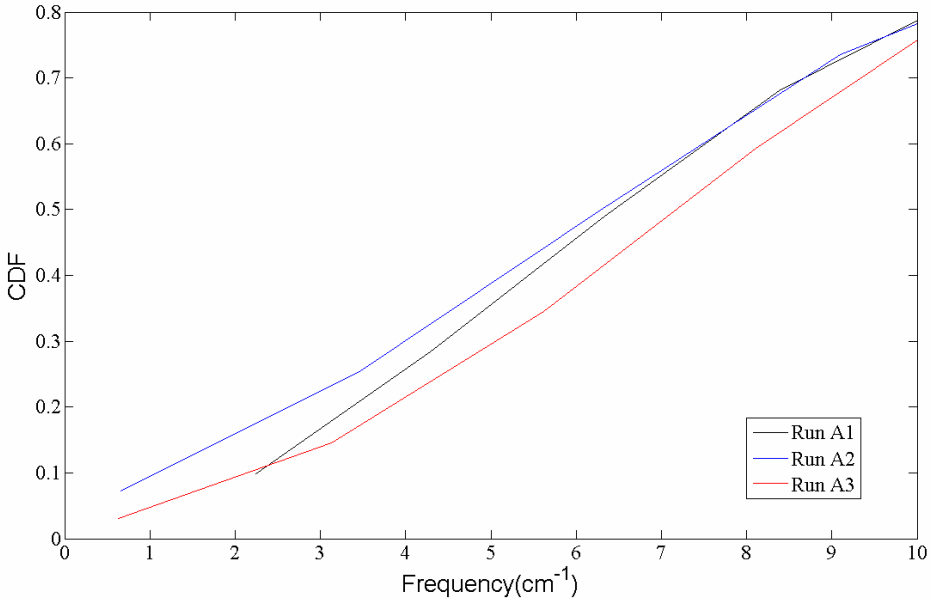


Figure 5.28. CDF of DHFR frequencies (Case 2)

Figure 5.29 is the comparison of the CDFs of Case 1 and Case 2. In both cases, NADPH bound state has slower vibrational frequencies compared to those in the $NADP^+$ bound state.

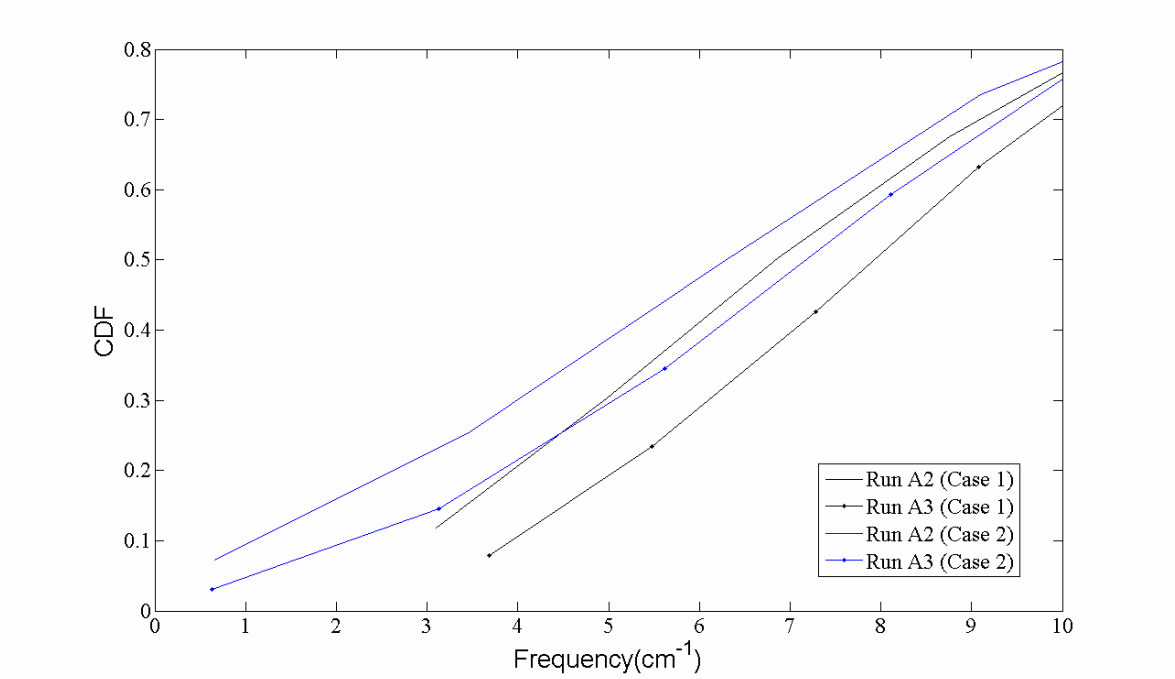


Figure 5.29. CDF comparison of Case1 – Case 2

5.2. Investigation of the Dynamics of TIM

Two different MD simulations are performed on TIM. Run B1 is of free state, Run B2 is of DHAP bound state. Runs are sampled twice from different 3.2 ns long parts; thus Run B1-1 and Run B1-2 comprise the samples of the free state, while samples of the bound state are named as Run B2-1 and Run B2-2 (Table 5.6). The difference between Run B1-1 and Run B1-2 is that the former is sampled from the open and the latter from the closed forms of active site loop 6 (both in the absence of ligand). The RMSD between the average conformations of Run B1-1 (open form) and B1-2 (closed form) is 1.758 Å, whereas the RMSD between the average conformations of Run B2-1 and B2-2 is 0.754 Å. The same procedure followed in DHFR analysis is applied on the data obtained from the simulations of TIM. Results are discussed below.

Table 5.6. Notations for TIM runs

	1st Sample	2nd Sample
Free State	Run B1-1	Run B1-2
DHAP Bound State	Run B2-1	Run B2-2

5.2.1. Results of PCA of Free and Bound States

Table 5.7 gives the sum of eigenvalues of run samples. Free state seems to be more flexible than ligand bound state.

Table 5.7. Sum of eigenvalues of run samples

Free State	Run B1-1	335.9 Å^2
	Run B1-2	332.5 Å^2
DHAP Bound State	Run B2-1	316.8 Å^2
	Run B2-2	292.2 Å^2

The first 5, 10 and 60 PCs of Run B1-1 explain 47%, 59% and 81% of C^{α} respectively. Eigenvalues and the percentage explanation of the modes of Run B1 (Run B1-1,2 averaged) and Run B2 (Run B2-1,2 averaged) can be seen in Figure 5.30a-b. It is interesting that the first two modes in the liganded state have higher explanation powers compared to the unliganded form, which suggests an increase of collectivity when ligand is bound to the protein.

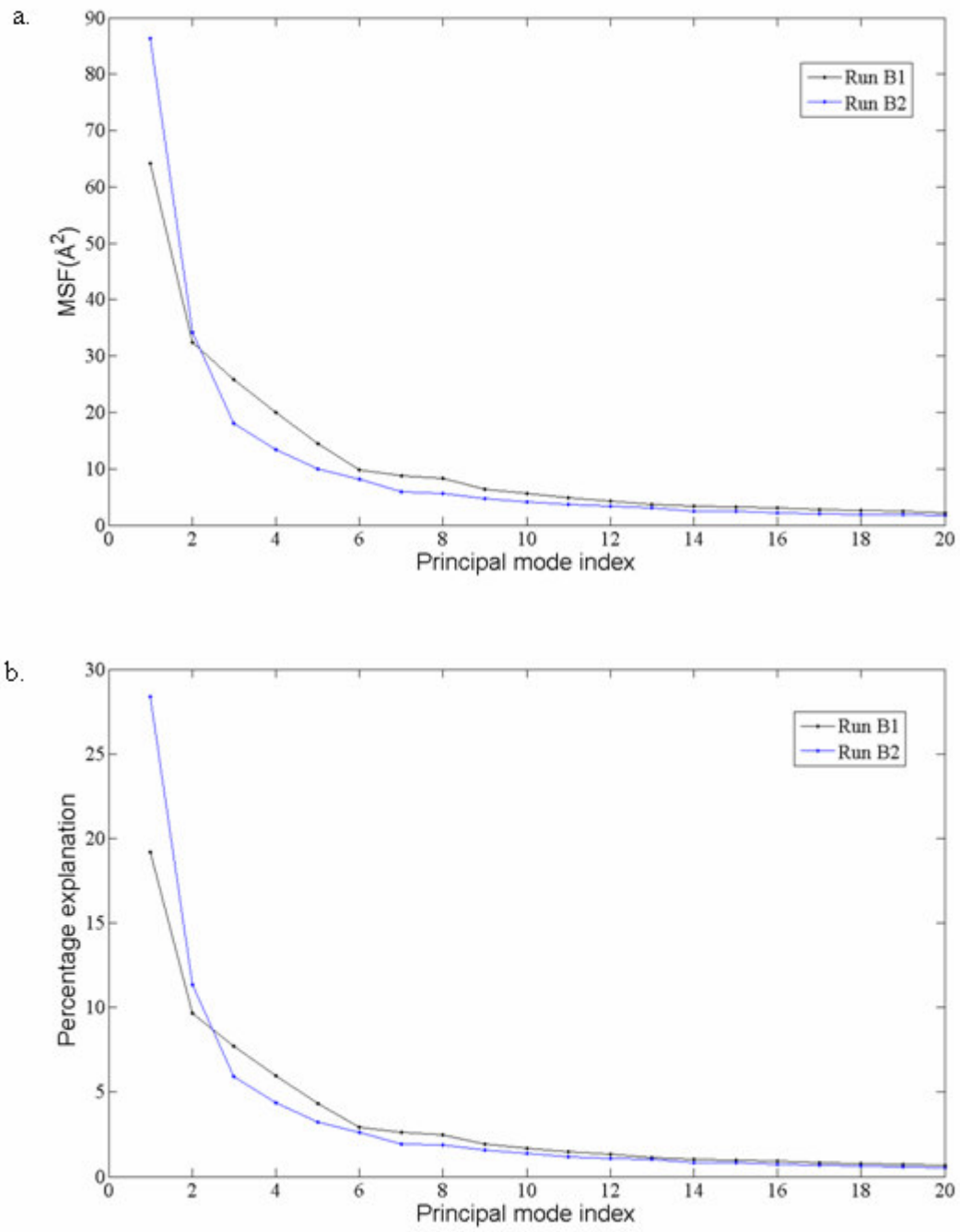
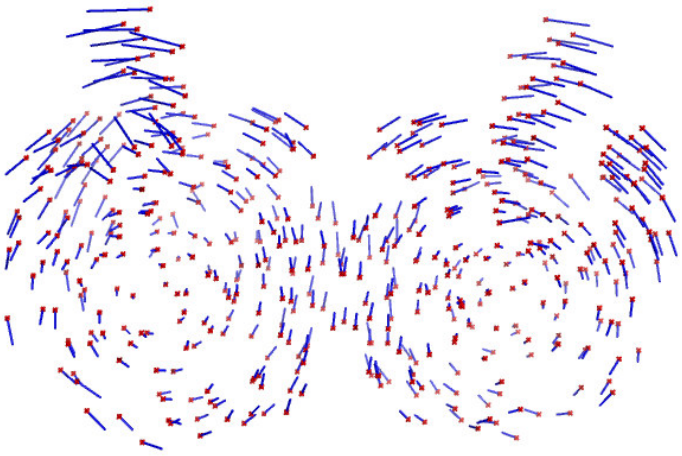


Figure 5.30. (a) Eigenvalues, (b) their percentage explanation of the total variance for the first 60 PCs (Run B1-1,2)

Displacement vector illustrations of PC 1 are illustrated in Figure 5.31. Rotating type of motion between two lobes is apparently seen, and ligand binding does not change this motion.

a.



b.

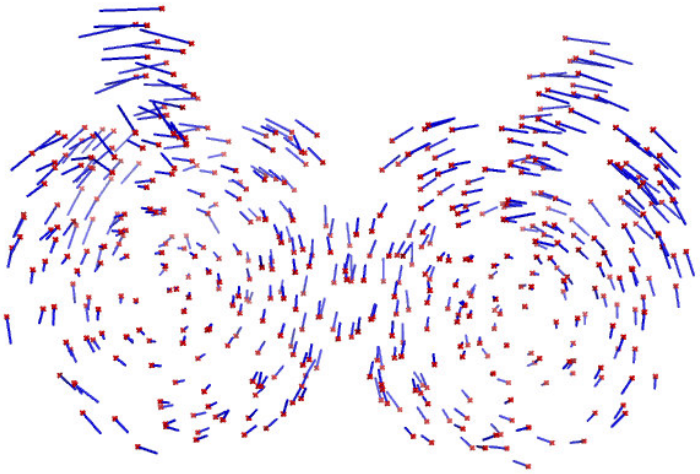


Figure 5.31. Vector field illustrations along PC1 of Run (a) B1-1, (b) B2-1

A number of snapshots are produced by moving the protein along PC 1. Figure 5.32 illustrates these superimposed snapshots of the two Run B1 samples, while Figure 5.33 shows the superimposed snapshots of the two Run B2 samples. The coordinated motion of the loop 6 with the rest of the protein is clearly seen in all figures. Contrary to what has

been observed for DHFR, it is seen that ligand binding on TIM, slightly increases collectivityof protein motions.

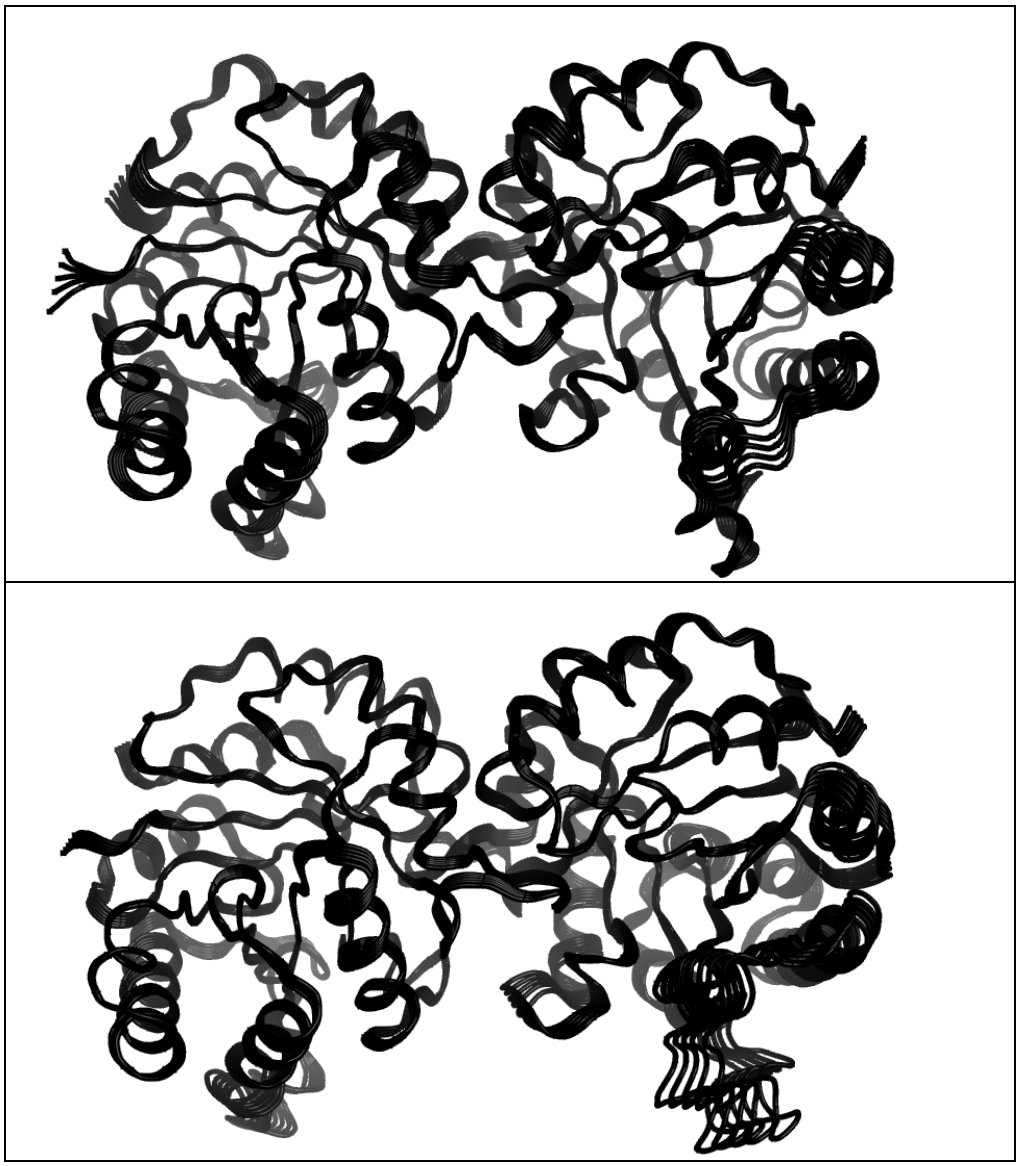


Figure 5.32. Projections of the free TIM conformations onto PC 1 for runs B1-1 and B1-2

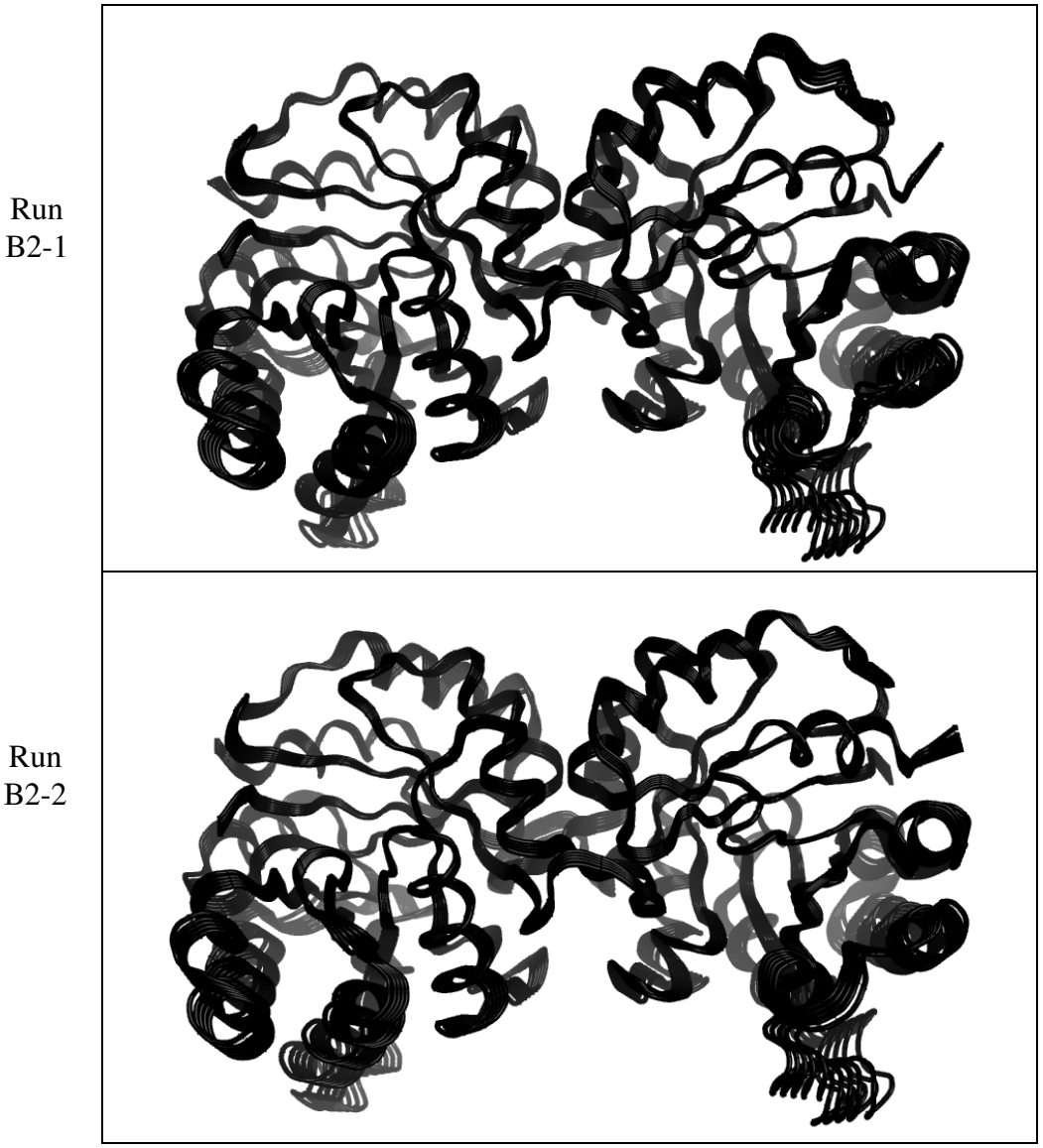


Figure 5.33. Projections of the liganded TIM conformations onto PC 1 for runs B2-1 and B2-2

5.2.2. Results of Time Series Analysis of Free and Bound States

A number of samples of time series models derived for Run B1-1 are shownbelow:

$t_1 : (1 - 0.952B)(1 - 0.257B - 0.019B^2)\nabla z_t = (1 - 0.989B)\, a_t\,,\ \sigma_a^2 = 0.6361$ (5.8)

$t_2 : (1 - 0.337B)(1 - 0.035B - 0.0568B^2)\nabla z_t = (1 - 0.0561B)\, a_t\,,\ \sigma_a^2 = 0.7229$ (5.9)

$t_{10} : (1 - 0.94B)(1 - 0.73B + 0.074B^2)\nabla z_t = (1 - 0.98B)(1 - 0.79B)\, a_t\,,\ \sigma_a^2 = 0.523$ (5.10)

$t_{40} : (1 - 0.9394B)(1 - 0.4246B + 0.0541B^2)\,\nabla z_t = (1 - 0.7334B)\, a_t\,,\ \sigma_a^2 = 0.4243$ (5.11)

t_{60} : $(1 - 0.9317B)\nabla z_t = (1 - 0.686B)(1 + 0.1816B)a_t$, $\sigma_a^2 = 0.3375$ (5.12)

Table 5.8 gives the number of ARIMA model types with respect to the modes. The most encountered models are ARIMA(3,1,1) and ARIMA(3,1,2) among nonstationary modes, ARMA(3,1) among stationary modes.

Table 5.8. Number of principal modes with respect to their model orders (TIM)

Type and order of the model	# of modes (B1-1)	# of modes (B1-2)	# of modes (B2-1)	# of modes (B2-2)
ARIMA (2,1,1)	2	0	0	0
ARIMA (2,1,2)	1	2	2	0
ARIMA (3,1,1)	*2*	*6*	*7*	*10*
ARIMA (3,1,2)	*14*	*6*	*1*	*2*
ARMA(1,2)	3	3	0	0
ARMA(2,1)	0	2	0	0
ARMA(2,2)	2	11	2	3
ARMA(3,1)	*32*	*25*	*46*	*43*
ARMA(3,2)	4	5	2	2

Figure 5.34 is the comparison of variance of model residuals for free and bound states. It is interesting to note that the first and third modes have higher σ_a^2 values for the liganded cases, indicating higher random walk step sizes, thus an increase in the anharmonic motions for the liganded case. Nevertheless, MSF of the unliganded TIM is higher than the liganded form, which shows that increase in anharmonic motions should be compensated by some other effect so that the flexibility of the liganded form is lower (see the comparison of low vibrational frequencies below).

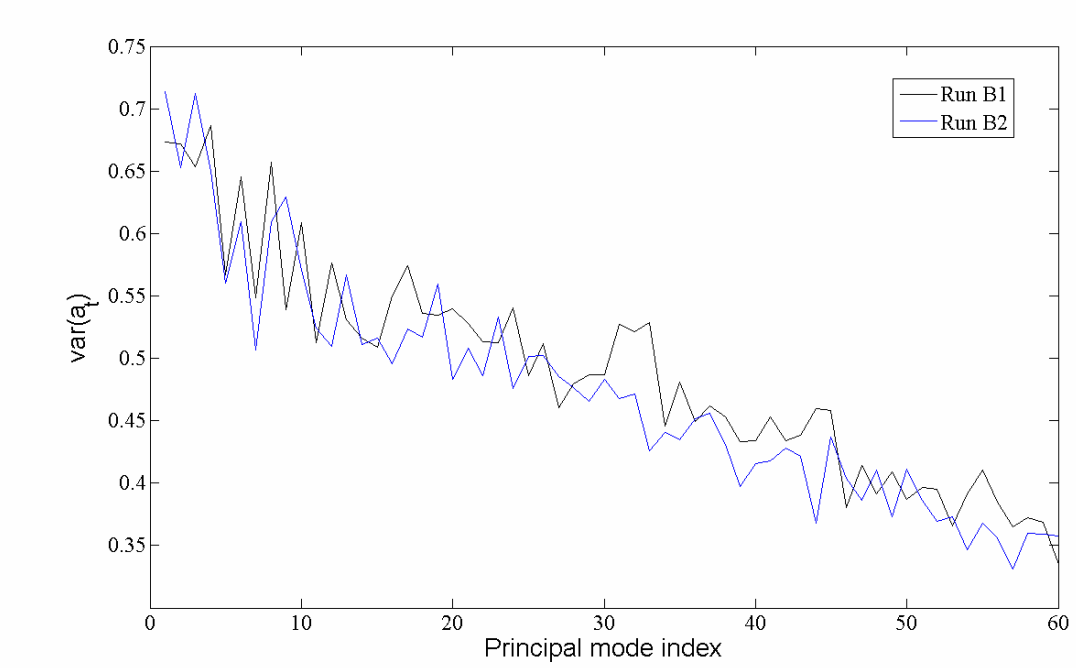


Figure 5.34. Comparison of σ_a^2 with respect to modes for TIM

Boxplot examination of MA root θ_2 gives similar results as DHFR. θ_2 values do not imply any differences neither in open/closed forms of Loop 6, (Figure 5.35a-b) nor in free-bound comparison (Figure 5.35c).

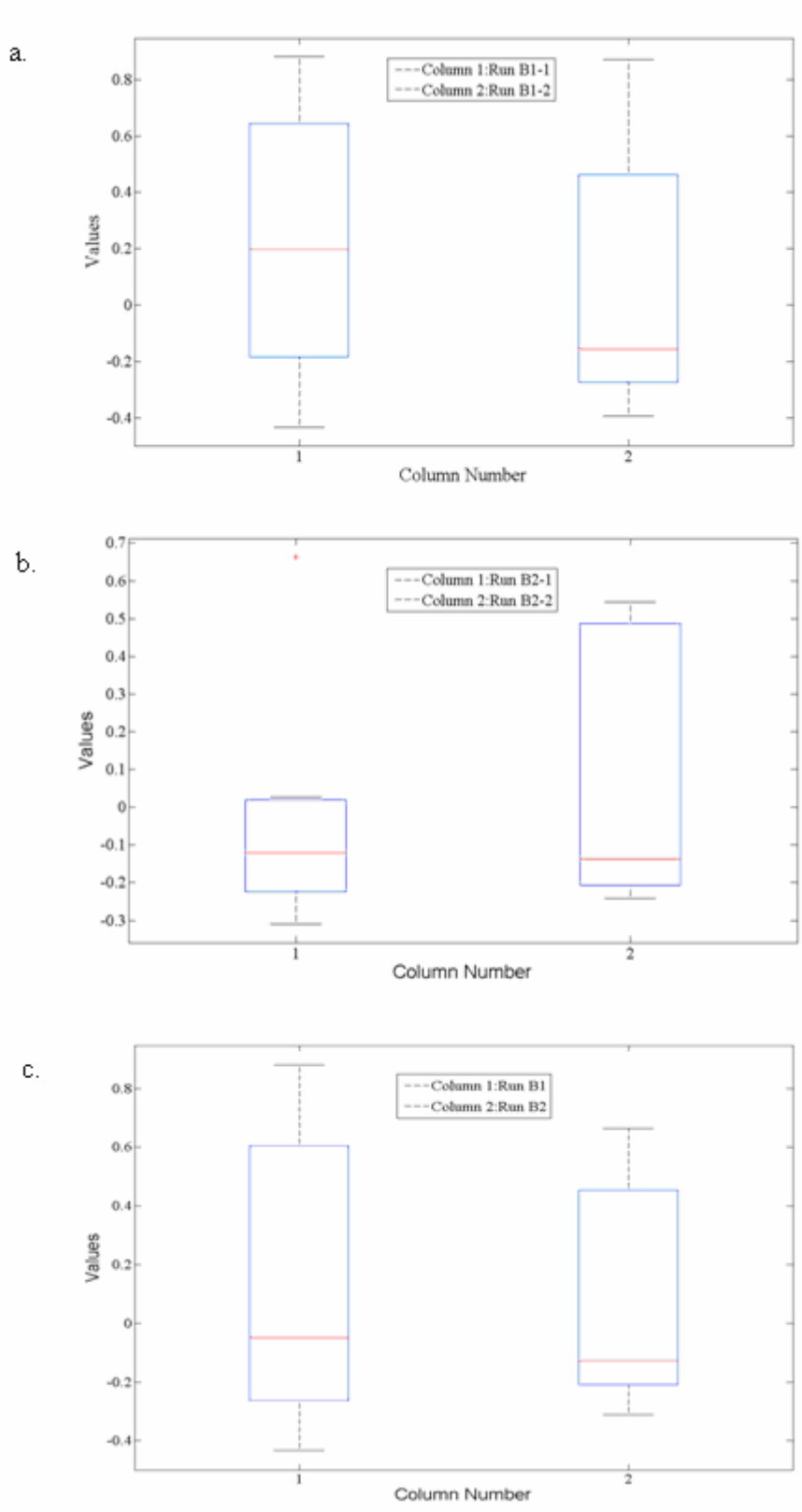


Figure 5.35. Boxplot of θ_2 roots of (a) Run B1, (b) Run B2, (c) Runs B1, B2

Table 5.9 gives the number of frequencies captured, that is, underdamped modes.

Table 5.9. The number of underdamped modes of TIM

RUN B1-1	37
RUN B1-2	25
RUN B2-1	43
RUN B2-2	44

Figure 5.36 illustrates the frequency histogram graphics of free and ligand bound TIM. It is apparent that binding shifts the the low frequencies to higher values.

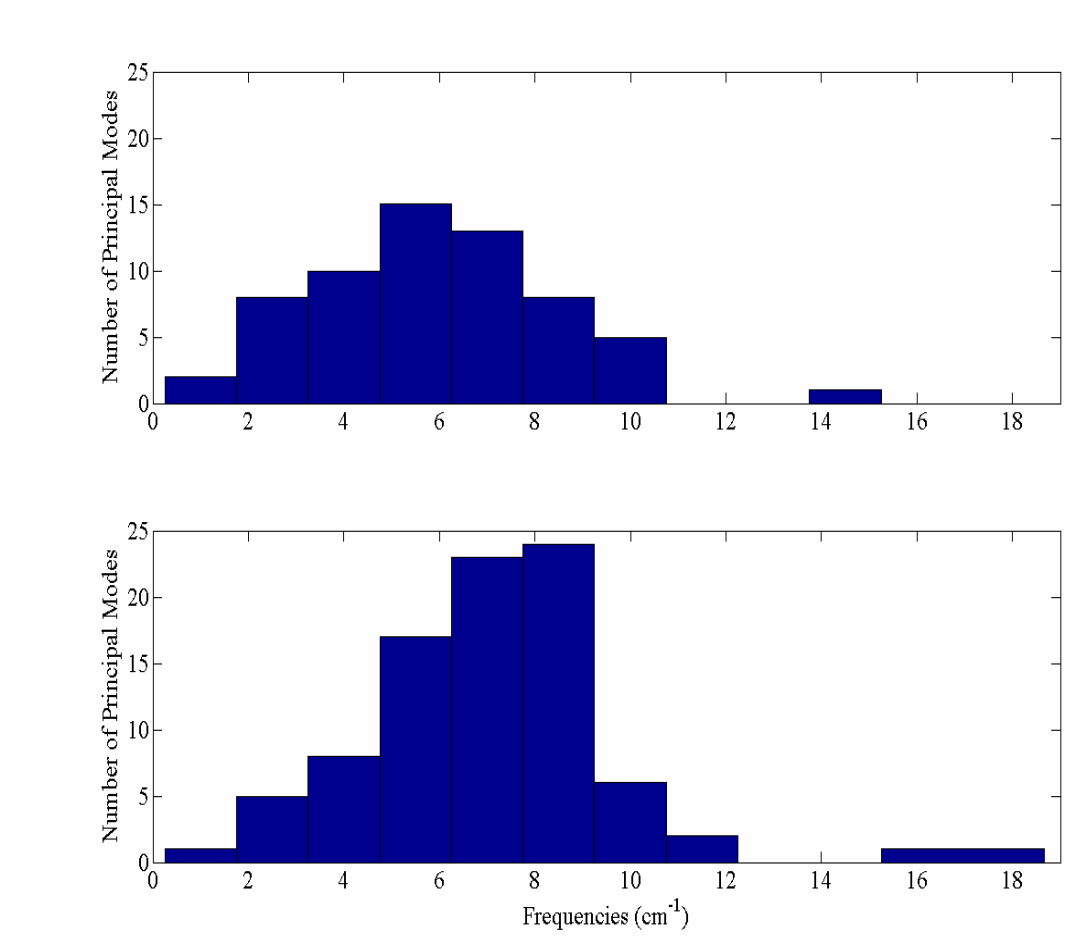


Figure 5.36. Histogram graphic of frequencies of (a) free TIM, (b) bound TIM

Since CDF of frequencies has revealed the differences between DHFR states, CDF of both states of the TIM are compared. Figure 5.37a shows that there is not significant

difference between open and closed forms of free TIM. This comment is made based on the "inherent variability" of DHFR (Figure 5.22a). The difference of CDF between ligand bound state samples (Figure 5.37b) is larger than that of free state samples.

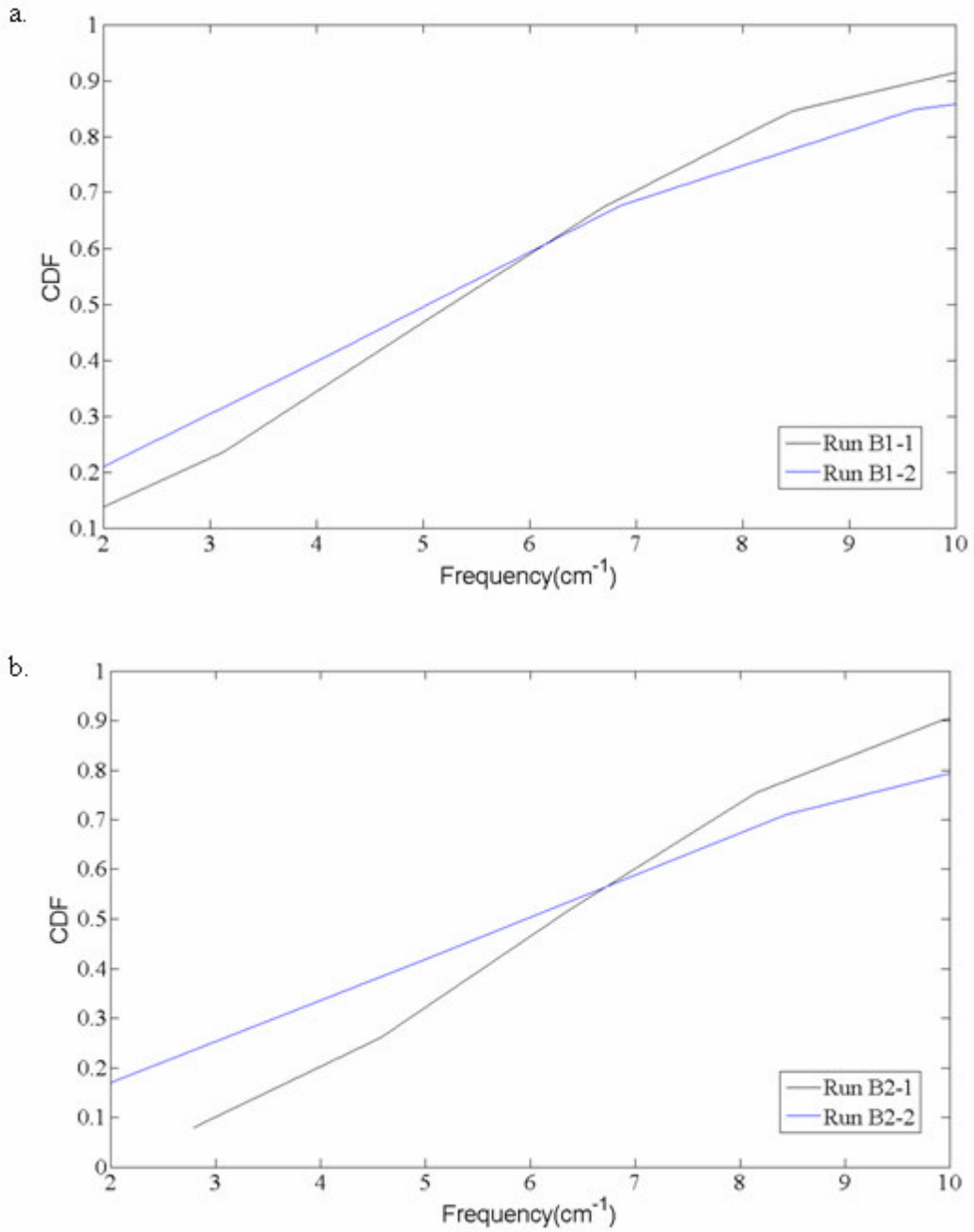


Figure 5.37. CDF of Run (a) B1 and (b) B2 frequencies

Figure 5.38 illustrates CDF comparison of all frequencies belonging to free and ligand bound TIM. As expected (on observing Figure 5.36), ligand bound state yields higher frequencies than free state.

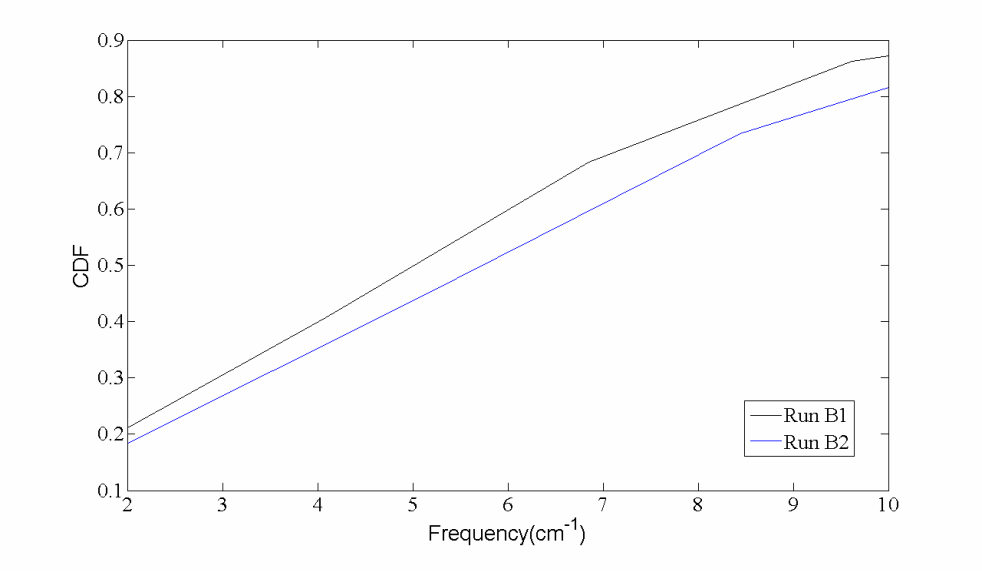


Figure 5.38. CDFs of TIM frequencies

5.3. Comparison of TIM and DHFR

As it is expected, the higher molecular weight protein TIM (residue number=494) has lower frequencies than DHFR having 159 residues[37]. The modes with low frequencies in free state are shifted to higher frequencies in ligand bound state for both DHFR and TIM (Figure 5.39).

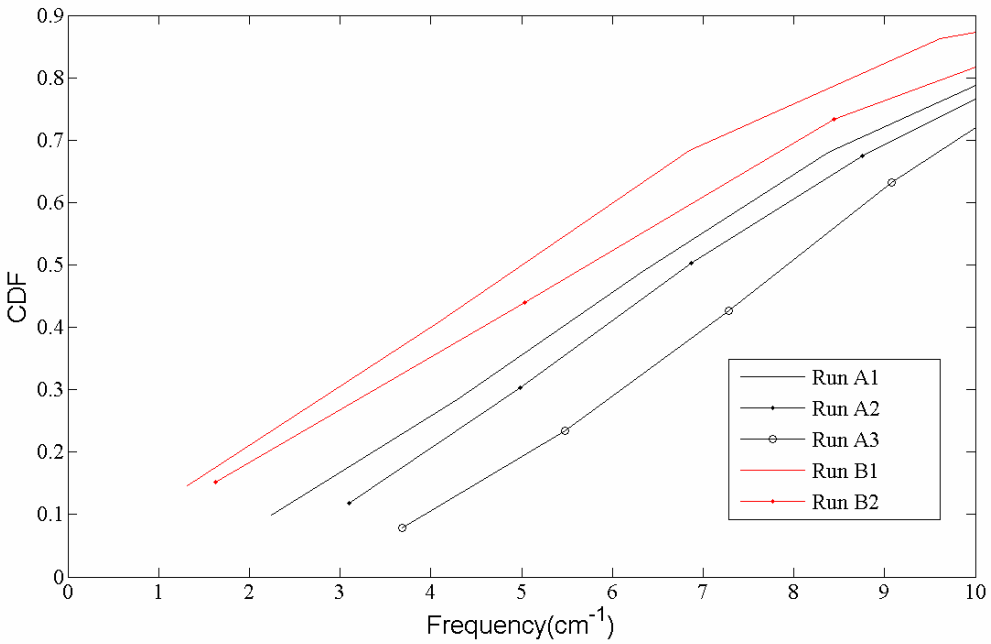


Figure 5.39. CDFs of all 5 runs' frequencies (Case 1 for DHFR)

For both DHFR and TIM, transition from nonstationary modes to stationary modes is observed at higher modes for the free states compared to ligand bound states. For TIM, ARIMA(3,1,1) and ARIMA(3,1,2) are the most encountered models among nonstationary modes, ARMA(3,1) is the most encountered one among stationary modes. These were ARIMA(3,1,2) and ARMA(3,1) respectively for DHFR. This indicates that different proteins have common time series models and parameters basis. The number of underdamped modes in TIM is considerably higher compared to those in DHFR.

In a previous study[1], protein could be modeled by ARIMA model with lower parameter number. This can be explained by the fact that the proteins of the current study are bigger in size than the one previously studied.

6. CONCLUSION

Proteins have been investigated in many studies and curiosity of scientists about protein activities have been increasing due to medical reasons. Proteins can be examined by experimental and computational studies. Molecular dynamics (MD) simulations technique is one of those techniques, by which important informations about protein dynamics can be acquired. In this thesis, PCA and time series analysis are applied on the C^{α} atomic trajectories extracted from the MD simulations of enzymes - DHFR and TIM - to investigate the collective dynamics of the free and ligand bound states of the proteins.

DHFR has been analyzed in apo form and with two different ligands, $NADP^+$ and NADPH. The collective motions of the DHFR in three states show considerable difference. Opening/closing motion of M20 loop with the twisting motion of the protein as a result of the concerted motions of the CD, FG, GH loops and helix C can be clearly observed in the unliganded form. In the $NADP^+$ bound state, flexibility of the M20 loop is reduced and the coordination of the distal sides in the DHFR is almost completely lost. It is interesting that the concerted motions in the $NADP^+$ bound DHFR can be observed at local regions. When NADPH is bound to DHFR, opening/closing type of motion is seen even more clearly, however some of the twisting type of motion with the collective character spreading over the whole protein is reduced. It is concluded that binding of ligands and even small differences in the ligands (only a hydride difference between $NADP^+$ and NADPH) may change the collective motions of the protein drastically. The examination with the ligand's atoms taken into consideration has shown that NADPH contributes to the collective motions more than $NADP^+$, which is an indication of the stronger binding of NADPH to DHFR.

The underlying time series model types and orders are found to be similar for both proteins and states. In addition, they are similar to the models acquired in the previous study[1]. ARIMA(2,1,1), ARIMA(2,1,2), ARMA(3,1) and ARMA(3,2) models were mostly detected in the previous study on proteins tendamistat (residue number = 74) and ImmE7 (residue number = 87). ARIMA(3,1,1), ARIMA(3,1,2), ARMA(3,1), ARMA(3,2) models have been mostly detected in the current study. This fact implies the generality of

the underlying dynamics of proteins. ARIMA(2,1,1) and ARIMA(2,1,2) models are not common for the current proteins DHFR and TIM. This seems to be caused by the fact that tendamistat and ImmE7 proteins are smaller proteins compared to DHFR (residue number = 159) and TIM (residue number = 494). That is, as the number of residues increases, number of variables increases, requiring a higher ordered model. The number of underdamped modes obtained for TIM are greater than DHFR, which implies that the dampening effect of water on the vibrational motions of the smaller protein DHFR are higher compared to that of TIM. Time series parameters obtained have been observed to be similar for free and bound states of both proteins. AR and MA parameters and variances of random shocks have been interpreted in terms of protein fluctuations and could be interpreted as pseudo-periodic oscillations, intramolecular forces in the protein and intermolecular forces between the protein and the water molecules, like they were identified in previous studies.

It has been reported in numerous studies that binding of a small molecule changes the flexibility of the protein, and so affects the binding free energy. Therefore, protein flexibility should be taken into account when the binding affinity of ligands is examined. On the other hand, the effect of ligand binding on the flexibility of the protein is in dispute: it was observed in some studies that ligand binding reduced the flexibility of the protein [33-36], while in others, ligand binding increased the flexibility [36,38,39].

Time series models of the collective coordinates along the principal components have given valuable information about the dynamical behavior of the collective motions. In both DHFR and TIM, the most important difference between the unliganded and liganded forms are in the vibrational frequency distribution. It should be pointed out that linear stochastic time series modeling is a much more reliable method, compared to quasi-harmonic analysis, which does not take the anharmonicity into account. Time series models, on the other hand, takes the nonstationarity of the protein trajectory, which is a consequence of the anharmonicity of the protein fluctuations, into consideration. Examination of DHFR with the ligand's atoms taken into consideration has shown that frequencies detected for the DHFR-$NADP^+$ and DHFR-NADPH binary complexes are slightly lower than unliganded DHFR. However, this shift of lowest frequencies should be due to the additional mass of the ligand, since frequencies obtained only for the DHFR in

three states have shown that unliganded states have the smallest vibrational frequencies, followed closely by $NADP^+$ bound DHFR, and NADPH bound DHFR. It is interesting that though the average conformations and the collective motions of the unliganded and NADPH bound DHFR are quite similar compared to $NADP^+$ bound DHFR, their densities of vibrational low frequencies are remarkably different. A similar observation may be made for the unliganded and liganded states of TIM, which leads to the conclusion that ligand binding may change the low vibrational frequencies of a protein without much affecting the structure and even the collective motions. This shift of lowest frequencies may have impact on the catalytic cycle.

Considering only the frequencies, one may have concluded that unliganded form of DHFR is the most flexible form, followed by $NADP^+$ bound form, and NADPH bound form, whereas MSF of $NADP^+$ bound DHFR have come out to be smallest. On the other hand, one should be cautious in using low frequency as a synonym for protein flexibility. Vibrational frequencies correspond to intraminimum motions, so they are only related to harmonic motion. It is interesting to note that the variance of random shocks, the driving force of the protein motions, are lower for $NADP^+$ bound DHFR in the low mode index region. The variance of random shock terms was shown to be proportional to the random walk step sizes[10], that is the interminimum motions. Therefore, the anharmonic motions of $NADP^+$ bound DHFR are smaller compared to other states, and this becomes the dominating factor in determining the flexibility of the protein backbone. This may be the reason why, though the vibrational frequencies of $NADP^+$ bound DHFR is considerably lower than NADPH bound DHFR, the total MSF of the $NADP^+$ bound DHFR are the smallest among the three states. This may be an important conclusion that in the calculation of binding affinities for ligands and the backbone flexibilities, considering only the harmonic vibrational frequencies and leaving out the anharmonic motions may be misleading.

Analysis of the open and closed forms of TIM, though their RMSD is as high as 1.76 Å, has shown similar frequency PDFs. This indicates that the change of vibrational frequencies due to distinct conformations visited by the protein is not high. On the other hand, the effect of ligand binding on TIM has given similar results to those obtianed from DHFR. Ligand bound TIM has been shown to have higher frequencies than free TIM.

Frequencies of TIM are lower than those of DHFR due to the fact that TIM is a larger protein than DHFR. It is interesting that, unlike DHFR, anharmonic motions of TIM are magnified in the presence of ligand, however the lowest frequencies become the dominating factor determining the flexibility of the liganded and unliganded states, thus the unliganded TIM is more flexible than the liganded form.

As recommendations for future study, different ligands, for example THF for DHFR, GAP for TIM, can be investigated in terms of their effects on proteins. Different temperatures can be tried for comparison with previous studies; for instance MD simulations can be performed at 120 K and obtained data can be analyzed by time series analysis to make comparisons with a previous experimental study[33]. It is found that the number of underdamped modes is considerably higher in TIM compared to DHFR, which indicate that the dampening effect of water on protein vibrational frequencies differ protein to protein. The dampening behavior of the modes and the time series model paramaters of the overdamped modes in different protein forms should be elaborated further.

APPENDIX : DETAILS OF RUNS

Run A1 : DHFR , free, starting X-ray coordinate 1RA1, at 300 K

A1-1 : 4 - 7.2 ns

A1-2 : 13 - 16.2 ns

Run A2 : DHFR bound with $NADP^+$, starting X-ray coordinate 1RX1, at 300 K

A2-1 : 4 – 7.2 ns

A2-2 : 13 – 16.2 ns

Run A3 : DHFR bound with NADPH, starting X-ray coordinate 1RX1, at 300 K

A3-1 : 4 – 7.2 ns

A3-2 : 13 – 16.2 ns

Run B1 : TIM, free, starting X-ray coordinate 8TIM, at 300 K

B1-1 : loop 6 is open: 5 – 8.2 ns

B1-2 : loop 6 is almost closed: 56.8 – 60 ns

Run B2 : TIM bound with DHAP, starting X-ray coordinate 1TPH, at 300 K

B2-1: 5 – 8.2 ns

B2-2: 9 – 12.2 ns

REFERENCES

1. Alakent, B., *Investigation of Protein Dynamics Using Time Series Analysis,* Ph.D. Thesis, Boğaziçi University, 2005.

2. Go, N., T. Noguti and T. Nishikawa, "Dynamics of a Small Globular Protein in Terms of Low-Frequency Vibrational Modes", *Proceedings of the National Academy of Sciences, USA*, Vol. 80, No. 12, pp. 3696-3700, June 1983.

3. Brooks B. and M. Karplus, "Harmonic Dynamics of Proteins: Normal Modes and Fluctuations in Bovine Pancreatic Trypsin Inhibitor", *Proceedings of the National Academy of Sciences USA*, Vol. 80, No. 21, pp. 6571-6575, November 1983.

4. Kitao A. and N. Go, "Investigating Protein Dynamics in Collective Coordinate Space", *Current Opinion in Structural Biology*, Vol. 9, pp. 164-169, 1999.

5. van Aalten, D. M. F., A. Amadei, A. B. M. Linssen, V. G. H. Eijsink, G. Vriend and H. J. C. Berendsen, "Essential Dynamics of Thermolysin: Confirmation of the Hinge-Bending Motion and Comparison of Simulations in Vacuum and Water", *Proteins*, Vol. 22, No. 1, pp. 45-54, May 1995.

6. Amadei, A., A. B. M. Linssen and H. J. C Berendsen, "Essential Dynamics of Protein", *Proteins*, Vol. 17, pp. 412-425, 1993.

7. Kitao, A., S. Hayward and N. Go, "Energy Landscape of a Native Protein: Jumping-Among-Minima Model", *Proteins*, Vol. 33, No. 4, pp. 496-517, 1998.

8. Kitao, A., F. Hirata and N. Go, "The Effects of Solvent on the Conformation and the Collective Motions of Protein: Normal Mode Analysis and Molecular Dynamics Simulation of Mellitin in Water and in Vacuum", *J. of Chemical Physics*, Vol. 158, pp. 447-472, 1991.

9. Hayward, S., A. Kitao, F. Hirata and N. Go, “Effect of solvent on collective motions in globular proteins”, *J. Mol. Biol.*, Vol. 234, pp. 1207-1217, 1993.

10. Alakent, B., P. Doruker and M. C. Çamurdan , “Time series analysis of collective motions in proteins”, *J. of Chemical Physics*, Vol. 120, pp. 1072-1088, 2004.

11. Alakent, B., P. Doruker and M. C. Çamurdan, “Application of time series analysis on molecular dynamics simulations of proteins: A study of different conformational spaces by principal component analysis”, *J. of Chemical Physics*, Vol. 121, pp. 4759-4769, 2004.

12. Alakent, B., M. C. Çamurdan and P. Doruker, “Hierarchical structure of the energy landscape of proteins revisited by time series analysis. I. Mimicking protein dynamics in different time scales”, *J. of Chemical Physics*, Vol. 123, 144910, 2005.

13. Alakent, B., M. C. Çamurdan and P. Doruker, “Hierarchical structure of the energy landscape of proteins revisited by time series analysis. II. Investigation of explicit solvent effects”, *Chemical Physics*, Vol. 123, 144911, 2005.

14. Cansu S. and P. Doruker, “Dimerization Affects Collective Dynamics of Triosephosphate Isomerase”, *Biochemistry*, Vol. 47, No. 5, pp. 1358-1368, February 2008.

15. Case, D.A., T.E. Cheatham, III, T. Darden, H. Gohlke, R. Luo, K.M. Merz, Jr. A. Onufriev, C. Simmerling, B. Wang and R. Woods, “The Amber biomolecular simulation programs”, *J. Computat. Chem.* Vol. 26**,** pp. 1668-1688, 2005.

16. Petsko, G. A. and D. Ringe, *Protein Structure and Function*, Blackwell Publishing, London, 2004.

17. Shimon A. B. and M. Eisenstein, “Looking at Enzymes from the Inside out: The Proximity of Catalytic Residues to the Molecular Centroid can be used for Detection of

Active Sites and Enzyme–Ligand Interfaces", *J. Mol. Biol.*, Vol. 351, pp. 309–326, 2005.

18. Cummins, P. L., K. Ramnarayan, U. C. Singh and J. E. Gready , "Molecular Dynamics/Free Energy Perturbation Study on the Relative Affinities of the Binding of Reduced and Oxidized NADP to Dihydrofolate Reductase", *J. Am. Chem. Soc.*, Vol. 113, pp. 8247-8256, 1991.

19. Chen J., R. I. Dima and D. Thirumalai, "Allosteric Communication in Dihydrofolate Reductase: Signaling Network and Pathways for Closed to Occluded Transition and Back", *J. Mol. Biol.*, Vol. 374, pp. 250-266, 2007.

20. Sawaya M. R. and J. Kraut, "Loop and Subdomain Movements of Escherichia coli Dihydrofolate Reductase: Crystallographic Evidence", *Biochemistry*, Vol. 36, pp. 586–603, 1997.

21. Available: http://www.scielo.br/scielo.php?script=sci_arttext&pid

22. Radkiewicz J. L. and C. L Brooks, "Protein Dynamics in Enzymatic Catalysis: Exploration of Dihydrofolate Reductase", *J. Am. Chem. Soc.,* Vol. 122, No. 2, pp. 225-231, 2000.

23. Available: http://en.wikipedia.org/wiki/Triose_phosphate_isomerase

24. Rozovsky S. and A. E McDermott, "The time scale of the catalytic loop motion in triosephosphate isomerase", *Journal of molecular biology*, Vol. 310, No. 1, pp. 259-270**,** 2001.

25. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, "The Protein Data Bank", *Nucleic Acids Research*, Vol. 28, No. 1, pp. 235-242, January 2000.

26. Duan, Y., C. Wu, S. Chowdhury, M.C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R.Luo and T. Lee, "A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins", *J. Comput. Chem.*, Vol. 24**,** pp. 1999-2012, 2003.

27. Essman, U., L. Perera, M. L. Berkowitz, T. A. Darden, H. Lee and L. G. Pedersen, "A smooth Particle Mesh Ewald method", *J. Chem. Phys.*, Vol. 103, pp. 8577-8593, 1995.

28. Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L Klein.,"Comparison of simple potential functions for simulating liquid water", *J. Chem. Phys.*, Vol. 79, pp. 926, 1983.

29. Berendsen, H. J. C., J.P.M. Postma, W.F. van Gunsteren, A. DiNola and J.R. Haak, "Molecular dynamics with coupling to an external bath", *J. Chem. Phys.*, Vol. 81, pp. 3684-3690, 1984.

30. Zhao, Y. and H. Ke, "Crystal Structure Implies That Cyclophilin Predominantly Catalyzes the *Trans* to *Cis* Isomerization", *Biochemistry*, Vol. 35, No. 23, pp. 7356-7361, 1996.

31. Chatfield, C., *The Analysis of Time Series: Theory and Practice*, Chapman and Hall, 1975.

32. Box, G. E. P. and G. M. Jenkins, *Time Series Analysis Forecasting and Control*, Holden-Day, San Francisco, 1970.

33. Cheng, JW., CA Lepre, JM. Moore, "15N NMR relaxation studies of the FK506 binding protein: dynamic effects of ligand binding and implications for calcineurin recognition", *Biochemistry*, 33(14):4093–4100, 1994.

34. Rischel, C., J.C. Madsen, K.V. Andersen and F.M. Poulsen, "Comparison of Backbone Dynamics of *apo-* and hob-Acyl-Coenzyme A Binding Protein Using 15N Relaxation Measurements", *Biochemistry*, 33, 13997–14002, 1994.

35. Schmid, F.F. and M. Meuwly, "All-atom Simulations of Structures and Energetics of c-di-GMP-bound and free PleD" **,** *J. Mol. Biol.*, 374, 1270–1285, 2007.

36. Balog, E., T. Becker, M. Oettl, R. Lechner, R. Daniel, J. Finney and J. C. Smith, "Direct Determination of Vibrational Density of States Change on Ligand Binding to a Protein", *Physical Review Letters*, Vol. 93, No. 2, 028103-1-4, 2004.

37. Ben-Avraham, D., "Vibrational Normal-Mode Spectrum of Globular Proteins", *Physical Review B*, Vol. 47, No. 21, pp. 559-560, June 1993.

38. Zidek, L., M. V. Novotny and M. J. Stone, "Increased protein backbone conformational entropy upon hydrophobic ligand binding", *Nature Structural Biology* ,6, 1118 (1999).

39. Fischer, S., C. S. Verma, "Binding of buried structural water increases the flexibility of proteins**",** *Proc. Nat. Acad. Sci*., Vol. 96, pp. 9613-9615, 1999.