EFFECT OF TEMPERATURE ON COLLECTIVE DYNAMICS OF PROTEINS: A TIME SERIES ANALYSIS

by

Özlem Türe

B.S., Chemical Engineering, Yıldız Technical University, 2005

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

Graduate Program in Chemical Engineering

Boğaziçi University

2008

EFFECT OF TEMPERATURE ON COLLECTIVE DYNAMICS OF PROTEINS: A TIME SERIES ANALYSIS

APPROVED BY:

Prof. Pemra Doruker Turgut (Thesis Supervisor)	.....
Assist. Prof. Burak Alakent (Thesis Co-Supervisor)	.....
Assist. Prof. Elif Özkırımlı	.....
Assoc. Prof. Can Özturan	.....
Assoc. Prof. Ramazan Yıldırım	.....

DATE OF APPROVAL: 08.09.2008

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisors Prof. Dr. Pemra Doruker Turgut and Assist. Prof. Burak Alakent for their invaluable help and guidance for the preparation of this thesis.

I am grateful to Assoc. Prof. Ramazan Yıldırım, Assist. Prof. Elif Özkırımlı and Assoc. Prof. Can Özturan for the time they devoted in reading and commenting on my thesis.

It was a great pleasure for me that this work has been supported by TÜBİTAK (104M247), the Turkish Academy of Sciences.

My special thanks are to Prof. Dr. Türkan Haliloğlu for being kind and giving moral support. I also would like to thank to all PRC members for their friendliness. I am very grateful to my family who was always with me with the endless support and love.

ABSTRACT

EFFECT OF TEMPERATURE ON COLLECTIVE DYNAMICS OF PROTEINS: A TIME SERIES ANALYSIS

Molecular Dynamics (MD) simulations are used to analyze the internal motions of proteins. In this thesis, the molecular dynamics trajectories of the apo form of dihydrofolate reductase (DHFR), and both apo and holo forms of triosephosphate isomerase (TIM) at three different temperatures, 200 K, 300 K and 400 K are examined. Analysis mainly consists of utilization of two methods: principal components analysis (PCA) to determine the collective protein fluctuations with high mean square fluctuations, and linear time series analysis to examine the collective vibrational motions in detail. Time series model parameters obtained for the free states of DHFR and TIM are similar, indicating the reliability of the analysis. It is found that at high temperatures collectivity reduces and global twisting motion seen in both proteins remarkably diminishes. At low temperatures, the important loop motions are reduced. Vibrational frequencies of the first 40 principal modes are extracted by time series analysis, and probability density functions of these frequencies are plotted to compare different MD runs. It is seen that simulations at higher temperatures have lower frequency distributions. Nevertheless, the difference between 300 K and 400 K is very small compared to the frequency shift from 200 K to 300 K. For its ligand bound form, TIM has higher frequencies than the free form at 200 K, as seen at 300 K in a previous study. However, ligand binding reduces the global twisting motion of the two monomers of TIM remarkably, which is opposite to what has been observed at 300 K. This shows that ligand binding may have different effects on the collective motions at different temperatures.

ÖZET

SICAKLIĞIN PROTEİNLERİN KOLEKTİF DİNAMİKLERİ ÜZERİNE ETKİSİ: BİR ZAMAN SERİLERİ ANALİZİ

Moleküler Dinamik simülasyonları proteinlerin iç hareketlerini analiz etmek için kullanılır. Bu tezde, dihidrofolat redüktazın (DHFR) apo formunun ve triozfosfat izomerazın (TIM) hem apo hem de holo formlarının moleküler dinamik verileri üç farklı sıcaklıkta, 200 K, 300 K ve 400 K'de, incelenmiştir. Analizler çoğunlukla iki metodun kullanımını kapsamaktadır: kolektif protein salınımlarını yüksek ortalama kareler dağılımı ile tespit etmek için ana bileşenler (esas modlar) analizi ve kolektif titreşimsel hareketleri ayrıntılı olarak incelemek için doğrusal zaman serileri analizi. DHFR ve TIM için elde edilen zaman serileri model parametreleri birbirine benzerdir. Bu da analizin güvenilir olduğunu göstermektedir. Yüksek sıcaklıklarda kolektivitenin ve bütünsel burulma hareketinin azaldığı tespit edilmiştir. Düşük sıcaklıklarda, önemli bölgelerin hareketlerinde azalma olmuştur. Zaman serileri analizi ile ilk 40 moda ait titreşimsel frekanslar çıkarılmış ve bu frekansların farklı moleküler dinamik çalışmaları için olasılıksal yoğunluk fonksiyonları karşılaştırmak amacıyla figür olarak çizilmiştir. Yüksek sıcaklıklardaki simülasyonların daha düşük frekans dağılımları olduğu görülmüştür. Bununla beraber, 300 K ve 400 K arasındaki frekans geçiş aralığı 200 K ve 300 K' dekine göre daha dardır. TIM'in bağlı formunun frekansları serbest formuna göre 200 K' de daha önceki çalışmada bulunduğu gibi daha yüksektir. Fakat ligand bağlanması 200 K'de TIM'in her iki monomerinin de bütünsel burkulma hareketini bozmuştur ki bu 300 K için daha önceki yapılan bir çalışmada bulunanın tam tersidir. Bu durum, ligand bağlanmasının proteinin kolektif hareketleri üzerinde farklı sıcaklıklarda faklı etki yapabileceğini göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

C_α	Carbon Alpha
Å	Angstrom
ns	nano second
ps	pico second
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 components analysis
PDB	Protein data bank
PDF	Probability distribution function
RMSD	Root mean square deviation
TIM	Triosephosphate isomerase

1. INTRODUCTION

Proteins are one of the essential components of living systems. They constitute the macromolecules that have important roles in biology. Proteins contribute to the structure of an organism and execute most of the tasks required for it to function. These functions are determined by their internal motions and accompanying conformational changes.

Proteins are made up of amino acid residues that are connected by peptide bonds. They are long-chain polymers, but unlike most polymers they have a well-defined native state. This state is defined by weak, non-covalent interactions among residues, as a result of which large fluctuations in the atoms are expected. While the proteins fluctuate about their native state, they make transitions between a large number of conformational sub-states [1]. Normal Mode Analysis (NMA) and Molecular Dynamics (MD) simulations are helpful to investigate the protein dynamics [2-5]. These methods have shown that collective motions are directly related with the function of proteins.

Molecular dynamics (MD) is a potential energy surface-based computer simulation technique that can provide the details of the motion of a protein. In MD simulations, it is possible to generate a sequence of points in the phase space that are connected in time, which correspond to the successive conformations of the protein. Newton's equation of motion is solved for each atom in the protein with respect to the chosen force field, and one can obtain a molecular trajectory. Data obtained from the MD simulations comprise the Cartesian coordinates of atoms sampled during the simulation and are high-dimensional due to frequent sampling and large number of atoms. Principal Component Analysis (PCA) can be used to reduce the dimensions of the data, which captures the essential dynamics of the protein [4,6,7-13]. In literature, different names such as effective normal mode analysis and essential dynamics are used for this method. Alakent *et al* [10-13,14] used stochastic linear time series models to model the collective protein fluctuations and related the model parameters to different aspects of protein motions. Alakent *et al* [10,11-14] showed that, PCA is a reliable method in extracting the important dynamics of proteins and linear stochastic time series models can be used in explaining the collective dynamics of a protein along its principal components.

In this thesis, the time evolution of the MD simulations of dihydrofolate reductase (DHFR) and triosephosphate isomerase (TIM) are analyzed by applying time series analysis to the projections of atomic displacements onto the principal components. DHFR is pharmacologically and clinically important as it is the target of the "antifolate" drugs, which are useful anti-neoplastic, antibacterial and antimalarial agents [15]. TIM is an important enzyme in glycolysis. It catalyzes the interconversion between dihydroxyacetone phosphate (DHAP) and D-glyceraldehyde 3-phosphate (GAP) [16]. In its deficiency, susceptibility to infections and various neurological dysfunctions occur [17]. In this work, free (unliganded) forms of DHFR, and both free and ligand bound forms of TIM are analyzed at different temperatures. PCA is applied to the data obtained from each simulation to extract the essential protein dynamics. Time series analysis is used to derive linear stochastic time series models at different temperatures for the free and liganded forms. Parameters of the time series models are used to examine the dynamic behavior of these proteins under different conditions. MD simulations are performed by using AMBER 8.0 commercial software package [18]. MATLAB 7.0 is used for PCA and for obtaining time series models. PYMOL (version 0.99rc6) is used for protein illustrations.

In the second section of this thesis, protein structure, function and dynamics are explained in the context of DHFR and TIM. In the third section, MD simulation method is explained in detail. In the fourth section, statistical methods used in the analysis of MD trajectories are explained. Section five is the Results and Discussion, where the analyses of the MD results are examined in detail. In the final section, conclusions and recommendations about the thesis are presented.

2. PROTEIN STRUCTURE, DYNAMICS AND FUNCTION

2.1. Protein Structure and Dynamics

Proteins are biopolymers. They are formed by the linkage of numerous combinations of 20 residues. Proteins we observed in nature have evolved, through selective pressure, to perform specific functions. The functional properties of proteins depend upon their three dimensional structures. The specific three-dimensional structure of a protein is called the native state. Native state arises because its particular linear sequence of amino acids folds to generate compact domains.

All of the 20 amino acids have in common a central carbon atom (C_α). A hydrogen atom (H), an amino group (NH_2), and a carboxyl group (COOH) are attached to the C_α atom. The side chain attached to the C_α through its fourth valence distinguishes one amino acid from another. Depending on this side chain, each amino acid has different chemical properties [19].

The amino acid sequence of a protein's polypeptide chain is called its primary structure. Different regions of the sequence form local regular secondary structures, such as alpha (α) helices or beta (β) strands. The tertiary structure is formed by packing such structural elements into one or several compact globular units called domains. The final structure may contain several polypeptide chains arranged in a quaternary structure. As shown in Figure 2.1, by formation of such tertiary and quaternary structure, amino acids far apart in the sequence are brought close together in three dimensions to form a functional region, an active site.

The three dimensional structure determines the protein function. The folded domains can act as modules forming large assemblies such as virus particles or muscle fibers, or sometimes they can serve as specific catalytic or binding sites. Two major methods, namely X-ray crystallography and NMR, are used to determine the three dimensional structures of proteins. Lately electron microscopy (EM) is used for the structure determination of supramolecules.

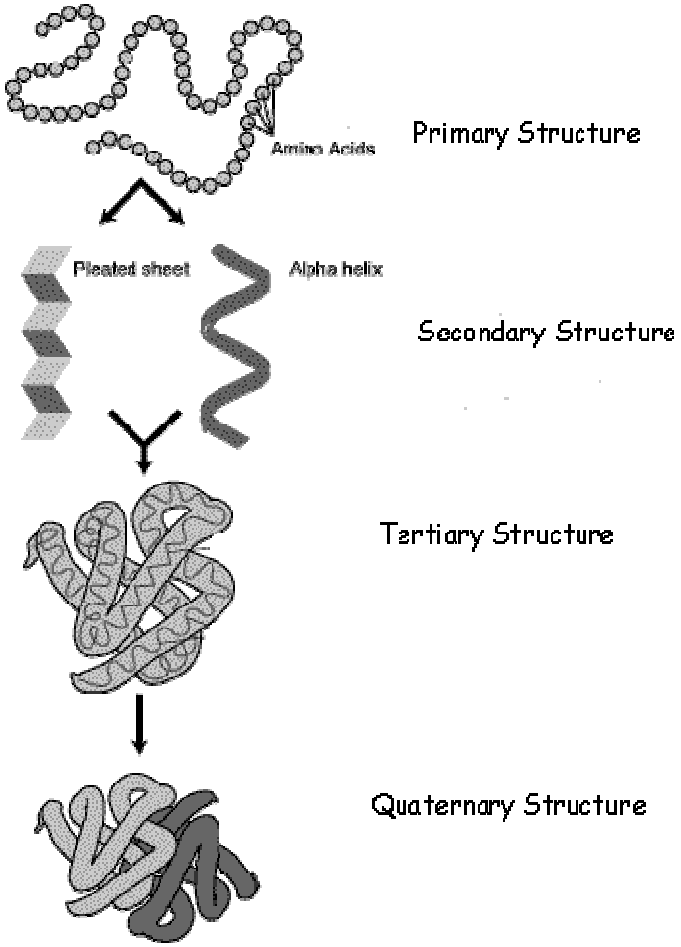


Figure 2.1. The formation of quaternary structure

Proteins are not rigid bodies. They sample a large ensemble of conformations around the native state. Since proteins are very important biological molecules for the cellular function, researchers have sought to elucidate how these complex macromolecules execute various functions. Even though the static structures are known for many proteins, the functions of proteins are governed not only by their static structure but also by their dynamic character. For a detailed description of a protein's multidimensional energy landscape, the relative probabilities of the conformational states (thermodynamics) and the energy barriers between these states (kinetics) are required [20].

2.2. Functions of Proteins

Proteins are very important molecules in our cells. They are involved in virtually all cellular functions. Each protein within the body has a specific function or several functions. Some proteins are involved in structural support, while others are involved in bodily movement, or in defense against germs. Proteins have numerous types that they can act as enzymes, antibodies etc. [21].

In enzymatic reactions, the reactant molecules are called substrates, and enzymes convert them into products. Most of the processes occurring in the cell require enzymes to have significant rates. Enzymatic proteins act as catalysts in cellular reactions. They accelerate the reaction by decreasing the activation energy barrier to products (Figure 2.2). Enzyme reaction rates are nearly millions of times faster than the uncatalyzed reactions.

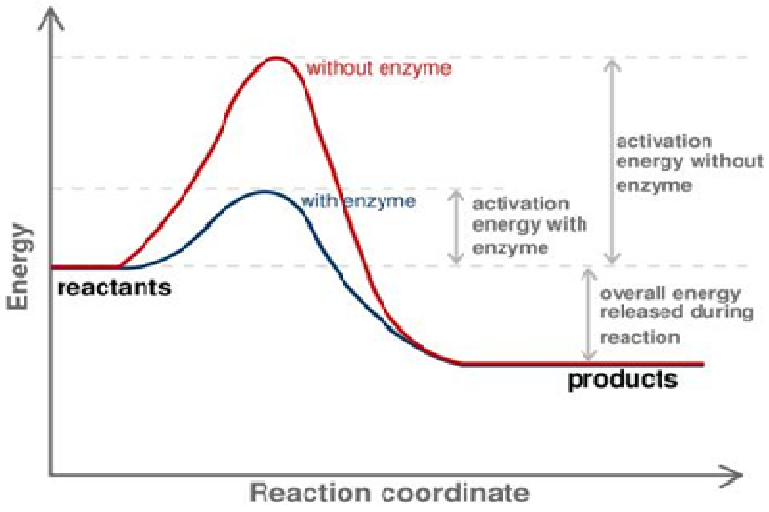


Figure 2.2. Effect of enzymes on reaction kinetics [22]

In catalytic reactions, ligand binds to a specific part of an enzyme, named binding site, by means of intermolecular forces. Interactions between the protein and the ligand are essential for the protein to function properly [23]. Ligands mostly bind to a small number of residues, but it affects the other parts of the protein structure leading to significant

changes in loops [23]. It is shown in previous studies that ligand binding shifts the vibrational frequencies of proteins [24].

2.3. Effect of Temperature on Enzymes

Like most chemical reactions, temperature has a remarkable effect on enzyme-catalyzed reactions. A 10^{o}C rise in temperature will increase the activity of most enzymes by 50 to 100 per cent, however above normal temperatures (about 60^{o}C) heat has an irreversible effect on enzymes. This denaturation occurs due to the change in the structure of proteins, especially in the active site, which leads to the inactivation of proteins. As the positive and negative effects of temperature are taken into consideration then enzymes may be said to have an optimum temperature for their action [25]. In this section, the dynamics of DHFR and TIM are examined in free and ligand bound forms at temperatures 200 K, 300 K and 400 K, respectively. It is not possible for an enzyme to execute its functions properly at 200 K. At such a low temperature the motion of the parts of the enzyme is restricted related with the small amount of kinetic energy. Due to the limited mobility, the catalytic mechanism in which the enzyme takes part spoils. At 300 K the enzyme can execute its function properly where at 400 K the enzyme is not functional related with the conformational changes in the structure. In a previous study, vibrational density of states (frequency distribution) of free and ligand bound forms of proteins are examined at 120 K by using molecular dynamics and normal mode analysis [26]. The results of this study show that the free energy change, which is determined by frequency change, contributes to the binding equilibrium [26].

2.4. Dihydrofolate Reductase (DHFR)

Dihydrofolate reductase (DHFR), reduces dihydrofolate (DHF) to tetrahydrofolate (THF), using NADPH (reduced form of $NADP^{+}$, Nicotinamide Adenine Dinucleotide Phosphate) as the hydride donor. DHF can be converted to the kinds of tetrahydrofolate cofactors used in 1-carbon transfer chemistry. In the catalytic cycle of DHFR (Figure 2.3), the active site loops are in closed conformation in the forms of holoenzyme (E^{NH}) and Michaelis complex (E^{NH}_{DHF}), while the ternary (E^{N+}_{THF}), binary (E_{THF}) and product release

complexes (E^{NH}_{THF}) are in the occluded conformation. In the figure, NH and N^+ refer to NADPH and $NADP^+$, respectively [27].

Dihydrofolate reductase deficiency has been linked to megaloblastic anemia. DHFR is clinically important and it has been recognized as a drug target for inhibiting DNA synthesis in rapidly dividing cells such as cancer cells.

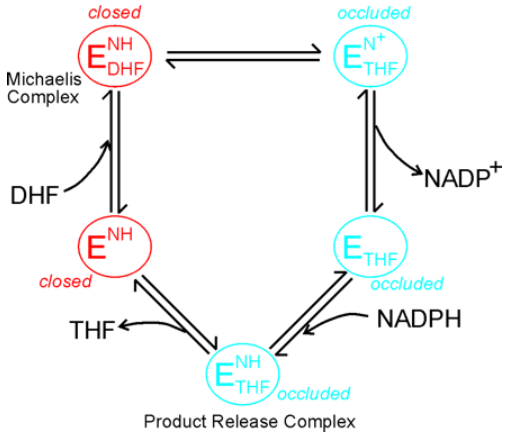


Figure 2.3. Catalytic cycle of DHFR [27]

Because tetrahydrofolate is the active form of folate in humans, inhibition of folate can cause functional folate deficiency. As folate is needed by rapidly growing cells to make thymine, this may have a therapeutic effect.

DHFR consists of 159 residues. The important regions are M20 (residues 9-24), FG (residues 117-131) and GH (residues 142-149) loops (Figure 2.4). M20 loop can be in the open, closed or occluded conformation. In Figure 2.4, the green, blue and orange colored parts are M20, FG and GH loops, respectively. The pink colored parts are the helices in the structure where blue colored parts are the blue parts correspond to the sheets. In the figure, DHFR with the ligand NADPH and substrate DHF is shown.

2.5. Triosephosphate Isomerase (TIM)

Triosephosphate isomerase (TIM) is an enzyme that catalyzes the reversible inter-conversion of the triosephosphate isomers dihydroxyacetone phosphate (DHAP) and D-

glyceraldehyde 3-phosphate (GAP) (Figure 2.5). Triosephosphate isomerase is a highly efficient enzyme, performing the reaction billions of times faster than it would occur naturally in solution.

TIM plays an important role in glycolysis and is essential for efficient energy production. In humans, deficiencies in TIM are associated with a progressive, severe neurological disorder called triosephosphate isomerase deficiency. TIM has been found in nearly every organism, including animals such as mammals and insects as well as in fungi, plants and bacteria. However, some bacteria that do not perform glycolysis, lack TIM.

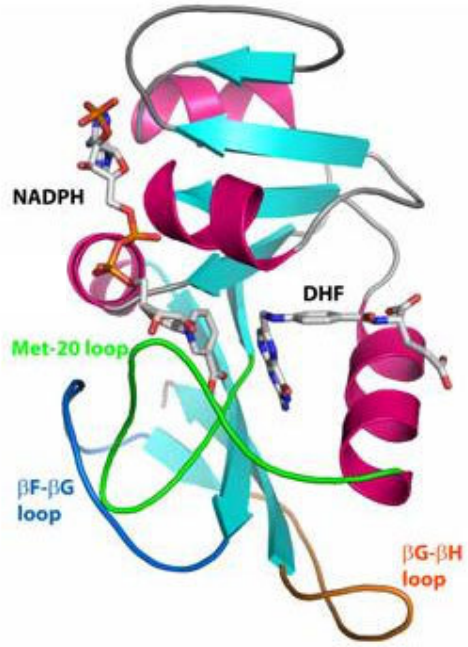


Figure 2.4. Cartoon representation of DHFR with NADPH and DHF shown in sticks [28]

Triosephosphate isomerase consists of 494 residues and is a dimer of identical subunits (Figure 2.6). The three-dimensional structure of a subunit contains eight α-helices on the outside (red colored) and eight parallel β-strands (yellow colored) on the inside. This structural motif is called an αβ-barrel, or a TIM-barrel, and is by far the most commonly observed protein fold. The active site of this enzyme is in the center of the barrel. Loop 6 (blue colored), which comprises residues 166 to 176; closes over the active site and protects the ligand from the solvent.

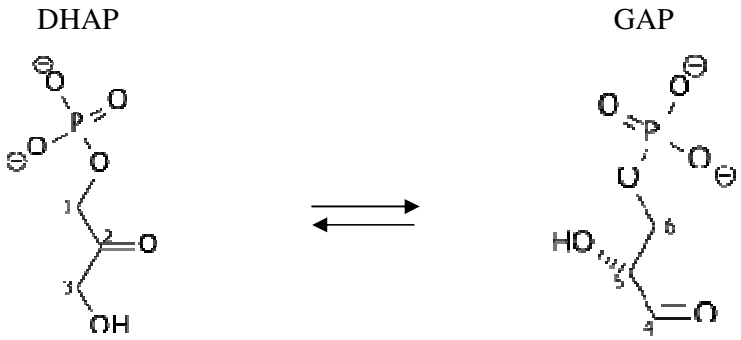


Figure 2.5. The inter-conversion of DHAP and GAP [29]

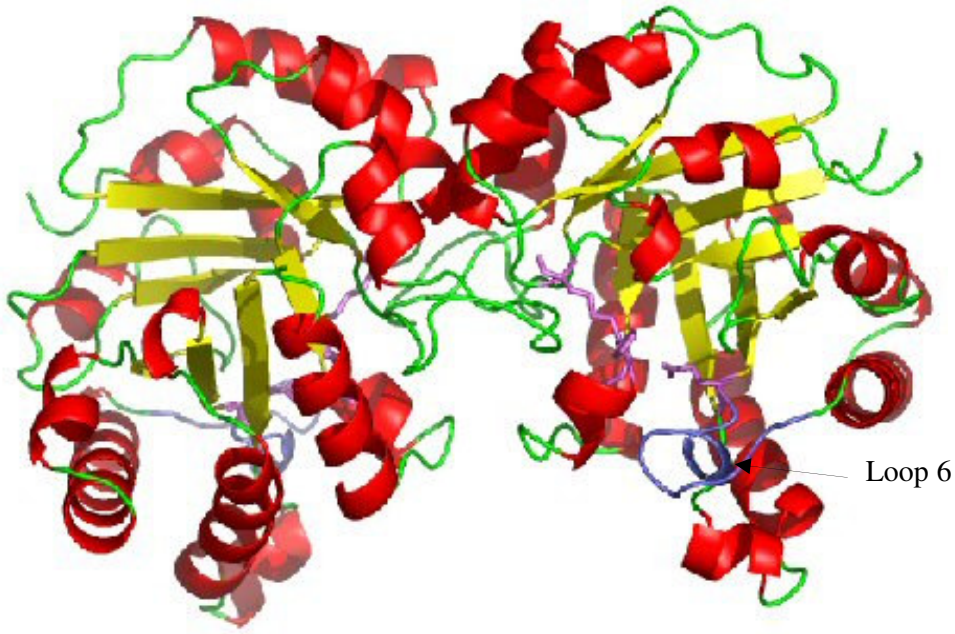


Figure 2.6. Cartoon representation of TIM [30]

3. MOLECULAR DYNAMICS SIMULATIONS OF PROTEINS

3.1. Molecular Dynamics (MD) Simulations of Biomolecules

Molecular dynamics (MD) is a computer simulation technique used for understanding the physical basis of the structure-function relation of biological macromolecules. The early view of proteins as relatively rigid structures has been replaced by the idea that the internal motions and resulting conformational changes play a vital role in their function [10,31]. MD makes it possible to study these motions and conformational changes.

In MD simulations, a trajectory of the protein is obtained by determining the positions and velocities of the particles. This trajectory is obtained by solving the differential equations embodied in Newton's second law for each particle:

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i} \tag{3.1}$$

This equation describes the motion of a particle of mass m_i along one coordinate (x_i) with F_{x_i} being the force on the particle in that direction. In MD simulations, force fields are used to describe the potential energy surface of a system as a function of the atomic positions. The quality of a force field determines the validity of the results obtained. Force fields are based on an empirical model of interactions involving bond stretching, bond angle deformations and rotation of bonds, as well as non-bonded interactions, namely van der Waals and electrostatic interactions within a system. In this thesis, an "all-atom" force field, AMBER ff03, is used [1,17].

The success of an MD simulation depends not only on a high quality force field but also on the computational limitations regarding the number of particles. For this reason, additional algorithms should be used to increase the efficiency of MD simulations. These algorithms are based on realistic and reasonable assumptions, such as the application of

non-bonded cutoffs. In the case of non-bonded interactions, the distances between atoms larger than the cutoff distance (r_{cut}) are neglected.

It is impossible to cover the molecule with infinitely many solvent molecules. Therefore, in order to simulate a protein in solvent to mimic the real situations, periodic boundary conditions (PBC) are used. A cubic box is implemented in all three dimensions, forming a lattice of identical cubes. The surrounding imaged atoms exert forces on the real atoms in the interior cube, while motions of imaged atoms are calculated by symmetry operations. Thus, energy calculations need to be performed and stored only for real atoms [1,17].

3.2. Steps of an MD Simulation

There are three stages in an MD simulation. These are minimization, equilibration and data collection. The initial native folded structure used in the MD simulations is obtained from Protein Data Bank (PDB) [32], which is either determined by X-ray crystallography or NMR. In energy minimization stage, the structure is brought to a conformation where its potential energy is close to an energy minimum with respect to the force field being used. Since it is difficult to find a global minimum due to the nonlinear nature of the protein interactions, a local minimum close to the starting structure is found [24]. If the simulation medium is explicit solvent, the protein is placed in a periodic box along with the water molecules before minimization. Widely used energy minimization techniques are steepest descent and conjugate gradient algorithms.

The equilibration stage is an adjustment period, during which the temperature, energies and other parameters are equilibrated. Initial velocities are assigned according to the Maxwell-Boltzmann distribution at the specified temperature to start the integration of the equation of motion and the target temperature is maintained by adjusting the velocity of the atoms. If there is no control over temperature, fluctuations or drifts in temperature may occur during the simulation.

In the data collection stage, M (sample number) snapshots are obtained from MD simulations. Each snapshot file comprises the Cartesian coordinates of all the atoms

sampled during the simulation. So, the first data file has the coordinates of the initial conformation, and the last one has the coordinates of the finally sampled conformation [24].

3.3. MD Simulation Protocols for DHFR and TIM

The initial X-ray conformations in MD simulations are taken from PDB. The PDB codes are 1RA1 for the free forms of DHFR and 8TIM and 1TPH for the free and ligand bound forms of TIM, respectively. Explicit solvent condition is applied in the simulations which are held at 200 K, 300 K and 400 K for both DHFR and TIM with durations of 3.2 to 32 ns. Data are collected in MD simulations for a period of 3.2 ns with fixed sampling intervals of 0.8 ps. Simulation condition is constant number of moles, temperature and pressure (constant NTP). A 9 Å cutoff is used in the solvated system appointed by the Ewald summation technique with particle-mesh electrostatics. The structure is solvated in TIP3P water by using a truncated octahedron periodic box (AMBER manual) of 58 Å dimensions [33]. Isotropic position scaling (NTP) is used with constant pressure conditions. Energy minimization is applied with steepest descent algorithm of 50 cycles. The initial velocities are assigned according to the Boltzmann distribution at 10 K, and the temperature is increased to 200 K, 300 K and 400 K respectively. The temperatures are kept constant by weak coupling algorithm. Verlet algorithm is used to make the integrations of Newton's equation of motion [34]. An integration time step of 2 fs is used by applying SHAKE algorithm for the bonds with hydrogen [14], to satisfy bond geometry constraints during the simulation.

4. ANALYSIS OF MD TRAJECTORIES WITH STATISTICAL METHODS

4.1. Principal Components Analysis (PCA)

Principal components analysis (PCA) is a technique used to reduce multidimensional data sets to lower dimensions for analysis. PCA is mathematically defined as an orthogonal linear transformation that transforms data to a new coordinate system such that the projection of data on the first latent coordinate (called the first principal component) has the highest variance; projection to the second on the second principle component has the second greatest variance, and so on.

In this study the data matrix X, with dimensions M x q is reduced to smaller matrices. First a covariance matrix (C), a nonsingular q x q symmetric matrix constructed from the data matrix, is obtained by PCA.

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

Here, $\overline{X}$ denotes the matrix of averages. C can be decomposed to an orthonormal "loadings" matrix P and a diagonal matrix Λ (Equation 4.2)**.** Nonzero diagonal elements of Λ correspond to the variation along each principle component. Any number of principal components, r can be taken into consideration. If r is equal to q, then all the variables have been taken into consideration and there is no dimensional reduction.

$$C = P \Lambda P^T \tag{4.2}$$

Projection of the data matrix onto the principal axes is obtained as,

$$T = (X - \overline{X})P. \tag{4.3}$$

Each column of T is named as scores vector and shown by t_i. T matrix has M x r size.

Trajectory of the Cartesian coordinates of the C_α atoms is considered to be sufficient for capturing the essential variation in a protein's fluctuations. All the trajectories obtained from the MD simulations have equal sample sizes of 4000 snapshots sampled at 0.8 ps intervals. These trajectories form the X data matrix, and PCA is applied to this X matrix. Finally, time series analysis is applied to the collective coordinates (scores) obtained by PCA.

4.2. Time Series Analysis

A time series is a set of observations where each observation is recorded at a specific time. In discrete time series, the observations are made as a discrete set. In this case the observations are made at fixed time intervals. Most of discrete time series are constructed by sampling of continuous time series. The time series is said to be deterministic, if the future values of a time series are exactly determined by some mathematical function, while statistical time series are results of stochastic processes and governed by an underlying probabilistic mechanism [35].

Stationary processes are a very special class of stochastic processes and based on the assumption that the process is in a particular state of statistical equilibrium. Stochastic processes are called "strictly stationary processes" if the probability distribution does not change with respect to time. A process is "weakly stationary" if the mean (μ) and variance (σ^2) are constant.

4.2.1. Autoregressive (AR) and Moving Average (MA) Processes

The notation AR(p) refers to the autoregressive model of order p. The AR(p) model is written as,

$$z_t = C + \emptyset_1 z_{t-1} + \cdots + \emptyset_p z_{t-p} + a_t \quad (4.4)$$

where, z_1, z_2, z_3, …, z_p denote a set of p observations. In equation (4.4), $\emptyset$ is the autoregressive parameter, and a_t is a random variable (named random shock), assumed to have zero mean and constant variance in all time periods [35]. It is further assumed that a_t is not autocorrelated. Thus, the a_t terms have the properties of the error terms in a regression equation when the standard assumptions for the regression model hold. In addition, the parameter C is included in the model to allow for the fact that the time series can have non-zero mean.

A time series is a moving-average MA (q) process of order q if it is a linear function of current and past random shocks as shown below:

$$\check{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{4.5}$$

where θ_i (i=1,2,3,…,q) are moving average parameters and a_t is white noise as usual.

4.2.2. Autoregressive Integrated Moving Average Processes

Practical experiences suggest that it is worthwhile to generate model equations that involve both autoregressive (AR) and moving average (MA) terms, as autoregressive moving average (ARMA) models can provide excellent representations of actual stationary time series using relatively few unknown parameters. Frequently adequate representation of a series with pure auto regressions or pure moving averages can only be achieved with high-order models, while a model involving both autoregressive and moving average terms requires a relatively small total number of parameters.

The stationary time series z_t, with mean μ, is said to be generated by autoregressive moving average model of order (p,q), denoted ARMA (p,q), if

$$\check{z}_t - \emptyset_1 \check{z}_{t-1} - \cdots - \emptyset_p \check{z}_{t-p} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{4.6}$$

where,

$$\check{z}_t = z_t - \mu. \tag{4.7}$$

In backward shift operator notation, the ARMA (p,q) model can be shown as

$$\emptyset(B)(z_t - \mu) = \theta(B)a_t. \quad (4.8)$$

Here, the autoregressive (AR) and moving average (MA) parameters are respectively,

$$\emptyset(B) = 1 - \emptyset_1 B - \cdots - \emptyset_p B^p \text{ and } \theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q.$$

Autoregressive moving average models have been successfully used to represent the behavior of stationary time series over a very wide field of practical applications. For the time series whose levels may not be stationary, frequently period-to-period changes, or first differences, of the series will be stationary. If the observed time series is X_t, differencing is defined as follows:

$$w_t = \check{z}_t - \check{z}_{t-1} = (1 - B)\check{z}_t = \nabla\check{z}_t \quad (4.9)$$

If an observed time series X_t has been differenced sufficiently to yield a stationary series w_t, the possibility of fitting to w_t a stationary autoregressive moving average model to w_t can be considered. After differencing, it will be reasonable to assume that the differenced series has mean zero, so that an ARMA (p,q) model can be written as:

$$\emptyset(B)w_t = \theta(B)a_t. \quad (4.10)$$

An autoregressive integrated moving average model of order (p,d,q) which is ARIMA (p,d,q) is shown below, where d denotes the degree of differencing:

$$\emptyset(B)(1 - B)^d X_t = \theta(B)a_t. \quad (4.11)$$

4.2.3. Autocorrelation Functions of AR, MA and ARMA Processes

For a real stochastic process z_t, autocovariance is the covariance of the signal against

a time-shifted version of itself at lag k. When each state of the series has a mean, $E(z_t) = \mu$, then it is defined by [1,35]

$$\gamma_k = \mathrm{cov}(z_t, z_{t+k}) = E[(z_t - \mu)(z_{t+k} - \mu)] \tag{4.12}$$

Since autocovariances are difficult to interpret, it is preferable to use correlations. The correlations provide a scale-free measure of the strength of linear association. The correlations between the observed values of a time series separated by k lags are called the autocorrelations of the process, and denoted by ρ_k,

$$\rho_k = \frac{\gamma_k}{\sigma_z^2} \tag{4.13}$$

When the series is assumed to be stationary then certain restrictions are imposed on the values that can be jointly taken by the autoregressive parameters $\emptyset_1, \emptyset_2, \ldots, \emptyset_p$. In that case, it can be shown that z_t has mean μ, and the autocorrelations obey,

$$\rho_k = \emptyset_1\rho_{k-1} + \emptyset_2\rho_{k-2} + \cdots + \emptyset_p\rho_{t-k}, \; k=1,2,3,\ldots \tag{4.14}$$

which can be written as

$$\emptyset(B)\rho_k = 0 \tag{4.15}$$

In the equation above, B operates on k. The general solution is,

$$\rho_k = A_1G_1^{\,k} + A_2G_2^{\,k} + \ldots + A_pG_p^{\,k} \tag{4.16}$$

where, $G_1^{-1}, G_2^{-1}, \ldots, G_p^{-1}$ are the roots of the equation $\emptyset(B) = 0$. If the roots are real, each term $A_iG_i^{\,k}$ decays to zero as k increases, which referred to as a damped exponential. A pseudo-periodic behavior is encountered as $Ad^k\sin(2\pi f_0 + F)$, if the pair of roots are complex. In this equation, d is the damping factor, f_0 is the frequency in the cycles and F is the phase. For any AR (2) or ARMA (2, i) model having complex roots, with i denoting any integer, frequency and damping factors can be found by using the relations below:

$$d = \sqrt{-\emptyset_2} \tag{4.17}$$

$$f_0 = \frac{1}{2\pi}\cos^{-1}\left(\frac{\emptyset_1}{2\sqrt{-\emptyset_2}}\right) \tag{4.18}$$

Below is the example of autocorrelation distributions for the processes with real and complex roots, respectively.

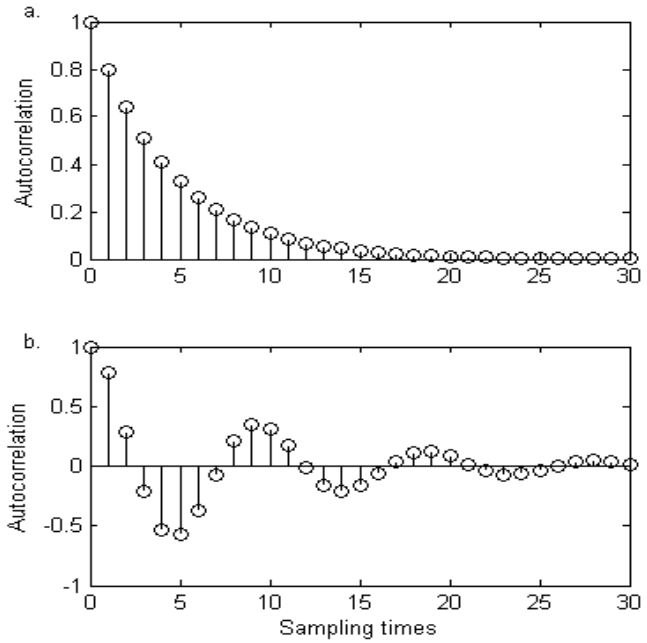


Figure 4.1. Autocorrelation function of processes with (a) real and (b) complex roots

Unlike an AR process, the autocorrelation function is zero beyond order q for a MA process (Figure 4.2a).

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \ldots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \ldots + \theta_q^2}, & k \le q \\ 0, & k > q \end{cases} \tag{4.19}$$

In an ARMA process with an order of (p, q), the autocorrelations for the first q lags depend both on the AR and MA parameters however, after lag q it only depends on the AR

parameter. Figure 4.2b is an example of an autocorrelation function plot for an ARMA (1, 1) process.

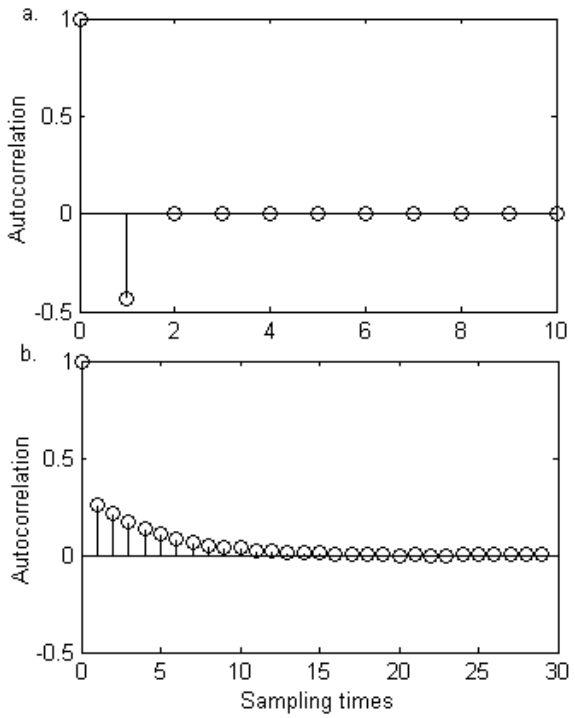


Figure 4.2. ACF distributions for (a) MA (1) and (b) ARMA (1, 1) processes

The underdamped behavior is observed with a graph of $\emptyset_1$and $\emptyset_2$ parameters of the autoregressive (AR) part of the model equation. In Figure 4.3, the grey colored part shows the region where under damped data are encountered according to the inequality, and the triangle shows the stationarity limits.

$$\emptyset_1^2 + 4\,\emptyset_2 < 0 \tag{4.20}$$

Application of time series on the scores obtained from PCA is as follows. Time series models are fitted to the scores to elucidate the mechanism of internal motions. Models are identified according to the statistical properties of residuals (random shock estimates). In choosing the most suitable model among many models generated for the scores data, the following steps are followed:

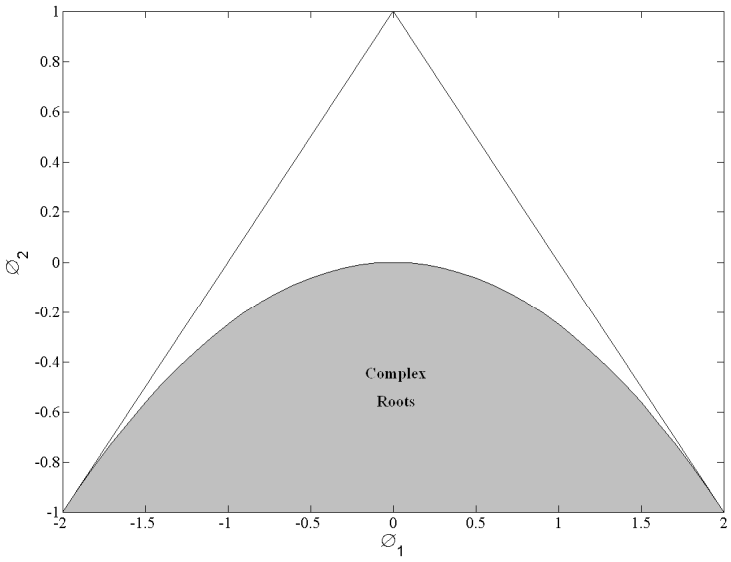


Figure 4.3. Boundaries for stationary and under damped behavior

- The simplest possible model with the smallest number of parameters (low ordered model) is chosen according to the principal of parsimony.

- The autocorrelation function of the residuals of the model should be between the confidence limits. If the autocorrelation function of the residuals at lags greater than zero follows a pattern, the order of the model should be increased.

- If the roots of the AR and MA equations are close to each other, then they will cancel each other. So, lower order models are used instead.

- In order to make the comparison of the models of different scores easier, models with similar orders are selected.

- Complex roots imply an underdamped behavior which yields frequencies. When the autocorrelation function of the residuals shows similar behavior for different models, the model with the complex roots is preferred.

In this study, ARIMA (3,1,2), ARMA (3,1) and ARIMA (3,1,1) models are frequently encountered.

5. RESULTS AND DISCUSSION

In this part, MD simulations of DHFR and TIM are analyzed. Average conformations and mean square fluctuations (MSF) of proteins at different temperatures are calculated and compared. PCA is applied to DHFR and TIM at different temperatures for their free and ligand bound forms, and 40 PCs are taken into consideration in the time series analysis afterwards.

5.1. Investigation of the Dynamics of DHFR

A total of six runs are performed for the free forms of DHFR, two at each temperature of 200 K, 300 K and 400 K, respectively. Short simulations at each temperature are extracted from their longer simulations. The simulation lengths and temperatures of DHFR are shown in Table 5.1:

Table 5.1. DHFR simulation lengths

Temperature	Duration	
200 K	3.2 ns	8 ns
300 K	3.2 ns	32 ns
400 K	3.2 ns	8 ns

The longer simulations are used to extract domain motions by PCA, because longer simulations give more reliable information about the collective motions of the protein. Abbreviations that will be used for DHFR runs are shown in Table 5.2. S denotes short simulations which are all 3.2 ns, and L denotes long simulations. The simulation temperatures of 200 K, 300 K and 400 K are also encoded by 2, 3, and 4, respectively, and "D" represents runs for DHFR.

Time series analysis is only applied to short simulations. The periods of vibrational motions of the modes to be extracted by time series analysis are smaller than 10 ps; therefore a simulation length of 3.2 ns is adequate. In all the runs, a constant sampling interval of 0.8 ps is used for easier comparison of the time series model parameters.

Table 5.2. Notations for DHFR runs

	200 K	300 K	400 K
3.2 ns	D2S	D3S	D4S
8 ns	D2L		D4L
32 ns		D3L	

5.1.1. Comparison of Average Conformations and MSF Analyses of DHFR

The root mean square distance (RMSD) between the average conformations of runs D2L and D3L is 0.726 Å. RMSD between runs D2L and D4L is 1.378 Å and between runs D3L and D4L is 1.214. So the average conformations of runs at 200 K (D2L) and 300 K (D3L) are closer than any other runs. Figure 5.1 shows the average conformation comparisons of runs D2L, D3L and D4L (illustrated in blue, black and red, respectively). In the average structure of run D2L, M20 loop is in its most open conformation (Figure 5.1a, Figure 5.1c and Figure 5.1d). Conformation of M20 loop is similar at 300 K and 400 K. GH loop seems to adopt different average conformations for different runs. On the other hand, deformation is observed especially in the conformation of helix C for the run at 400 K.

In Figure 5.2, MSF of residues are shown. It is clearly seen from the figure that, the mobility of all the regions in DHFR increases as temperature increases. The fluctuations of helix C (residues 44-50), GH loop (residues 142-149), CD loop (residues 64-71) and residues between 80 and 90 are prominent in run D4L. In run D3L and run D4L, the regions with the highest MSF are the CD and, the GH loops, respectively. In run D2L, M20 loop has the highest mobility. Although MSF are the smallest in all the regions of the protein in run D2L, the mobile regions (loops) of the protein can still be identified even at the low temperature. As temperature is increased, energy barriers are overcome especially in those regions, increasing the MSF.

5.1.2. PCA Results of DHFR

The sums of eigenvalues (the total MSF of all residues) for all runs are given in Table 5.3. According to the values, DHFR at 400 K is far more flexible than it is at 200 K

and 300 K, as seen.

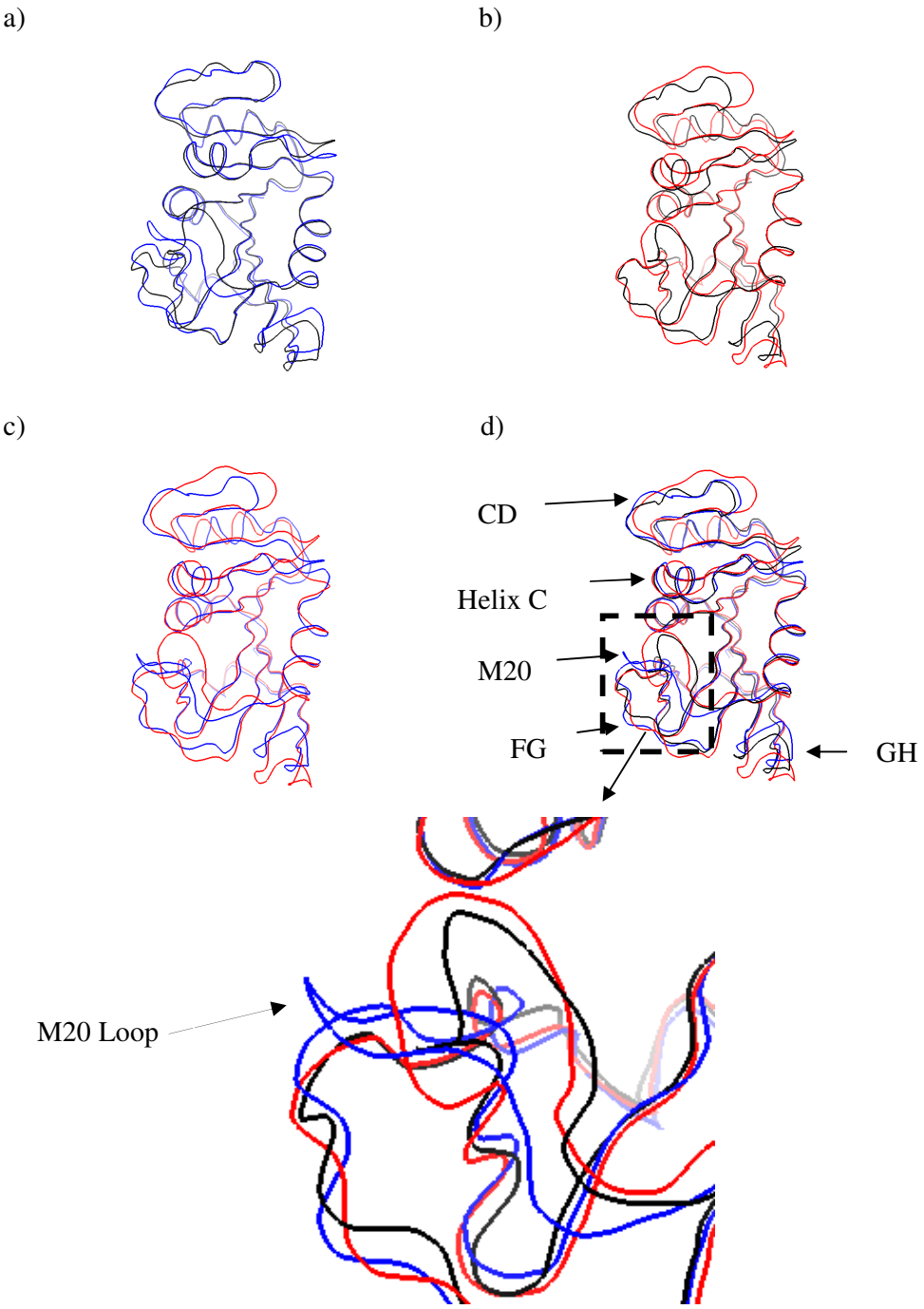


Figure 5.1. Average conformations of DHFR for runs a) D2L-D3L, b) D3L-D4L, c) D2L-D4L and d) D2L-D3L-D4L (blue – 200 K, black – 300 K, red – 400 K)

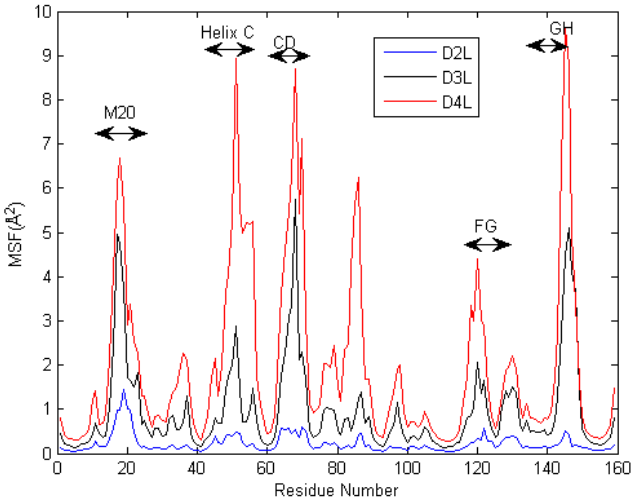


Figure 5.2. MSF of residues for DHFR runs

Table 5.3. Sum of the eigenvalues for DHFR

200 K	35.32 Å²
300 K	152.48 Å²
400 K	320.22 Å²

In the case of eigenvalues, which are the total MSF of residues along the PCs, it is observed that DHFR is in its most mobile state at 400 K and fluctuations are reduced at lower temperatures (Figure 5.3). This is an expected result, since the kinetic energy of the protein is higher at high temperatures, which makes it possible to overcome energy barriers and sample a large range of conformations.

A more reliable measure of the collectivity of the motions at different temperatures may be acquired by examining the percentage of the variance explained by each eigenvalue. Figure 5.4 shows the percentage variance explanation of the modes for all runs. The first 5, 10 and 40 PCs of run D4L explain 55, 68 and 88 per cent of C_α fluctuations, respectively. As it can be seen from the figure, the percentage variance explanation value is higher in the first mode of run D2L and lower in the remaining modes than other runs. In run D4L, for the first three modes, that explain the collective motion,

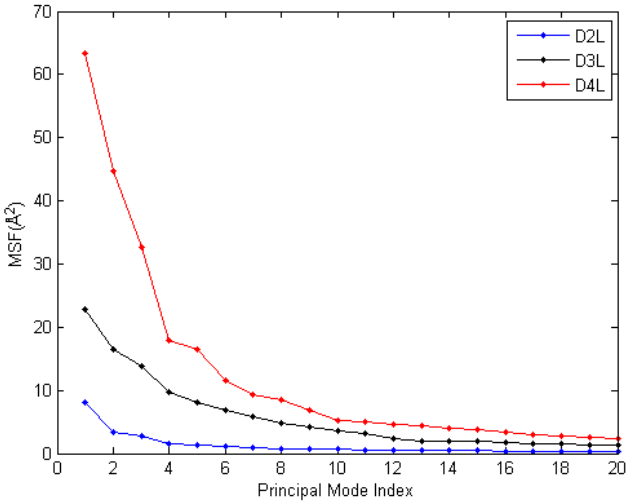


Figure 5.3. Eigenvalue obtained for runs D2L, D3L and D4L

eigenvalues are higher. On the other hand, if the rest of the low indexed modes are also taken into consideration, it is seen that collective motions at 300 K are better pronounced compared to runs at the two other temperatures, which may be related with the functionality of the protein at the optimum temperature. These results support another study [38], whose more multiminimum principal components are found to appear as temperature is increased, and the "excess" MSF from 210 K to 300 K is found to be due to these multiminimum dynamics. In the current study, the intermediate modes (with indices 2 to 12) have a low anharmonicity at 200 K, but a higher anharmonicity at 300 K, and therefore contributing more to the MSF as temperature is increased.

To observe the collective motions more clearly, the displacement vectors along the first principle component of DHFR are examined for each temperature. To see these motions more clearly, only the displacement vectors (without the secondary structure of the protein) are viewed in two different perspectives of the protein. In Figure 5.5, these two perspectives of DHFR are shown: the side and top views of the protein.

Figure 5.6 gives the displacement vector representation of runs D2L, D3L and D4L, where red dots denote the C_α atoms, and black lines correspond to the direction of motion of C_α atoms along PC 1. The motion of M20 loop can easily be observed in the first perspective, while the second perspective more clearly indicates a global twisting type of

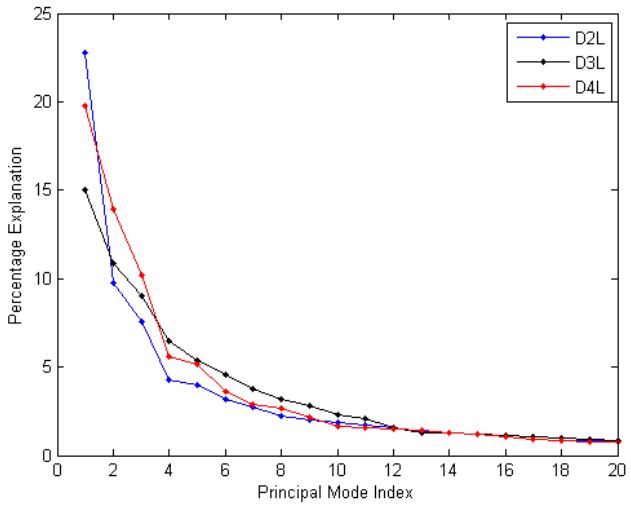


Figure 5.4. Percentage variance explanation of the eigenvalues of DHFR

motion, in which the upper parts of M20 loop of the protein move in the opposite direction of the lower parts.

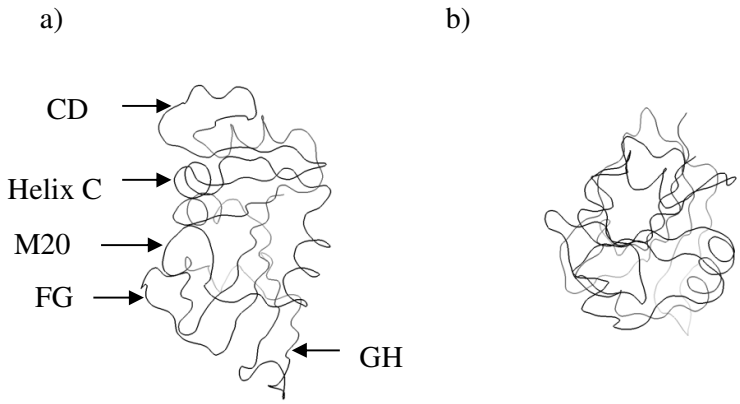


Figure 5.5. Native state illustrations for run D3L from the a) side and b) top views

In Figure 5.6a, the twisting type of motion at 200 K can easily be observed however opening/closing motion of M20 loop is less pronounced compared to other temperatures. The global twisting motion and the motion of M20 loop are more explicit at 300 K, especially seen in the second perspective (Figure 5.6b). The closing motion of M20 loop is clear while helix C moves in the opposite direction. In Figure 5.6c, a better pronounced opening/closing motion of M20 loop is seen, while, the global twisting motion is diminished at 400 K. These results hint the importance of the coordination of the

opening/closing motion of the M20 loop with the global collective motions of DHFR for its function, which can only be attained at an optimum temperature.

a)

b)

c)

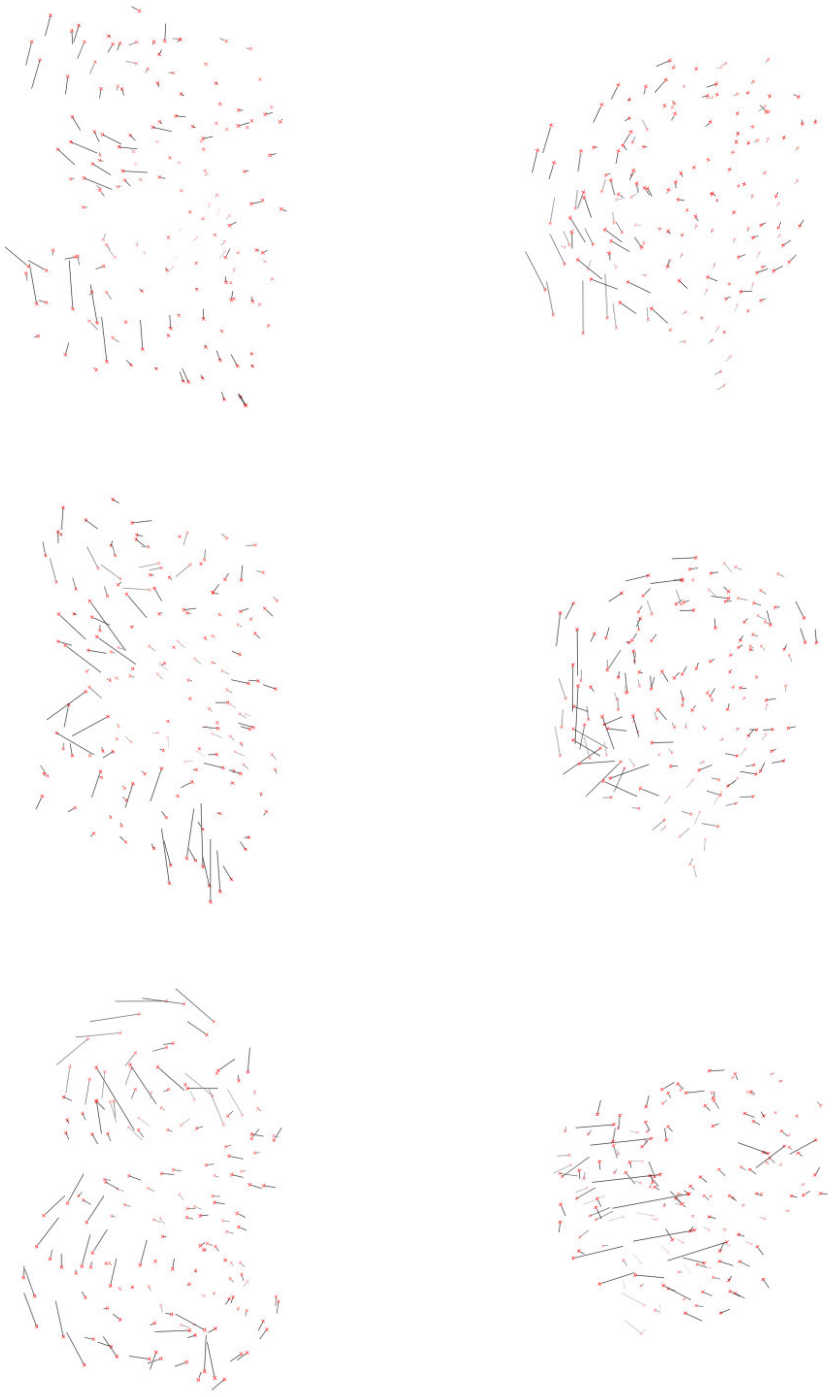


Figure 5.6. Vector field representations of PC 1 in runs a) D2L, b) D3L and c) D4L

Collective motions represent simultaneous movement of the different parts of protein. In order to have a better insight about the contribution of different regions in DHFR to its collective motions, its C_α trace is moved along PC 1 and PC 2 and the resulting snapshots are superimposed on each other. The results are shown in Figures 5.7 to 5.9 for three different temperatures. In Figure 5.7, M20 loop of DHFR at 200 K is in its closest conformation to the FG loop for the first principal component. The motion is not the opening/closing type. A lateral motion exists. On the other hand, along PC 2, M20 loop's opening/closing motion onto the NADPH binding site is more prominent. M20 loop moves in the opposite direction of Helix C where it moves in correlation with FG loop. The movement of GH loop is parallel to M20 loop along both of the principle components. Along PC 2, the movement of CD loop is more evident. Its motion is in the opposite direction of M20 loop. In the first principle component the slight movement of helix C is seen however it is more stationary along PC 2. These results show that M20 loop cannot overcome the energy barriers to make the opening/closing type of motion at 200 K.

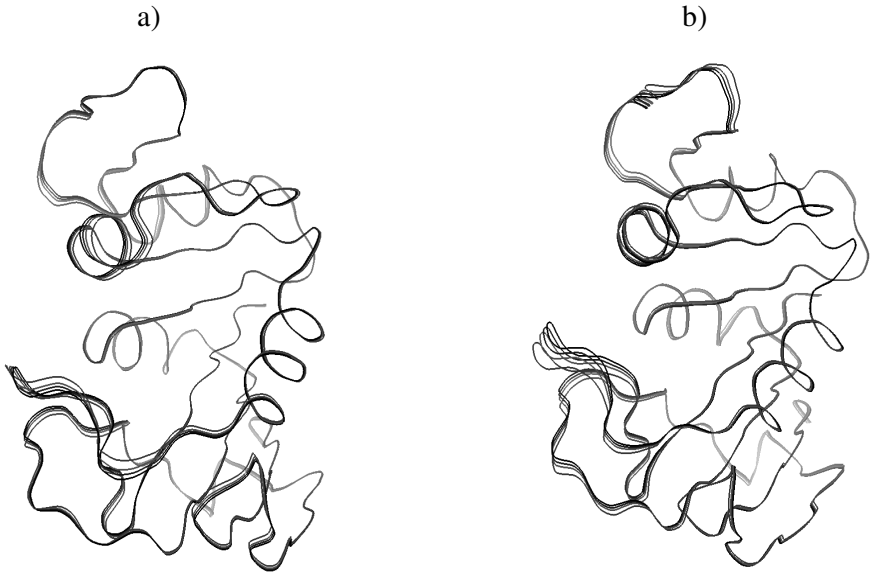


Figure 5.7. Motion of DHFR in run D2L along the a) PC 1 and b) PC 2

In Figure 5.8, the snapshots of DHFR along PC 1 and PC 2 at 300 K are shown. Along PC 1, the opening/closing motion of M20 loop is clear. It moves in parallel with GH loop and in the opposite direction of helix C. A cooperative movement of FG loop can also

be seen. Along PC 2, CD and GH loops dominate the motion. A slight movement of M20 loop is present. There is a collective character along both PCs.

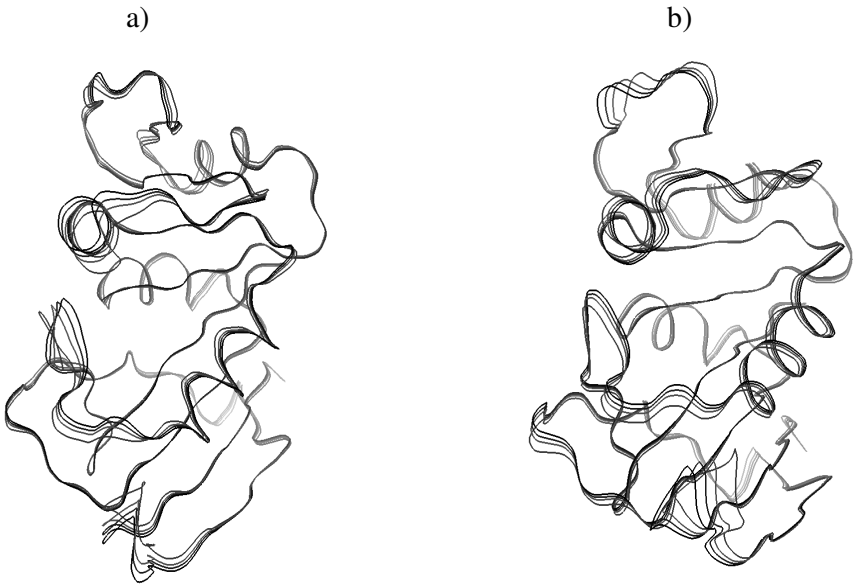


Figure 5.8. Projections of DHFR conformations for run D3L onto a) PC 1 and b) PC 2

Figure 5.9 gives the illustrations of superimposed conformations of DHFR at 400 K along PC 1 and PC 2. In this case, the opening /closing motion of M20 loop is obvious for PC 1 and PC 2. The contributions of CD and GH loops are evident. However, it is also seen that the structure of the protein (especially helices) is distorted, and the fluctuations seem less ordered (especially CD loop).

5.1.3. Derivation of Time Series Models

The projection of the MD trajectory on PC 1 (t_1 scores) of run D2S is shown in Figure 5.10. The t_1 scores have a variance of 5.55 $Å^2$ and variance explanation percentage of 18.94. It should be remarked that the percentage explanation of PC 1 in run D2S (3.2 ns) is different from that of run D2L (8 ns). In this figure, the levels of the scores fluctuate in a wide range during the run showing that the time series is nonstationary.

There are two more parameters that give clues about nonstationarity. These are probability density function (PDF) and autocorrelation function (ACF) of the series. For

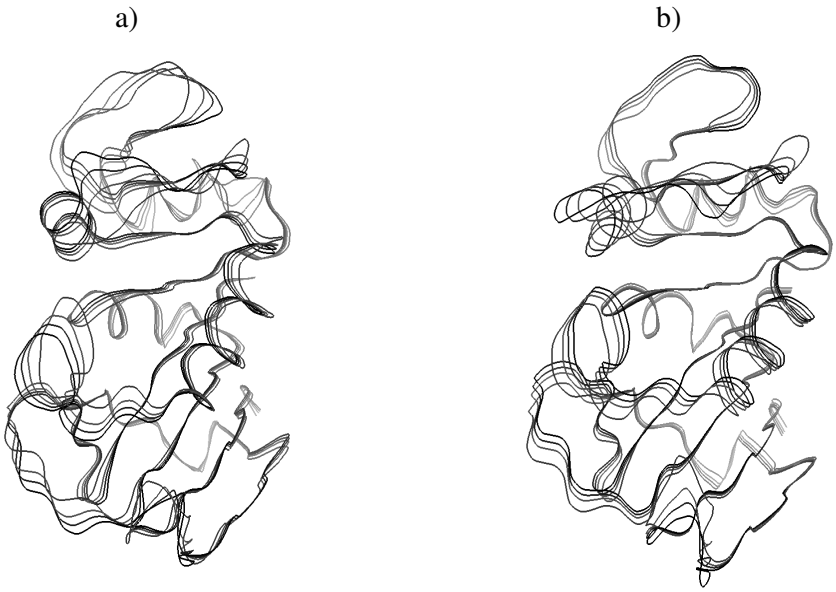


Figure 5.9. Projections of DHFR conformations for run D4L onto a) PC 1 and b) PC 2

t_1 of run D2S, the PDF of the scores shows that the distribution is non-Gaussian (Figure 5.11a). The multiminimum character of the first principal mode encountered even at 200 K should not be surprising, since a double-well principal mode of myoglobin has been found at 180 K in a previous study [36]. Besides, the ACF of t_1 scores does not die out to zero at moderate time lags implying that t_1 is nonstationary (Figure 5.11b).

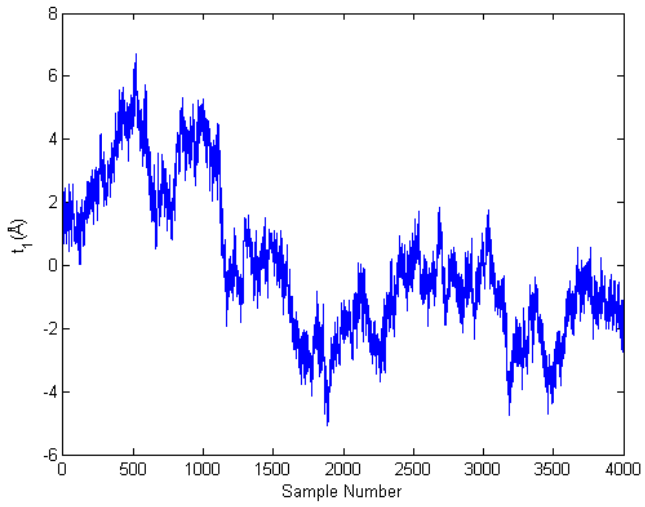
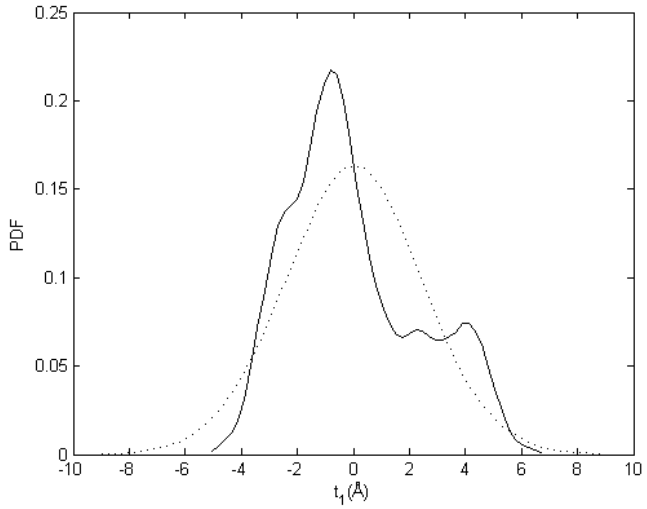


Figure 5.10. Trajectory of t_1 scores of run D2S

a)



b)

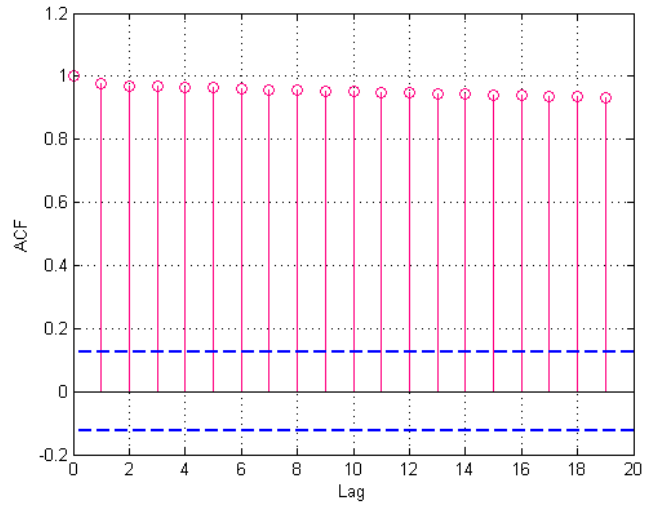


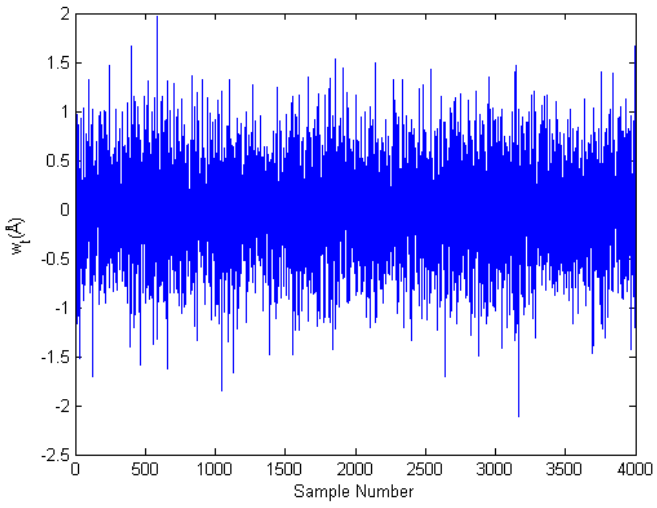
Figure 5.11. a) PDF and b)ACF of t_1 scores of run D2S

To make the time series stationary, difference operator (d=1) is used once on the t_1 scores. Taking the difference by $w_t = \nabla z_t$ (z_t represents the time series, here it is t_1), w trajectory and its PDF plots are shown in Figure 5.12. As a result of differencing, the trajectory w_t has a constant zero mean and variance, and its PDF is Gaussian.

A previous study on time series analysis of a relatively smaller protein provides models of ARIMA (2,1,1) and ARIMA (2,1,2) to explain the motion of the protein in water

[1].

a)



b)

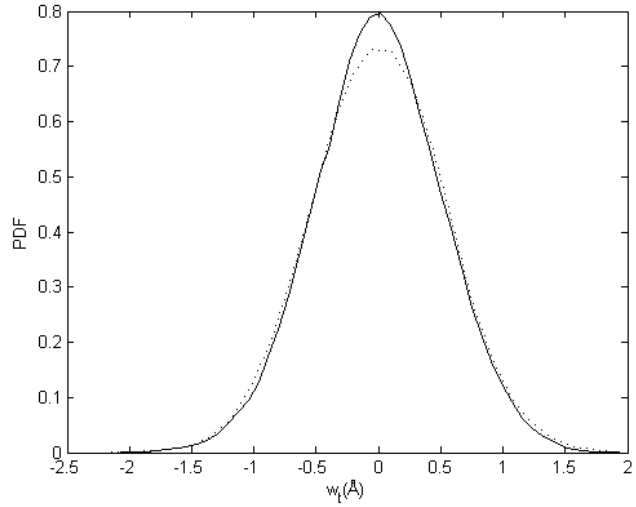


Figure 5.12. (a) w_t trajectory obtained by differencing the t_1 scores (b) PDF of w_t

In these models, AR (2) corresponds to the intraminimum oscillations with a pseudo-periodic character, MA part represents the autocorrelation between the random forces acting on the protein, and I (1) denotes the interminimum motions. In this study, the orders of the derived stochastic time series models are higher, possibly due to the complexity of the fluctuations of DHFR.

In Figure 5.13, the autocorrelation of w_t is shown. One can see that the autocorrelation function cannot belong to an AR (1) or MA (1) process and, the lowest order model can be an ARMA (1,1) process.

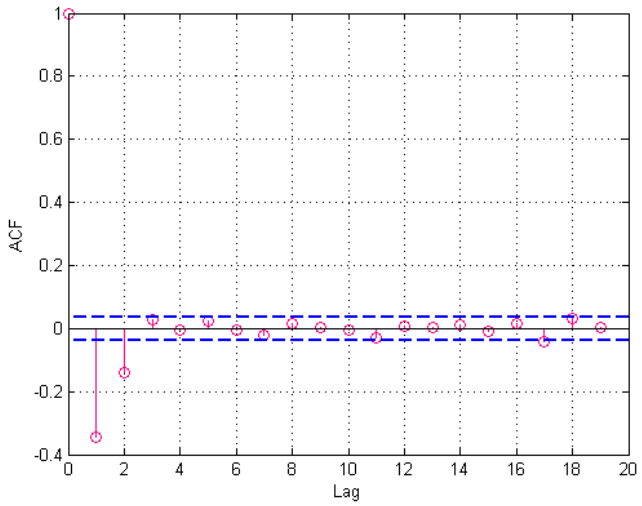


Figure 5.13. Autocorrelation function of w_t

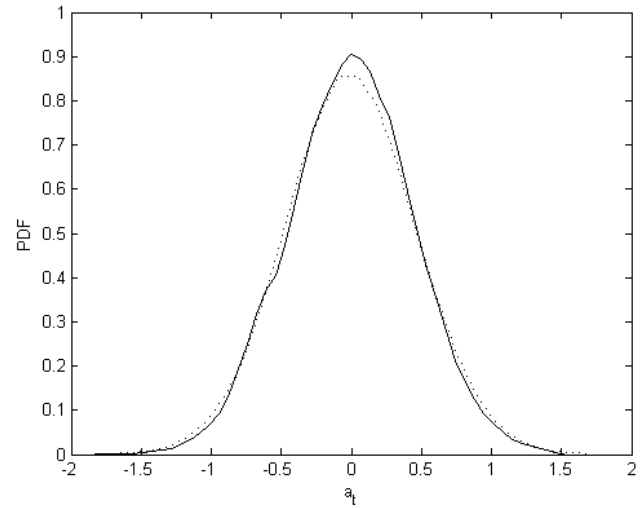
MATLAB's System Identification Toolbox is used to derive the models and find their least square estimates of the model parameters. When the autocorrelation function of the residuals of time series model predictions and other parameters are taken into consideration, ARIMA (3,1,2) is found to be the most adequate model for low-ordered modes of DHFR. For PC 1, an ARIMA (3,1,2) process has been identified with the following parameters and variance of residuals (σ_a^2):

$$(1\text{-}0.9595B)(1\text{-}0.2022B+0.1079B^2)\nabla z_t = (1\text{-}0.9749B\text{-}0.6077B^2)a_t \quad (5.1)$$

$$\sigma_a^2 = 0.1989\ \text{Å}^2 \quad (5.2)$$

In Figure 5.14a, the PDF of residuals (a_t) show that they have a Gaussian distribution. In Figure 5.14b, the ACF of the residuals are shown, and the 95 per cent confidence limits for zero autocorrelation are represented with dotted lines. ACF of residuals mostly fluctuate between the confidence limits for the first 50 lags.

a)



b)

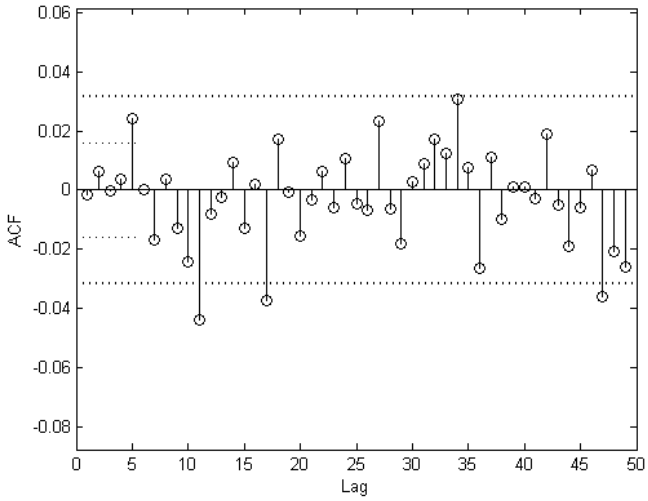


Figure 5.14. a) PDF and b) ACF of the residuals

The AR characteristic equation has the second order polynomial term ($1-0.2022B+0.1079B^2$) with complex roots. Roots of this equation give insight about the pseudo-periodic (underdamped) oscillatory character. If the roots were real, then the behavior would be overdamped.

MA part of the polynomial equation is factorized as (1-9749B) and (1-6077B).The (1-0.9595B) term in AR part and the (1-9749B) term in MA part are very close. If they

canceled each other, a lower ordered ARIMA (2,1,1) process would be obtained. However, the autocorrelation function of the residuals of this ARIMA (2,1,1) process is found to be non satisfactory. Therefore, these two terms do not cancel each other, and the ARIMA (3,1,2) process is chosen as the model to represent the t_1 scores. This behavior of the time series parameters is also encountered in many other modes (see below), and is likely to be a result of the highly nonlinear character of protein dynamics.

Time series analyses for the first 40 modes of run D2S are done. The first 16 modes except a few modes are nonstationary. The remaining modes are mostly stationary; with a few nonstationary ones. ARIMA (3,1,2) is convenient for the nonstationary modes, and ARMA (3,1) is found to be the most adequate model for the stationary modes. These stationary and nonstationary models are shown below, respectively:

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

$$(1-\emptyset_3 B)(1-\emptyset_1 B-\emptyset_2 B^2)z_t=(1-\theta_1 B)a_t \quad (5.4)$$

t_2, t_{20} and t_{40} scores of run D2S are chosen as representative examples for the time series models obtained. These scores are deliberately selected to explain the relationship between time series model parameters and principal components.

t_2: $(1-9653B)(1-0.2255B+0.07501B^2)\nabla z_t=(1-0.973B)(1-0.6682B)a_t$, $\sigma_a^2=0.2033$ Å^2 (5.5)

t_{20}: $(1-0.976B)(1+0.156B-0.0546B^2)z_t=(1-0.7048B)a_t$, $\sigma_a^2=0.1098$ Å^2 (5.6)

t_{40}: $(1-0.952B)(1+0.2101B-0.0514B^2)z_t=(1-0.7264B)a_t$, $\sigma_a^2=0.071$ Å^2 (5.7)

Two of the models above are ARMA (3,1) and one is ARIMA (3,1,2). The second mode (t_2) is a nonstationary one expressed by an ARIMA (3,1,2) model. The $(1-0.2255B+0.07501B^2)$ term of the autoregressive characteristic equation denotes the intraminimum motions. t_2 has an underdamped behavior with a frequency of 8.5714 cm^{-1}. The difference operator represents the interminimum motions. For the second mode,

although the $(1 - 9653B)$ term of the AR part and $(1 - 0.973B)$ term of the MA part are very close they do not cancel out. In previous studies these additional AR and MA terms were not encountered, because the case was to define the dynamics of a smaller protein in water [10-13]. More terms are required to explain the nonlinear characteristic of larger proteins. For this reason, it is assumed that these additional AR and MA parameters are not significant in terms of protein motions but used to derive time series models in a statistically acceptable way. Figure 5.15 shows that for nonstationary modes θ_1 term is always higher than $Ø_3$ term where it is the vice versa for stationary modes due to the absence of a difference operator. The θ_1 versus $Ø_3$ distribution for other runs give similar results. Overall, the additional AR and MA terms are to make a better linear approximation.

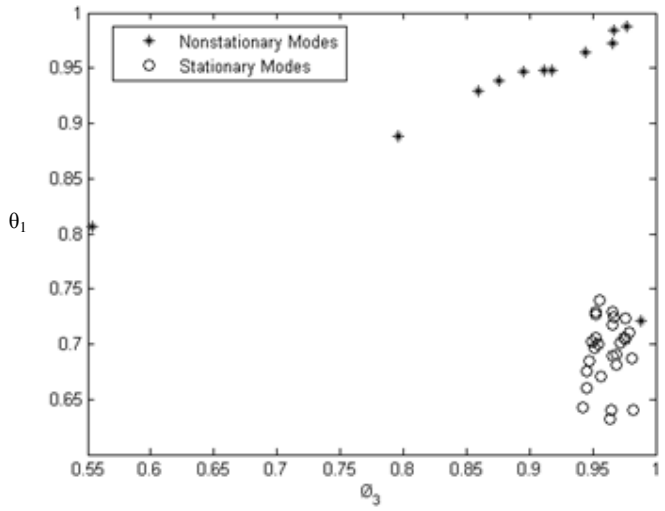


Figure 5.15. $Ø_3$ versus θ_1 for the stationary and nonstationary modes of run D2S

5.1.4. Comparison of Time Series Models of DHFR at Different Temperatures

ARIMA (3,1,2) model is fit to the nonstationary modes of runs D2S and D4S, for the stationary modes ARMA (3,1) model is used. The comparisons of model parameters at 200 K, 300 K and 400 K are done by referring to a previous study [24]. The models and corresponding number of modes are shown in Table (5.4) below, where it is seen that the number of nonstationary modes is smaller at 200 K:

Table 5.4. Types of models and corresponding number modes (DHFR)

Type and Order of the Model	Run D2S	Run D4S
ARIMA (3,1,2)	12	16
ARMA (3,1)	28	24

The compared parameters in the time series models are i) the variance of random shocks (σ_a^2), ii) the θ_2 term of the MA part, and iii) $Ø_1$ and $Ø_2$ terms of the AR part of the characteristic equation. Differences in these parameters make it easier to analyze the effect of temperature on the dynamics of DHFR. Furthermore, $Ø_1$ and $Ø_2$ terms are used to determine the low vibrational frequencies and the corresponding damping factors.

Figure 5.16 shows the variances of residuals at three different temperatures with respect to modes of DHFR. The variance of random shocks is higher as the temperature increases, which is expected due to the direct effect of temperature.

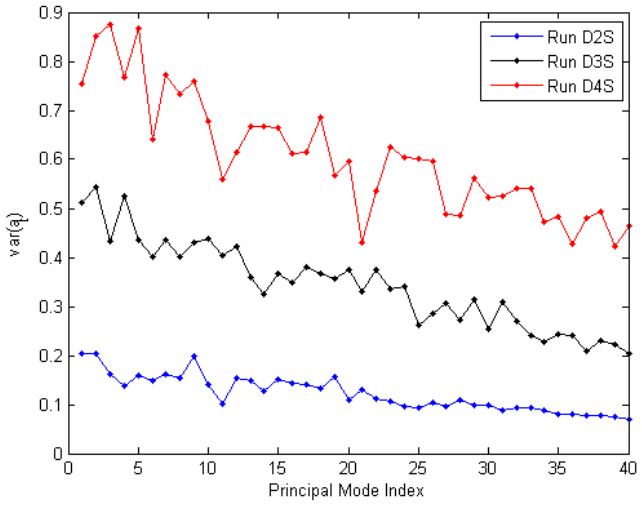


Figure 5.16. Variances of residuals with respect to modes for all runs of DHFR

The correlation between successive random force terms is dictated by the θ_2 terms in the time series models. Boxplot analysis of θ_2 parameters is made to see the differences between the dynamics of DHFR at different temperatures. A boxplot is a graphical summary of the data, where data are described by the smallest observation, lower quartile, median, upper quartile, and the highest observation. In Figure 5.17, the horizontal lines in

the middle of the boxplot diagrams show the median of samples, and each box contains 50 per cent of all observations. Previous studies have shown a wide variability of the θ_2 values, therefore the results obtained from the analysis of this parameter must be examined with caution. In this study the θ_2 parameter for the two proteins shows opposite trends with respect to temperature. In the simulations of DHFR, θ_2 parameter goes from positive to negative values as temperature is increased (Figure 5.17). This means that the random force terms in DHFR are highly negative correlated at low temperatures, whereas this correlation can take a larger range of values, with a median close to zero, as temperature is increased. This, actually, makes sense, as at low temperatures, protein does not have the necessary energy to overcome the barriers. Thus, the random force terms are such to prevent the protein move away from its energy minimum conformation. On the other hand the results obtained for TIM (Figure 5.32) shows that this above interpretation may be insufficient, or cannot be generalized to all proteins, or there are other factors that should be taken into consideration.

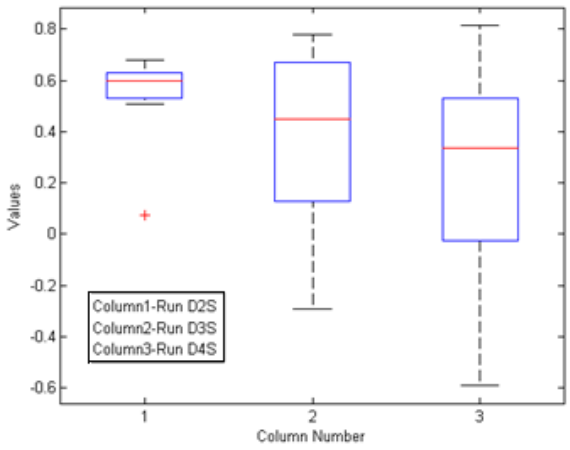


Figure 5.17. Boxplot of θ_2 roots of all DHFR runs

By using the $(1 - \emptyset_1 B - \emptyset_2 B^2)$ term in the AR polynomial equation the frequencies of pseudo-periodic intraminimum motion can be obtained. Complex roots show that the mode is underdamped. Since water has a dampening effect, overdamped motions are encountered in lower indexed modes. In Table 5.5, the number of underdamped modes among 40 modes is shown. It is to be noted that the highest number

underdamped modes, though with a small difference, is obtained at 200 K. Nevertheless, this difference is much better pronounced for TIM (see Table 5.8).

Table 5.5. The number of underdamped modes of DHFR for the first 40 PCs

Run D2S	24
Run D3S	22
Run D4S	21

Figure 5.17 is the comparison of frequencies for all runs by histograms. It is seen that, the lowest vibrational frequencies shift to lower values as the temperature increases. This shows that the frequencies of the collective vibrational motions in DHFR are lowered as temperature increases. It is also interesting to note that the shift in frequencies seen between 200 K and 300 K is higher than those between 300 K and 400 K. This may be attributed to the observation that the collective motions are not distorted, but amplified, when temperature increases from 200 K to 300 K, whereas the collective motions are distorted at 400 K. Therefore frequencies observed at 400 K, unlike those observed at 200 K and 300 K, should not be interpreted as functional collective vibrational fluctuations, but frequencies of the fluctuations on the unfolding path of the protein.

Another way of examining the frequencies is to compare the cumulative probability density functions (CDFs) of the vibrational frequencies at three different temperatures. CDFs for all runs are shown in Figure 5.19. It is observed that, CDF of the vibrational frequencies of DHFR shift to lower values as the temperature increases, as seen in Figure 5.19.

Damping factors are proportional to the maintenance of the vibrational motion of the protein in a minimum, and inversely proportional to the friction that the protein experiences. Figure 5.20 gives the boxplot representation of damping factors for the runs D2S, D3S and D4S. It is seen that the vibrational motions of DHFR at 200 K and 300 K have almost the same damping factors, which shows that the tendency of the modes to maintain their vibrational fluctuations are similar at both temperatures. The relatively higher damping factors at 400 K show that the frequencies of the fluctuations on the

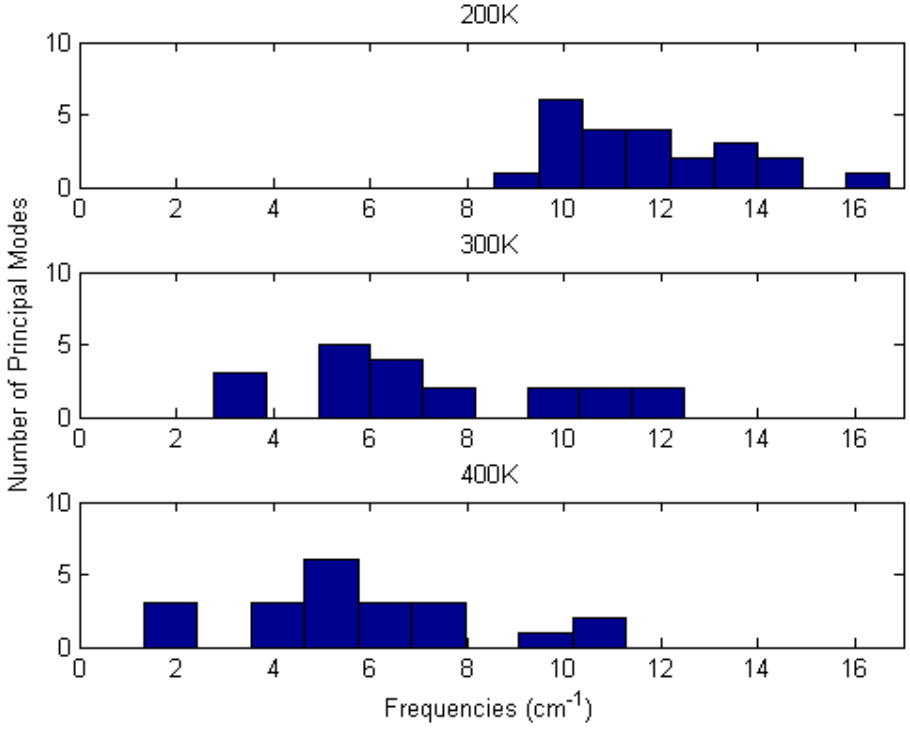


Figure 5.18. Histograms of DHFR frequencies at three different temperatures

unfolding pathway tend to be maintained for a longer period of time, which may be attributed to the higher kinetic energy of the modes.

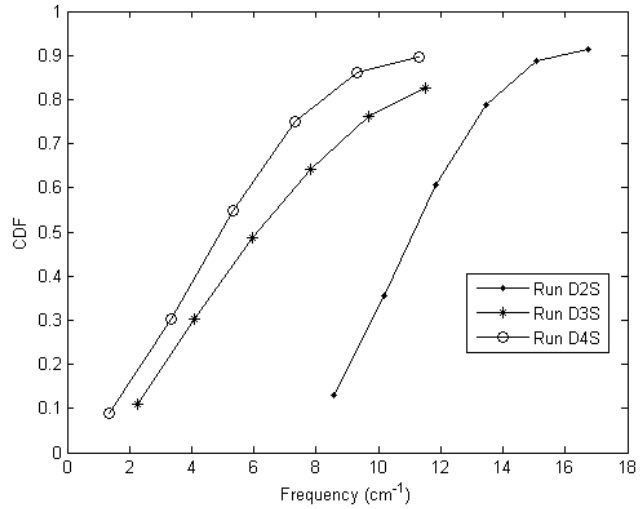


Figure 5.19. CDFs of frequencies (cm^{-1}) for all runs

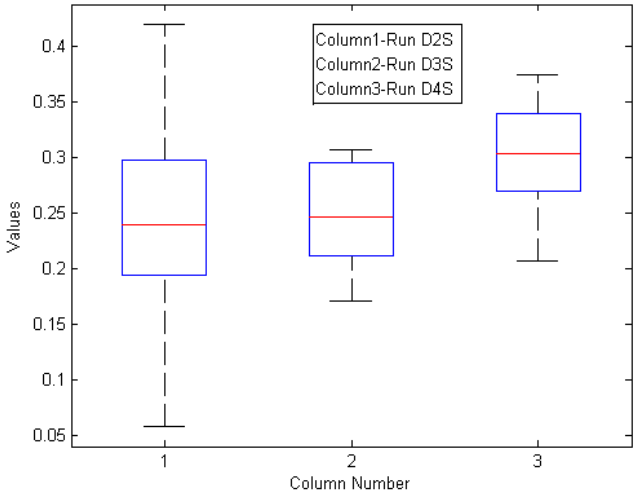


Figure 5.20. Boxplot of damping factors for the runs of DHFR at different temperatures

The frequencies of modes have been discussed so far in previous paragraphs. In order to compare the overdamped modes in all runs, the $\emptyset_1$ and $\emptyset_2$ parameters of the AR polynomial equation which are responsible from the intraminimum motions should be compared. Figure 5.21 shows the autoregressive parameters of runs D2S, D3S and D4S. As the previous analyses suggested, the AR parameters in absolute value are higher as temperature is increased, meaning that protein conformations have a higher correlation as temperature increases.

5.2. Investigation of the Dynamics of TIM

MD simulation data for TIM are obtained for both free and DHAP bound forms of TIM. The free form simulations are performed at 200 K (Run T2f) and 400 K (Run T4f), while the simulation data of the free form of TIM at 300 K (Run T3f) and bound form (Run T3b) are taken from a previous study [24]. The simulation of the bound form is performed at 200 K (Run T2b). All runs have 3.2 ns durations with fixed sampling intervals of 0.8 ps. The same methodology used for DHFR is applied to the analysis of data obtained from the MD simulations of TIM.

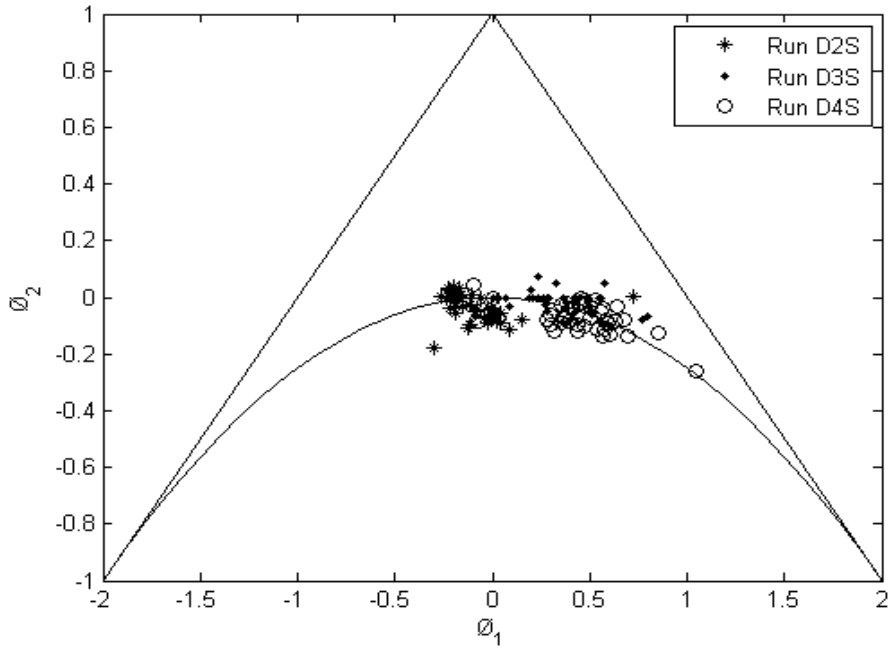


Figure 5.21. $\emptyset_1$ and $\emptyset_2$ parameters of runs D2S, D3S and D4S

5.2.1. PCA Results of TIM

5.2.1.1. Effect of Temperature on Free States

First, mobilities of the residues at three different temperatures are examined. Figure 5.22 shows the MSFs of the residues of free TIM. As the temperature increases, the flexibility increases due to a rise in the kinetic energy of TIM. The MSF of the residues in TIM shows a less uniform change at different temperatures, as compared to those of DHFR in Figure 5.2. MSF of many regions seem to be affected little from an increase of 200 to 300 K, but severely affected to a temperature increase to 400 K. It is also seen that the opening/closing motion of loop 6 (residues 166-176) is restricted at 200 K. The mobility of loop 6 increases remarkably, as temperature is increased, which is similar to the behavior of the M20 loop in DHFR.

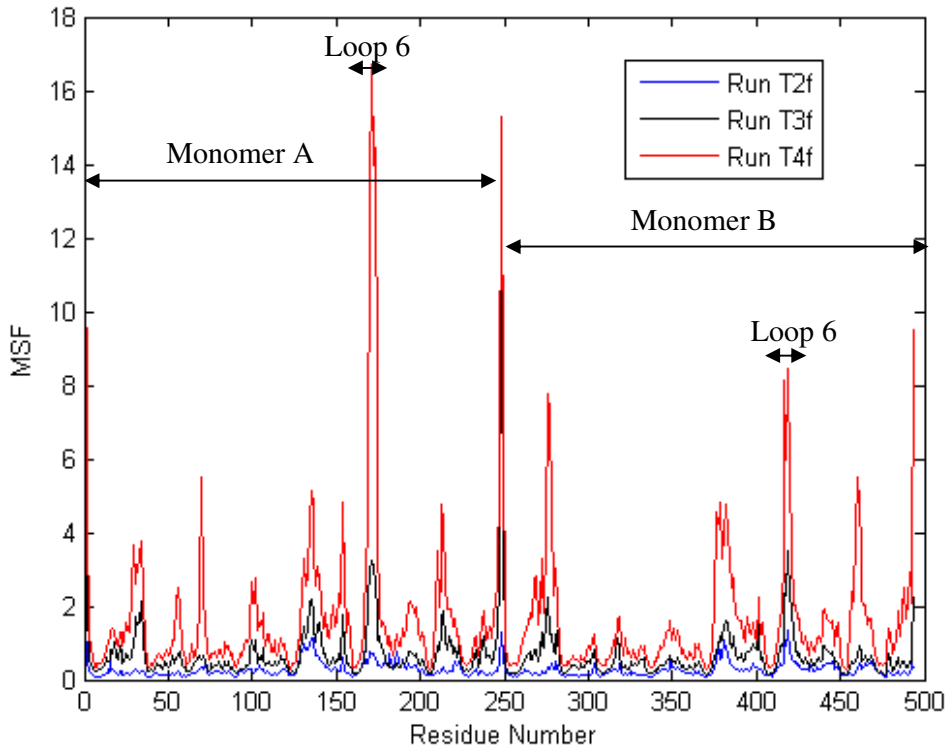


Figure 5.22. MSFs of residues along runs of free TIM at three different temperatures

PCA of the MD simulation of free forms at 200 K and 400 K are performed. Table 5.6 shows the sums of eigenvalues of all runs of TIM.

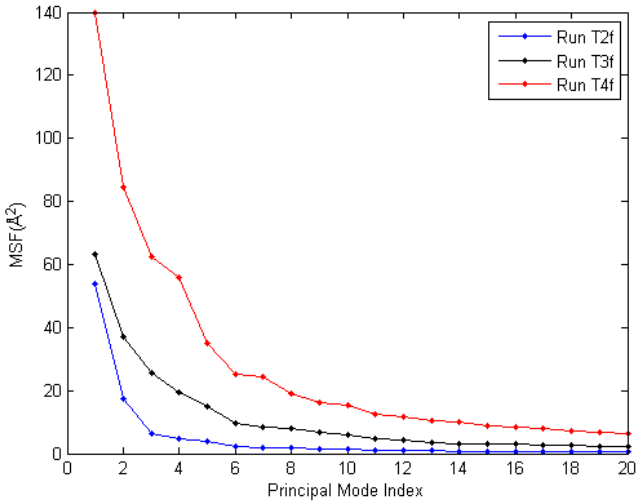
Table 5.6. Sum of the eigenvalues of the free forms of TIM at three temperatures

Free TIM	Run T2f	138.53 Å^2
	Run T3f	335.9 Å^2
	Run T4f	815.16 Å^2

As it is expected, eigenvalues increase as the temperature increases (Figure 5.23a). Figure 5.23b shows the percentage variance explanation of the modes of runs T2f, T3f and T4f. It is seen from the figure that the percentage variance explanation value is higher in the first two modes of run T2f. The eigenvalues are very close for runs T3f and T4f. If all modes are considered the collective motions at 300 K and 400 K are better pronounced compared to run at 200 K. These results are similar to those found for DHFR, to a certain extent. The lowest indexed modes at 200 K have a higher explanation power, while the rest

of the modes have a remarkable low explanation power, compared to those at 300 K. The difference from the results of DHFR is that 300 K and 400 K have very similar eigenvalue explanation power for TIM, which shows that the collectivity of PCs alone cannot account for the functionality of these motions.

a)



b)

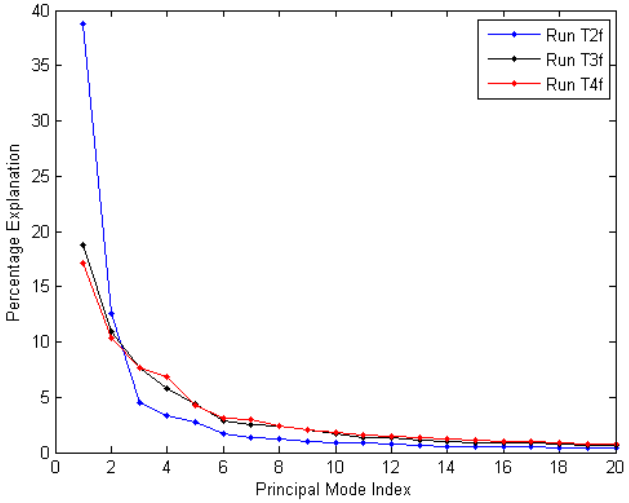


Figure 5.23. a) Eigenvalue distribution and b) percentage variance explanation of the first 20 PCs of TIM for runs T2f, T3f and T4f

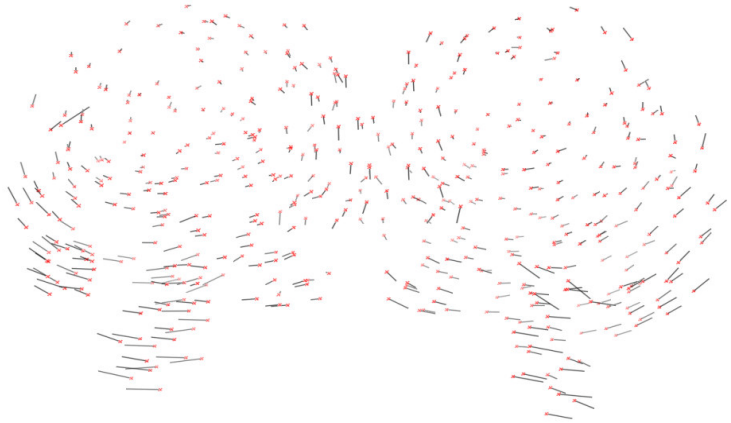
To observe the collective motions in unliganded TIM, the vector field illustrations of the protein at three different temperatures are given in Figure 5.24 in the same perspective with Figure 2.6. In this figure red dots denote the Cα atoms and black lines represent the direction of motion along PC 1. In Figure 5.24a, a global twisting motion of two monomers in opposite directions is clearly seen, however the opening/closing motion of loop 6 is reduced at 200 K. In Figure 5.24a, the global twisting motion, though a little dampened, is still observed evident, and the motion of loop 6 onto the active site is well pronounced at 300 K. In Figure 5.24c, the opening/closing motion of loop 6 is more apparent, yet the global twisting motion is distorted at 400 K. These results are very similar to those found for the motion of M20 loop and global twisting motion of DHFR at different temperatures: a fine balance in between the opening/closing of a certain loop onto the active and the coordinated motion of the whole protein is required for its function, which is achieved at the physiological temperature.

The snapshots resulting from the movement of free TIM along PC 1 at different temperatures are superimposed in Figure 5.25. It is seen that at low temperature the movement of TIM is restricted. In Figure 5.25a, the motion of loop 6 is limited as compared to the other runs. It is in the most stable state at 200 K. The collective behavior is seen at 200 K and 300 K. At 300 K, the opening/closing motion of loop 6 can be seen. In Figure 5.25c, the motion of loop 6 is explicit and TIM is more mobile than it is at 200 K and 300 K. In Figure 5.25c it is clear that loop 6 is more mobile and flexible compared to the other runs. At 400 K, as loop 6 closes over the ligand, helix in the upper part of the ligand binding site moves in the opposite direction and becomes a little bit distorted. The movement of loop 6 is reduced along PC 2 and no any significant change in the motion of other parts of the protein is seen.

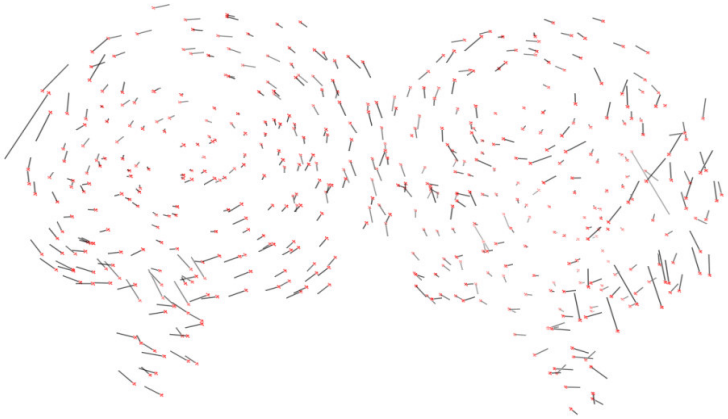
5.2.1.2. Effect of Binding on TIM at Different Temperatures

The total MSF of all residues in the ligand bound TIM at 200 K is found to be 74.00 $Å^2$, as opposed to 138.53 $Å^2$ of the unliganded form, as shown in Table 5.6, which means that the free state of TIM is more mobile than the bound state at 200 K, This is similar to what has been observed at 300 K [24], however the difference in MSF between two states at 300 K is not as high as it is at 200 K.

a)



b)



c)

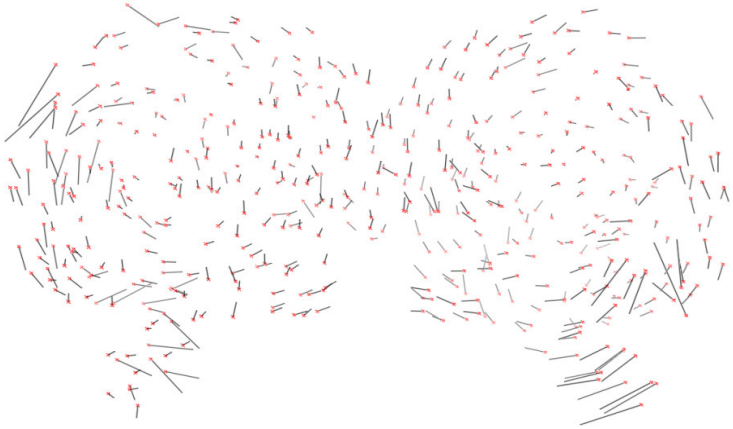


Figure 5.24. Vector field illustrations of runs a) T2f, b) T3f and c) T4f for PC 1

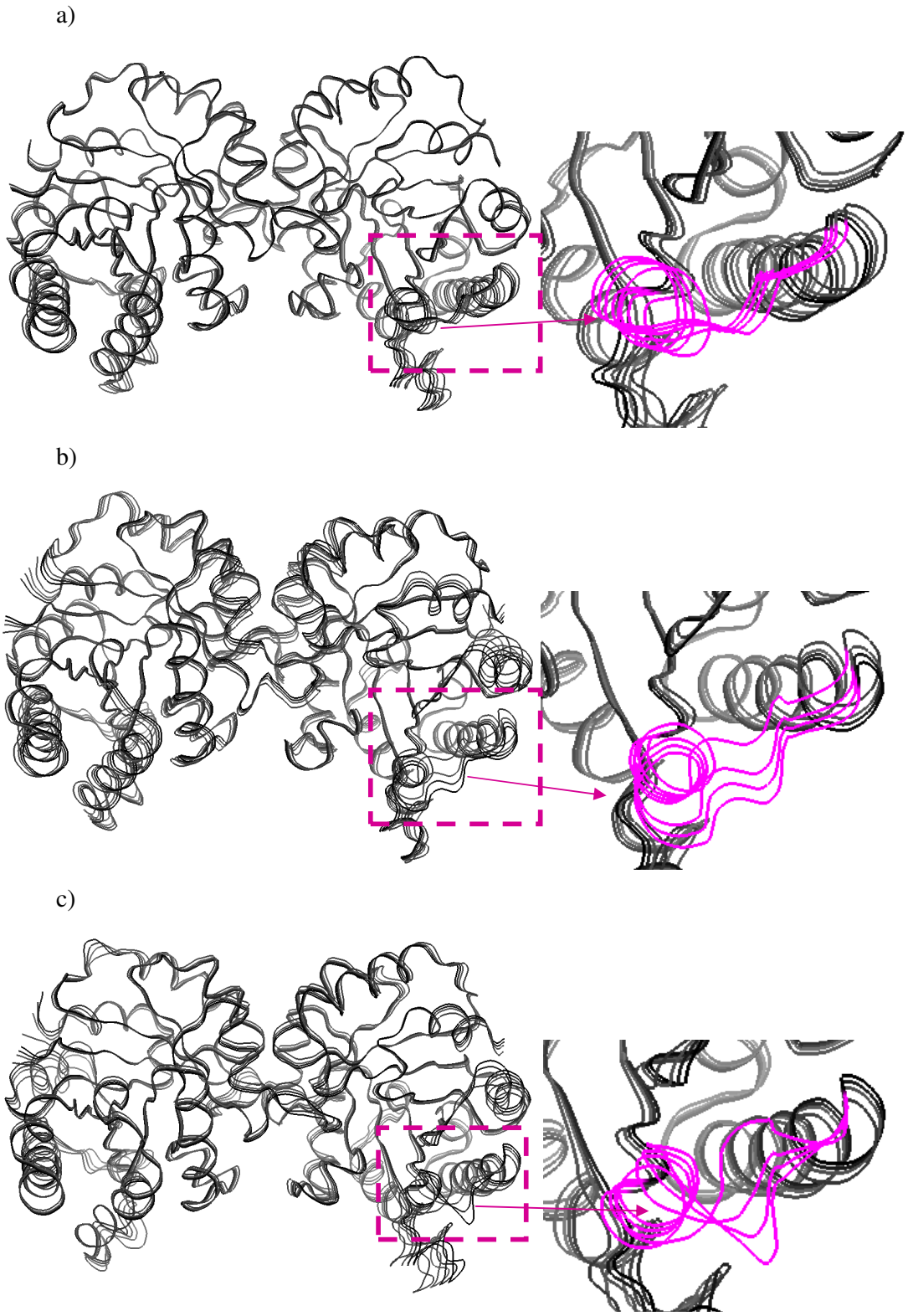


Figure 5.25. Projections of the conformations of PC 1 for runs a) T2f, b) T3f and c) T4f

The first 5, 10 and 40 PCs of Run T2f explain 62, 68 and 79 per cent of C_α fluctuations, respectively. To see the differences in the case of ligand binding, the eigenvalues and percentage explanation of the PCs of runs T2f and T2b are given in Figure 5.26a and Figure 5.26b, respectively. It is seen in Figure 5.26a that the low indexed principal modes have higher MSF in the free form. In Figure 5.26b, it is seen that the only the first two principal components of the free state of TIM yield higher explanation power as compared to the ligand bound. This is just the opposite of what has been observed at 300 K [24]. At 300 K, the first two modes have a higher percentage of variance explained for the ligand bound state. This result hints that ligand binding may have different effects on the protein dynamics at different temperatures.

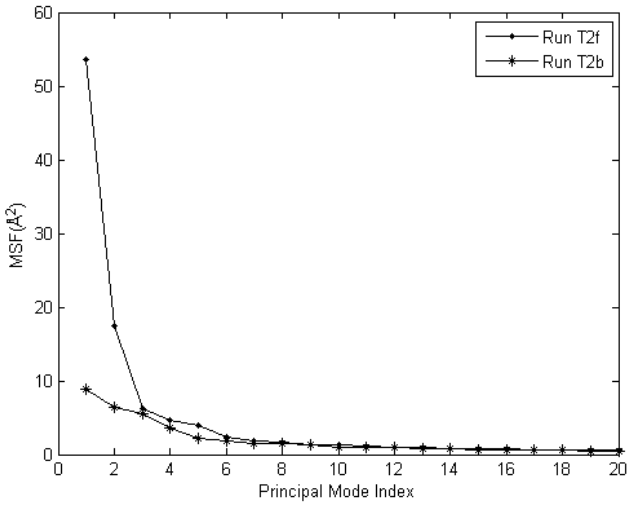
The displacement vector illustration of PC 1 and PC 2 for the free form of TIM at 200 K is shown in Figure 5.27. For the first mode, twisting motion of two monomers is clear in TIM where it is remarkably diminished for the second mode. The motion of loop 6 is less pronounced in PC 1, and cannot be seen in PC 2 at 200 K.

In Figure 5.28, the vector fields of the motion of the liganded TIM along its first two principal modes are shown. It is seen that the global twisting motion of TIM is remarkably reduced in PC 1, and loop 6 is stationary. On the other hand, the second mode shows the twisting motion very clearly, however with very low amplitude. This means that the global twisting motion is dampened as a result of ligand binding.

That the collective twisting motion of the two monomers is reduced at 200 K upon ligand binding is a different result from what was obtained for TIM at 300 K. In the previous study little change in the collective character of the protein was found between the free and bounded forms of TIM (Figure 5.29) [24].

The superimposed snapshots of the free and bound TIM along PC 1 at 200 K are presented in Figure 5.30. These superimposed conformations give idea about the ligand binding effect on the collective motions of TIM. According to the illustrations, the ligand binding decreases the collectivity along PC 1. Adding to this, loop 6 is in the closed conformation on DHAP in the liganded form where it is open in the free form of TIM at 200 K.

a)



b)

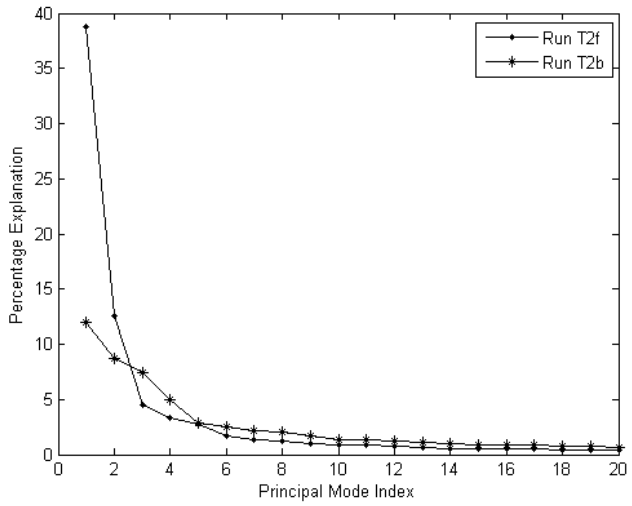
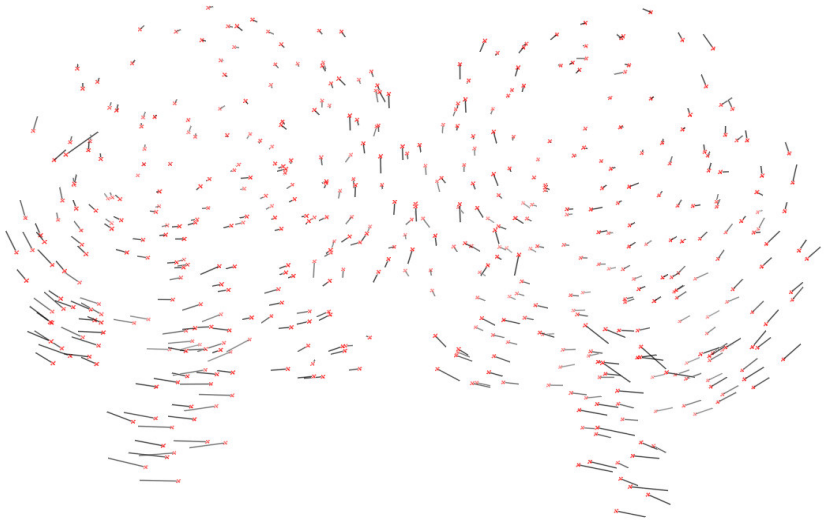


Figure 5.26. a) Eigenvalue distribution and b) percentage variance explanation of the first 20 PCs of TIM for runs T2f and T2b

It is also seen that loop 6 in the open conformation in the free form tends to move in a closed position on PC 1, while loop 6 in the closed conformation in the liganded form does not seem to make an opening motion. The opening/closing motion is not seen in the superimposed conformations of the free and ligand bound forms of TIM at 200 K along PC 2.

a)



b)

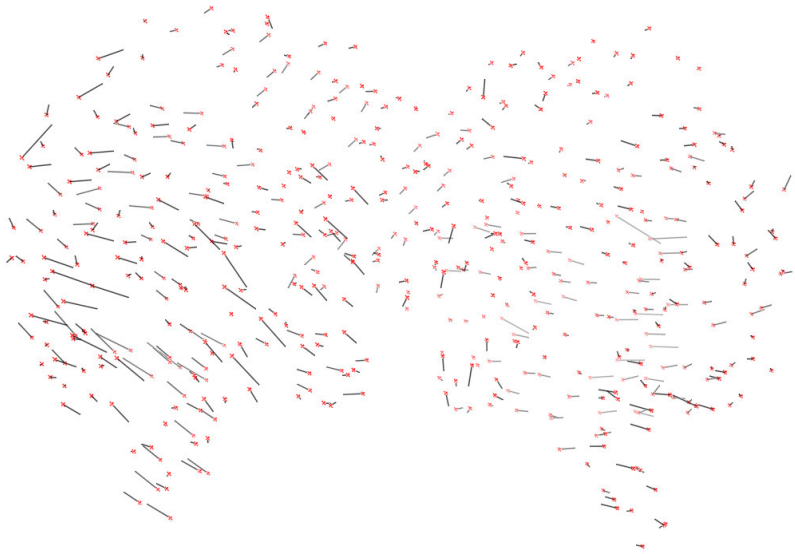


Figure 5.27. Vector field illustration of run T2f for a) PC 1 and b) PC 2 at 200 K

5.2.2. Time Series Analysis Results of TIM

5.2.2.1. Time Series Analysis Results at Different Temperatures

Models generated for the modes of TIM at different temperatures are ARIMA (4,1,2), ARIMA (4,1,1), ARIMA (3,1,1), ARIMA (3,1,2), ARMA (3,1) and ARMA (3,2).

The results of time series analysis of run T2f are shown as representatives for t_1, t_2, t_{10}, t_{20} and t_{40}:

a)

b)

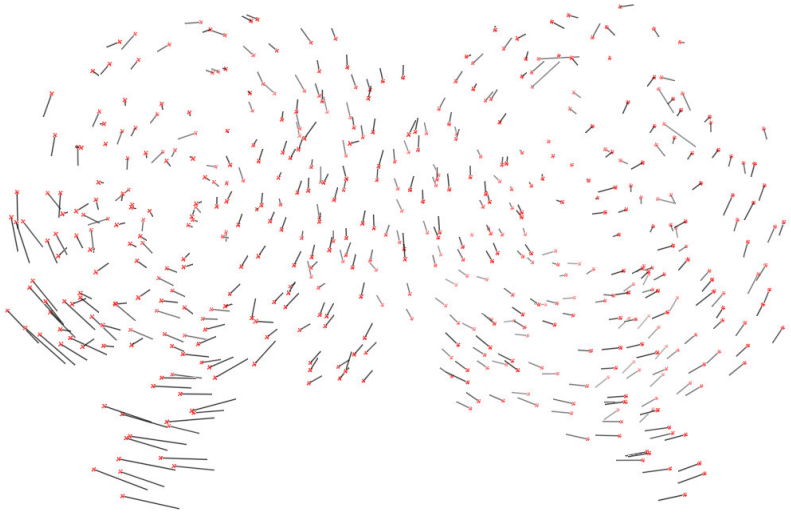


Figure 5.28. Vector field representation of run T2b for a) PC 1 and b) PC 2 at 200 K

$t_1 : (1-0.4945B-0.1321B^2)(1-0.1231B+0.0452B^2)\nabla z_t = (1-0.1139B)(1-0.7865B)a_t$, $\sigma_a^2 = 0.4256$ Å^2 (5.8)

t_2 : $(1-0.633B)(1-0.3228B+0.1187B^2)\nabla z_t = (1-0.8416B)(1-0.1124B)a_t$, $\sigma_a^2 = 0.4256$ $Å^2$ (5.9)

t_{10} : $(1-0.7643B)(1-0.2534B-0.07536B^2)\nabla z_t = (1-0.7461B)a_t$, $\sigma_a^2 = 0.2585$ $Å^2$ (5.10)

a)

b)

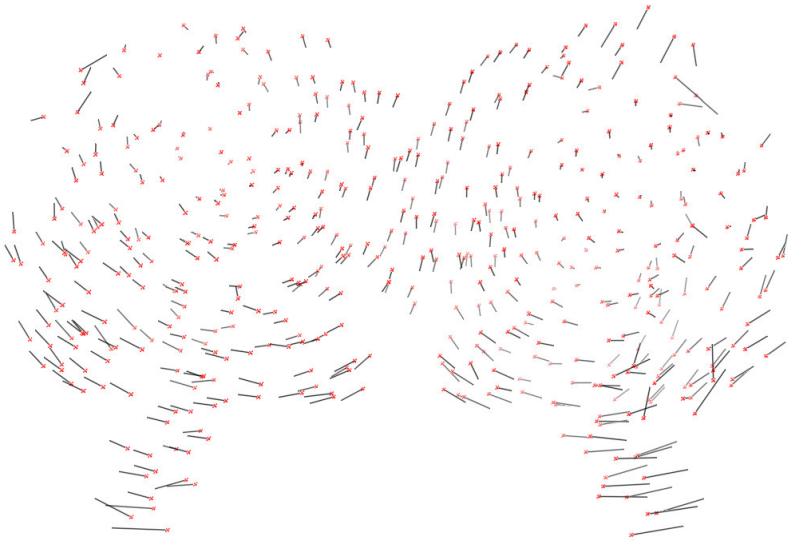


Figure 5.29. Vector field illustration of a) free and b) ligand bound TIM for PC 1 at 300 K

a)

b)

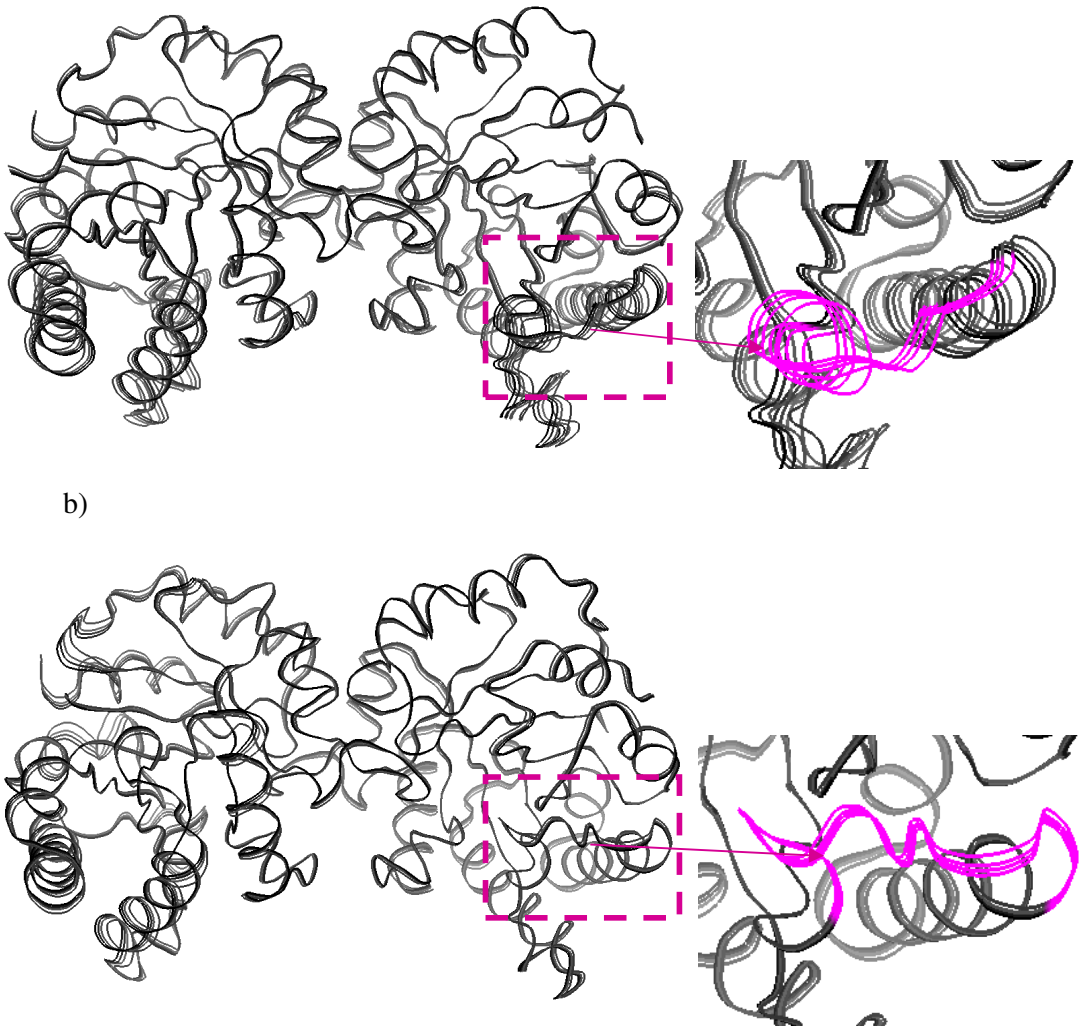


Figure 5.30. Projections of the a) free and b) bound conformations of TIM along PC 1

t_{20} : $(1\text{-}9897B)(1\text{-}0.7783B+0.1342B^2)z_t = (1\text{-}0.8457B)(1\text{-}0.603B)a_t$, $\sigma_a^2 = 0.1665$ Å^2 (5.11)

t_{40} : $(1\text{-}0.9574B)(1\text{-}0.09813B\text{-}0.09951B^2)z_t = (1\text{-}0.6326B)a_t$, $\sigma_a^2 = 0.1751$ Å^2 (5.12)

As the models of both DHFR and TIM are taken into consideration, it is seen that the changeover between the stationary and nonstationary modes is observed in higher modes in all cases. The most frequently encountered models are ARIMA (3,1,1) for nonstationary and ARMA (3,1) for stationary modes. Table 5.7 gives the number of modes for runs T2f, and T4f with respect to the model orders. It is seen that the number of nonstationary modes are smaller at 200 K, similar to a result obtained for DHFR. This result confirms the

previous finding that the number of anharmonic modes increase as temperature is increased [36].

Table 5.7. Number of modes with respect to model types and orders of TIM

Type and Order of Models	Number of Modes For Run T2f	Number of Modes For Run T4f
ARIMA (4,1,1)	1	5
ARIMA (4,1,2)	1	0
ARIMA (3,1,1)	9	11
ARIMA (3,1,2)	5	9
ARIMA (2,1,1)	0	1
ARMA (3,1)	17	13
ARMA (3,2)	7	1

Figure 5.31 gives the residual variances of TIM for runs T2f, T3f and T4f with respect to the principal modes. The variance of random shocks yields higher values for high temperature, as expected.

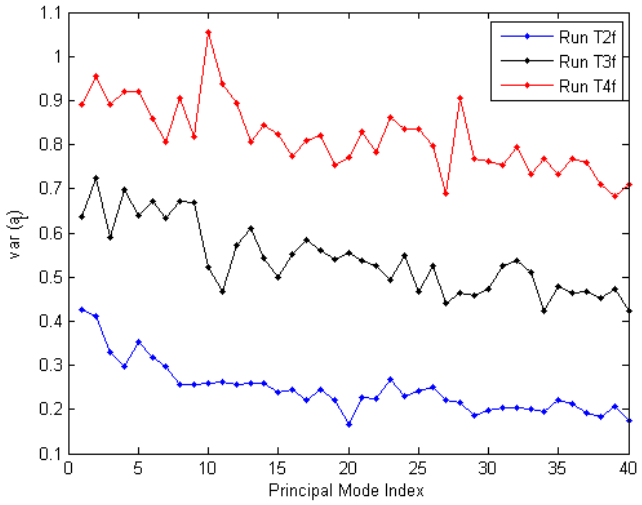


Figure 5.31. Residual variances with respect to modes of TIM at 200 K, 300 K and 400 K

Figure 5.32 gives the boxplot analysis of θ_2 parameters to see the differences between the dynamics of TIM at different temperatures. It is seen that θ_2 parameter goes from negative to positive values as temperature increases. This is completely opposite to what has been observed for DHFR in Figure 5.17.

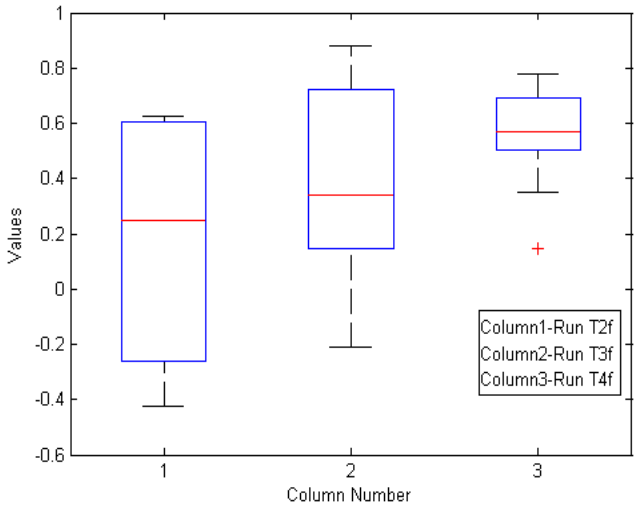


Figure 5.32. Boxplot of θ_2 parameters of TIM for runs T2f, T3f and T4f

Table 5.8 gives the number of underdamped modes of TIM. The number of underdamped modes of TIM is significantly higher than those of DHFR (Table 5.5), which shows that the dampening effect of water on the vibrational motions change from protein to protein. It is also seen in Table 5.8 that the number of underdamped modes in TIM is definitely higher at 200 K: 39 out of 40 principal modes are underdamped. This finding conforms to a previous study [37], where no overdamped modes have been found for myoglobin at a low temperature (120 K), and it has been indicated that dynamic fluctuation of the water molecules determines the friction on the vibrational frequencies.

Table 5.8. Number of underdamped modes of TIM for the first 40 modes

Free TIM	T2f	39
	T3f	27
	T4f	31

Figure 5.33 shows the histogram representation of the frequencies of TIM for runs T2f, T3f and T4f. It is seen that the vibrational frequencies shift to smaller values as temperature is increased from 200 K to 300 K, as observed for DHFR (Figure 5.18). While frequency distribution is similar at 300 K and 400 K for TIM, there are three principal modes having frequencies smaller than 2 cm^{-1} at 400 K, compared to a single mode having a smaller frequency than 2 cm^{-1}at 300 K. It should be recalled that frequencies of DHFR have shifted to slightly lower values as temperature is increased from 300 K to 400 K (Figure 5.18), thus one can say that the effect of temperature on vibrational frequencies of both proteins are similar.

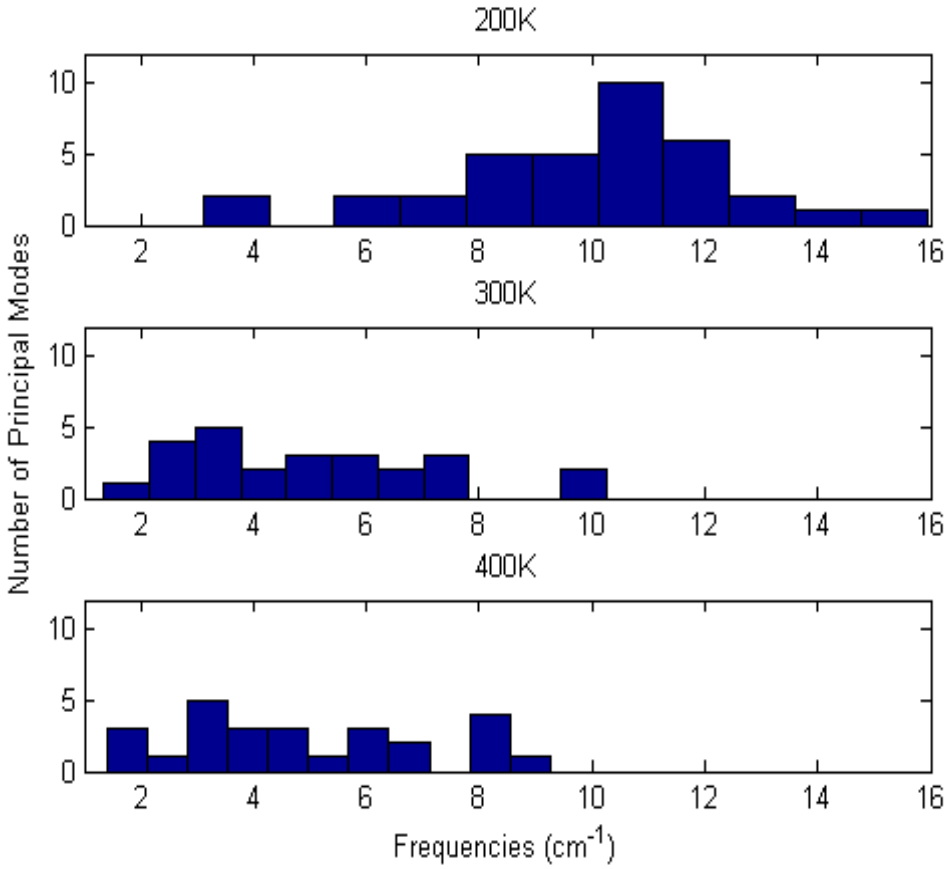


Figure 5.33. Histograms of TIM frequencies for the runs at different temperatures

Figure 5.34 is the comparison of the CDFs of frequencies at different temperatures. It can be seen that the CDF of the frequencies at 300 K and 400 K give similar values, which are considerably higher than those at 200 K.

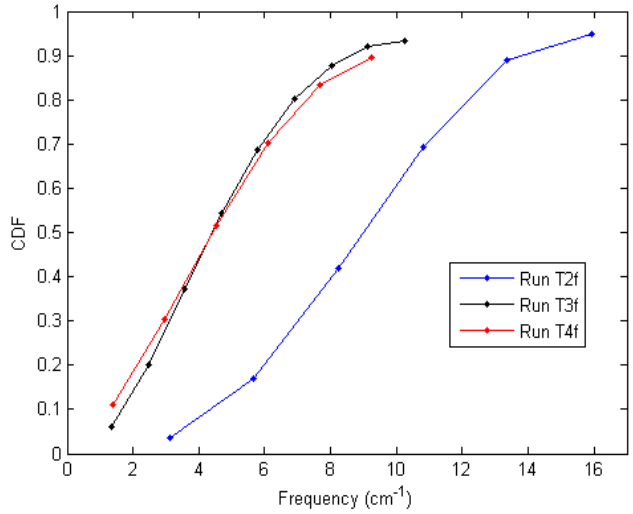


Figure 5.34. CDF comparison of TIM frequencies at 200 K, 300 K and 400 K

Figure 5.35 gives the boxplot representation of damping factors for the runs D2S, D3S and D4S. Damping factors for TIM at 200 K and 300 K are very similar, while those at 400 K seem to be slightly higher, which support the findings for DHFR (Figure 5.20).

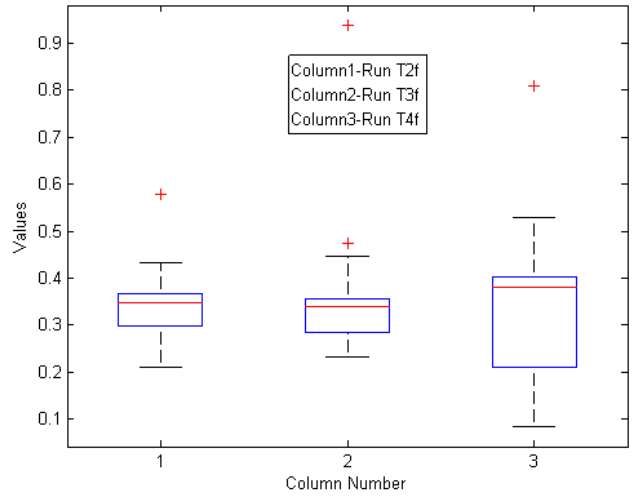


Figure 5.35. Boxplot of damping factors for the runs of free TIM at different temperatures

Figure 5.36 shows that TIM has lower frequencies than DHFR at each temperature. This is an expected result due to the size of the protein.

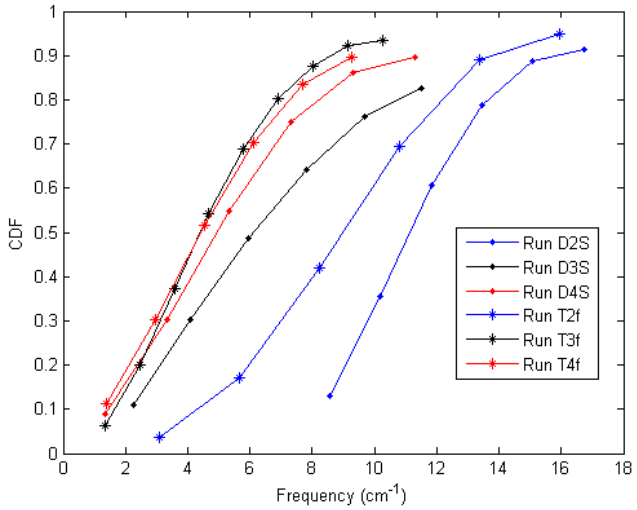


Figure 5.36. CDFs of the runs at different temperatures for DHFR and TIM

5.2.2.2. Time Series Analysis Results for the Free and Bound Forms of TIM

Table 5.9 gives the number of modes for runs T2b with respect to the model orders, where it is seen that ARIMA (3,1,1), ARMA (3,1) and ARMA (3,2) processes are the most common models.

Table 5.9. Number of modes with respect to model types and orders of ligand bound TIM at 200 K

Type and Order of Models	Number of Modes For Run T2b
ARIMA (4,1,1)	0
ARIMA (4,1,2)	0
ARIMA (3,1,1)	22
ARIMA (3,1,2)	2
ARIMA (2,1,1)	0
ARMA (3,1)	10
ARMA (3,2)	6

Figure 5.37 shows the variances of the residuals with respect to modes of TIM. The variance of random shocks (σ_a^2) for run T2f is higher for the first mode. The lowest indexed principal modes correspond to the inharmonic motions in the protein, where interminimum motions are prominent. For this reason it can be concluded that, the interminimum motions are more pronounced in the free form of TIM. It is interesting that this behavior is not observed at 300 K, which means that the change in the energy barriers due to binding can be overcome at 300 K, but not at 200 K.

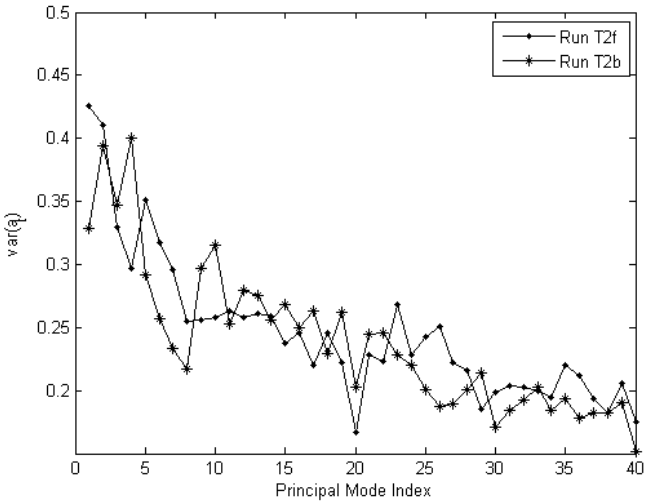


Figure 5.37. Residual variances with respect to the modes of TIM for runs T2f and T2b

As discussed before, the differences seen in the boxplot of θ_2 parameter should be interpreted with caution (Figure 5.38). It is seen that this parameter takes a wide range of values for both the free and liganded forms.

In Figure 5.39, the histograms of the low vibrational frequencies are given for the unliganded and liganded TIM. It is seen that the lowest vibrational frequencies increase when the ligand binds to the protein at 200 K.

Figure 5.40 shows the comparison of the CDFs of the vibrational frequencies of TIM for the free and bound forms. As it was obtained for TIM at 300 K [24], the CDF of the liganded form yields higher frequencies compared to that of the free form.

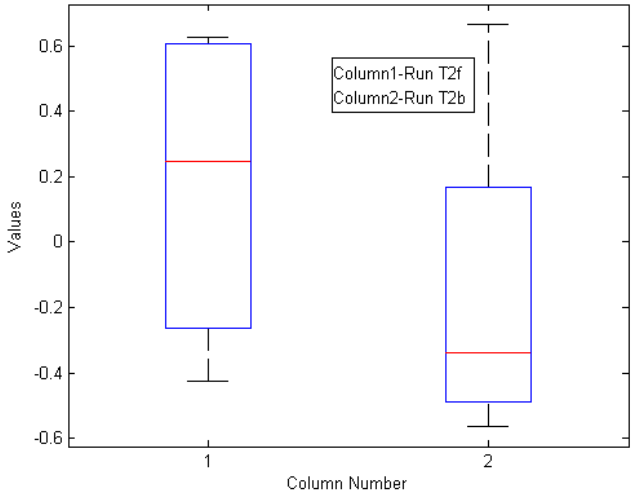


Figure 5.38. Boxplot of θ_2 parameters of TIM for runs T2f and T2b

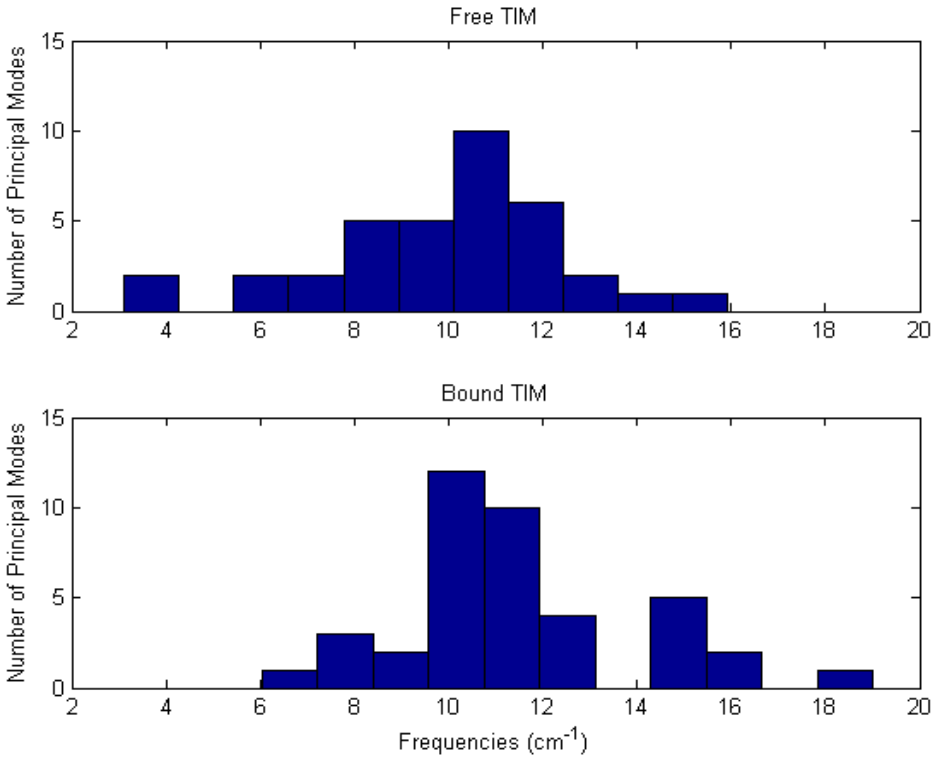


Figure 5.39. Histograms of the low vibrational frequencies for the free and liganded forms of TIM at 200 K

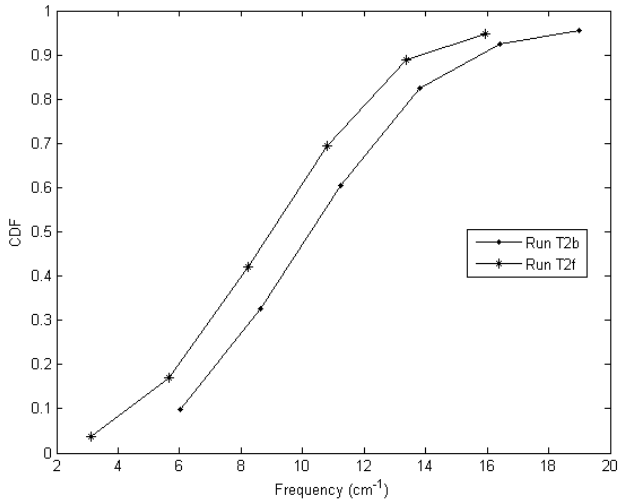


Figure 5.40. CDF comparison of TIM frequencies for the free and bound forms

Figure 5.41 shows the boxplot of damping factors of the free and ligand bound forms of TIM at 200 K. There is no significant difference between the free and liganded forms of TIM.

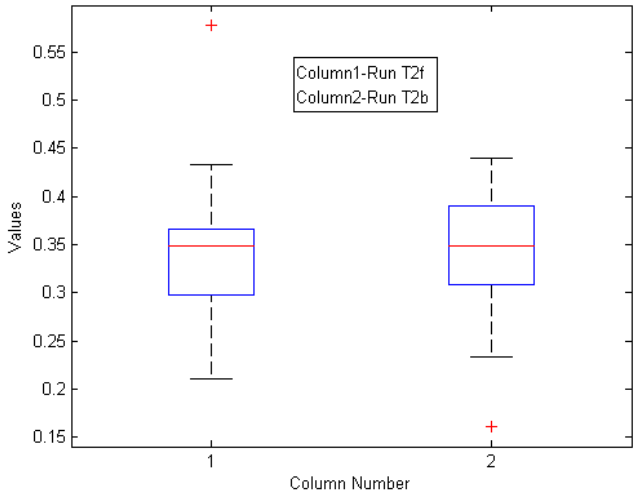


Figure 5.41. Boxplot of damping factors of the free and ligand bound TIM

Figure 5.42 is the CDF of the frequencies of TIM at 200 K with the frequencies obtained in a previous study for 300 K [24]. The effect of temperature and ligand binding on vibrational frequencies can be clearly seen: as the temperature increases the CDFs shift

to lower values both for the free and bound forms of TIM, and the free forms have lower frequencies.

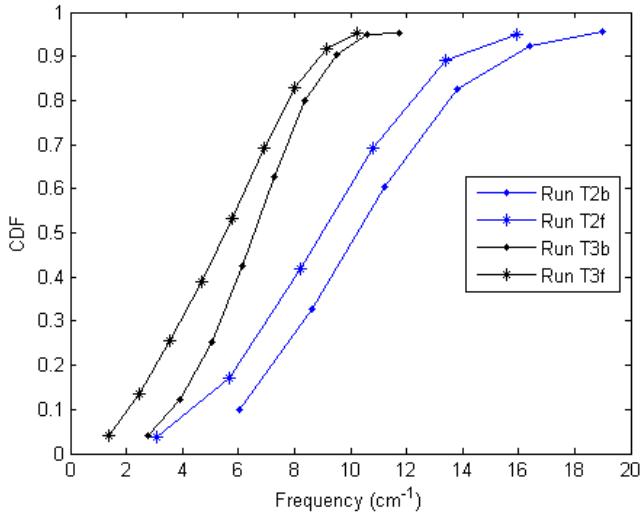


Figure 5.42. CDF comparison of free and bound TIM frequencies at 200 K and 300 K

6. CONCLUSION

For many years, researchers have sought to uncover the secrets of proteins and how they execute various functions due to their high importance both in pharmacological and medical sciences. There are experimental and computational methods to examine proteins. In this thesis, a simulation method, namely molecular dynamics (MD) simulation, is used to extract information about the dynamics of DHFR and TIM. PCA and time series analysis are used to analyze the C_α atomic trajectories obtained from MD simulation to examine the dynamics of DHFR and TIM for their free and ligand bound states at different temperatures.

The analyses of unliganded DHFR and TIM have been done at 200 K, 300 K and 400 K. Longer simulations are used for PCA analysis. In terms of the relation of the global motions of the protein with the important loop fluctuation, similar trends have been obtained for both proteins at different temperatures. At 200 K, the opening/closing motion of M20 loop in DHFR and loop 6 in TIM are reduced, while global twisting motion can be clearly observed. At 300 K, the collective twisting type of motions of both proteins are well pronounced, and the opening/closing type of motion of M20 loop in DHFR and loop 6 in TIM can be observed. At 400 K these two loops are highly mobile; however the global twisting motion in both proteins is nearly lost. It is seen that helix C in DHFR gets distorted at this high temperature, which shows that DHFR is in an unfolding pathway. It may be concluded that the global and the important loop motions in proteins both contribute to the function as a result of their coordinated motion, which only occurs at the physiological temperature. Loops cannot overcome the energy barriers at low temperatures, while high kinetic energy distorts the concerted motions at high temperatures.

In addition to the analysis of collective motions seen in the lowest indexed principal modes, comparison of the percentage variance explained by eigenvalues at 200 K and 300 K for both proteins show an interesting result. The percentage explanation of the first (or the first two) eigenvalues at 200 K are higher than that at 300 K, while eigenvalues with indices in between 3 to 10 are higher at 300 K (Figures 5.4b and 5.22b). One may take the

percentage explanation of eigenvalues as a measure of collectivity. Considering the similarity of collective motions in the first PC at both temperatures, it is possible to say that the global twisting motion at 200 K contributes higher to the protein motions than the same contribution at 300 K. Furthermore, the contribution of intermediate modes to the protein motions at 200 K is smaller than those at 300 K. It can be argued that these results show the importance of the collectivity of the intermediate modes: the flexibilities of the principal modes up to 10 or even more are important in protein function.

The time series models of each principal component give information about the collective motions in each case. Changes in the ligand bound and free forms of DHFR and TIM at different temperatures are examined by means of vibrational frequency distribution. Since linear stochastic time series modeling takes the anharmonicity into account, it is a more reliable method than quasiharmonic analysis. For instance, time series models consider the nonstationarity of the protein trajectory, which is a consequence of anharmonicity in protein motions. For both proteins, the number of nonstationary modes is found to be smaller at 200 K, showing that the number of anharmonic modes is reduced at low temperatures. The analyses of unliganded DHFR and TIM at 200 K and 300 K show that vibrational frequencies clearly shift to lower values as the temperature increases. This shows the superiority of stochastic time series models to Langevin models, which are insufficient to describe the diffusional motion of the modes, and thus the shifting of the especially lowest frequencies with respect to temperature are not clearly seen [37]. It is also seen that damping factors at these two temperatures are similar. The lowering of frequencies when temperature is increased from 300 K to 400 K, on the other hand, is not well pronounced for DHFR and even negligible for TIM. This result should be interpreted with caution, because unlike the fluctuations at 200 K and 300 K, the fluctuations at 400 K are on the unfolding pathway of the protein, and the collective character of the motion is considerably lost. These results, nevertheless, show that protein continues to make its vibrational motion with slightly lower frequencies and slightly higher damping factors compared to physiological temperature.

It is known that ligand binding affects the flexibility of the protein and binding free energy. For this reason, when the binding affinity of ligands is in case the flexibility of protein should be taken into consideration. There is a dispute on ligand binding effect. In

some studies scientists observed that ligand binding reduces flexibility of the protein [26,36] where others observed the vice versa [36-38]. In this thesis, ligand binding is found to increase the flexibility of TIM at 200 K. In a previous study, ligand binding effect to TIM at 300 K has been examined [24]. It has been found that the collective twisting motion of the two monomers is not changed, while low vibrational frequencies have been shifted to higher values. In this study, ligand binding effect on the collective dynamics of TIM at 200 K is examined. Contrary to what has been observed at 300 K, it is observed that ligand binding considerably dampens this twisting motion of the monomers and MSF of the protein is reduced. The variances of random shocks of the first principal mode in both states are different, which shows that the anharmonic motions in the liganded state are dampened compared to the unliganded state. On the other hand, it should be recalled that simulation at 200 K is 3.2 ns length, considerably shorter than the simulations at 300 K. This brings the necessity to check the reliability of this result by extending the MD simulation time. It is found that, similar to what has been observed at 300 K, vibrational frequencies of the liganded TIM shift to higher frequencies at 200 K.

As future studies, MD simulations starting at temperatures lower than 200 K to 300 K with smaller temperature intervals can be performed to extract the relation between the frequencies and damping factors to temperature more clearly. It should be clarified whether damping factors of the modes will be increased below a certain temperature, where interminimum motions are reduced. A detailed examination on the collective motions of both enzymes can be done by specifically focusing on the loop motions. The finding that ligand binding has different effects on the global motions of TIM at 200 K and 300 K can be checked by elongating the simulation length of MD at 200 K.

APPENDIX A: DETAILS OF RUNS

Run D2L: DHFR, free, starting X-ray coordinate 1RA1, at 200 K

D2L: 8 ns.

Run D3L: DHFR, free, starting X-ray coordinate 1RA1, at 300 K

D3L: 3 – 35 ns.

Run D4L: DHFR, free, starting X-ray coordinate 1RA1, at 400 K

D4L: 8 ns.

Run D2S: DHFR, free, starting X-ray coordinate 1RA1, at 200 K

D2S: 3.2 ns.

Run D3S: DHFR, free, starting X-ray coordinate 1RA1, at 300 K

D3S: 4 – 7.2 ns. (Taken from a previous study [24] – Run A1-1)

Run D4S: DHFR, free, starting X-ray coordinate 1RA1, at 400 K

D4S: 3.2 ns.

Run T2f: TIM, free, starting X-ray coordinate 8TIM, at 200 K

T2f: 1.8 – 5 ns.

Run T2b: TIM, bound, starting X-ray coordinate 1TPH, at 200 K

T2b: 1.8 – 5 ns.

Run T3f: TIM, free, starting X-ray coordinate 8TIM, at 300 K

T3f: 5 – 8.2 ns. (Taken from a previous study [24] – Run B1-1)

Run T3b: TIM, bound, starting X-ray coordinate 1TPH, at 300 K

T3b: 5 – 8.2 ns. (Taken from a previous study [24] – Run B2-1)

Run T4f: TIM, free, starting X-ray coordinate 8TIM, at 400 K

T4f: 1.8 – 5 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. Chandra, S. V., L. S. D. Caves, R. E. Hubbard and G. C. K. Roberts, "Domain Motions in Dihydrofolate Reductase: A Molecular Dynamics Study", *J. of Molecular Biology*, Vol. 266, pp. 776-796, 1997.

16. Kürkçüoğlu, Ö., R. L. Jernigan and P. Doruker, "Loop Motions of Triosephosphate Isomerase Observed with Elastic Networks", *Biochemistry*, Vol. 45, pp. 1173-1182, 2006.

17. Cansu, S., Collective Dynamics and Conformational Sampling of Triosephosphate Isomerase, M.S. Thesis, Boğaziçi University, 2007.

18. 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.

19. Branden, C. and J. Tooze, *Introduction to Protein Structure*, Garland Publications, 1999.

20. Henzler-Wildman, K. and D. Kern, "Dynamic Personalities of Proteins", *Nature*, Vol. 450, pp. 964-972, 2007.

21. Phillips, T., *Proteins*, http://biotech.about.com/od/technicaltheory/g/Proteins.htm, 2008.

22. Available: http://en.wikipedia.org/wiki/Image:Activation2_updated.svg.

23. 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.

24. Başkan, S., Effect of Ligand Binding on Protein Dynamics: A Time Series Analysis, M.S. Thesis, Boğaziçi University, 2008.

25. Available: http://www.biotopics.co.uk/other/enzyme.html.

26. 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.

27. McElheny, D., J. R. Schnell, J. C. Lansing, H. J. Dyson, and P. E. Wright, "Defining the Role of Active-Site Loop Fluctuations in Dihydrofolate Reductase Catalysis", *PNAS*, Vol. 102, pp. 5032-5037, 2005.

28. Available: http://research.chem.psu.edu/sjbgroup/projects/dihydro.htm.

29. Available: http://en.wikipedia.org/wiki/Triosephosphate_isomerase.

30. Available: http://chemistry.umeche.maine.edu/CHY431/Conformation5b.html.

31. Karplus, M. and J. A. McCammon, "Molecular Dynamics Simulations of Biomolecules", *Natural Structural Biology*, Vol. 9, pp. 646-788, 2002.

32. 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.

33. Kerrigan, J. E., *AMBER 8.0 Introductory Tutorial*, 2004.

34. 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.

35. Brockwell, P. J. and R. A. Davis, *Introduction to Time Series and Forecasting*, Springer, 2002.

36. Tournier, A. L. and J. C. Smith, "Principal Components of Protein Dynamical Transition", *Physical Review Letters*, Vol. 91, No. 20, 208106, 2003.

37. Moritsugu, K. and J. C. Smith, "Langevin Model of the Temperature and Hydration Dependence of Protein Vibrational Dynamics", *Journal of Physical Chemistry B*, Vol. 109, 12182-12194, 2005.

38. 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.

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

40. 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.