ABSTRACT

A MOLECULAR DYNAMICS SIMULATION BASED PRINCIPAL COMPONENT ANALYSIS FRAMEWORK FOR COMPUTATION OF MULTI-SCALE MODELING OF PROTEIN AND ITS INTERACTION WITH SOLVENT

by
Tao Wu

This dissertation presents a new computational framework for calculating the normal modes and interactions of proteins, macromolecular assemblies and surrounding solvents. The framework employs a combination of molecular dynamics simulation (MD) and principal component analysis (PCA). It enables the capture and visualization of the molecules' normal modes and interactions over time scales that are computationally challenging. It also provides a starting point for experimental and further computational studies of protein conformational changes.

A protein's function is sometimes linked to its conformational flexibility. Normal mode analysis (NMA) and various extensions of it have provided insights into the conformational fluctuations associated with individual protein structures. In traditional NMA, each protein requires a customized model for such analysis due to its mechanical complexity. The methodology presented here is applicable to any protein with known atomic coordinates. Because of its computational efficiency and scalability, it facilitates the study of slow protein conformational changes (on the order of milliseconds), such as protein folding.

PCA reduces the dimensionality of MD atomic trajectory data and provides a concise way to visualize, analyze, and compare the motions observed over the course of a simulation. PCA involves diagonalization of the positional covariance matrix and identification of an orthogonal set of eigenvectors or "modes" describing the direction of maximum variation in the observed conformational distribution. Consequently, slow conformational changes can be identified by projecting these dominant modes back to original trajectory data.

In this work, the new multiscale methodology was first applied to a relatively small mutant T4 phage lysozyme, establishing its equilibrium atomic thermal fluctuations and its inter-residue fluctuation correlations. These results were compared with published data obtained by NMA, by finite element methods, and by experiment. The eigenmodes captured are in quantitative agreement with previously published results. With this success on a small protein, the method was applied to the interaction of mutated hemoglobin molecules that cause sickle cell anemia and the atomic level details of which are unknown. The new methodology reveals slow motion processes of the hemoglobin-hemoglobin interaction.

MD based PCA is computationally expensive. Thus, this dissertation work also includes a widely-applicable parallel programming implementation of the modeling framework to improve its performance.

A MOLECULAR DYNAMICS SIMULATION BASED PRINCIPAL COMPONENT ANALYSIS FRAMEWORK FOR COMPUTATION OF MULTI-SCALE MODELING OF PROTEIN AND ITS INTERACTION WITH SOLVENT

by
Tao Wu

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

January 2011

APPROVAL PAGE

A MOLECULAR DYNAMICS SIMULATION BASED PRINCIPAL COMPONENT ANALYSIS FRAMEWORK FOR COMPUTATION OF MULTI-SCALE MODELING OF PROTEIN AND ITS INTERACTION WITH SOLVENT

Tao Wu

Dr. Usman W. Roshan, Dissertation Co-Advisor Date
Associate Professor of Computer Science, NJIT

Dr. Xiaodong (Sheldon) Wang, Dissertation Co-Advisor Date
Professor and Chair of McCoy School of Engineering
Midwestern State University

Dr. Barry Cohen, Dissertation Co-Advisor Date
Associate Dean of the College of Computing Sciences, NJIT

Dr. Carol A. Venanzi, Committee Member Date
Distinguished Professor of Chemistry, NJIT

Dr. Narain Gehani, Committee Member Date
Professor of Computer Science, Dean of the College of Computing Sciences, NJIT

Dr. Hongya Ge, Committee Member Date
Associate Professor of Electrical and Computer Engineering, NJIT

BIOGRAPHICAL SKETCH

Author:	Tao Wu
Degree:	Doctor of Philosophy
Date:	January, 2011

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, New Jersey, 2011
- Master of Science in Computer Science,
  University of Detroit Mercy, Detroit, Michigan, 2004
- Bachelor of Engineering in Construction,
  University of Xi'an Architecture Technology, Xi'an, China, 1996

Major: Computer Science

Publications and Presentations:

Tao Wu, X. Sheldon Wang, Barry Cohen, and Hongya Ge, *Molecular Modeling of Normal and Sickle Hemoglobins*, Journal for Multiscale Computational Engineering, 8(2), 237–244 (2010).

Tao Wu, Ye Yang, X. Sheldon Wang, Barry Cohen, and Hongya Ge. *A Hierarchical Coarse Grain Modeling of Hemoglobin*, The Second International Symposium on Computational Mechanics, 2–18, December 2009.

Tao Wu, X. Sheldon Wang, Hongya Ge, and Barry Cohen, *Multi-scale and multi-physics modeling of sickle-cell disease part I molecular dynamics simulation*, International Mechanical Engineering Congress & Exposition, 66418, November 2008.

To My Parents and Wife Wen

ACKNOWLEDGMENT

I am greatly thankful to my advisors, Dr. Barry Cohen, Dr. Xiaodong (Sheldon) Wang and Dr. Usman W. Roshan, whose encouragement, supervision and support from the preliminary to the concluding level enabled me to complete my research and write the dissertation. Secondly, I offer my regards and blessings to my committee members, Dr. Carol A. Venanzi, Dr. Narain Gehani and Dr. Hongya Ge, who supported me in many respects during the completion of the project. I also would like to thank Dr. Ye Yang who worked with me and offered a lot of help.

This research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

Lastly, it is a pleasure to thank those who made this dissertation possible and most importantly my wife Wen, who gave me the moral support I required.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

Computational and mathematical models help biologists and medical researchers to understand the causes of disease, molecular level processes, and much more. These models uncover a multitude of biological facts, such as genome sequences and protein properties. Experimental and computational methods are integrated into biological research to better understand complex biological systems. Computational biology provides powerful methods to address critical scientific questions outside a laboratory.

Understanding a protein's mechanics is a prerequisite for insight into its biological function, since most proteins perform their functions through structural deformation, also called conformational change. Such conformational change has been modeled by atomic level simulation such as molecular dynamics (MD) simulation.

It is difficult to theoretically explain and predict the conformational changes in protein binding or interaction, though there are standard methods to experimentally probe this phenomenon. These methods include fluorescence resonance energy transfer, X-ray crystallography (Rossmann *et al.* 2005), cryoelectron microscopy (cryo-EM) (Saibil 2000), NMR (Ishima and Torchia 2000), etc. Conformational change is not accessible to MD simulation available today due to its computational cost. It happens on millisecond scales (Elber 2005; Schlick *et al.* 1997) compared to MD simulation on nanosecond scale. Therefore other computational methods must be applied. One of these methods is normal mode analysis (NMA), which has proven useful for studying collective motion of biological macromolecules. In addition, NMA with coarse-grained modeling of protein structures of large proteins has been a computational alternative to atomic level simulation. Studies (Marques and Sanejouand 1995; Perahia and Mouawad 1995; Tama *et al.* 2000) show that some of the lowest-frequency normal modes of several proteins, including hemoglobin, are strongly correlated with their large amplitude conformational changes.

Proteins and other biological macromolecules are large and chemically inhomogeneous. Their dynamic behaviors or conformational changes are in timescales from femtoseconds (high-frequency local bond vibration) to milliseconds (protein folding) (Harris and Laughton 2007). This complex system is difficult to study theoretically. Although NMA provides information about slow protein motion (Marques and Sanejouand 1995; Perahia and Mouawad 1995), its harmonic approximation limits its use to small amplitude motion in the potential energy surface associated with a local minimum (Barton *et al.* 2002). It is difficult, using NMA, to analyze such large scale effects as the hydrophobic effect (Ma 2005), which plays an important role in protein-protein interaction. A combination of MD simulation and principal component analysis (PCA) provides a systematic way to study protein slow motion. This method overcomes some of the difficulties of NMA.

Protein-protein interaction such as macromolecular docking has been studied extensively. The ultimate goal is the prediction of the three dimensional structure of the macromolecular complex as it would occur in a living organism. If the bond angles, bond lengths and torsion angles of the components are not modified at any stage of formation of the complex, it is known as rigid body docking. In general, hemoglobin-hemoglobin interaction caused by hydrophobic interaction could be treated as rigid body docking. Docking processes that include conformational change, or flexible docking, are much more complicated. Flexible docking could, in theory, be studied by quantum mechanics and other methods. Slow motion analysis is suitable for some rigid body docking.

Sickle cell anemia is the first disease whose cause was pinpointed at a genetic level. The switch of a single DNA base pair in the hemoglobin gene from A to T changes an amino acid in hemoglobin from glutamic acid to valine (Herrick 1910). Red blood cells (RBCs) are composed largely of hemoglobin. In normal RBCs, hemoglobin is globular. Normal hemoglobin tends to form a protective layer of surrounding water molecules. This protective water coating keeps hemoglobin molecules separated from each other. RBCs have a flexible membrane and can easily squeeze through capillary vessels. In the mutated hemoglobin molecule, one normally hydrophilic spot becomes slightly hydrophobic and, in deoxygenated states, tends to lose its protective layer of water molecules. Consequently, the hemoglobin molecules tend to stick together and form a chain of hemoglobin beads. In a cascading event, such chains form bundles, eventually altering the red blood cell membrane from flexible to stiff. As a result, the affected RBCs switch from their normal dumbbell shape to a sickled shapes. Stiff sickle cells tend to block capillary vessels and directly contribute to sickle cell anemia.

Hydrophobic interaction is the main cause for sickle hemoglobin (hemoglobin S) sticking to itself. Interaction of water with hydrophobic objects plays a major role in molecular self-assembly processes (Israelachvili and Wennerstrom 1996; Lum *et al.* 1999; Tanford 1980), as well in the process of aggregation of hemoglobin S. In this study, slow motion analysis was performed to study the process of hemoglobin-hemoglobin interaction.

This dissertation presents a new computational framework for calculating the normal modes and interactions of proteins, macromolecular assemblies and surrounding solvents. The framework employs a combination of MD simulation and PCA. It enables the capture and visualization of the molecules' normal modes and interactions over time scales that are computationally challenging, providing a starting point for experimental and further computational studies of protein conformational changes. The details of this framework are introduced in Chapter 2.

The new multiscale methodology was first applied to a relatively small mutant T4 phage lysozyme. The methods and results are presented in Chapter 3. These results are compared with published data obtained by NMA and finite element methods. The eigenmodes captured are in quantitative agreement with previously published results.

With this success on a small protein, the method was applied to the interaction of mutated hemoglobin molecules that cause sickle cell anemia, which is described in Chapter 4. The results suggest an explanation of hemoglobin aggregation at the atomic level.

MD based PCA is computationally expensive. This work includes a widely applicable parallel programming implementation of the modeling framework to improve its performance. These results are introduced in Chapter 5.

1.1 Classical Molecular Dynamics

In classical molecular dynamics (MD), atomic trajectories are computed by solving Newton's equation of motion Eq(1.1), which is a system of second order ordinary differential equation (ODE),

$$M\frac{d^2}{dt^2}q(t) = -\nabla U(q(t)), \tag{1.1}$$

where q is the position vector, M is the diagonal mass matrix, $U(q)$ is the potential energy, and $-\nabla U(q)$ is the force.

In practice, the following system of first order ODEs are solved,

$$\dot{q} = M^{-1}p, \quad \dot{p} = -\nabla U(q), \tag{1.2}$$

where p is the momentum vector.

The potential energy, U, is typically given by

$$U = U^{bo} + U^{nbo}, \tag{1.3}$$

$$U^{bo} = U^b + U^a + U^d + U^i, \tag{1.4}$$

$$U^{nbo} = U^{LJ} + U^e. \tag{1.5}$$

Note that U^{bo} terms are bonded potential, U^{nbo} terms are unbonded potential, U^b and U^a terms are harmonic (linear and angular spring) potentials that model covalent bond interaction, U^d terms are dihedral potential, U^i terms are improper potential, the U^e term is the coulomb potential, and U^{LJ} term models a van der Waals attraction and hard core repulsion. The coefficients in these potentials are determined experimentally, often aided by theoretical approaches such as *abinitio* quantum mechanical calculation.

It is unrealistic to expect that accurate trajectories can be computed for long time intervals in MD simulation because MD trajectories are chaotic. One can only expect that the trajectories will have the correct statistical properties, which can be verified by using the initial velocities that are randomly generated from a Maxwell distribution.

The computational difficulty of molecular dynamics simulation resides in the computation of the force f_i that is a gradient of an anharmonic potential field prescribed to all atoms. Further, the time step for integrating the equation of motion is typically on the order of femtoseconds, whereas protein function occurs at much larger time scale, from nanoseconds to seconds. The accessible time scale for molecular dynamics commonly in the order of nanoseconds (Elber 2005; Schlick *et al.* 1997). However, Shaw *et al.* (2009) reported that millisecond-scale MD simulation have been performed on Anton, which is a recently completed special-purpose supercomputer designed for MD simulation of biomolecular systems. Even with this huge improvement, MD simulation may be computationally inhibited for large protein mechanics, where large spatial and temporal scales are required.

1.2 Coarse-grained Structural Model of Protein Molecules

Coarse-grained models have been widely used for large scale protein analysis, which enormously reduces the degrees of freedom. Since the dominant motion of a protein structure is represented by a carbon backbone chain (Amadei *et al.* 1993), in coarse-grained models protein structure is represented by α-carbon atoms for the protein backbone chain. Moreover, the high computational costs usually arise from the complicated anharmonic potential field. Consequently, the simplification of such a potential field for a protein structure described by α-carbon atoms is the key issue for the coarse-grained modeling of proteins.

1.2.1 Normal Mode Analysis

Normal mode analysis (NMA) is a computational alternative to atomic simulation such as MD for understanding large protein mechanics (Brooks and Karplus 1985; Gibrat and Gõ 1990; Harrison 1984; Tama *et al.* 2000). The principle of NMA is similar to that typically employed for structural mechanics. Once the stiffness matrix (Hessian matrix) for a structure is constructed, the modeling analysis provides the vibration information of such a structure. The stiffness matrix for a protein structure is usually established based on the computation of second gradients of anharmonic potential field prescribed to all atoms. In general, calculation of a stiffness matrix is implemented at equilibrium position, which is obtained by minimization of anharmonic potential. This implies that, for large proteins, the computation of a stiffness matrix along the minimization process is a computationally expensive process.

NMA is also referred to as quasi-harmonic analysis (Teeter and Case 1990), since the modal analysis is implemented with harmonic approximation to potential energy V for small displacement,

$$V \approx V_0 + \frac{1}{2}\sum_{i,j} K_{i,j}(|r_{i,j}| - |r^0_{i,j}|)^2, \tag{1.6}$$

where $r_{i,j}$ is the displacement between atom i and j, $r^0_{i,j}$ is the initial displacement, and $K_{i,j}$ is the Hessian matrix (stiffness matrix) for a protein structure given by $K_{ij} = \partial^2 V/\partial r_i \partial r_j$. Quasi-harmonic analysis is used to solve the eigenvalue problem such as $K_{ij}v_j = \omega^2 m_i v_i$, where ω is the natural frequency, v_i is the normal mode corresponding to natural frequency ω, and m_i is the atomic mass for i^{th} atom. The cross-correlation matrix representing the thermal fluctuation motion can be computed from equilibrium statistical mechanics theory,

$$L_{ij} = \sum_{n=7}^{3N} \frac{k_B T}{m_i} {\omega_n}^2 (v_i \otimes v_j), \tag{1.7}$$

where k_B is the Boltzmann's constant, and T is the absolute temperature. The subscript n for natural frequency and normal mode represents the mode number (Weiner 1983). It should be noted that summation goes from 7 to $3N$, where N is the total number of atoms, and rigid body modes correspond to six zero eigenmodes.

1.2.2 Gõ Model

Hayward and Gõ (1995) introduced a more simplified potential field for α-carbon atoms such that α-carbon atoms are prescribed by a potential field consisting of covalent bonds for consecutive α-carbon atoms and non- bonded interaction (i.e. van der Waal's interaction) for native contacts. The thermal fluctuation behavior of protein structures has been well described by Gõ model.

Low-frequency normal modes relevant to protein dynamics is insensitive to details of potential field (Teeter and Case 1990). Hayward and Gõ (1995) pointed out that short-range interaction may govern the protein dynamics. Furthermore, the motion of protein structure fits well with a backbone chain represented by α- carbon atoms. Gõ potential can be addressed as,

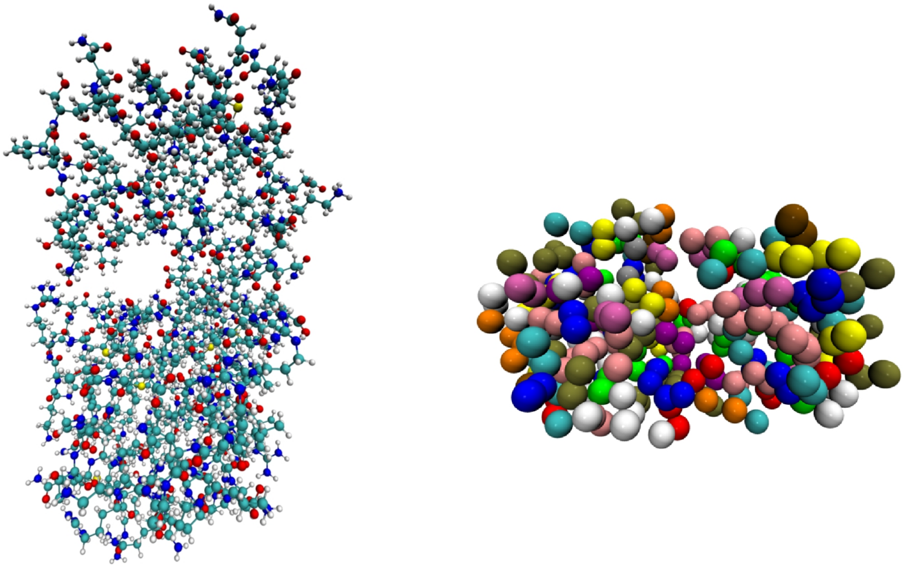
$$V \approx \sum_{i}[\frac{k_1}{2}(r_{i,i+1}-r^0_{i,i+1})^2+\frac{k_2}{4}(r_{i,i+1}-r^0_{i,i+1})^4]+\sum_{i,j}4\varepsilon(\frac{1}{r^6_{i,j}}-\frac{1}{r^{12}_{i,j}}), \tag{1.8}$$

where $r_{i,j}$ is the distance between i^{th} and j^{th} α-carbon atoms.

Superscript zero indicates the equilibrium state. The first summation represents the nonlinear elastic energy for covalent bonds, while the last summation shows the non-bonded interaction for native contact. Native contact is defined so that α-carbon atoms i and j are in the native contact if $r_{i,j}$ is less than a specific cut-off distance, such as 10 Å.

1.2.3 Elastic Network Models

In an elastic network model (ENM), the system is represented by a network of beads connected by elastic springs. Generally one bead represents one amino acid (Fig. 1.1). However, elastic network models can also be used with an all-atom description as well. The ENM is widely used for its simplicity. Although ENM only includes harmonic fluctuation, it correctly represents the topology of the system and yields the right pattern of the principal modes, which normally associate with protein functions.



(a) Molecular structure with CPK view, colored by atom type: H white; O red; N blue; C cyan; S yellow; P tan.

(b) Elastic network model, each bead who stands for one residue is connected with abstract springs within a cutoff distance.

Figure 1.1 Model T4 lysozyme (PDB: 3lzm) with elastic network model

ENM was first studied by Tirion (1996). She assumed the harmonic approximation to potential field prescribed to α-carbon atoms. The harmonic potential field can be presented only for native contacts and covalent bonds with identical force constant,

$$V \approx \sum_{i,j} \frac{\gamma}{2}(r_{i,j} - r^0_{i,j})^2 H(r_c - r^0_{i,j}), \tag{1.9}$$

where γ is a force constant, $r_{i,j}$ is the distance between i^{th} and j^{th} α-carbon atoms, superscript 0 indicates the equilibrium state, r_c is the cut-off distance defining a native contact, and $H(x)$ is the Heaviside unit step function defined as $H(x) = 0$ if $x < 0$; otherwise $H(x) = 1$.

1.3 Principal Component Analysis

As a linear combination of all state variables, for instance, nodal unknowns in finite element analysis, the general coordinate $\xi(t)$ can be introduced as

$$\xi(t) = \sum_{i=1}^{m} \phi_i x_i(t) = \boldsymbol{\phi}^{\mathrm{T}} \mathbf{x}(t), \tag{1.10}$$

where m represents the total number of channels or state variables and $x_i(t)$ is the i^{th} time dependent variable.

If the variance of $\xi(t)$ within the time interval $[t_0, t_1]$ is maximized, the data spreads mostly in this general direction. Consequence, this direction, represented as a linear combination of all state variables, will be called the principal direction, as illustrated in Fig. 1.2. Of course, it is anticipated that the total number of principal direction r should be much smaller than the spatial dimension m. Denote the variance of $\xi(t)$ can be expressed as,

$$\xi(t) = \int_{t_0}^{t_1} (\xi(t) - \bar{\xi})^2 dt, \tag{1.11}$$

with time averaged value,

$$\bar{\xi} = \int_{t_0}^{t_1} \xi(t) dt / (t_1 - t_0), \tag{1.12}$$

It is often more practical to use temporal discretization instead of dealing with infinite dimensional problems in the time domain. For example, there are n time snapshots within the time interval $[t_0, t_1]$. For a typical k^{th} channel or variable, denoted as $\bar{x}_k = \sum_{j=1}^{n} x_{kj}/n$, the variance of this channel can be expressed as $\sum_{j=1}^{n} (x_{kj} - \bar{x}_k)^2$. Consequently, the variance based on the general coordinate ξ in Eq. (1.10) can be expressed in terms of $\sum_{j=1}^{n} (\xi_j - \bar{\xi})^2$, where the mean value of ξ is defined as $\bar{\xi} = \sum_{j=1}^{n} \xi_j / n$.

The procedure to maximize the variance ξ in $\mathscr{R}^n$ leads to an optimization problem: find $\phi \in \mathscr{R}^m$ such that $\boldsymbol{\phi}^{\mathrm{T}}(\mathrm{x}_k - \bar{x}_k)(\mathrm{x}_k - \bar{x}_k)^{\mathrm{T}}\boldsymbol{\phi}$ is maximized subject to the constraint $\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi} = 1$. It is then clear that the principal component $\boldsymbol{\phi}$ is in fact the eigenvector of the

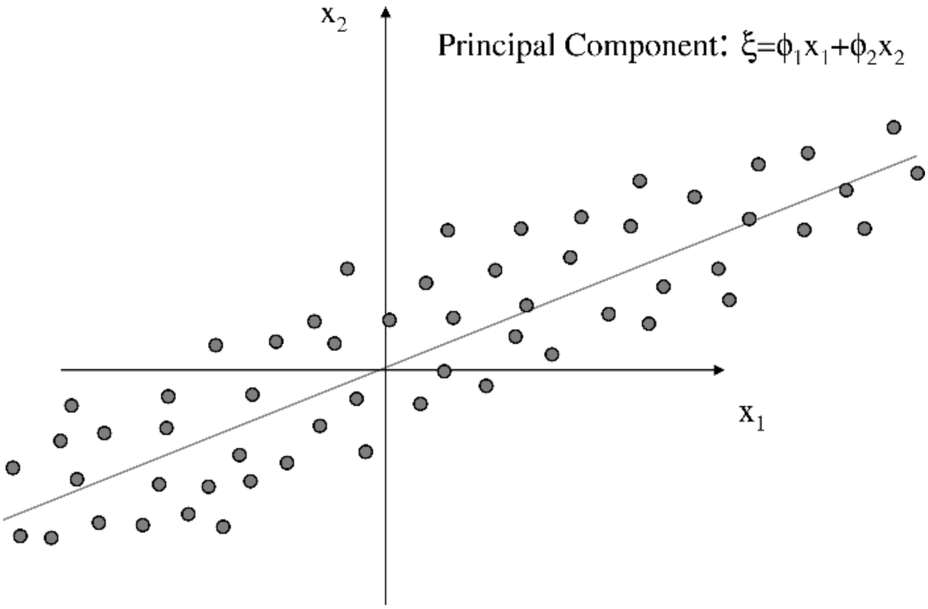


Figure 1.2 Illustration of principal component

covariance matrix,

$$\boldsymbol{A} = (\mathrm{x}_k - \bar{x}_k)(\mathrm{x}_k - \bar{x}_k)^{\mathrm{T}}, \tag{1.13}$$

and the number of principal components r depends on the rank of the covariant matrix $\boldsymbol{A}$.

Principal component analysis involves diagonalization of the positional covariance matrix A to identify an orthogonal set of eigenvectors or “modes” describing the direction of maximum variation in the observed conformational distribution. Covariance matrix A is calculated with Eq. (1.13). PCA diagonalization of the covariance matrix involves the following eigenvalue problem:

$$\boldsymbol{A}\mu = \lambda\mu, \tag{1.14}$$

where μ are the eigenvectors and λ are the eigenvalues of covariance matrix $\boldsymbol{A}$.

One motivation for PCA is to reduce the dimensionality of the MD trajectory data and provide a concise way to visualize, analyze, and compare large-scale collective motion observed over the course of the simulation. Particularly eigenvectors with the largest eigenvalues provide the biggest contribution to the observed covariance (Cheng *et al.* 2007; Rod *et al.* 2003). The so-called "essential" modes from a PCA are usually a selection of these eigenvectors and associated eigenvalues that collectively account for a large percentage of the total observed motion.

In the following example, PCA will be used to analyze a simple dynamical system. This example illustrates the PCA procedure and its relationship with the original data. Consider the three body system shown in Fig. 1.3, where the edge points are fixed and cannot move. According to the mathematical model (Wang 2008), there are three degrees of freedom in the system. Therefore, the dynamics equation of the system is:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{R}(t), \tag{1.15}$$

$$\mathbf{M} = \begin{pmatrix} \mathbf{m} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{2m} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{3m} \end{pmatrix}, \quad \mathbf{K} = \begin{pmatrix} \mathbf{2k} & -\mathbf{k} & \mathbf{0} \\ -\mathbf{k} & \mathbf{2k} & -\mathbf{k} \\ \mathbf{0} & -\mathbf{k} & \mathbf{2k} \end{pmatrix},$$

$$\mathbf{K}\boldsymbol{\phi} = \omega^2\mathbf{M}\boldsymbol{\phi}, \tag{1.16}$$

with $\mathbf{x} = [x_1 \;\; x_2 \;\; x_3]^T$, m for mass and k for stiffness of the spring, ω for natural frequency of this system.

With the standard procedure, the natural frequency can be calculated by solving the generalized eigenvalue problem. The eigenvalues equal to ω^2, where ω denotes natural frequency. Table 1.1 shows the eigenvalues and the corresponding eigenvectors for the calculation result.

Table 1.1: Eigenvalues and Eigenvectors for the Spring System

number	eigenvalue	eigenvector
1	1.18161	0.40298, 0.46837, 0.27219
2	2.23607	-0.40825, 0.40825, 0.40825
3	3.45502	0.06539, -0.33758, 0.87135

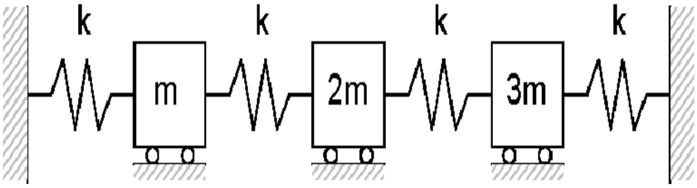


Figure 1.3 Illustration of three oscillator system. Three bodies (not affected by gravity), with mass m; $2m$; and $3m$, attached to four springs, each with spring constant k.

To build a dynamical system based on these three bodies, Verlet integration is used to generate trajectories for each integration or timestep. A trajectory matrix $\boldsymbol{A}$ is built where each row holds a body's displacement data and each column records displacements at one time step. PCA analysis will be performed as introduced.

Verlet integration is a numerical method used to integrate Newton's equation of motion. It is frequently used to calculate trajectories of particles in MD simulation. Newton's equation of motion for conservative physical systems is,

$$\mathbf{M}\ddot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) = -\nabla V(\mathbf{x}(t)), \tag{1.17}$$

where t is the time, $\mathbf{x}(t)$ is the ensemble of the position vector of N objects, V is the scalar potential function, $\mathbf{F}$ is the negative gradient of the potential giving the ensemble of forces on the particles,

Typically, an initial position and initial velocity are given. To discretize and numerically solve this initial value problem, a time step $\Delta t > 0$ is chosen and the sampling point sequence $t_n = n\Delta t$ is considered. The task is to construct a sequence of points that closely follow the points on the trajectory of the exact solution.

Verlet integration can be seen as using the central difference approximation to the second derivative,

$$\boldsymbol{a}_n = \frac{\Delta^2 \boldsymbol{x}_n}{\Delta t^2} \approx \frac{\boldsymbol{x}_{n+1} - 2\boldsymbol{x}_n + \boldsymbol{x}_{n-1}}{(\Delta t)^2}. \tag{1.18}$$

The Verlet algorithm uses Eq. (1.18) to obtain the next position vector from the previous two as

$$\boldsymbol{x}_{n+1} = 2\boldsymbol{x}_n - \boldsymbol{x}_{n-1} + \boldsymbol{a}_n(\Delta t)^2. \tag{1.19}$$

Two systems have been built. The first system is running in ideal condition; the second system is running with noise. To add noise to this system, random disturbing displacements are added at every integration. To stimulate the system, the initial position of every mass is set corresponding to one eigenvector of the natural frequencies, shown in Table. 1.1, namely as k, where $k = 1,2,3$; the initial velocities are set to zero; every time step Δt is set with $\Delta t = 0.05$ and total integration or time steps $nstep = 200$.

Table 1.2: Eigenvector and Rayleigh-Ritz Results with Verlet Algorithm

k	Eigenvector	Rayleigh-Ritz quotient
1	-0.50870, -0.67489, -0.53455	1.53942
2	-0.21210, 0.71023, 0.67125	1.93758
3	-0.01360, 0.16764, -0.98575	3.36764

Once the displacement matrix has been calculated, the covariance matrix is produced as Eq. (1.13). Eigenvectors for this covariance matrix are listed in Table.1.2. The new trajectory is produced by projecting data back corresponding each of the eigenvector.

The Rayleigh-Ritz method is used to find the approximate real resonant frequencies of systems with multiple degrees of freedom, such as spring mass systems (Chakraverly 2009). Figure. 1.4, 1.5, and 1.6 show the frequency computed by Rayleigh-Ritz with initial position corresponding to natural frequency. Based on linear algebra, the Rayleigh-Ritz quotients must lie between the minimal and maximal eigenvalues. It can be easily observed that the dominant eigenmode looks like the corresponding eigenvectors and Rayleigh-Ritz quotients are approximately equal to the corresponding eigenvalues. Most importantly, if one imposes an excitation with frequency around a generic natural frequency on the system, the result is resonance. In addition, the Rayleigh-Ritz quotient will tend to converge with the natural frequency as the number of snapshot $nstep$ increases based on the computing result.

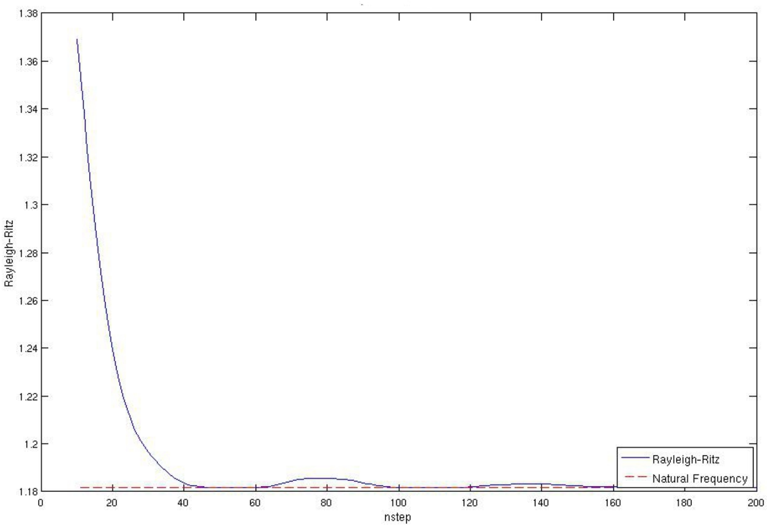


Figure 1.4 Frequency computed by Rayleigh-Ritz with initial position corresponding to first natural frequency; dashed line denotes the corresponding natural frequency.

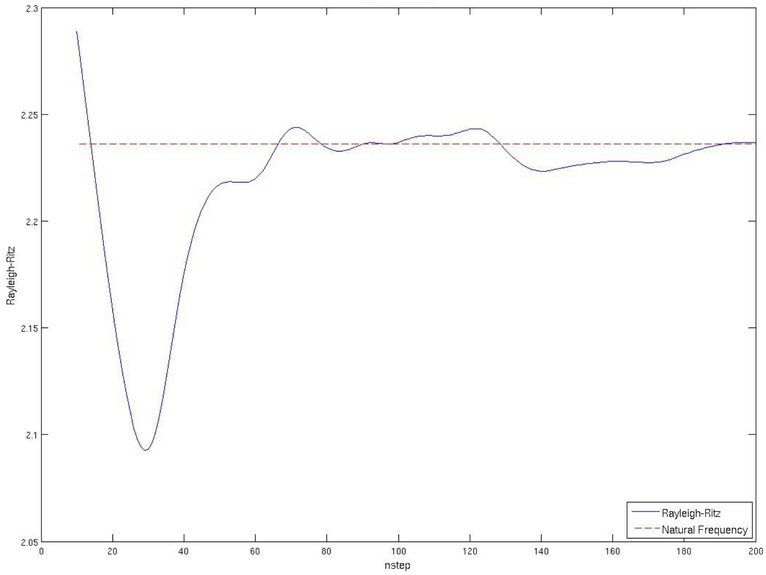


Figure 1.5 Frequency computed by Rayleigh-Ritz with initial position corresponding to second natural frequency; dashed line denotes the corresponding natural frequency.

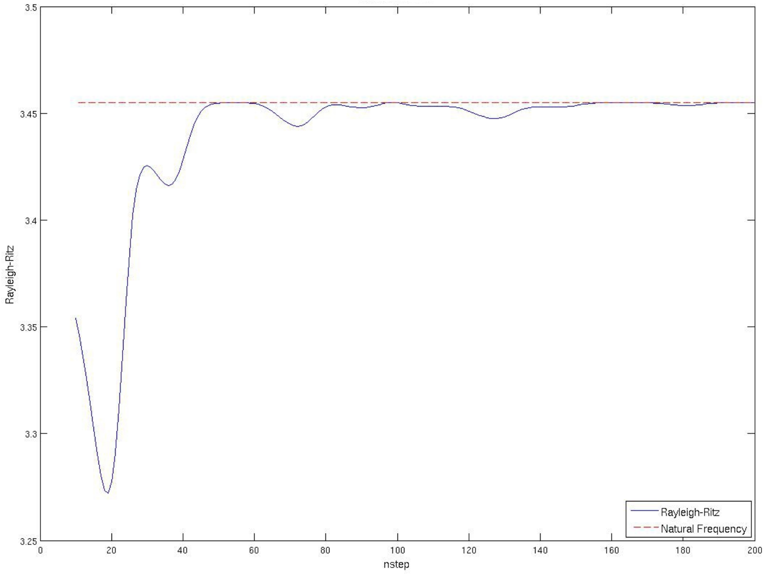
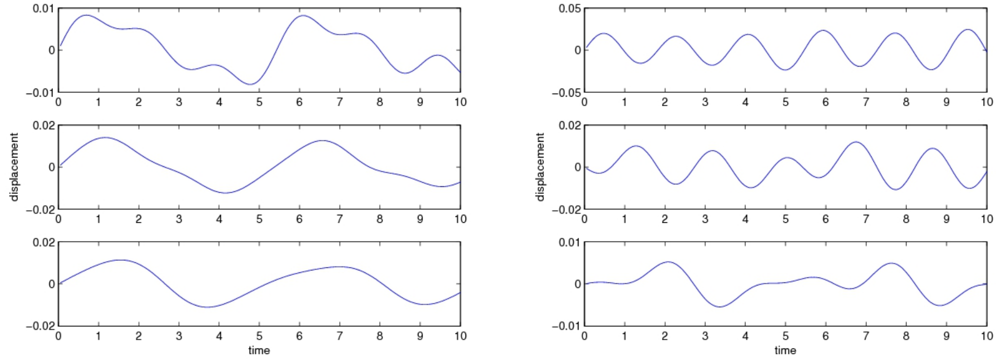


Figure 1.6 Frequency computed by Rayleigh-Ritz with initial position corresponding to third natural frequency; dashed line denotes the corresponding natural frequency.

The projected trajectories present the motion of the original set. Eigenvalues for the covariance matrix show how important each mode's contribution is to the whole motion. The percentages of the variance for each mode are: $97.1969, 2.0035, 0.7996$. In other words, one mode dominates the entire motions of the system. This result is true for any type of initial displacements proportional to the eigenmode.

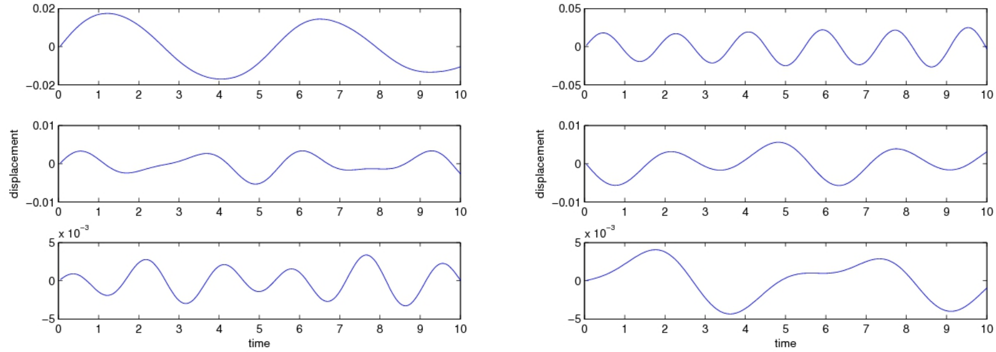
The original displacements are showed in Fig. 1.7. Projected displacements after PCA are shown in Fig. 1.8. After the PCA projection, the modes of displacements are close to the original motion. Similar phenomena (Figs. 1.9 and 1.10) can be discovered after applying noise in the spring system. In the latter case, the motion reinstated using PCA are between the system eigenmode and the dominant mode.

PCA is also applicable to the study of complex systems, for example the conformation (slow motion) analysis of biological macromolecules. Biological macromolecules such as proteins are large and have thousands of degree of freedoms. It is theoretically difficult to explain and predict the conformational changes in proteins. Slow motions such



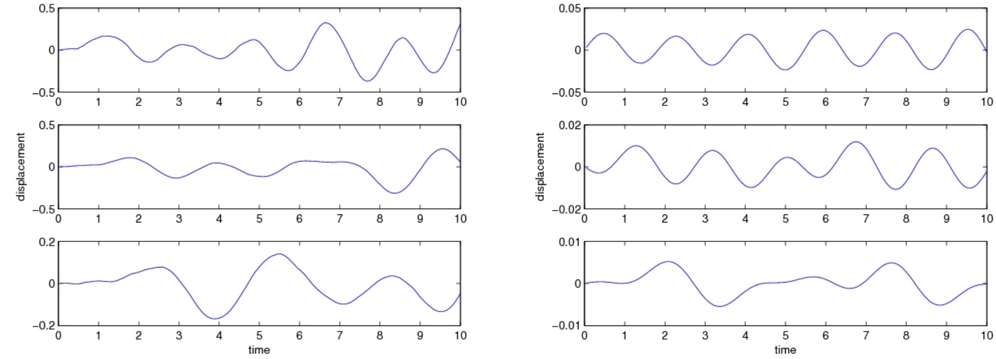
(a) Excited by initial displacement with $k = 1$ (b) Excited by initial displacement with $k = 3$

Figure 1.7 Time *vs.* displacement without noise



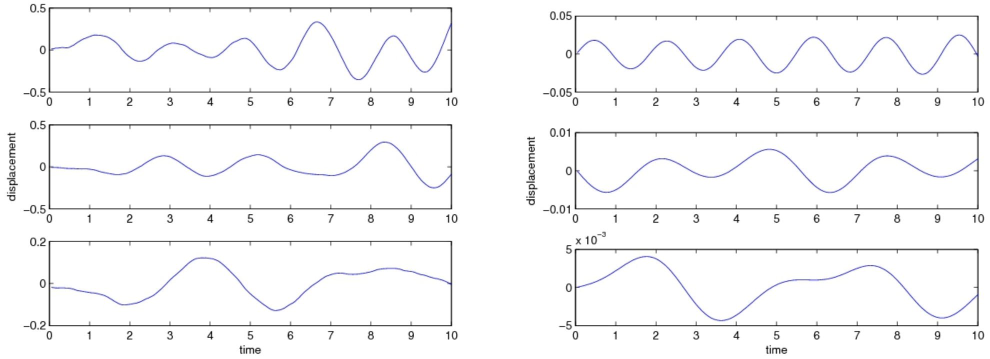
(a) Excited by initial displacement with $k = 1$ (b) Excited by initial displacement with $k = 3$

Figure 1.8 Time *vs.* projected displacement without noise



(a) Excited by initial displacement with $k = 1$ (b) Excited by initial displacement with $k = 3$

Figure 1.9 Time *vs.* displacement with noise



(a) Excited by initial displacement with $k = 1$ (b) Excited by initial displacement with $k = 3$

Figure 1.10 Time *vs.* projected displacement with noise

as protein folding can be dominant motions. The projection of dominant motion provides an effective way to study protein slow motion.

1.4 Quasi-harmonic Analysis

Quasi-harmonic analysis (QHA) (Harris and Laughton 2007; Karplus and Kushick 1981) and related principal component analysis techniques, based on the eigenvalue decomposition of an ensemble of protein conformation, have provided a useful method for identifying motion, particularly at long time scales (Ramanathan and Agarwal 2009). QHA captures the large-scale conformational fluctuation in a protein by diagonalizing the mass-weighted covariance matrix, known as the atomic fluctuation matrix ($F_{\alpha\beta}$). For a system with N atoms, $F_{\alpha\beta}$ is a $3N \times 3N$ symmetric matrix, defined as:

$$F_{\alpha\beta} = \langle m_{\alpha}^{1/2}(x_{\alpha} - \langle x_{\alpha}\rangle) m_{\beta}^{1/2}(x_{\beta} - \langle x_{\beta}\rangle)\rangle, \tag{1.20}$$

where α and β represent the $3N$ degrees of freedom in Cartesian space; m is the mass of the atom and the quantity within $\langle\rangle$ denotes an average over the ensemble of structures in MD simulation.

The inverse square roots of the eigenvalues determined by diagonalizing $F_{\alpha\beta}$ represent the frequencies associated with the protein eigenmode. The eigenvectors represent the displacement vectors of the individual atoms. The lowest frequencies correspond to large-scale cooperative motion in the protein; the higher frequencies represent localized motion. For a system with N atoms, there are $3N-6$ internal modes. However, due to computational costs, typically only a limited number of slow modes (lowest frequencies) are computed.

QHA allows identification of protein motion at a variety of time scales that are related to the length of the MD simulation. The atomic fluctuation matrix can be computed from protein conformation sampled during a single MD simulation that is of limited duration. It also can be computed with a collection of MD trajectories that represent a long time scale.

QHA is computationally expensive, especially with a large number of atoms and the long MD simulation. Various methods have been developed to minimize the computational cost, such as time-averaged normal coordinate analysis (TANCA) (Noid *et al.* 2000). Approximation methods lose information. An all-atom, all- time-frame computing approach with parallel/distributed computing is presented in this dissertation and details can be found in Chapter 5.

A significant advantage of QHA over other methods is that it allows identification of slow conformational fluctuations that span distant areas of the protein energy landscape, such as the reaction pathway during enzyme catalysis (Agarwal 2006; Agarwal *et al.* 2002). Overcoming some of the limitation of NMA based approaches, QHA provides a methodology to explore protein motions with longer time scales.

The difference between QHA and PCA is that in computing the atomic fluctuation matrix ($F_{\alpha\beta}$) QHA takes mass to amplify the motion of heavy atoms (that is, atoms other than hydrogen). Hydrogen atoms moves faster and are more likely to relate to high frequency motion. QHA and PCA also study protein conformational changes by exploring of location of each residue. These locations are usually represented by the location of the alpha-carbon of each residue. Motion of hydrogen atoms is related with localized motion. Therefore QHA is better at describing slow motion than PCA. The research on these phenomena is presented in Section. 3.4.

1.5 High Performance Computing

The implementation of PCA and other analytical methods in this dissertation involve operations on very large matrices and other computationally expensive tasks. There are few tools to do such analysis efficiently. Therefore, this research involved designing and implementing a parallel computing package for the operations performed, as presented in Chapter 5.

High-performance computing (HPC) uses supercomputers and computer clusters to solve advanced computational problems. In this research, Franklin, a massively parallel processing (MPP) system, was used to test the methods and produce the results.

Franklin belongs to National Energy Research Scientific Computing Center. It has (in October 2010) $9,572$ nodes, each with a quad core processor core. It has a theoretical peak performance of 9.2 GFlop/sec per core (4 flops/cycle if using $SSE128$ instruction). More information about the Franklin system is listed in Table 1.3.

Table 1.3: Franklin System Specification

Franklin Specification	
Number of compute nodes	9,572
Processor cores per node	4
Number of compute processor cores	38,288
Number of spare compute nodes	20
Processor core type	Opteron 2.3 GHz Quad Core
Processor core theoretical peak	9.2 GFlop/sec
System theoretical peak (compute nodes only)	352 TFlop/sec
Physical memory per compute node	8 GB
Memory usable by application per node	7.38 GB
Number of login nodes	10
Switch interconnect	SeaStar2
Measured MPI point-to-point bandwidth	1.6 GB/sec
Measured MPI point-to-point latency	$6.5 - 8.5\mu s$
File system	Lustre
Usable disk space	436 TB
Theoretical IO bandwidth	32 GB/sec aggregate
Batch system	Torque/Moab

Source: `http://www.nersc.gov/nusers/systems/franklin/about.php`, accessed November 2, 2010.

CHAPTER 2

EQUILIBRIUM CONFORMATIONAL FLUCTUATION ANALYSIS

Equilibrium conformational fluctuation plays an essential role in the biological function of proteins. These fluctuation range from subtle rearrangements of a few atoms to large-scale movement involving the entire protein. Protein motion timescales occupy a correspondingly large range, from femtosecond to microseconds and even longer (Agarwal 2006; Cannon and Benkovic 1998). The lower end of this range, or fast motion, is commonly known as vibration, which includes bond stretching and angle bending. The upper end of this range, or slow motion, involves large conformational changes (Elber 2005; Harrison 1984; Rossmann *et al.* 2005; Saibil 2000; Schlick *et al.* 1997; Tama *et al.* 2000), such as rotation of side chains and movement of flexible loops. Slow fluctuations have been studied in relation to possible involvement with protein function (Agarwal 2005, 2006; Agarwal *et al.* 2002; Benkovic and Hammes-Schiffer 2003; Cannon and Benkovic 1998; Caratzoulas *et al.* 2002; Eisenmesser *et al.* 2002, 2005; Hammes 2002; Hammes-Schiffer 2002; Henzler-Wildman *et al.* 2007). However, proteins are chemically inhomogeneous (Ma 2005). The complexity of protein dynamic behavior makes it difficult to obtain a theoretical description of thermal fluctuation. In this research, a combination of classical molecular dynamics (MD) simulation and principal component analysis (PCA) is introduced to study equilibrium conformational fluctuation of proteins.

Protein motions and their possible role in protein function are studied with normal-mode analysis (NMA) and its various extensions. Conformational fluctuation and individual protein structures are associated with protein structure and function. NMA is based on harmonic approximation of the potential energy minimum, which is computed by diagonalizing the Hessian matrix. However, recent research shows that slow conformational changes of proteins are largely anharmonic (Tournier and Smith 2003). Furthermore, NMA ignores solvents, but water plays an important role in protein conformation. Nevertheless

NMA is not well suited for protein conformational change that covers widely separated areas of the conformational energy landscape, such as enzyme catalysis (Ramanathan and Agarwal 2009).

The increase in available computing power provides a systematic way to study protein conformational changes with MD simulation. A combination of MD simulation with PCA and its extensions have provided insights into protein and DNA dynamics. These methods have been successfully applied to the thermodynamics of a number of nucleic acid interactions including sequence selective drug-DNA association (Harris *et al.* 2005), large conformational transition in DNA (Noy *et al.* 2007), and protein-DNA complexes (Dixit *et al.* 2005). A generalized package of this methodology has been built in this research and is demonstrated in this chapter.

2.1 Framework

One motivation of this research is to build a generalized package for protein equilibrium conformational fluctuation analysis. This method may be applied to any protein with known atomic coordinates. The standard procedure is as follows:

1. Select an appropriate PDB file from Protein Data Bank and run an MD simulation. Save the trajectories matrix $A_{3m\times n}$, where m is number of atoms and n is the number of snapshots from simulation shown as follows:

$$A^T = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{m1} & y_{11} & \cdots & y_{m1} & z_{11} & \cdots & z_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} & y_{12} & \cdots & y_{m2} & z_{12} & \cdots & z_{m2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{mn} & y_{1n} & \cdots & y_{mn} & z_{1n} & \cdots & z_{mn} \end{pmatrix}^T$$

2. Clean the data to remove translation and rotation. Details can be found in Section 2.4.1.

3. Compute the covariance matrix with Eqs. (1.13) and (1.20).

4. Solve the eigen problem with covariance matrix to get eigenvectors and eigenvalues.

5. Assess the number of modes required to accurately represent the conformational change by analyzing overlap coefficients and eigenvalues. Overlap coefficients analysis was introduced in Section 2.2.

6. Calculate a new "mode matrix" by projecting selected eigenvectors or "modes" back to covariance.

7. Analyze conformational changes based on the results of previous steps.

8. Do correlation analysis, introduced in Section 2.3.

2.2 Mode Selection and Evaluation with Overlap Coefficients

The overlap coefficients (OCs) of the α-carbon atoms are calculated to measure how well each PCA mode represents all motion (Bradley *et al.* 2008). PCA modes (v_i) are compared with a vector (Δx), that describes the displacement of α carbon atoms from the minimized starting simulation position. The structure is experimentally determined by X-ray crystallography. This vector is calculated following alignment of the minimized and crystal structures with the MD average structure to remove rotation and translation of the center of mass.

The weight factor ($\alpha^{(i)}$) for each PCA mode is calculated as follows:

$$\alpha^{(i)} = \frac{v_i \cdot \Delta x}{\| v_i \|^2}. \tag{2.1}$$

It measures whether or not Δx can reasonably be represented in the vector space described by various subsets of PCA modes. A subset S of modes is calculated as a reconstructed vector ($\tilde{V}$) as follows:

$$\tilde{V} = \sum_{i \in S} \alpha^{(i)} v_i. \tag{2.2}$$

Then the overlap coefficients, which show the similarity between Δx and each PCA mode (v_i or $\tilde{V}$) are calculated by the angle between the two vectors:

$$cos(\theta) = \frac{v \cdot \Delta x}{\| v \| \cdot \| \Delta x \|}. \tag{2.3}$$

The relative error in the reconstructed vectors can be computed as:

$$\varepsilon = \frac{\| \Delta x - \tilde{v} \|}{\| \Delta x \|}. \tag{2.4}$$

This error simply provides a measure of how well the displacement due to the recalculated vector recapitulates the observed crystallographic Δx and thus allows users to assess the number of modes required to accurately represent the conformational change.

2.3 Introduction of Correlation Analysis Method

Correlations among the fluctuations of residues, measured at alpha-carbon atoms, provide additional insight into protein function (Brooks *et al.* 1983; Lange and Grubmüller 2006; Waight *et al.* 2010). An unweighted pair group method with arithmetic means (UPGMA) clustering algorithm is used to define groups of residues with similar correlation patterns under equilibrium fluctuation. The correlation is measured by constructing correlation matrices. The details of correlation matrices can be found in Section 2.4.3.

A correlation analysis procedure similar to Bradley *et al.* (2008), is followed:

1. Construct a positional correlation matrix S_{ij} (Section 2.4.3).

2. Compute the effective "distance" d_{ij}, based on the correlation measure:

$$d_{ij} = \sqrt{1.0 - S_{ij}^2} \tag{2.5}$$

The distances between residues represent the strength of the relationship between them, with the closest residues having the strongest correlation.

3. Generate a "treelike" representation of correlated residues by the UPGMA method, based on the calculated effective distances. The UPGMA method is presented in Section 2.4.4.

4. Do a depth-first search to choose the smallest clusters that contain at least one residue from each of the groups defined by the selection criteria. Experimental information plays an important role in the definition of these clusters.

5. Complete the cluster selection.

   (a) Figure out other important residues of the selection criteria that were not found in the initial cluster selection.

   (b) Find the clusters that contain those important residues identified in step 5a but do not contain any residue selected in the previous depth first step.

Correlation matrices provide insight into the pairwise correlation of residues. While residue clusters are likely to involve inter-domain communication, they do not show individual residue interaction that could be important for the protein functional mechanism. Although the correlations identified by clustering do not provide direct information about the series of events that generate the correlations, they do provide a starting point for experimental and further computational studies designed to model conformational changes in the protein.

2.4 Methods and Theory

2.4.1 MD Simulation Data Cleaning and Preparing for PCA

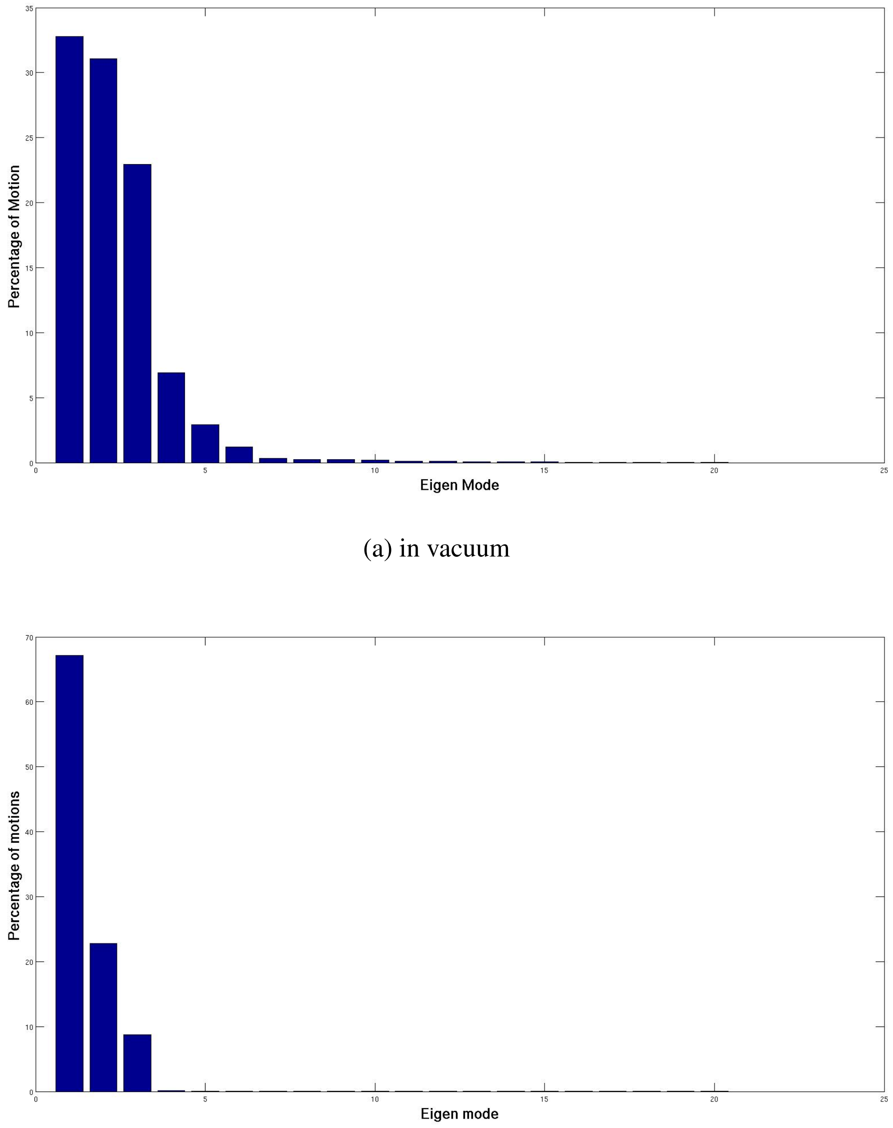
Based on the eigenvalues from PCA analysis presented in Chapter 3 (Fig 2.1), the dominant motions were the first three motions for each system. Each mode was projected back corresponding with its eigenvector. The motions are shown in Figs. 2.2 and 2.3. Rotation is dominant in the third motion as is evident from Fig. 2.3. The other two motions they are a combination of rotation and translation. This is generally true for any simulation of a protein system because translation and rotation are usually the dominant motions of the original data set. This is one evidence that PCA captures slow conformational changes. However, further significant motions lie behind the dominant rotation and translation. To bring these motions to the fore, alignments are used to remove translation and rotation before PCA.

A reference structure is defined for alignment. The coordinates of each snapshot is aligned with the reference structure. The reference structure is defined by the user and usually is the initial position or the averaged position of all snapshots. Translation alignment can easily be done by moving all coordinates by a vector $\boldsymbol{v}$ which equals to $\boldsymbol{v}_o - \boldsymbol{v}_r$, where $\boldsymbol{v}_o$ and $\boldsymbol{v}_r$ hold the coordinates of the center of mass of the original and reference systems.

The center of mass is traced by $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$. At each time step, the center of mass can be evaluated as follows: for $k = 1, 2, 3$,

$$\bar{x}_k = \sum_{i=1}^{N} \frac{m_i x_k^i}{m}, \tag{2.6}$$

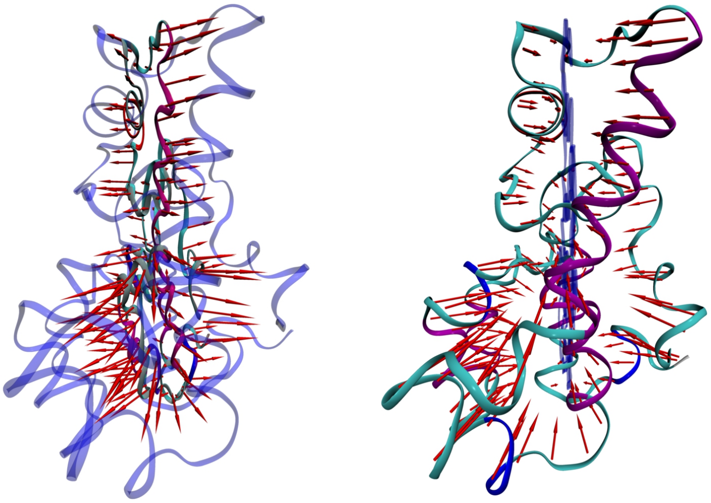
where N is the total number of atoms, x_k^i represents the position of the i^{th} atom in the k direction, and m is the total mass expressed as $\sum_{i=1}^{N} m_i$.



(a) in vacuum

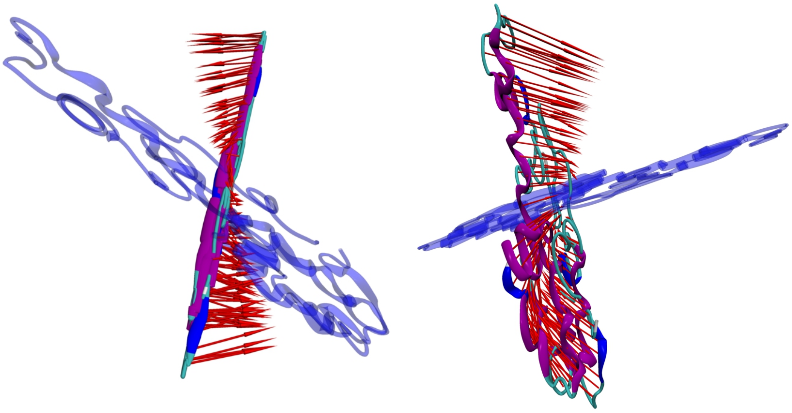
(b) in solvent

Figure 2.1 PCA eigenvalues with original data for T4 lysozyme MD simulation



(a) Expansion (b) Contraction

Figure 2.2 Non-aligned PCA mode 1 displacement visualization of MD simulation without solvent. The starting structure is shown in new-cartoon representation colored by structure. The destination structure is shown in blue ribbon. The red arrows indicate the direction of displacement based upon 1^{st} PCA mode of MD simulation without solvent.



(a) (b)

Figure 2.3 PCA mode 3 displacement visualization of MD simulation without solvent. The initial structure is shown in new-cartoon representation colored by structure. The destination structure is shown in blue ribbon. The red arrows indicate the direction of displacement based upon the third PCA mode of the MD simulation without solvent.

Alignment for rotation is calculated by the Kabsch algorithm, which is introduced in Section 2.4.2.

2.4.2 Kabsch Algorithm

The Kabsch algorithm (Kabsch 1976, 1978) is a method for calculating the optimal rotation matrix that minimizes the RMSD between two sets of points, namely rotation alignment.

The algorithm calculates only the rotation matrix. Both sets of coordinates must be transformed to their centroid first. Note that, in general, the "rotation matrix" represents a transformation to a different orthonormal basis rather than a rotation about a single axis.

The algorithm starts with two sets of paired points, P and Q. Each set of points can be represented as an $N \times 3$ matrix. The first row consists of the coordinates of the first point, the second row of the coordinates of the second point, the n^{th} row of the coordinates of the n^{th} point:

$$\textbf{P or Q} = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{pmatrix}$$

The algorithm works by calculating a covariance matrix, $\boldsymbol{A}$,

$$\boldsymbol{A} = \boldsymbol{P}^T\boldsymbol{Q}. \tag{2.7}$$

The optimal rotation $\boldsymbol{R}$ is based on the matrix formula $\boldsymbol{R} = (\boldsymbol{A}^T\boldsymbol{A})^{1/2}\boldsymbol{A}^{-1}$. However, in the case that $\boldsymbol{A}$ does not have an inverse, Kabsch (1976, 1978) propose another approach with singular value decomposition (SVD). First, calculate the SVD of the covariance matrix $\boldsymbol{A}$,

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \tag{2.8}$$

Next, decide whether you need to correct the rotation matrix to insure a right-handed coordinate system

$$d = sign(det(\boldsymbol{A})). \tag{2.9}$$

Finally, calculate the optimal rotation matrix $\boldsymbol{R}$, as

$$\mathbf{R} = \boldsymbol{V} \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{d} \end{pmatrix} \boldsymbol{U}^T. \tag{2.10}$$

2.4.3 Positional Correlation Matrix

A scalar correlation matrix was defined in Cartesian space by equation

$$S_{ij} = \frac{\langle(\boldsymbol{r}_i - \langle\boldsymbol{r}_i\rangle)(\boldsymbol{r}_j - \langle\boldsymbol{r}_j\rangle)\rangle}{\sqrt{\langle(\boldsymbol{r}_i - \langle\boldsymbol{r}_i\rangle)(\boldsymbol{r}_i - \langle\boldsymbol{r}_i\rangle)\rangle\langle(\boldsymbol{r}_j - \langle\boldsymbol{r}_j\rangle)(\boldsymbol{r}_j - \langle\boldsymbol{r}_j\rangle)\rangle}}, \tag{2.11}$$

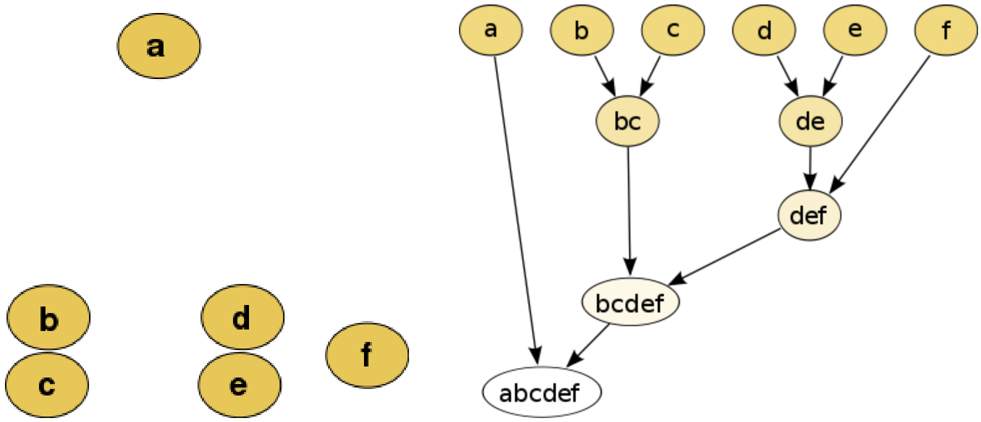
to denote the correlation between atoms i and j (Bradley *et al.* 2008). The value of S_{ij} ranges from -1.0 to 1.0, with positive and negative values indicating correlated and anti-correlated motion. $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ are the position of atoms i and j and $\langle\rangle$ denotes the trajectory average over snapshots. The matrix is calculated across all α carbon atoms of the protein.

2.4.4 The Unweighted Pair Group Method with Arithmetic Mean

The unweighted pair group method with arithmetic mean (UPGMA) (Durbin *et al.* 1999) is a simple agglomerative or hierarchical clustering method. The algorithm examines a pairwise distance matrix (or similarity matrix) to construct a rooted tree. At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters A and B is the average of all distances between pairs of objects x in A and y in B ($d(x,y)$):

$$\frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x,y). \tag{2.12}$$

UPGMA algorithm are illustrated in Fig 2.4.



(a) Row data

(b) Clustering with UPGMA

Figure 2.4 Illustration of UPGMA algorithm
Source: `http://en.wikipedia.org/wiki/Cluster_analysis`,
accessed November 2, 2010.

CHAPTER 3

CASE STUDY ON T4 LYSOZYME

This chapter presents a case study designed to evaluate the framework developed in Chapter 2 by analyzing the equilibrium conformational fluctuation of a T4 lysozyme. This study helps us to understand protein function by local conformational flexibility and provide a quantitative evaluation of the proposed MD-PCA framework. The fluctuation and dominant motion analysis generated by this research produce comparable results with published results arrived at by other methods.

The use PCA on MD simulation data is computationally expensive due to the size of proteins and large sets of time snapshots generated. Most previous work, therefore, analyzed only models in which α- carbon positions were used to represent residue position. This reduces the computational cost of PCA. This research expands these published methods (Bradley *et al.* 2008; Yang *et al.* 2010) from the residue level to the atomic level.

A question about all-atom PCA/QHA analysis is whether it produces accurate (slow motion) protein equilibrium conformational fluctuation. This research, employing analysis with eigenvalues, overlaps, correlation and modes, answers this question in the affirmative. The results suggest that QHA produces more precise slow motion than PCA. The details of this analysis are presented in the following section.

3.1 Molecular Dynamics Simulation of T4 Lysozyme

The T4 phage lysozyme molecular dynamics simulation utilizes the NAMD parallel molecular dynamics package version 2.6 (Phillips *et al.* 2005) with the CHARMM force field (Brooks *et al.* 1983). The initial structure was determined by X-ray crystallography (Matsumura *et al.* 1989), retrieved from the Protein Data Bank (PDB ID 3LZM; residue numbers 1-164).

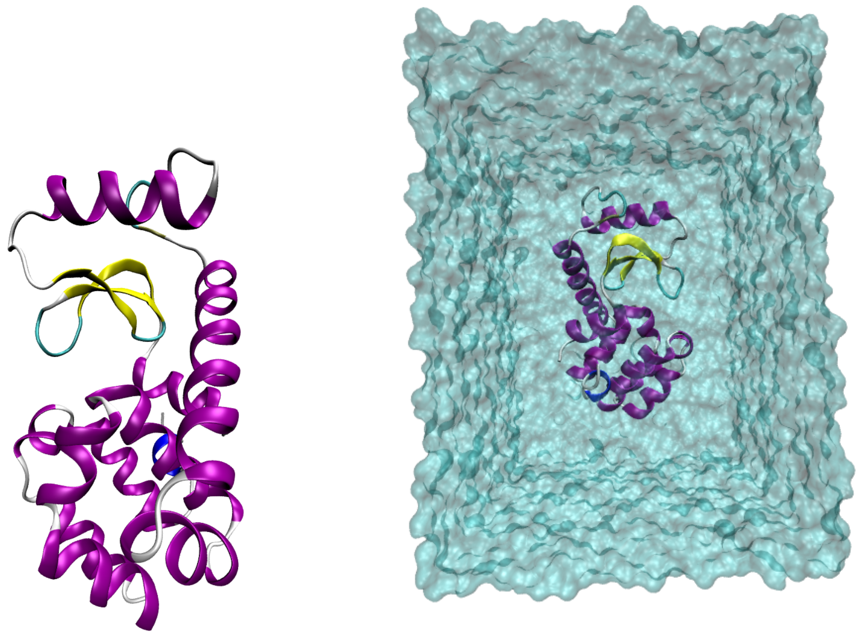
Two systems, one with solvent and another in vacuum, were built for MD simulation. The PSF package in VMD (Humphrey *et al.* 1996) is used to generate the structure file (PSF file) for MD simulation. First, the coordinates missing in the crystal structure are reconstructed. For the solvent system, the minimized structure is solvated in a periodic truncated cubic simulation box of $29,712$ TIP3P water molecules, providing a minimum of 15Å of water between the protein surface and any box edge. The water box is rotated to minimize the system size. Sodium and chloride ions are added to neutralize the total system and achieve a salt concentration of 0.05 mol/L. Ion placement is random; minimum distances between ions and molecules as well as between any two ions are set to 5Å. The parameter setting was shown in Table 3.1.

Table 3.1: MD Simulation Parameter Setting for T4 Lysozyme

Item	Setting
periodic boundary conditions	yes
cutoff distance	12Å
switch distance	10Å
pair list distance	13.5Å
timestep	2.0
rigid bonds	all
non-bonded frequency	2

Periodic boundary conditions are selected for simulation with solvent. The periodic boundary condition involves surrounding the system under study with identical virtual unit cells. The atoms in the surrounding virtual systems interact with atoms in the real system. These modeling condition are effective in eliminating surface interaction of the water molecules and creating a more faithful representation of the in *vivo* environment than what a water sphere surrounded by vacuum provides (Bhandarkar *et al.* 2008).

In the solvent simulation, energy minimization was performed in $20,000$ steps. There were no energy changes during last $1,000$ time-steps. This stage is shown in Fig. 3.1b Following minimization, the entire system was heated in 10 K increments up to 300 K with micro-canonical ensemble (NVE) equilibration per temperature step. The simulation was



(a) Molecular structure in vacuum

(b) Molecular structure in solvent

Figure 3.1 Schematic representation of the energy-minimized molecular structure analyzed for T4 lysozyme with structure view, colored by structure type: Alpha-Helix purple, 3_10_Helix blue, Pi_Helix red, Extended_Beta yellow, Bridge_Beta tan, Turn cyan, Coil white.

conducted at 300 K. Atomic trajectories were calculated in one femtosecond (fs) time steps using SHAKE (Ryckaert *et al.* 1977) constraints on hydrogen-heavy atom bonds. The total production simulation length is 15 nanosecond (*ns*). The first 9 *ns* were discarded as an "equilibration period." This extensive relaxation/equilibration period was necessary due to drift in the potential energy,

A similar procedure was performed in vacuum (shown in Fig. 3.1a), except that during the heating stage each step is 50 K. The total simulation length is 40 *ns*. In both simulations, snapshots were recorded every picosecond (*ps*) for analysis.

3.2 Equilibration Measures

It is important to assess whether or not the simulation achieved equilibrium or steady-state sampling before analyzing conformational fluctuation. Analysis of the molecular dynamics of the T4 lysozyme was performed as described in Section 3.1. The root mean-squared deviation (RMSD) of atomic position relative to the starting structure was calculated to measure equilibration and simulation stability.

RMSD is a numerical measure of the difference between two structures defined as,

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{atoms}}(r_i(t_1) - r_i(t_2))^2}{N_{atoms}}}, \tag{3.1}$$

where N_{atoms} is the number of atoms whose positions are being compared and $r_i(t)$ is the position of atom i at time t.

Figure 3.3 shows the results of the RMSD calculation. There is an initial rapid rise in RMSD that levels off around 10 *ns* in vacuum. By contrast, the RMSD in solvent was stable except during the first 0.2 *ns*. The solvent simulation was heated slowly to create a stable simulation result. The system was relaxed during the heating stage. The rapid rise reflects a significant conformational change during the simulation as demonstrated in Fig. 3.2. Without the "protection" of water molecules, the alpha-helix on the top bends

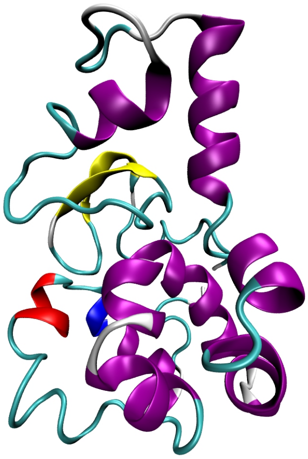


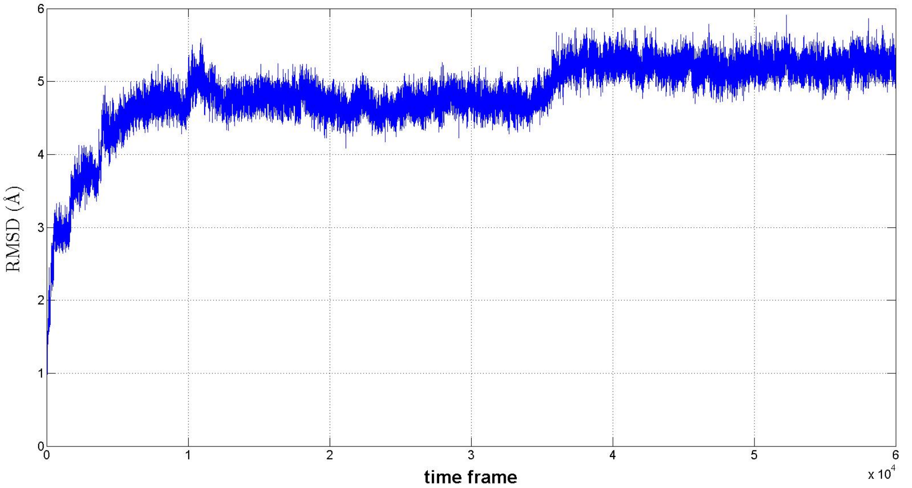
Figure 3.2 Illustration of T4 lysozyme structure after 40 *ns* simulation in vacuum

toware the lower part. This is one of the natural modes (Kundu *et al.* 2002). Based on these results, last 5 *ns* simulation with solvent was mainly used for fluctuation analysis.

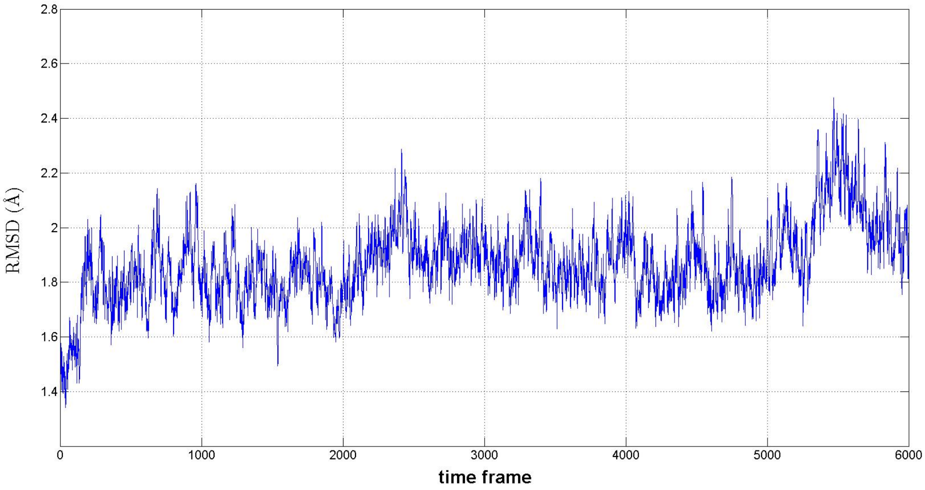
3.3 Fluctuation Analysis

Equilibrium thermal fluctuation of residues provides the first quantitative evaluation of the proposed MD- PCA framework and helps to understand protein function by local conformational flexibility. The root mean square fluctuation (RMSF) of α-carbon atoms is plotted as a function of residue number to show local conformational change. Atomic alpha-carbon fluctuations are compared to published results such as the All-Atom-Method (ATM), Rotational-Translational Block (RTB), Gaussian Network Model (GNM), and Finite Element Method (FEM) (shown in Fig. 3.4).

The T4 lysozyme shows flexibility at the solvent- exposed β-hairpin turn between residues 20 and 24, also at the short loop between residues 35 and 38. The solvent-exposed α-helix at residues 39-51 which can found in Fig. 3.4. These RMSF results are in quantitative agreement with other methods except at residue 115. This suggests that the α-helix containing residue 115 tends to bend at that spot. In the simulation without water, that α-helix does bend at residue 115 without solvent-exposed “protection”.

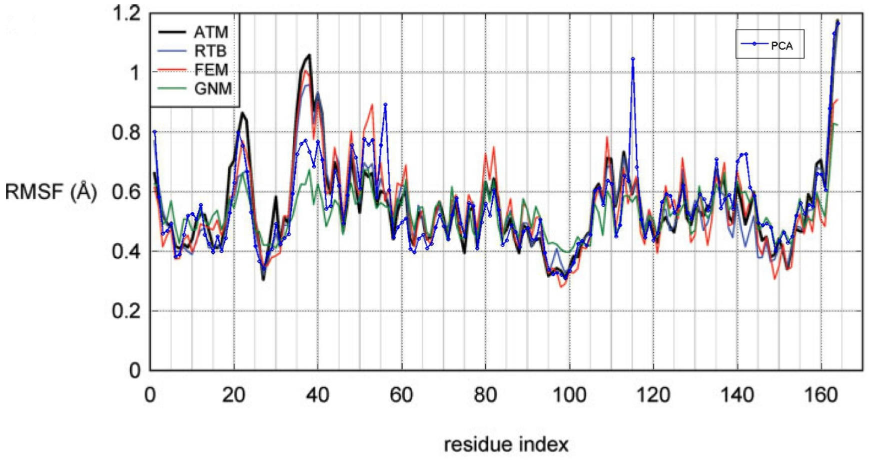


(a) RMSD value with T4 lysozyme simulation in vacuum

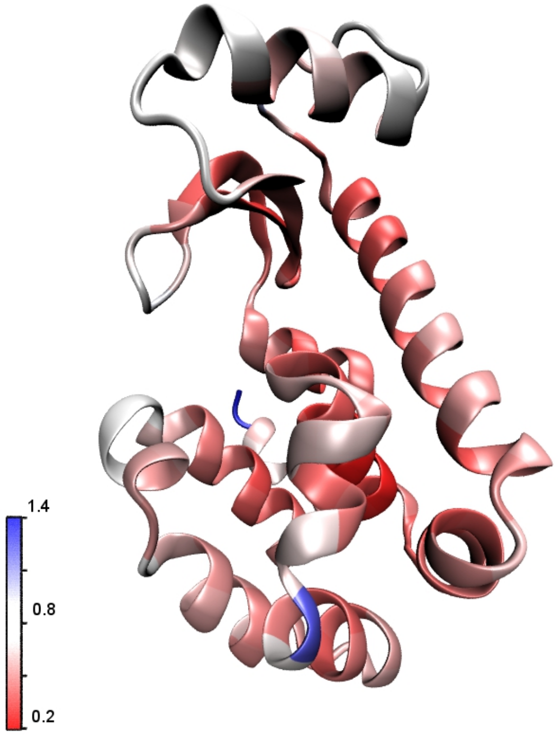


(b) RMSD value with T4 lysozyme simulation with solvent

Figure 3.3 Root Mean-Squared Deviation (RMSD) of atomic position relative to the starting structure with T4 lysozyme simulation.



(a) Methods used are: All-Atom-Method (ATM), Rotational-Translational Block (RTB), Gaussian Network Model (GNM), Finite Element Method (FEM), Data generated from simulation (PCA)



(b) Illustration RMSF, colored by RMSF value

Figure 3.4 RMSFs of the α-carbon atoms at 300 K for T4 lysozyme

3.4 Mode Selection and Evaluation with PCA and QHA

This research analyzes the motion of all atoms in the system, including the hydrogen atoms in the protein, which form a large part of the protein atoms and which have high frequency modes. In this section, eigen analysis and overlap analysis are used to evaluate the dominant modes of motion. Both PCA and QHA are used in the analysis.

Three models are tested with data taken from the same simulation. Model one is PCA with α-carbon atoms only; model two is PCA with all atoms; model three is QHA with all atoms. PCA with α-carbon model represents the benchmark solution. The purpose of this test is to evaluate the all atom methods.

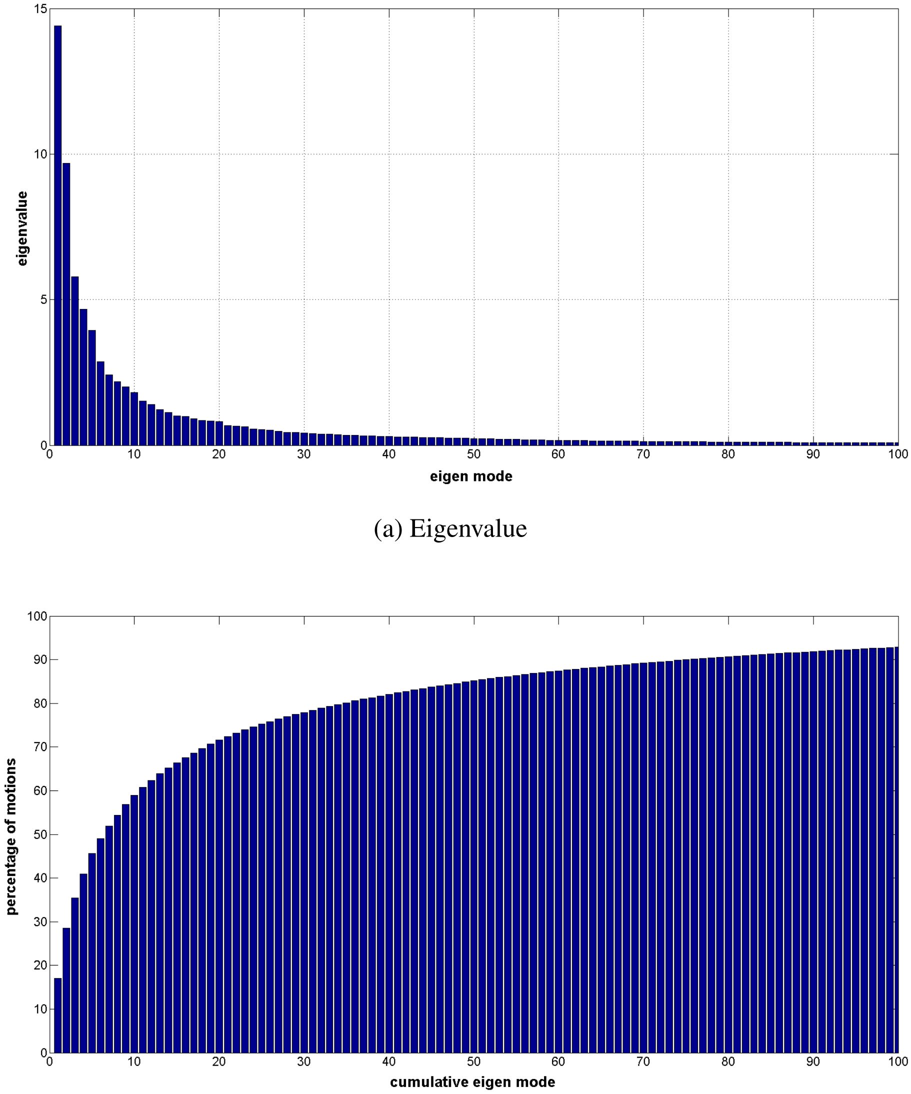
PCA transforms a number of possibly correlated variables into a number of uncorrelated variables called principal components. Its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. PCA is sensitive to the relative scaling of the original variables. In the following, the similarities and differences among these three models are presented.

3.4.1 Eigenvalue Analysis

Eigenvalues and eigenvectors are calculated according to the standard procedure in the three models. The results are shown in Figs. 3.5, 3.6, 3.7, and Tab. 3.2.

The screen test first proposed by Cattell (1966) yields the best results in practice to decide how many eigenmodes should be taken. Cattell suggests finding the place where the smooth decrease of eigenvalues appears to level off to the right of the eigenvalue plot. Following this suggestion, the first 10 eigenmodes are selected in all three models.

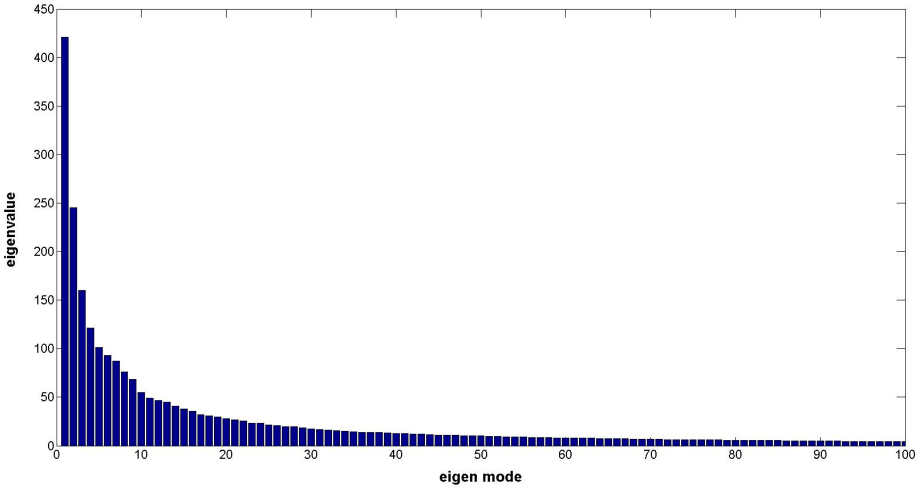
Although all atom PCA discovers the same number of dominant modes (slow motion) as α-carbon PCA does, the slow motions are disturbed. The cumulative energy of the first 10 is 59.1% in α-carbon PCA, 42.6% in all atom PCA, 46.2% in QHA. Quasi-harmonic analysis magnifies each atom's motion by its mass. QHA is a mediated method between all atom PCA and α-carbon PCA.



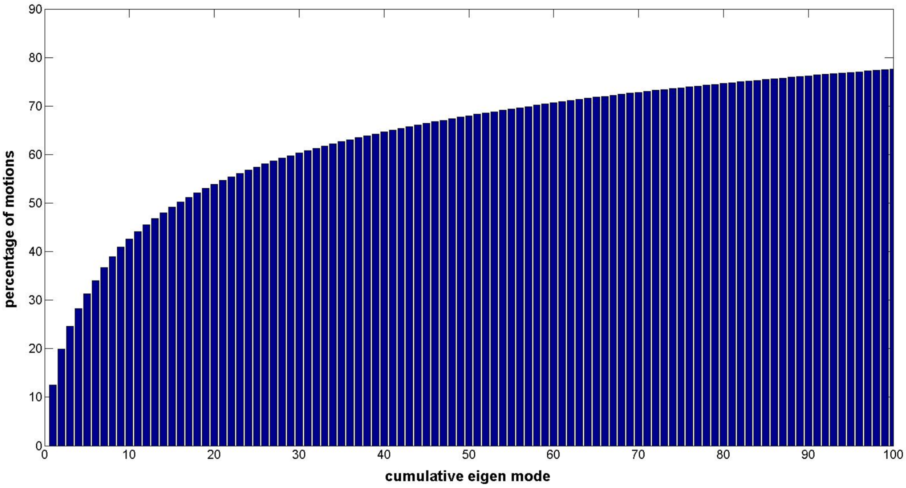
(a) Eigenvalue

(b) Cumulative energy content for each eigenvector

Figure 3.5 PCA eigenvalues calculated with α-carbon for T4 lysozyme MD simulation

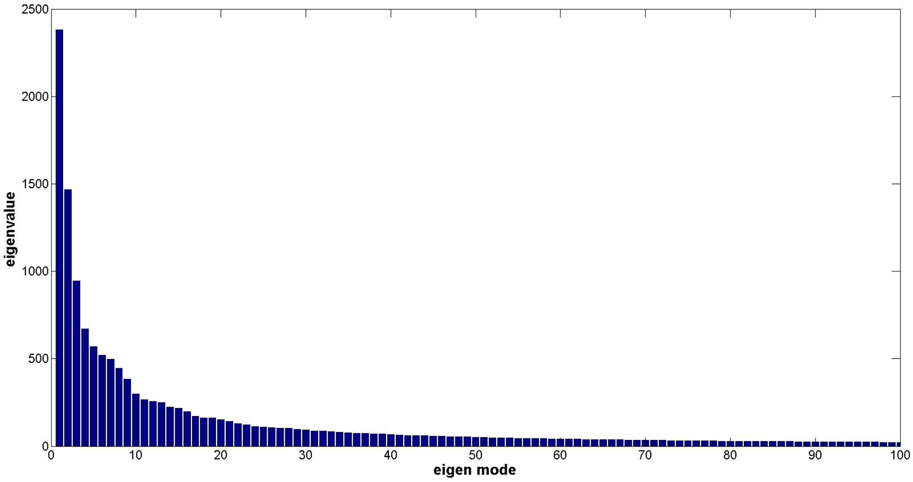


(a) Eigenvalue

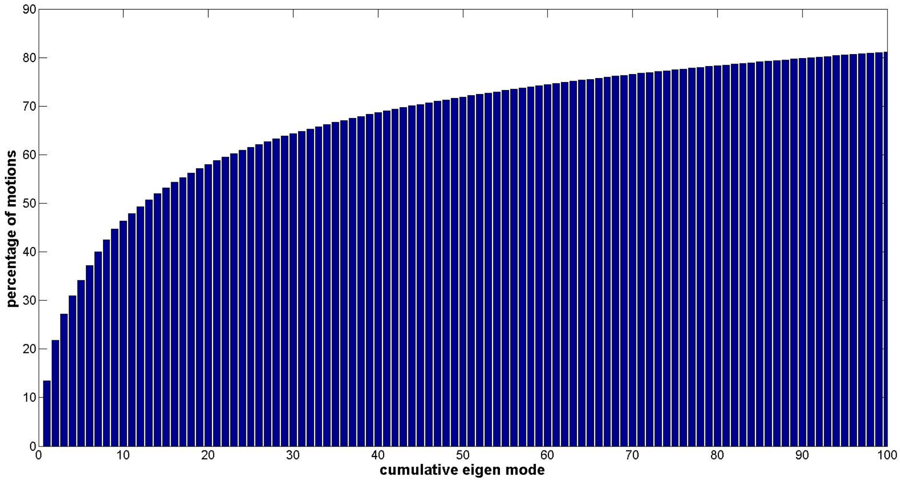


(b) Cumulative energy content for each eigenvector

Figure 3.6 PCA eigenvalues calculated with all-atoms for T4 lysozyme MD simulation



(a) Eigenvalue



(b) Cumulative energy content for each eigenvector

Figure 3.7 QHA eigenvalues calculated with all-atoms for T4 lysozyme MD simulation

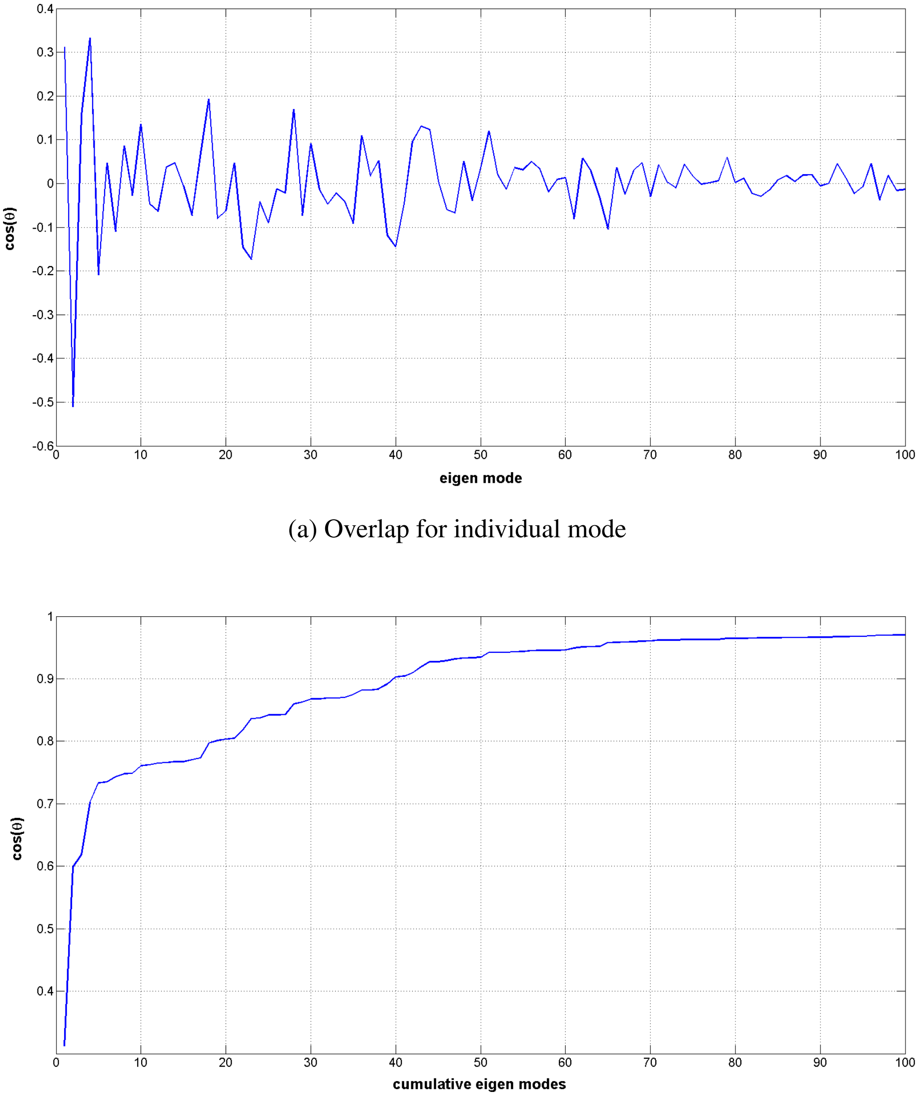
Table 3.2: Eigenvalues and Cumulative Energy Content

index	α-carbon PCA		PCA		QHA	
	eigenvalue	cumulative energy	eigenvalue	cumulative energy	eigenvalue	cumulative energy
1	14.411	0.171	421.061	0.126	2381.532	0.135
2	9.679	0.286	245.015	0.199	1468.108	0.218
3	5.786	0.355	160.080	0.247	944.536	0.271
4	4.667	0.410	121.069	0.283	672.259	0.309
5	3.954	0.457	101.190	0.313	570.720	0.341
6	2.865	0.491	93.356	0.341	520.547	0.370
7	2.409	0.520	87.318	0.367	498.236	0.398
8	2.192	0.546	75.949	0.390	444.936	0.423
9	2.003	0.570	68.630	0.410	384.959	0.445
10	1.810	0.591	54.685	0.426	297.665	0.462

3.4.2 Overlap Evaluation

With the calculated PCA/QHA modes placed into experimentally determined structural rearrangements, the overlap coefficients (OCs) were calculated as shown in Section 2.2. The results are shown in Figs. 3.8, 3.9 and 3.10. The full basis set of 492 α-carbon PCA modes gives a very high overlap ($cos(\theta) = 1.000$), as do all atom PCA/QHA, while carrying low error at 0.00001 for α-carbon PCA, 0.00006 for all atom PCA, and 0.0006 for all atom QHA. From this PCA basis set, a minimal set of modes was identified that represented the observed displacement. The first 10 modes were selected based on eigenvalue analysis.

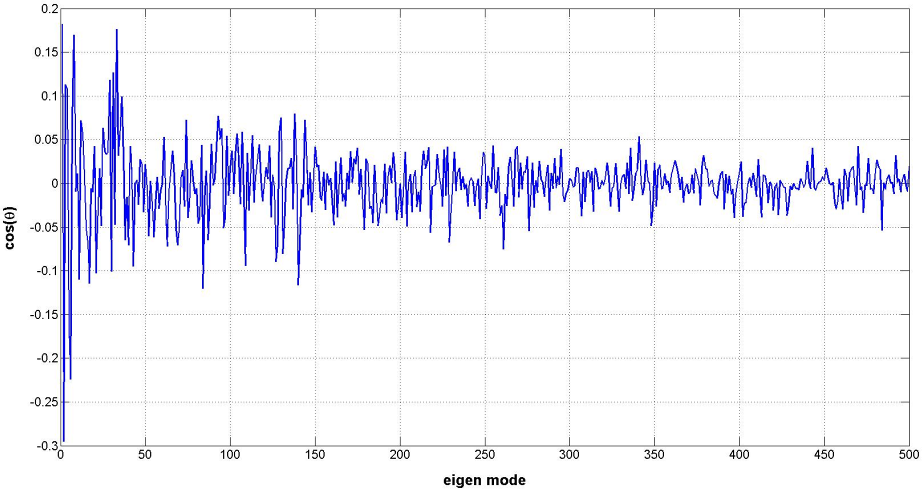
The first and largest all atom PCA mode accounts for 12.6% of the total motion, but has only a small overlap ($cos(\theta) = 0.17$) with the changes observed in the X-ray structure. This largest mode represents an asymmetric "twisting-like" mode (Fig. 3.13). The second mode accounts for 7.3% of the total motion but has a greater overlap ($cos(\theta) = -0.29$). This is a "flapping-like" motion (Fig. 3.14). A similar result is obtained with the all atom QHA method. Figure 3.11 shows different distribution of the errors with the QHA method. Errors tended to occur at the begin and end modes. This is evidence that QHA can "separate" slow motion and high frequency motion.



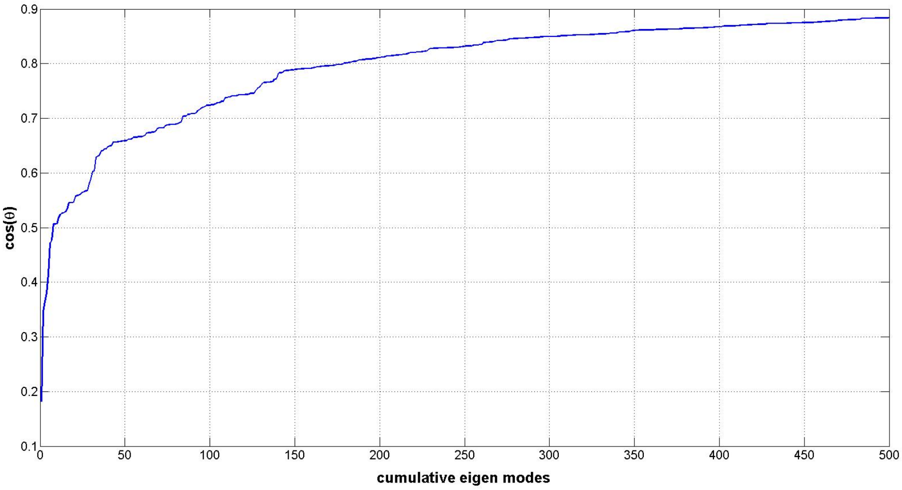
(a) Overlap for individual mode

(b) Overlap for comculative modes

Figure 3.8 PCA motion overlap measurement calculated with α-carbon for T4 lysozyme MD simulation

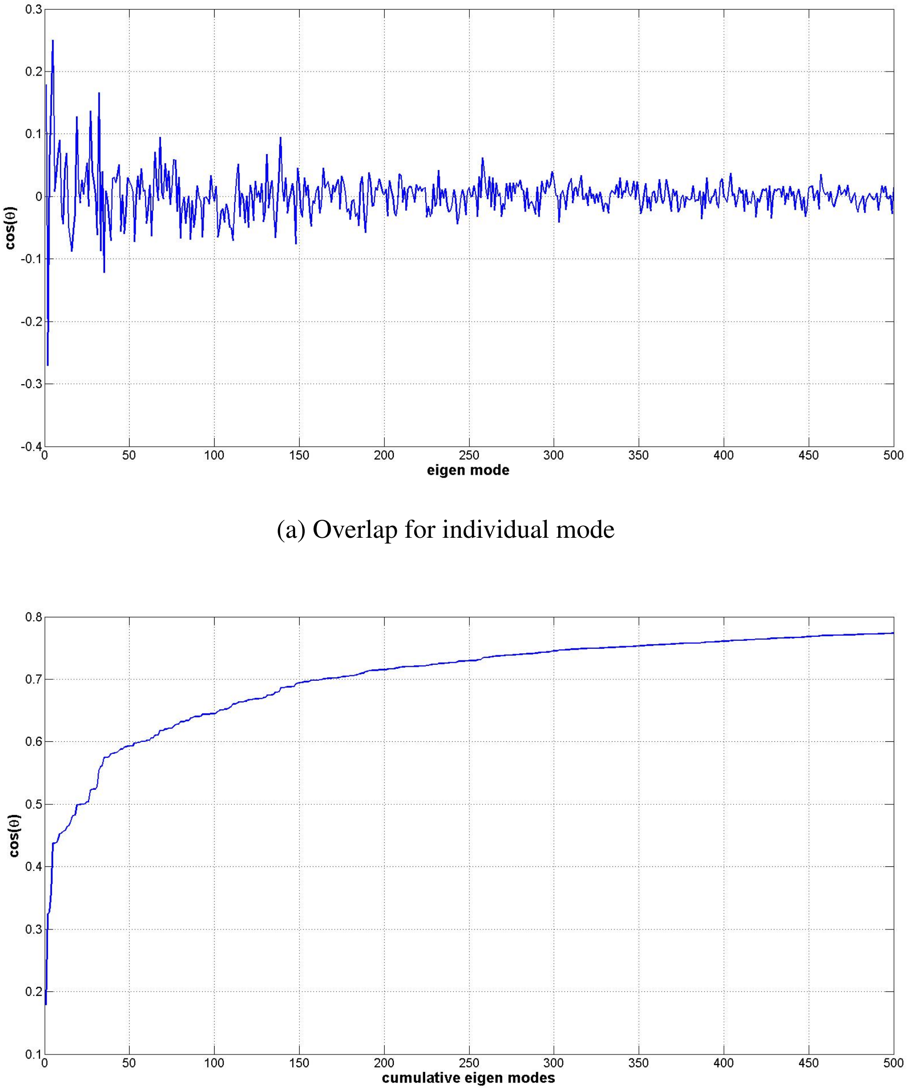


(a) Overlap for individual mode



(b) Overlap for comculative modes

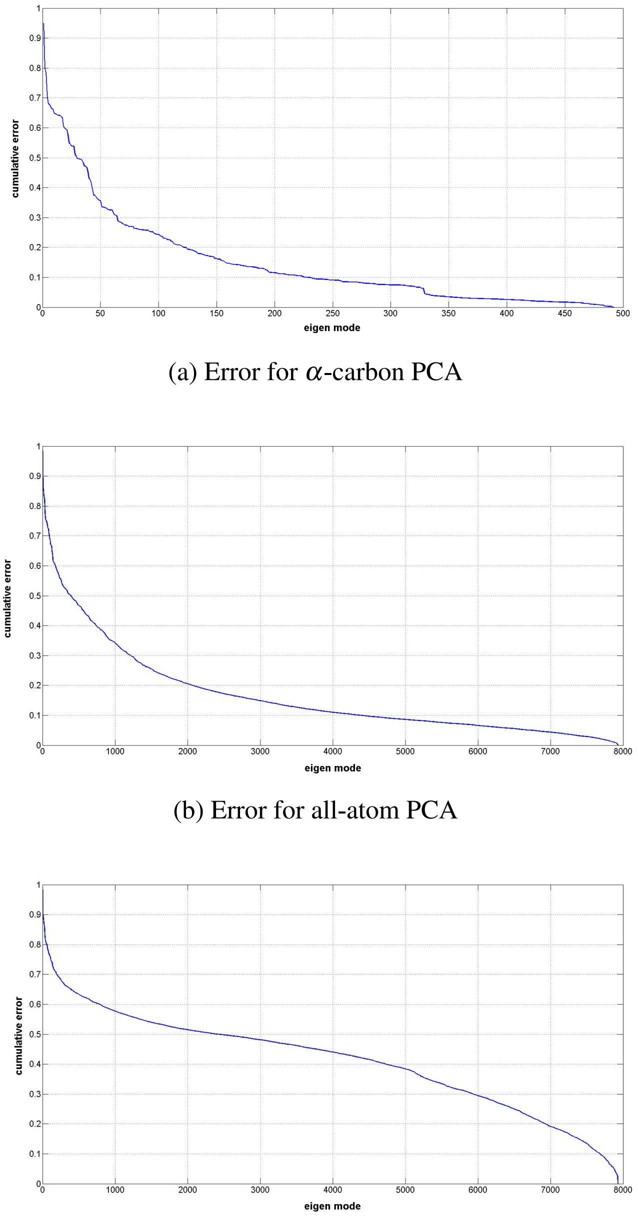
Figure 3.9 PCA motion overlap measurement calculated with all-atoms for T4 lysozyme MD simulation



(a) Overlap for individual mode

(b) Overlap for comculative modes

Figure 3.10 QHA motion overlap measurement calculated with all-atoms for T4 lysozyme MD simulation



(a) Error for α-carbon PCA

(b) Error for all-atom PCA

(c) Error for all-atom QHA

Figure 3.11 Motion error measurement

3.5 Identification of Slow Conformational Flexibility

The lowest modes of proteins describe their most characteristic conformational changes, which are often directly related to their function. The lowest mode of the T4 lysozyme is a hinge-bending mode similar to that of native hen egg lysozyme. Three lowest modes are presented in Figs. 3.13, 3.14, and 3.15, They are compared with the results from Bathe (2008). The second mode shows the hinge-bending mode.

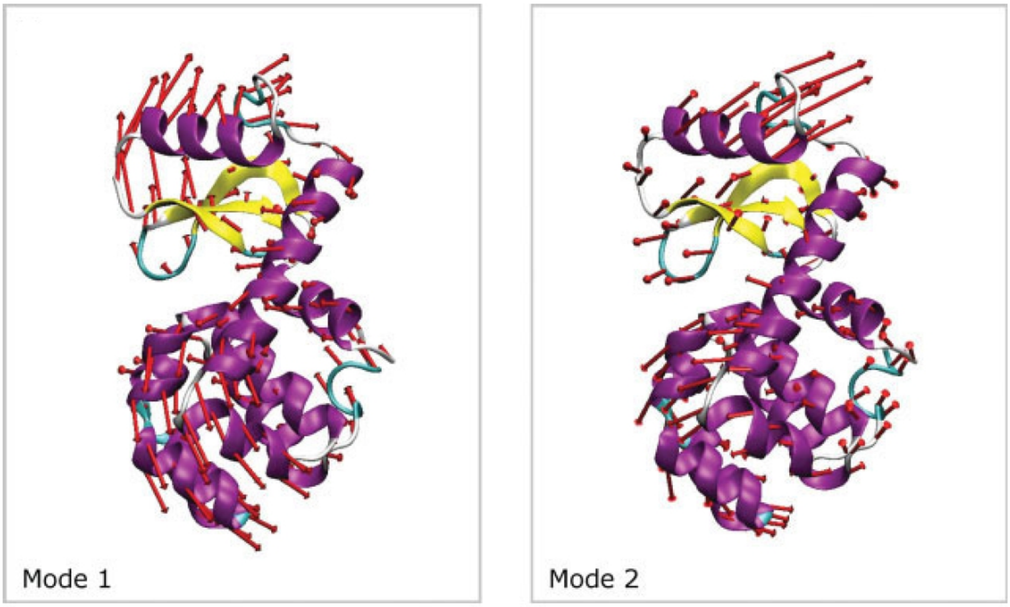
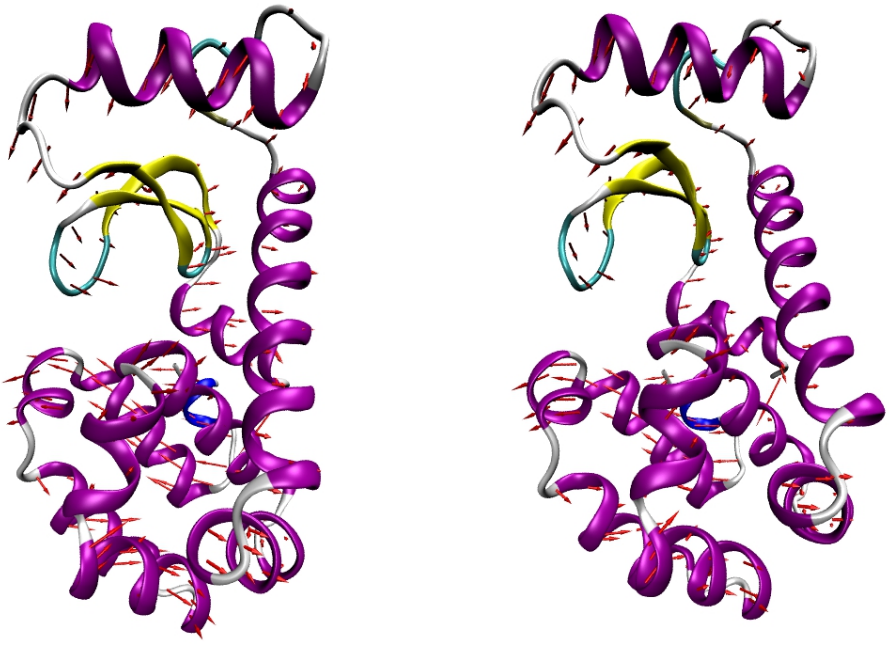


Figure 3.12 Schematic representation of the two lowest eigenmodes of T4 lysozyme computed using the FEM

3.6 Correlation Analysis of T4 Lysozyme

A cluster analysis of resideues was produced from the correlation matrix and is shown in Section 2.3. It depicts possible interactions among groups of residues. This method provides a starting point for experimental and computational studies designed to determine more fine-grained conformational changes in the protein.

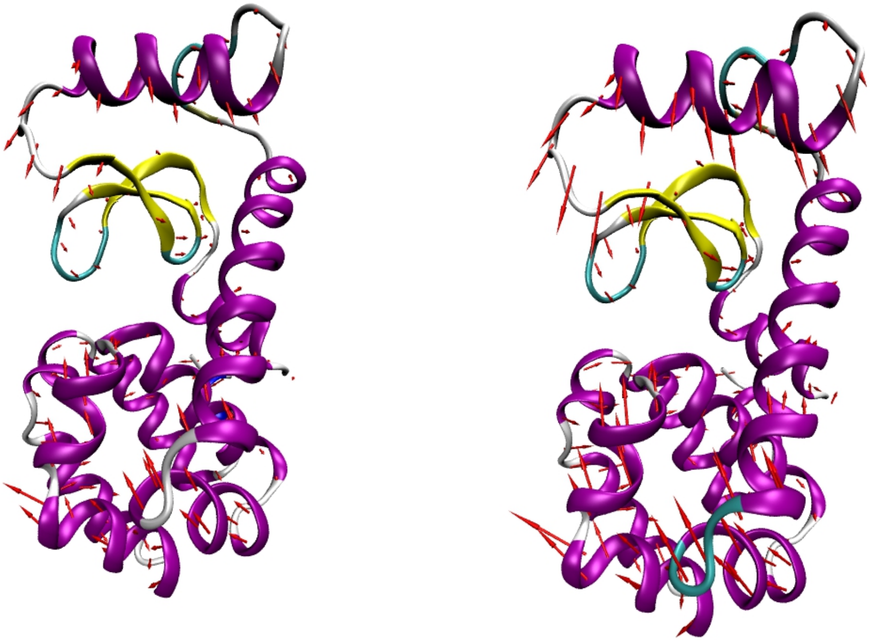
Figures 3.16a and 3.16b generated by Bathe (2008) were used as a basis for evaluation of these results. In Fig. 3.16c, where results were calculated from the raw simulation



(a) Generated from PCA

(b) Generated from QHA

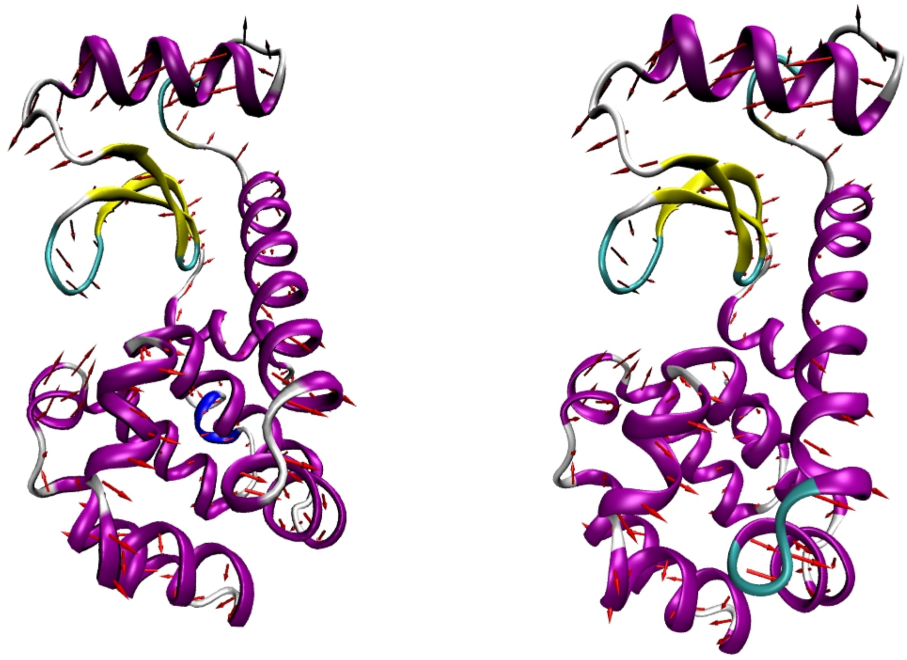
Figure 3.13 The 1^{st} eigenmode displacement visualization of T4 lysozyme MD simulation with solvent. The starting structure is shown in new-cartoon representation colored by structure. The red arrows indicate the direction of displacement based upon 1^{st} eigenmode of MD simulation with solvent.



(a) Generated from PCA

(b) Generated from QHA

Figure 3.14 The 2^{nd} eigenmode displacement visualization of T4 lysozyme MD simulation with solvent. The starting structure is shown in new-cartoon representation colored by structure. The red arrows indicate the direction of displacement based upon 2^{nd} eigenmode of MD simulation with solvent.



(a) generated from PCA

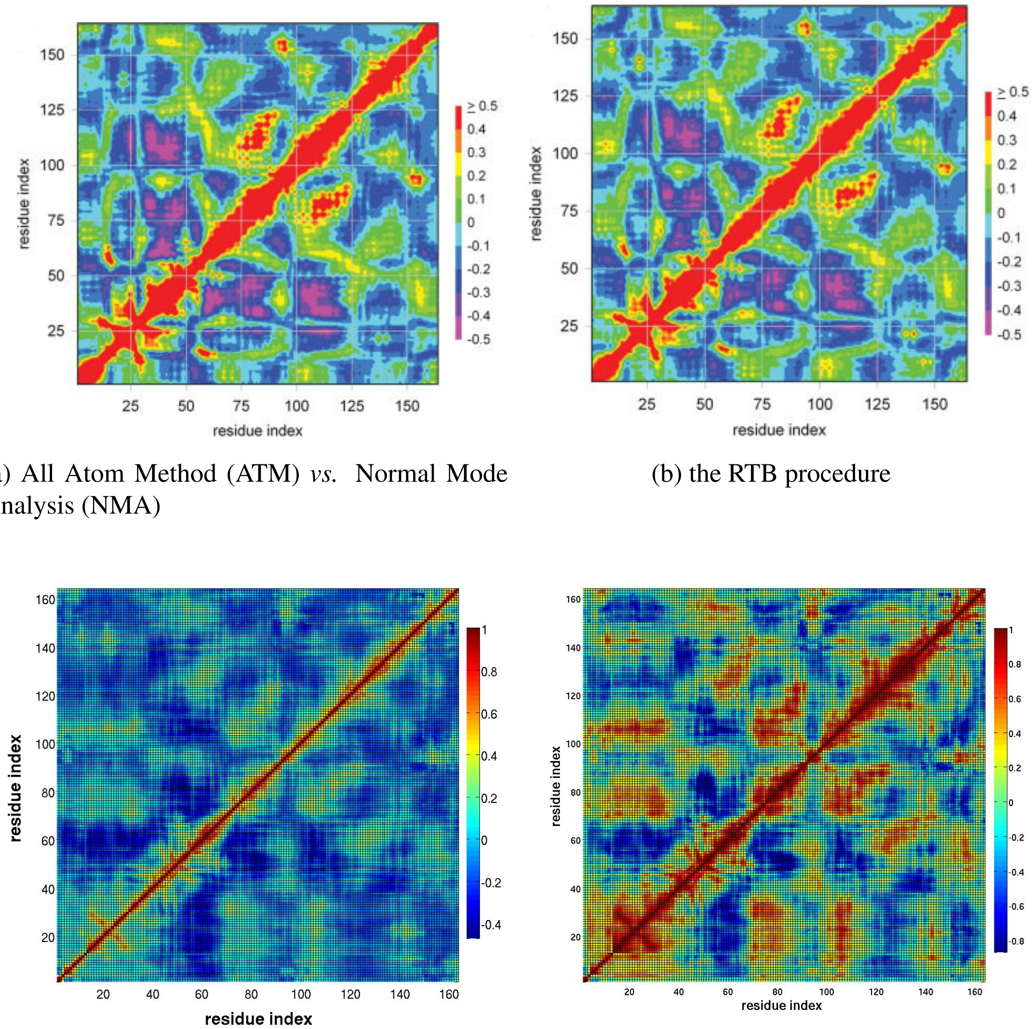
(b) generated from QHA

Figure 3.15 The 3^{rd} eigenmode displacement visualization of T4 lysozyme MD simulation with solvent. The starting structure is shown in new-cartoon representation colored by structure. The red arrows indicate the direction of displacement based upon 3^{rd} eigenmode of MD simulation with solvent.

data, the (anti-) correlation was weak. However, after projecting the first 10 eigenmodes to generate a new "cleaned" data set, the (anti-) correlation results were in closer agreement with Bathe's (Fig. 3.16d). However, more correlations among residue groups are found in this map, suggesting that more interaction have been identified with the PCA based method.

Clustering was performed using the UPGMA algorithm introduced in Section 2.3. In Fig. 3.17, the cluster showed only "contact based" relationships. This might be caused by the conformational changes in which a new harmonic balance has been reached. In Fig. 3.18, the red α-helix group shows the contribution of the "hinge-bending" mode. Furthermore, interaction between the two lower α-helix are strong too.

This chapter presents the MD-PCA framework of analysis applied to the equilibrium conformational fluctuation of the T4 lysozyme. This T4 lysozyme case study indicates the PCA/QHA analysis using all atoms is effective in extracting slow motions of the protein, yielding results in agreement with previous studies, as well as some novel and potentially useful new results.



(a) All Atom Method (ATM) *vs.* Normal Mode Analysis (NMA)

(b) the RTB procedure

(c) Original simulation data

(d) PCA method

Figure 3.16 Correlated fluctuation of α-carbon atoms computed for $T4$ lysozyme.

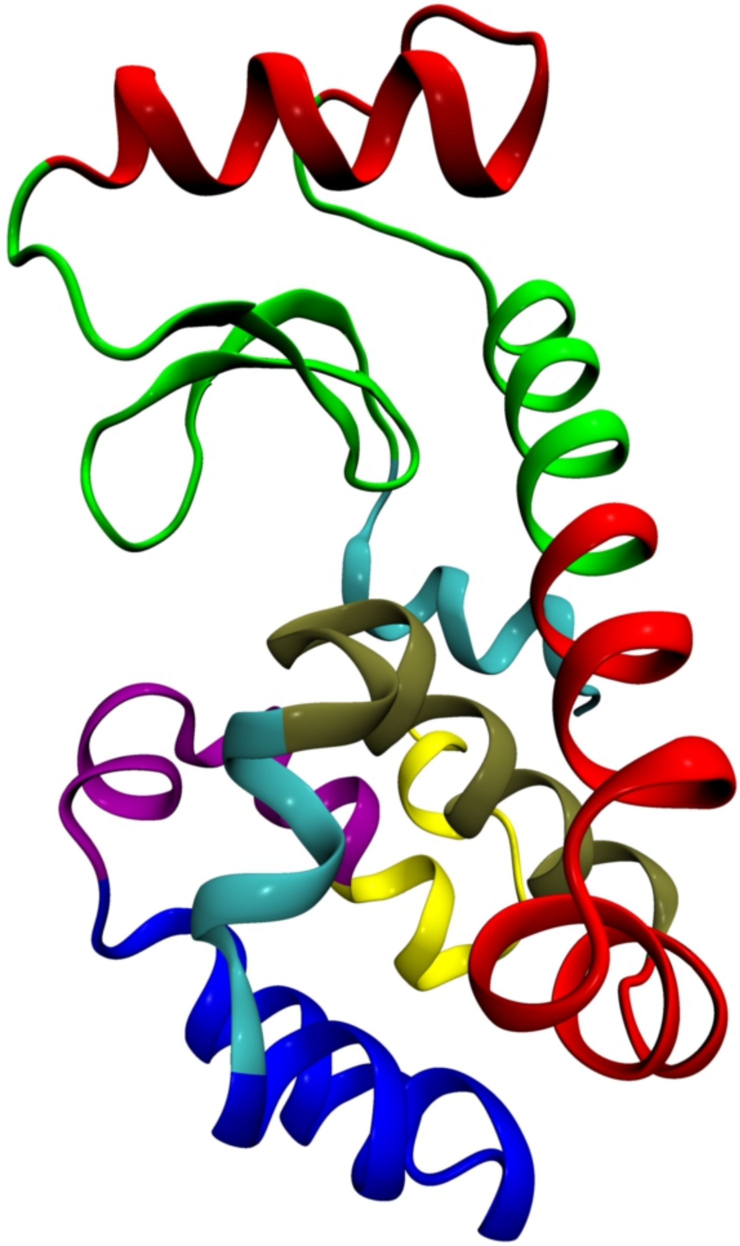


Figure 3.17 Residue clusters based on the correlation matrix for MD simulation without solvent. Clusters generated using the UPGMA algorithm were selected based on including binding site residues. These clusters imply groups of residues with concerted conformational fluctuation that could be important for inter-domain communication. Different clusters are indicated by color.

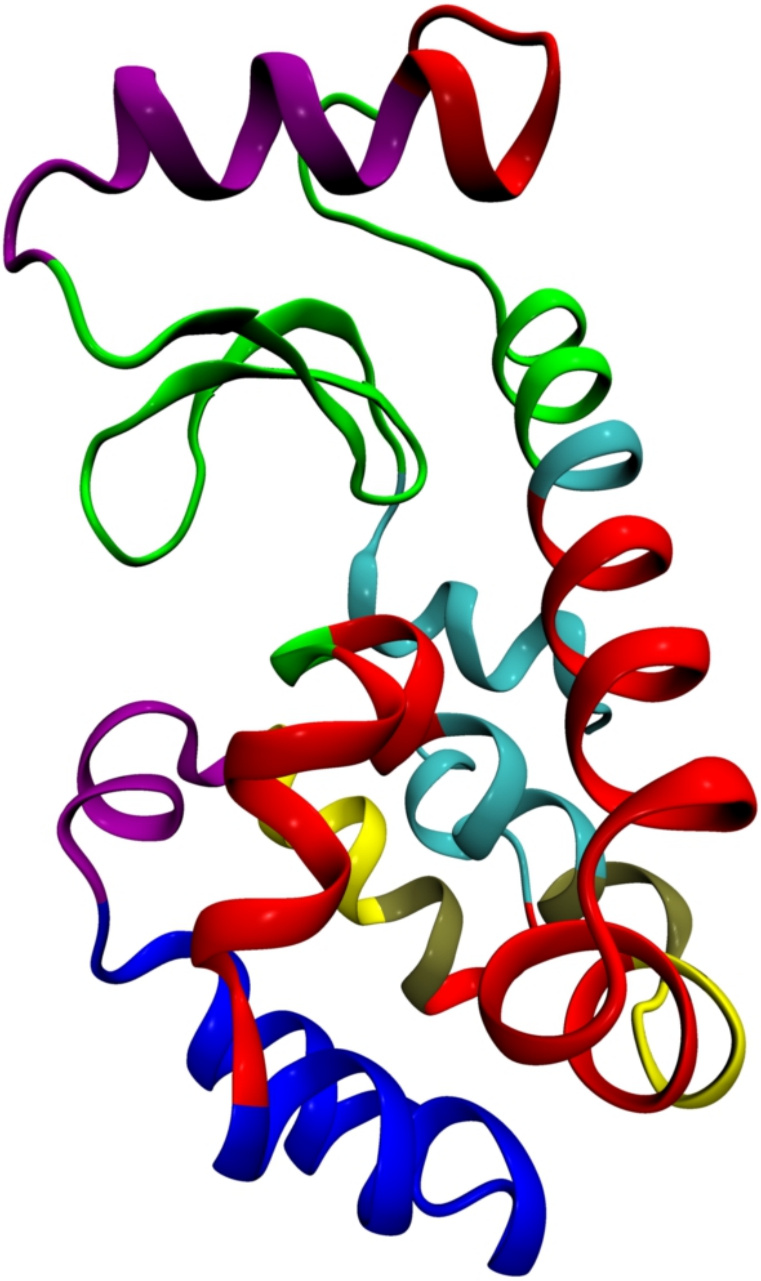


Figure 3.18 Residue clusters based on the correlation matrix first 10 PCA modes. Clusters generated using the UPGMA algorithm were selected based on including binding site residues. These clusters imply groups of residues with concerted conformational fluctuation that could be important for inter-domain communication. Different clusters are indicated by color.

CHAPTER 4

HEMOGLOBIN-HEMOGLOBIN INTERACTION ANALYSIS

It is difficult to theoretically explain and predict the conformational changes of proteins that occur during binding or interaction. Conformational change is not accessible to molecular dynamics simulations today due to their high computational costs. Therefore other computational methods must be applied. Previous studies (Harrison 1984; Tama *et al.* 2000) show that some of the lowest-frequency normal modes of several proteins, including hemoglobin, are strongly correlated with the large amplitude conformational change of these proteins. In this chapter, the MD based PCA method previously described is applied to large-scale (low-frequency) protein-protein interaction, ignoring local (high-frequency) motion. This method reveals that the dominant motion of aggregation of sickled hemoglobin molecules is caused by hydrophobic effects. This discovery demonstrates that it is feasible to model slow motions due to hydrophobic effects with MD based PCA.

The MD-PCA framework is used to analyze hemoglobin-hemoglobin interaction in this chapter. An artificial environment resembling sickled hemoglobin interaction is built. The simulation produces a noticeable “aggregation” process. The dominant modes calculated using PCA in this simulation show a damping-approach process during hemoglobin-hemoglobin interaction. Furthermore, this interaction displays a “softball-like” shape change in each hemoglobin molecule. That is, the atomic motions have greater amplitude opposite the point of interaction. The amplitude diminishes along with the distance to the “interaction-spot.” This suggests that there is not a single “attractive force bond” that causes the aggregation of hemoglobin, which matches the results of (Israelachvili and Wennerstrom 1996; Lum *et al.* 1999; Scatena *et al.* 2001; Wallqvist *et al.* 2001).

4.1 Hemoglobin and Sickle Cell Anemia

Hemoglobin is the main ingredient of red blood cells (RBC) in vertebrate blood. Hemoglobin is an iron-containing oxygen-transport metallo-protein. The primary structure of hemoglobin, like other protein molecules or polypeptides, is a long chain of many monomers or amino acids. The well known secondary structures are α helices and β sheets. The α- and β-globins are derived from gene clusters. The α-globin genes are on chromosome 16 and the β-globin genes are on chromosome 11. Both gene clusters contain not only the major adult genes, α and β, but other expressed sequences that are utilized at different stages of development. The orientation of the genes in both clusters is in the $5'$ to $3'$ direction, with the earliest expressed genes at the $5'$ end. These basic structures eventually form tertiary and quaternary structures. In humans, the hemoglobin molecule is most often an assembly of four globular protein sub-units, as shown in Fig. 4.1. Each sub-unit is composed of a protein chain tightly associated with a non-protein heme group. Each heme group contains one iron atom that is directly responsible for the transport of oxygen from the lungs to other tissues and transport of carbon dioxide from tissues to the lungs. The four polypeptide subunits are bound to each other by salt bridges, hydrogen bonds, and hydrophobic interaction. There are two kinds of contacts between the α and β chains: $\alpha_1\beta_1$ and $\alpha_2\beta_2$. The predominant hemoglobin in adulthood is HbA ($\pm 97\%$), which consists of two α and two β globin chains ($\alpha_2\beta_2$). Other hemoglobin types are HbA_2 (2 to 3.5%; $\alpha_2\delta_2$) and HbF ($< 2\%$; $\alpha_2\gamma_2$).

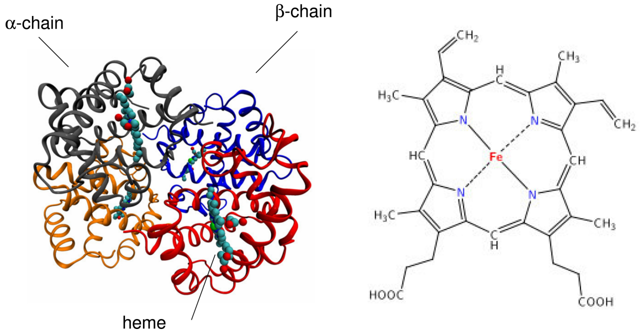


Figure 4.1 Quaternary structure of hemoglobin and its oxygen carrier heme

Although secondary and tertiary structures of various hemoglobin subunits are similar, reflecting extensive homology in amino acid composition, the variations in amino acid composition that do exist impart marked differences in hemoglobin's oxygen carrying properties. In addition, the quaternary structure of hemoglobin leads to physiologically important allosteric interaction between the subunits. One specific mutation in the genes for the hemoglobin protein results in a series of consequences – from the type of amino acid to the conformation or shape and the function of the complete hemoglobin molecule, which eventually leads to sickle cell anemia.

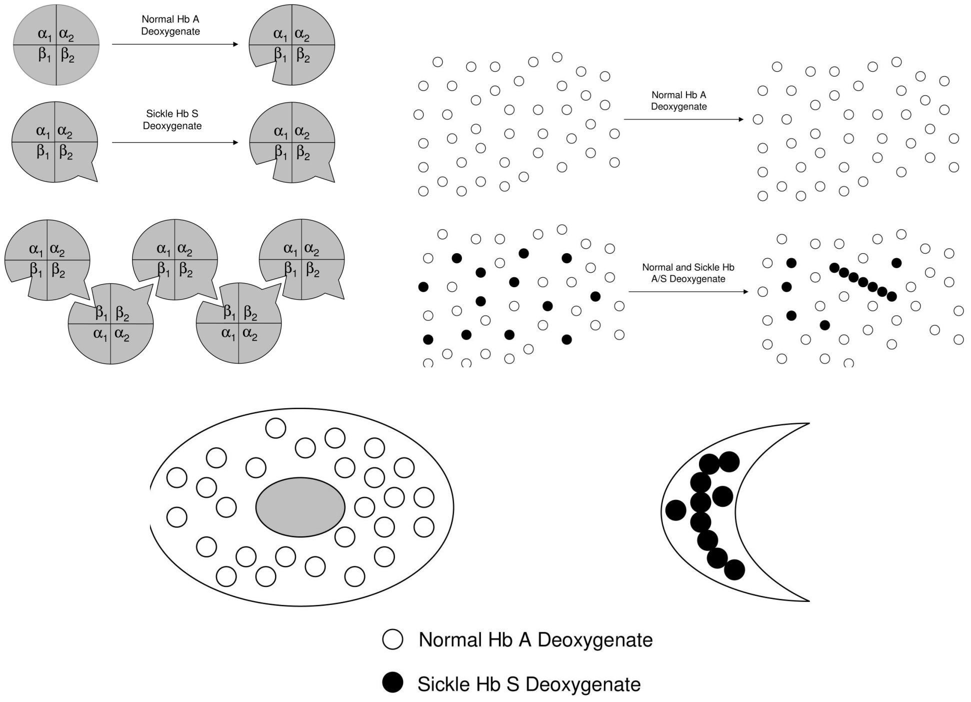

Normal Hb A Deoxygenate
Sickle Hb S Deoxygenate
Normal and Sickle Hb A/S Deoxygenate
α1 α2
β1 β2
○ Normal Hb A Deoxygenate
● Sickle Hb S Deoxygenate

Figure 4.2 An illustration of the genetic cause of macroscopic property changes of RBCs

This particular point mutation – β_6 – replaces glutamic acid with the less polar valine at the sixth position of the β chain, forming an abnormal globin: β^s. This results in the formation of "sickle hemoglobin" or HbS ($\alpha_2\beta_2^s$).

In Fig. 4.2, hemoglobin tetramers are represented as circles. Each quarter corresponds to one protein subunit. The β_6 mutation is indicated as a protrusion from the circle in the β_2 subunit and the hydrophobic pocket is a kink in the β_1 subunit. The interaction of different HbS tetramers and the formation of the fiber or chain in the deoxygenated state is also illustrated in Fig. 4.2. Upon deoxygenation, HbS has a certain probability of forming a hydrophobic interaction with an adjacent S globin, ultimately resulting in the polymerization of HbS. The hydrophobic residue provides a nucleating site for protein-protein interaction or “sticky patch.” Figure 4.3 provides an illustration of this docking phenomenon.

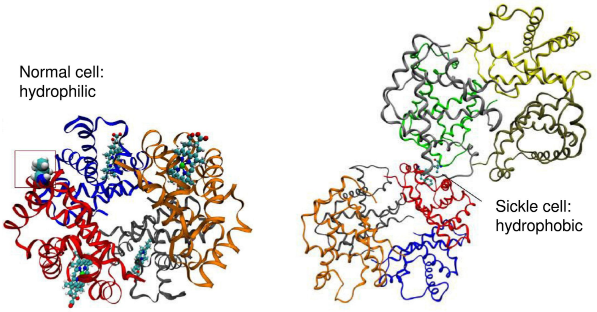
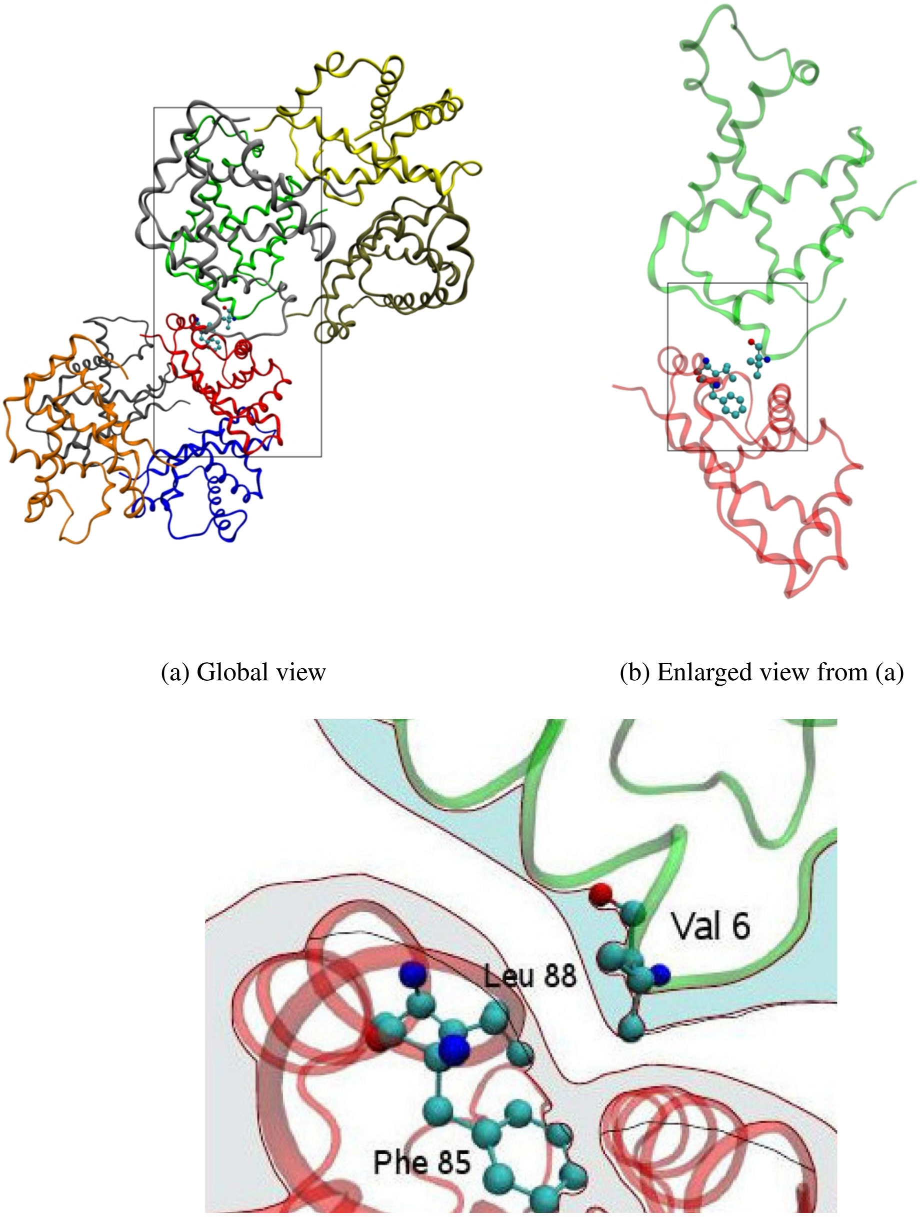


Figure 4.3 Illustration sickle mutation alters a hydrophilic β-globin site to a hydrophobic one

In essence, the underlying problem in sickle cell anemia is that the valine for glutamine substitution results in hemoglobin tetramers that aggregate into arrays upon deoxygenation. This aggregation leads to deformation of the red blood cell, making it relatively inflexible and unable to traverse the capillary beds. Repeated cycles of oxygenation and deoxygenation lead to irreversible sickling. The end result is clogging of the fine capillaries. Because bones are particularly affected by the reduced blood flow, frequent and severe bone pain results. This is the typical symptom of a sickle cell crisis. Long term recurrent clogging of the capillary beds leads to damage to the internal organs, in particular the kidneys, heart and lungs. The continual destruction of the sickled red blood cells leads to chronic anemia.

Under anaerobic conditions, sickle cell hemoglobin (HbS) polymerizes into highly elongated cables. Such polymers distort the shape and suppleness of RBCs, resulting in a sickle-like appearance as compared to the donut shape of normal RBCs. The rigid and



(a) Global view

(b) Enlarged view from (a)

(c) Donor and acceptor, enlarged view from (b)

Figure 4.4 Donor and acceptor during deoxgenated states

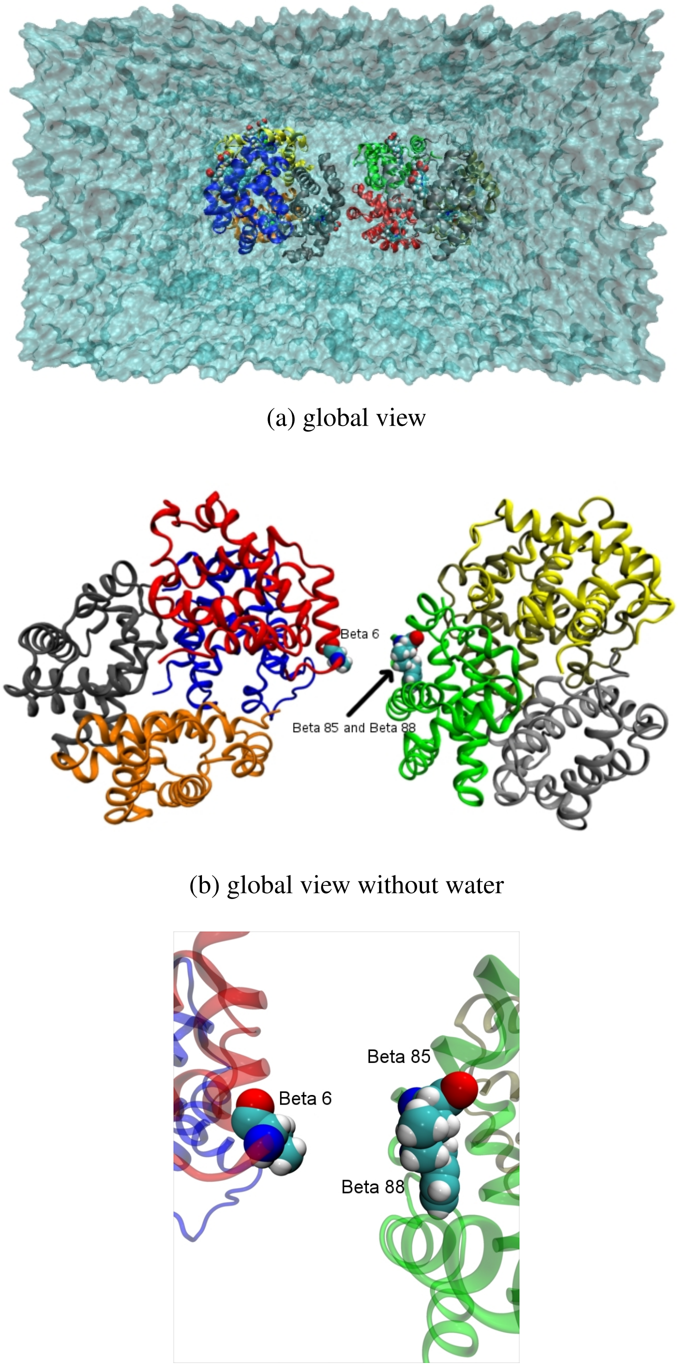
distorted sickled RBC has difficulty passing through the small capillary, thereby blocking blood flow. However, only deoxy-HbS, and not oxy-HbS, polymerizes. This is consistent with the fact that RBC sickling occurs in the capillaries where the O_2 concentration is relatively low and the deoxy-HbS concentration is relatively high. The sickle cell phenotype arises from a single mutation in the β-globin gene resulting in an amino acid substitution at the sixth residue of the β-chain with β-Val6 in HbS substituted for β-Glu6 in normal HbA. The hydrophobic β-Val6 side chains are exposed on the surface of two β-chains of HbS and they can fit into hydrophobic pockets created by the side chains of β-Phe85 and β-Leu88, also on the β-chain surface (Adachi *et al.* 1994; Jensen *et al.* 2004), as shown in Fig. 4.4. The spacing between these residues is such that deoxy-HbS molecules self-associate and polymerize. However, the geometrical spacing between these residues is different for the oxy-HbS conformer and it does not polymerize.

This red blood cell system (sickle or normal) was chosen as a test system for a multiscale and multi-physics modeling protocol because there is a wealth of clinical data for sickle cell diseases and, more importantly, because the sole cause of this hereditary disease, a mutation of a base pair in the β-globin gene, is known. Hydrophobic effect is the molecular level cause and it is a slow motion during protein-protein interaction. The test sought to determine whether this MD based PCA method aim to reveal this conformational change.

4.2 Molecular Dynamic Simulation with NAMD

The threshold for sickled hemoglobin aggregation is unknown, which makes it difficult to simulate this process. An artificial system was built by separating two docked mutated hemoglobin molecules a small distance apart (4Å ~5Å) along with the axis of their center of mass. The initial docked configuration was obtained from Kavanaugh *et al.* (1993) (PDB ID: 1BZ0). The initial system is illustrated in Fig. 4.5.

The PSF package in VMD (Humphrey *et al.* 1996) is used to generate a structure



(a) global view

(b) global view without water

(c) a closer look of the pocket

Figure 4.5 Initial simulation position for HbS

file (PSF file) for MD simulation. First, the coordinates missing in the crystal structure are reconstructed. The minimized structure is solvated in a periodic truncated cubic simulation box of $111,738$ TIP3P water molecules, providing a minimum of 15Å of water between the protein surface and any periodic box edge. The water box is large enough that the protein does not interact with its image in the next cell. The water box is rotated to minimize the system size.

Ions are placed in the water to represent a more typical biological environment. They are especially necessary because the protein being studied carries six excess negative charges. Six extra sodium ions are added to make the system neutral. The ions present shield the regions of the protein that carry the charge and make the entire system more stable. They are placed in regions of potential minima, since they will be forced into those regions during the simulation anyway. Sodium and chloride ions are added to neutralize the total system and achieve a salt concentration of 0.05 mol/L. Minimum distances between ions and the protein molecule as well as between any two ions are set to 5Å.

A periodic boundary condition was selected for simulation with solvent. The use of a periodic boundary condition involves surrounding the system under study with identical virtual unit cells. The atoms in the surrounding virtual systems interact with atoms in the real system. These modeling conditions effectively eliminate surface interaction of the water molecules and create a more faithful representation of the in *vivo* environment than a water sphere surrounded by vacuum provides (Bhandarkar *et al.* 2008).

Energy minimization involves searching the energy landscape of the molecule for a local minimum. MD minimization and equilibration simulation involve more than one minimization-equilibration cycle, often fixing and releasing molecules in the system.

Energy minimizes in $180\,fs$ equilibrates with the atoms in the protein fixed in space in $120\,fs$, minimizes the system again in $60\,fs$ and equilibrates in $120\,fs$ again, this time with the protein free to move. Fixing the protein allows the water, which typically responds much faster to forces than the protein, to do the relaxing in the first step. This saves computational effort and prevents the introduction of artifacts from an unstable starting structure. A final 100 *fs* energy minimization reaches the "minimized" position. The parameter settings are shown in Table 4.1.

Table 4.1: MD Simulation Parameter Setting for Hemoglobin Interaction

Item	Setting
periodic boundary conditions	yes
cutoff distance	12Å
switch distance	10Å
pair list distance	13.5Å
timestep	2.0
rigid bonds	all
non-bonded frequency	2

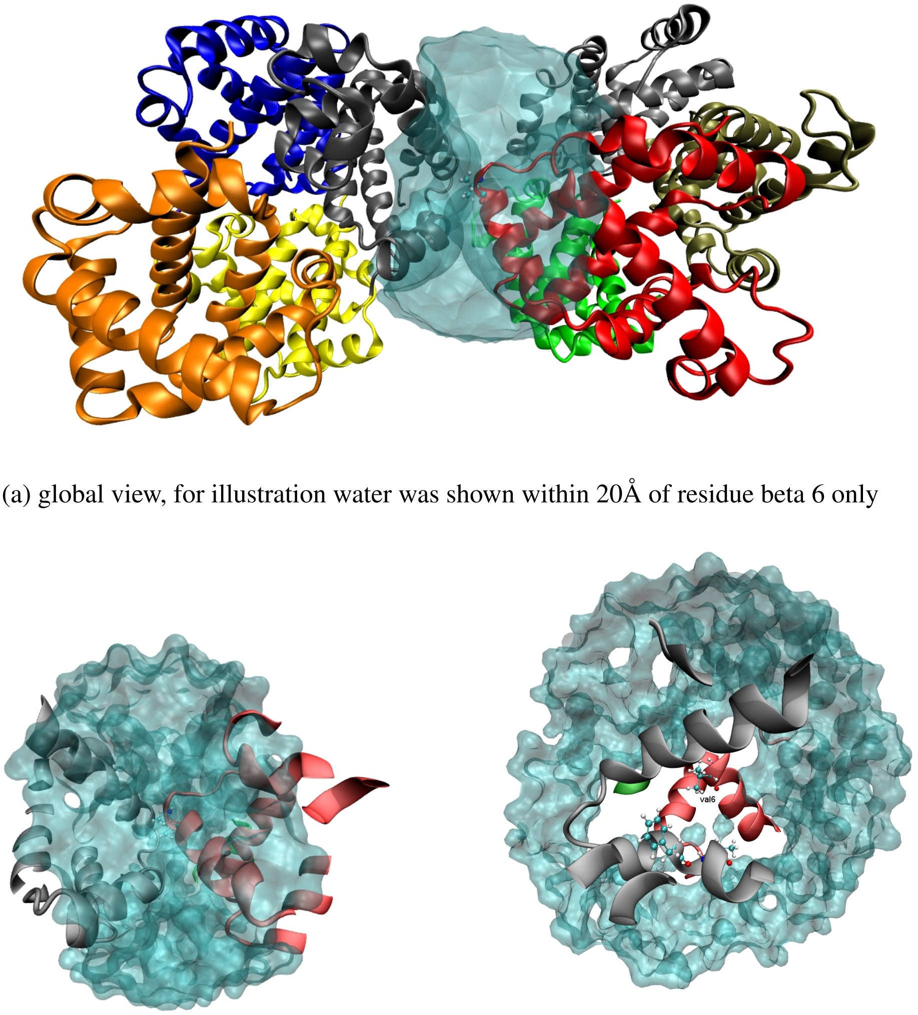
It is important to check if this system has the possibility of docking. After energy minimization, a vacuum space forms between β-Val6 side chains and the surface of another β-Phe85 and β-Leu88 (shown in Fig. 4.6). This leads to a docking process of the two hemoglobin molecules.

There were no energy changes during last $1,000$ timesteps in energy minimization. Following minimization, the entire system was heated in 5 K increments up to 310 K with micro-canonical ensemble (NVE) equilibration per temperature step. The production simulation was conducted at 310 K. The trajectory was calculated with one femtosecond (fs) timesteps using SHAKE (Ryckaert *et al.* 1977) constraints on hydrogen-heavy atom bonds.

In order to represent an environment more typical of the human body, a constant temperature of 310 K was set. The temperature is controlled by Langevin dynamics deploying a damping frequency of 5 picosecond^{-1} to all non-hydrogen atoms. Langevin dynamics are a means of controlling the kinetic energy of the system and so controlling the system temperature and/or pressure. The method uses the Langevin equation for a single particle:

$$m_i\frac{d^2x_i(t)}{dt^2} = F_ix_i(t) - \gamma_i\frac{dx_i(t)}{dt}m_i + R_i(t) \tag{4.1}$$

The second term on the right side represents a frictional damping that is applied to the



(a) global view, for illustration water was shown within 20Å of residue beta 6 only

(b) closer view

(c) Rotation by 90° from (b)

Figure 4.6 Illustration for energy minimization position with HbS simulation

particle with frictional coefficient $\gamma_i m_i$. The third term represents random forces that act on the particle (as a result of solvent interaction). These two terms are used to maintain particle kinetic energy to keep system temperature at a constant value.

The particle Mesh Ewald Method (PME) (Darden *et al.* 1993; Essmann *et al.* 1995) is used to compute the electrostatic forces. The grid spacing is kept below 1.0Å. Multiple time stepping is carried out using the Verlet impulse/r-RESPA integration scheme (Den Otter and Briels 2000). Van der Waals interaction is truncated at 13.5Å using a switching function starting at 12.0Å. Full periodic boundary conditions are imposed. The coordinates and energy are saved every 1 picosecond.

4.3 Results

The whole process of hemoglobin-hemoglobin interaction could be long. MD simulation works at a nanosecond scale and cannot reveal this whole process. In addition, microscopic details are missed by experiments, such as the behavior of hydrogen bonds and the role of water, could be studied and explored with slow motion analysis.

4.3.1 Conformational Change Analysis

The beginning approach between the two hemoglobin molecules occurs during the "heating" stage. Figure 4.7 shows the distance between the centers of mass of two HbS. The first 660 *ps* timesteps are the heating stage, followed by 60 *ps* equilibrating steps. A rapid drop can be noticed during the heating stage. After their initial displacement, these two molecules were pushed toward each other, and then reached a local equilibration. The distance during the last 9 *ns* shows a "damping" stage between two molecules.

Figure 4.8 shows the results of RMSD during regular simulation, that is, after the heating and equilibrating stages. RMSD is a measure of equilibration and simulation stabil-

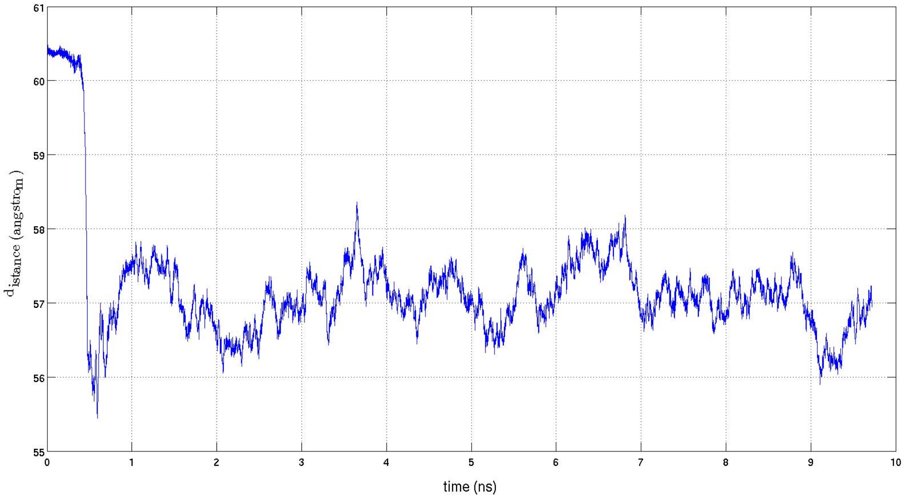


Figure 4.7 Distance of center of mass of two hemoglobin molecules

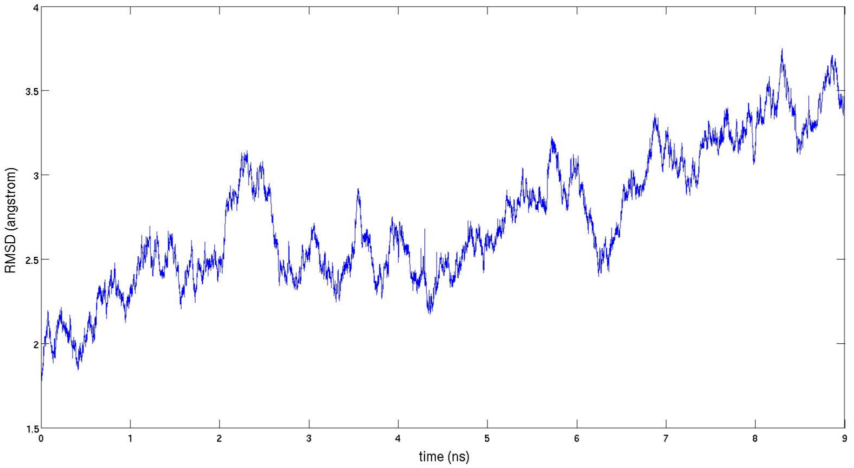


Figure 4.8 RMSD of HbS simulation

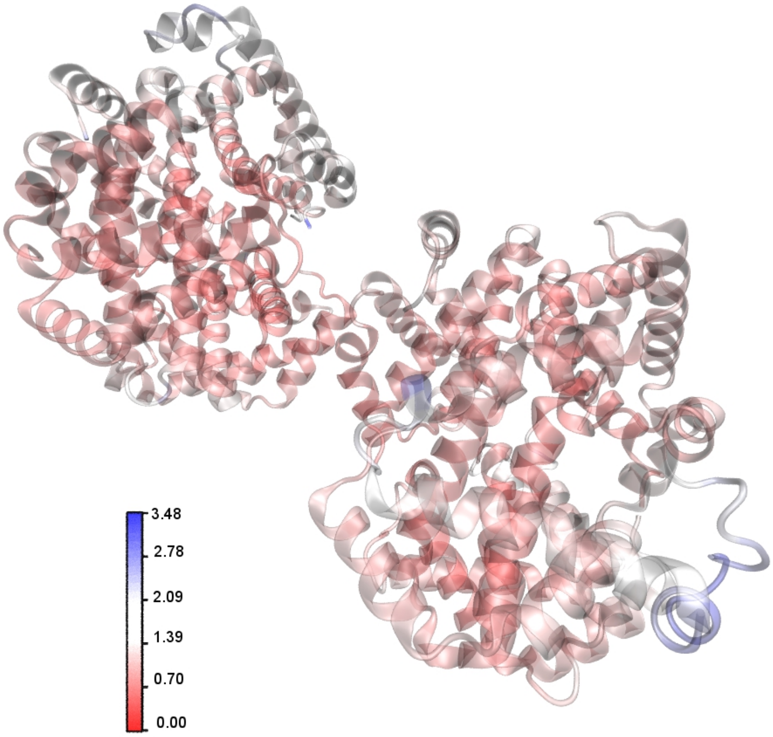
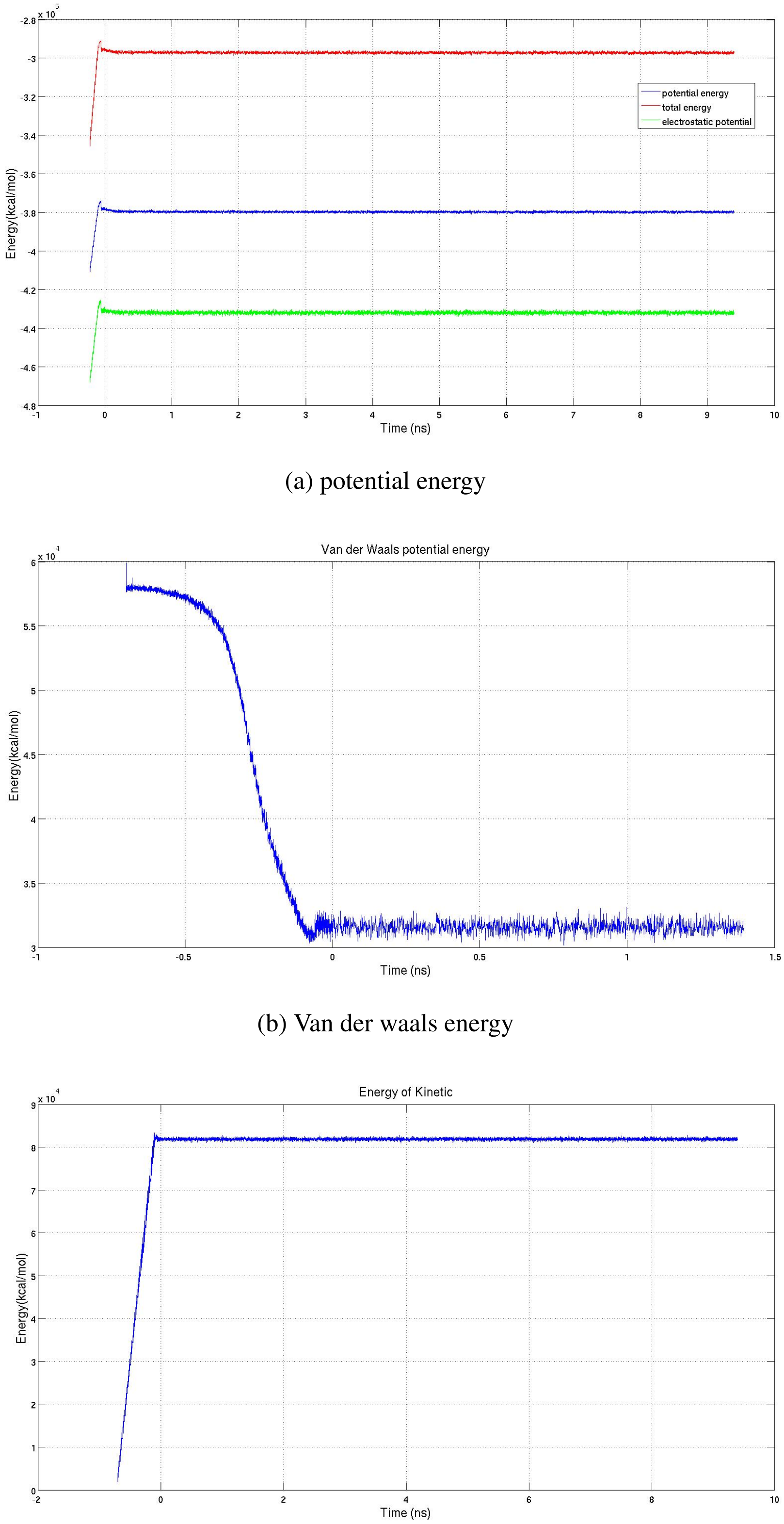


Figure 4.9 Illustration of RMSFs for hemoglobin interaction simulation, colored by RMSF value for each residue.



(a) potential energy

(b) Van der waals energy

(c) kinetic energy

Figure 4.10 Energy plotting, heating and equilibrating stage shown before timestep 0.

ity. Although the distance doesn't change during the last $9ns$, the RMSD changes indicates continued conformational changes.

Figure 4.9 illustrates the conformational changes for each residue. Blue areas indicate higher RMSF scores and red reflect lower RMSF scores. This shows that the most conformational fluctuation occurs on the surface at the far end of each hemoglobin. The insides of each hemoglobin are more rigid than the surface, except at the point of interaction. Therefore, the dynamic motion of each hemoglobin is not a rigid body motion, but more like the motion of a soft ball.

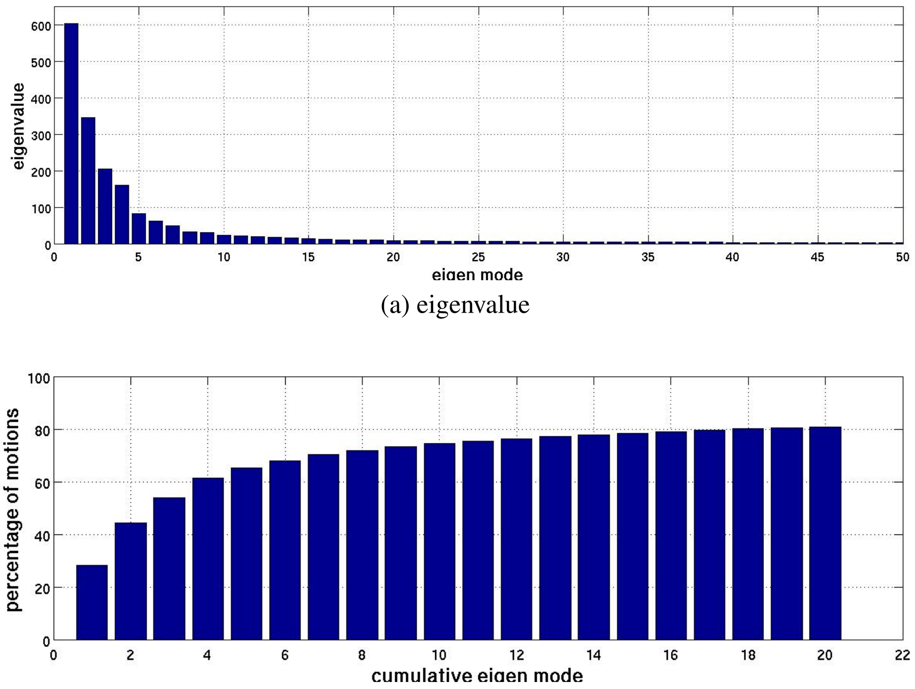
This method enable the study of conformational changes by analysis of energy as well. Figure 4.10 shows the energy calculation. To separate the heating stage from regular simulation, the start of regular simulation is set as time zero. During the heating stage, both kinetic energy and potential energy increase. During the heating stage, the velocities are reassigned so these energies do not reflect the conformational changes.

Van der Waals (VDW) forces are relatively weak compared to normal chemical bonds, but play a fundamental role. Figure 4.10b shows the calculation of VDW potential. As the two molecules approach, VDW increases until the so-called strong repelling position. VDW then decreases while the distance increases. After equilibration, VDW and other potentials remain stable within a local potential well. However, the conformational changes have not ended. These two molecules could reach another equilibration place and relative energies will continue to change as well.

4.3.2 PCA Analysis

The principle components of the MD results are calculated using the procedure explained in Chapter 2. The eigenvalues are shown in Fig. 4.11. In Fig. 4.11a, the "elbow" turn is at the $5th$ eigenvalue and the cumulative energy content of the first five eigenvalues is 63%. The first five eigenmodes are selected as the dominant modes for study.

The lowest mode shapes computed using PCA are projected onto the energy-mini-



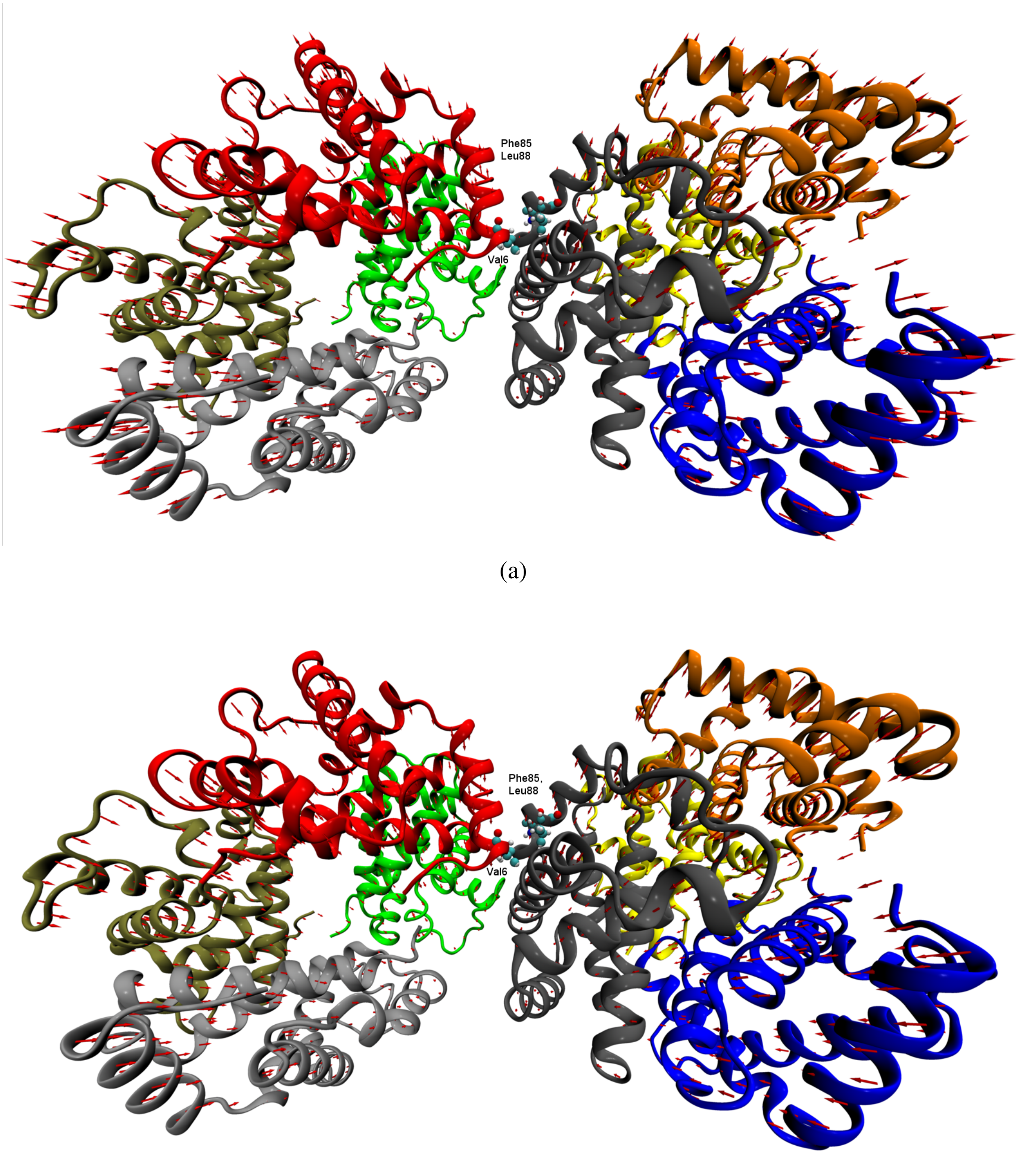
(a) eigenvalue

(b) cumulative energy content for each eigenvector

Figure 4.11 PCA eigenvalues calculated with α-carbon for hemoglobin interaction

mized structures of hemoglobin in Fig. 4.12 to visualize the directional behavior of their large length-scale collective motion. The lowest mode shapes of proteins describe their most typical conformational changes, which are often directly related to their function.

Based on the lowest mode shape, each hemoglobin molecule oscillates around the point of contact. The whole molecule is not rigid during this oscillation. The residues on the contact chain, namely chains B and H, shown in Fig. 4.12, move with smaller amplitude; the residues on the outer chains, namely chains A, C, E and F, move with larger amplitude. Furthermore, the farther the distance from the point of contact, the larger the amplitude of the motion of each residue. This result shows that the driving force for hemoglobin-hemoglobin aggregation is not a bonding force. Rather, the water environment plays an important role in the motions – which are a hydrophobic effect. The entire surface of the hemoglobin molecules are involved in this effect.



(a)

(b)

Figure 4.12 Hydrophobic motion captured by PCA analysis, colored by chains: chain A blue, chain B gray, chain C orange, chain D yellow, chain E tan, chain F silver, chain G green, chain H red. the red arrows indicate the direction of displacement based on 1^{st} eigenmode, the length of arrows indicate the amplitude.

4.3.3 Correlation Analysis with HbS Interaction

A correlation matrix showing the clustering of the motions of residues was produced, as shown in Section 2.3. This indicates a possible interaction between groups of residues. The correlation matrix is bases on Eq. (2.11). The clusters were identified by the UPGMA algorithm introduced in Section 2.3. Figure 4.13 shows the correlation map and Figure 4.14 illustrates the clusters, with different clusters indicated by color.

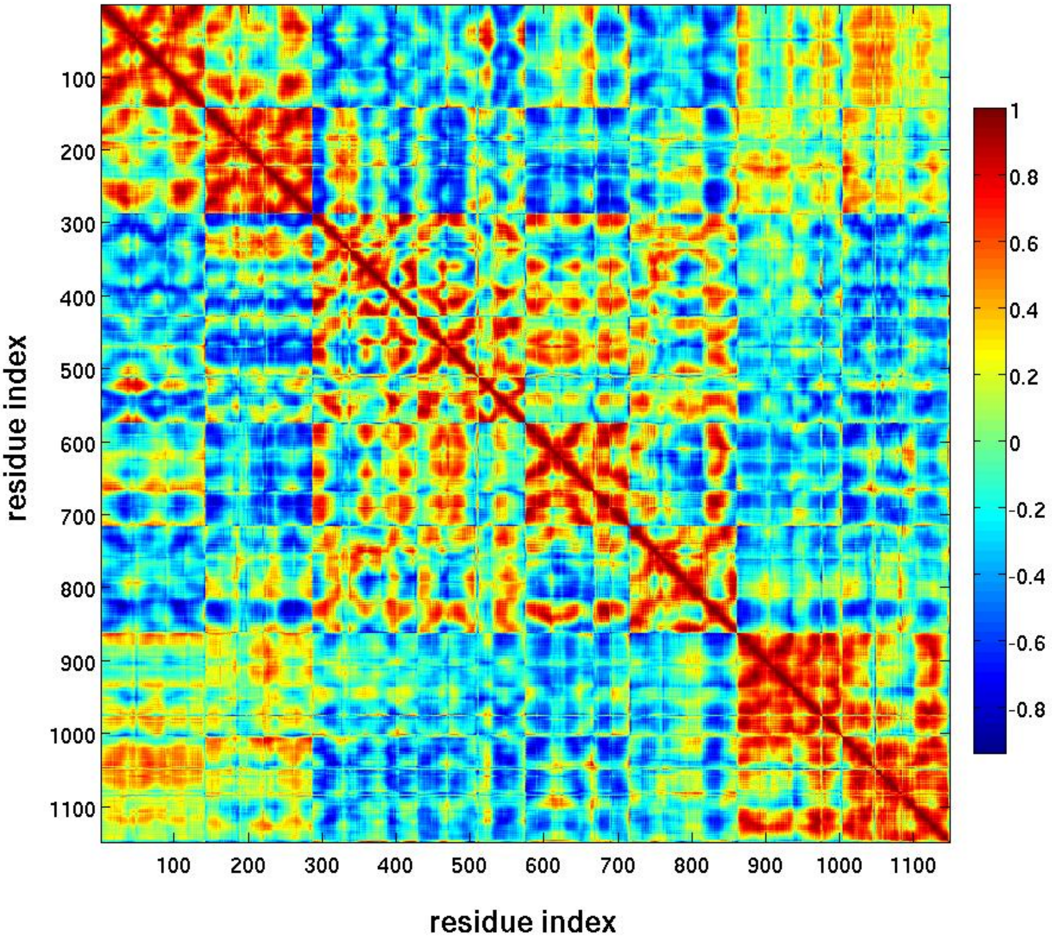


Figure 4.13 Correlation of HbS result

The hydrophobic "pocket" formed by residue β-Val6 on chain H and residue β-Phe85 and β-Leu88 on chain B, has been studied. The correlated residues with β-Phe85 and β-Leu88, which are in the same cluster, are in yellow; whereas the correlated residues

β-Val6 are in red. The yellow cluster indicates that the motion related to β-Phe85 and β-Leu88 come from outside α-helices. The driving "forces" are horizontal, which tend to "push" the protein molecules closer. On the other hand, the red cluster has vertical motion and contribute shear stress to the docking place. This stress may leat to the seperation of these two molecules. The heme near the docking spot plays an important role. It may contribute resistance to the shear stress. The formation of this heme is important. Hemoglobin aggregation occurs only during a deoxygenate state. The possible reason is the heme during deoxygenated stage which supports the stability of the docking pocket. Of course, this hypothesis needs further validation by experiments or other methods such as quantum mechanical calculation.

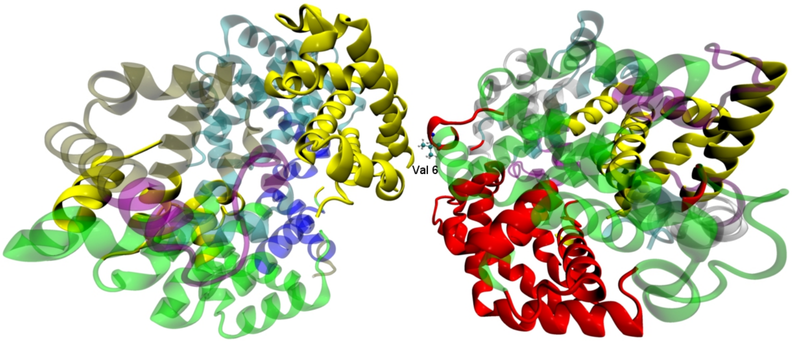


Figure 4.14 Residue clusters based on the correlation matrix first 5 PCA modes. Clusters generated using the UPGMA algorithm were selected based on including binding site residues. These clusters imply groups of residues with concerted conformational fluctuation that could be important for inter-domain communication. Different clusters are indicated by color.

4.4 Conclusion

In this chapter, the MD-PCA framework is used to study sickled hemoglobin-hemoglobin interaction. The fine details of protein motion are produced to predict the hemoglobin-hemoglobin interaction. A correlation matrix based on clustering of residue motion is produced. The clustering suggests a possible explanation for hemoglobin aggregation and points to the role of water during this intermediate state. These analyses provide a starting point for further study and for exploring protein conformational changes using the micromechanical and elastic properties of protein.

CHAPTER 5

PARALLEL PROGRAMMING

The MD-PCA framework is feasible for protein normal modes analysis. With increasing protein size and simulation time, this framework becomes increasingly computationally expensive. Therefore, in this work the computation intensive parts, such as calculation of covariance and eigenvalues, are implemented using parallel programming.

Parallel programming has become increasingly important as the design of processes has shifted to parallel computing in the form of multicore processors. Frequency scaling (or frequency ramping) was the dominant force in commodity processor performance increases from the mid-1980s until roughly the end of 2004. However increasing power consumption and heat generation of processors limits such development. To further improve performance, most computer programs must be implemented in parallel.

The main principle of parallel programming is to divide large problems into smaller ones that are not mutually dependent, which are then solved concurrently ("in parallel"). Parallel programming uses multiple processing elements simultaneously to solve a problem. This is accomplished by breaking the problem into independent parts so that the processing elements can execute their respective parts of the algorithm simultaneously. The processing elements can be diverse and include resources such as a single computer with multiple processors, several networked computers, specialized hardware, or any combination of the above. There are four modes for parallel computing:

- SIMD - Single Instruction Multiple Data. Processors are “lock-stepped”: each processor executes a single instruction synchronously on different data.

- SPMD - Single Program Multiple Data. Processors run asynchronously using a local copy of a program.

- MIMD - Multiple Instruction Multiple Data. Processors run asynchronously: each processor has its own data and its own instructions.

- MPMD - Multiple Program Multiple Data. Multiple autonomous processors simultaneously execute at least two independent programs.

In this work, parallel programming was implemented in programs written in C++ with Message Passing Interface (MPI). MPI is an application programming interface (API) specification that allows computers to communicate with one another. It is a language-independent communication protocol used to program parallel computers. It is also available in most computer clusters and supercomputers, such as the Franklin supercomputer in which the programs were tested.

These applications read and produce large amounts of data, making it very important to manage input and output (I/O) effectively. MPI-2 supports parallel I/O, in which multiple processes of a parallel program access data (reading and writing) from multiple disks. Parallel I/O mitigates the disadvantages of serial I/O, allowing larger datasets, scalability, and speedup.

The input data could be very large. For instance, the binary data file of original trajectories for hemoglobin-hemoglobin interaction, which records the displacement of $19,120$ atoms with $10,000$ snapshots of float data type, occupies approximately 2.13 gigabyte (GB) of memory. The intermediate data for solving eigen problems could be as large as 8 to 11 GB. These datasets are too large to be sent back to one node for serial I/O. Parallel I/O improves the application's scalability. The initial and intermediate data are distributed almost evenly among a cluster of nodes. Parallel I/O helps the application to solve a problem with very large datasets even though every node has limited memory size (2GB).

A math library was used for matrix operations to improve programming performance. BLACS and ScaLAPACK (Blackford *et al.* 1997) are the standard parallel numerical libraries because of their portability and design. They contains an extensive set of parallel routines for common linear algebra calculation that perform well and are scalable.

BLACS routines transfer local data from one processor's memory to another. These routines are wrapper routines, which call a lower-level message-passing library. Typically this is done by MPI itself. The BLACS routines include point-to-point communication, broadcast routines, and "combination" routines for doing "global" calculations (summation, maximum, and minimum) with data residing on different processors. An attractive feature of the BLACS communication routines is that they are array-based, meaning that the data being transferred always consist of an entire array or subarray.

The BLACS library also contains important routines for creating and examining the processor grid, which is expected and used by all ScaLAPACK routines. The processor grid is a 2-dimensional grid whose size and shape is programmer controlled. A processor is identified not by its traditional MPI rank, but rather by its row and column number in the processor grid, similar to MPI virtual topologies.

ScaLAPACK libraries contain routines for more sophisticated linear algebra calculations. These libraries tackle these three types of advanced problems: Solve a set of simultaneous linear equations; eigenvalue/eigenvector problems; linear least squares fitting. As with the basic libraries, many matrix types are supported, including real, double, complex, and double complex. Various algorithms for factorizing the matrices are also available.

The steps to use ScaLapack routines in the programs are listed below (detailed instructions for MPI programming can be found in (Pacheco 1997; Snir *et al.* 1998)):

1. Initialize the BLACS library for use in the program.

2. Create and use the BLACS processor grid.

3. Distribute pieces of each global array over the processors in the grid.

4. Have each processor initialize its local array with the correct values of the pieces of the global array it owns.

5. Call the ScaLAPACK routine.

6. Confirm/use the output of the ScaLAPACK routine.

7. Release the processor grid and exit the BLACS library.

ScaLAPACK uses a 2-dimensional block-cyclic distribution technique to parcel out global array elements onto the processor grid. The block-cyclic technique was chosen for this application because it gives the best load balance and maximum data locality for most of the ScaLAPACK algorithms (Blackford *et al.* 1997). The 2-dimensional block-cyclic distribution was accomplished by following five steps:

1. Divide the global array into blocks with *mb* rows and *nb* columns. Think of the global array as composed only of these blocks.

2. Present the first row of array blocks across the first row of the processor grid in order. If the processor grid columns are exhausted, cycle back to the first column.

3. Repeat with the second row of array blocks, with the second row of the processor grid.

4. Continue this method for the remaining rows of array blocks.

5. If processor grid rows are exhausted, cycle back to the first processor row and repeat.

The framework developed by this research is computationally intensive. It involves linear algebra calculations with very large matrices. It is not executable on most single computer configurations because of the memory limitation. Parallel computing provides a way to efficiently perform these kinds of computations on a large scale.

CHAPTER 6

CONCLUSION AND FUTURE WORK

An effective computational procedure based on MD-PCA framework is proposed to calculate the normal modes of proteins. QHA and PCA are compared, showing QHA to be better on slow motion analysis because it filters out the high frequency motion of hydrogen atoms. A method of clustering residue motions based on a correlation matrix is implemented. This provides a starting point for experimental and computational studies designed to further elucidate conformational changes in proteins.

This methodology was applied with a mutant T4 phage lysozyme. The equilibrium thermal fluctuation of all protein atoms and their inter-residue correlations established by this method are in agreement with published results arrived at by NMA and finite element methods. The eigenmodes captured by this method are in quantitative agreement with other computational results.

With this success on the analysis of a small protein, the methodology was applies to hemoglobin-hemoglobin interaction, the details of which at the atomic level are unknown. The method revealed a slow motion sickled hemoglobin-hemoglobin interaction, which is the molecular underpinning of sickle cell anemia. Parallel computing applications using C++ and MPI reduce the time costs of the computations.

This work estabishes the utility of the MD-PCA method to compute protein normal modes for a limited range of biomolecules. To evaluate its effectiveness fully, the present framework should be applied to a considerably broader set of proteins of variable structure and chemical composition. The study of hemoglobin-hemoglobin interaction suggests that the heme plays an important role in hemoglobin docking by stabilizing the docking pocket. Further study is called for, using the micromechanical and elastic properties of the protein, to confirm this claim.

APPENDIX A

INTRODUCTION OF NAMD

NAMD (NAnoscale Molecular Dynamics) package was selected for MD simulation (Bhandarkar *et al.* 2008; Phillips *et al.* 2005). Simulation of this complex system involves a large number of molecules. One way to create such simulation is to utilize parallel computers. In recent years, distributed memory parallel computers have offered cost-effective computational power. NAMD was designed to run efficiently on such parallel machines for simulating large molecules. NAMD is particularly well-suited to the increasingly popular Beowulf-class PC clusters, which are quite similar to the workstation clusters for which is was originally designed. Future versions of NAMD will also make efficient use of clusters of multi-processor workstation or PCs.

A.1 NAMD Feature

NAMD has several important features:

- **Force Field Compatibility** The force field used by NAMD is the same as that used by the programs CHARMM (Brooks *et al.* 1983) and X-PLOR (Brünger 1992). This force field includes local interaction terms consisting of bonded interaction between 2, 3, and 4 atoms and pairwise interaction including electrostatic and van der Waals forces. This commonality allows simulation to migrate between these three programs.

- **CMAP Correction** CMAP is an energy correction map based on quantum mechanical calculation and is one of the latest addition to the CHARMM force field. It improves protein backbone behavior and thus yields more accurate dynamic properties for the protein. For example, CHARMM22 with CMAP gives the experimentally observed α-helix while the CHARMM22 force field without CMAP gives a π-helix for certain model peptides.

- **Efficient Full Electrostatics Algorithms** NAMD incorporates the Particle Mesh Ewald (PME) algorithm, which takes the full electrostatic interaction into account. It is a useful method for dealing with electrostatic interaction in the system when periodic boundary condition are present. The Ewald sum is an efficient way of calculating long range forces in the periodic system. The particle mesh is a 3-D grid created in the system over which the system charge is distributed. From this charge, potentials and forces on atoms in the system are determined. This algorithm reduces the computational complexity of electrostatic force evaluation from $O(N^2)$ to $O(NlogN)$

- **Multiple Time Stepping** The velocity Verlet integration method is used to advance the position and velocity of the atoms in time (Allen and Tildesley 1989). To further reduce the cost of the evaluation of long-range electrostatic forces, a multiple time step scheme is employed. The local interaction (bonded, van der Waals and electrostatic interaction within a specified distance) are calculated at each time step. The longer range interaction (electrostatic interaction beyond the specified distance) are computed less often. This amortizes the cost of computing the electrostatic forces over several time steps. A smooth splitting function is used to separate a quickly varying short-range portion of the electrostatic interaction from a more slowly varying long-range component. It is also possible to employ an intermediate time step for the short-range non-bonded interaction, performing only bonded interaction at every time step.

A.2 NAMD File Format

- **PDB Files** The term PDB can refer to the Protein Data Bank http://www.rcsb.org/pdb/, to a data file provided there, or to any file following the PDB format. The PDB format is used to store coordinate or velocity data being input or output from NAMD. This is the standard format for coordinate data for most other molecular dynamics programs as well, including X-PLOR and CHARMM. Files in the PDB include information such as the name of the compound, the species and tissue from which is was obtained, authorship, revision history, journal citation, references, amino acid sequence, stoichiometry, secondary structure locations, crystal lattice and symmetry group, and finally the ATOM and HETATM records containing the coordinates of the protein and any waters, ions, or other heterogeneous atoms in the crystal. Some PDB files include multiple sets of coordinates for some or all atoms. Due to the limits of X-ray crystallography and NMR structure analysis, the coordinates of hydrogen atoms are not included in the PDB.
- **X-PLOR Format PSF Files** A PSF file, also called a protein structure file, contains all of the molecule specific information needed to apply a particular force field to a molecular system. The CHARMM force field is divided into a topology file, which is needed to generate the PSF file, and a parameter file, which supplies specific numerical values for the generic CHARMM potential function. The topology file defines the atom types used in the force field; the atom names, types, bonds, and partial charges of each residue type; and any patches necessary to link or otherwise mutate these basic residues. The parameter file provides a mapping between bonded and nonbonded interactions involving the various combinations of atom types found in the topology file and specific spring constants and similar parameters for all of the bond, angle, dihedral, improper, and Van der Waals terms in the CHARMM potential function.

  NAMD uses the same protein structure files that X-PLOR does. At this time, the easiest way to generate these files is using X-PLOR or CHARMM, although it is possible to build them by hand. CHARMM can generate an X-PLOR format PSF file with the command “write psf card xplor”.

- **CHARMM19, CHARMM22, and CHARMM27 Parameter Files** A CHARMM forcefield topology file contains all of the information needed to convert a list of residue names into a complete PSF structure file. It also contains internal coordinates that allow the automatic assignment of coordinates to hydrogens and other atoms missing from a crystal PDB file.

  NAMD supports CHARMM19, CHARMM22, and CHARMM27 parameter files in both X-PLOR and CHARMM formats. (X-PLOR format is the default, CHARMM format parameter files may be used given the parameter “paraTypeCharmm on”). For a full description of the format of commands used in these files, see the X-PLOR and CHARMM User’s Manual (Brünger 1992).

- **Parameter Files** A CHARMM forcefield parameter file contains all of the numerical constants needed to evaluate forces and energies, given a PSF structure file and atomic coordinates. The parameter file is closely tied to the topology file that was used to generate the PSF file, and the two are typically distributed together and given matching names.

- **DCD Trajectory Files** NAMD produces DCD trajectory files in the same format as X-PLOR and CHARMM. The DCD files are single precision binary FORTRAN files, so are transportable between computer architectures. They are not, unfortunately, transportable between big-endian (most workstations) and little endian (Intel) architectures. This same caveat applies to binary velocity and coordinate files. The utility programs flipdcd and flipbinpdb are provided with the Linux/Intel version to reformat these files. The exact format of these files is very ugly but supported by a wide range of analysis and display programs.

REFERENCES

Adachi, K.; Konitzer, P.; Paulraj, C. G.; and Surrey, S. (1994). Role of Leu-beta 88 in the hydrophobic acceptor pocket for Val-beta 6 during hemoglobin S polymerization. *Journal of Biological Chemistry*, *269*(26):17477–17480.

Agarwal, Pratul K. (2005). Role of protein dynamics in reaction rate enhancement by enzymes. *Journal of the American Chemical Society*, *127*(43):15248–15256. doi: 10.1021/ja055251s.

Agarwal, Pratul K. (2006). Enzymes: An integrated view of structure, dynamics and function. *Microbial Cell Factories*, *5*(1):2. ISSN 1475-2859. doi:10.1186/1475-2859-5-2.

Agarwal, Pratul K.; Billeter, Salomon R.; Rajagopalan, P. T. Ravi; Benkovic, Stephen J.; and Hammes-Schiffer, Sharon (2002). Network of coupled promoting motions in enzyme catalysis. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(5):2794–2799. doi:10.1073/pnas.052005999.

Allen, M. P. and Tildesley, D. J. (1989). *Computer Simulation of Liquids*. Clarendon Press, New York, NY, USA. ISBN 0-19-855645-4.

Amadei, Andrea; Linssen, Antonius B. M.; and Berendsen, Herman J. C. (1993). Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, *17*(4):412–425. ISSN 1097-0134. doi:10.1002/prot.340170408.

Barton, N. P.; Verma, C. S.; and Caves, L. S. D. (2002). Inherent flexibility of calmodulin domains: A normal-mode analysis study. *The Journal of Physical Chemistry B*, *106*(42):11036–11040. doi:10.1021/jp026692q.

Bathe, Mark (2008). A finite element framework for computation of protein normal modes and mechanical response. *Proteins: Structure, Function, and Bioinformatics*, *70*(4):1595–1609. ISSN 1097-0134. doi:10.1002/prot.21708.

Benkovic, Stephen J. and Hammes-Schiffer, Sharon (2003). A perspective on enzyme catalysis. *Science*, *301*(5637):1196–1202. doi:10.1126/science.1085515.

Bhandarkar, M.; Brunner, R.; Chipot, C.; Dalke, A.; Dixit, S.; Grayson, P.; Gullingsrud, J.; Gursoy, A.; Hardy, D.; Hénin, J.; Humphrey, W.; Hurwitz, D.; Krawetz, N.; Kumar, S.; Nelson, M.; Phillips, J.; Shinozaki, A.; Zheng, G.; and Zhu, F. (2008). *NAMD User's Guide*. `http://www.ks.uiuc.edu/Research/namd/2.6/ug/ug.html`.

Blackford, L. S.; Choi, J.; Cleary, A.; D'Azevedo, E.; Demmel, J.; Dhillon, I.; Dongarra, J.; Hammarling, S.; Henry, G.; Petitet, A.; Stanley, K.; Walker, D.; and Whaley, R. C. (1997). *ScaLAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA. ISBN 0-89871-397-8 (paperback).

Bradley, Michael J.; Chivers, Peter T.; and Baker, Nathan A. (2008). Molecular dynamics simulation of the Escherichia coli NikR protein: equilibrium conformational fluctuations reveal interdomain allosteric communication pathways. *Journal of Molecular Biology*, *378*(5):1155–1173. ISSN 1089-8638. doi:10.1016/j.jmb.2008.03.010.

Brooks, B and Karplus, M. (1985). Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proceedings of the National Academy of Sciences of the United States of America*, *82*(15):4995–4999.

Brooks, Bernard R.; Bruccoleri, Robert E.; Olafson, Barry D.; States, David J.; Swaminathan, S.; and Karplus, Martin (1983). Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, *4*(2):187–217. ISSN 1096-987X. doi:10.1002/jcc.540040211.

Brünger, Axel T. (1992). *X-PLOR, Version 3.1, A System for X-ray Crystallography and NMR*. The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University.

Cannon, William R. and Benkovic, Stephen J. (1998). Solvation, Reorganization Energy, and Biological Catalysis. *Journal of Biological Chemistry*, *273*(41):26257–26260. doi:10.1074/jbc.273.41.26257.

Caratzoulas, Stavros; Mincer, Joshua S.; and Schwartz, Steven D. (2002). Identification of a protein-promoting vibration in the reaction catalyzed by horse liver alcohol dehydrogenase. *Journal of the American Chemical Society*, *124*(13):3270–3276. doi:10.1021/ja017146y.

Cattell, Raymond B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2):245–276. doi:10.1207/s15327906mbr0102_10.

Chakraverly, Snehashish (2009). *Vibration of Plates*. CRC Press, New York. ISBN 1-4200-5395-7.

Cheng, Xiaolin; Ivanov, Ivaylo; Wang, Hailong; Sine, Steven M.; and McCammon, J. Andrew (2007). Nanosecond-timescale conformational dynamics of the human ±7 nicotinic acetylcholine receptor. *Biophysical Journal*, *93*(8):2622–2634. ISSN 0006-3495.

Darden, Tom; York, Darrin; and Pedersen, Lee (1993). Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics*, *98*(12):10089–10092. ISSN 00219606. doi:10.1063/1.464397.

Den Otter, W. K. and Briels, W. J. (2000). Free energy from molecular dynamics with multiple constraints. *Molecular Physics: An International Journal at the Interface between Chemistry and Physics*, *98*(12):773–781. ISSN 0026-8976.

Dixit, Surjit B.; Andrews, David Q.; and Beveridge, D. L. (2005). Induced fit and the entropy of structural adaptation in the complexation of CAP and lambda-repressor with cognate DNA sequences. *Biophysical Journal*, *88*(5):3147–3157. ISSN 0006-3495.

Durbin, Richard; Eddy, Sean R.; Krogh, Anders; and Mitchison, Graeme (1999). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, New York, NY, USA. ISBN 978-0521629713.

Eisenmesser, Elan Zohar; Bosco, Daryl A.; Akke, Mikael; and Kern, Dorothee (2002). Enzyme Dynamics During Catalysis. *Science*, *295*(5559):1520–1523. doi:10.1126/science.1066176.

Eisenmesser, Elan Zohar.; Millet, Oscar; Labeikovsky, Wladimir; Korzhnev, Dmitry M.; Wolf-Watz, Magnus; Bosco, Daryl A.; Skalicky, Jack J.; Kay, Lewis E.; and Kern, Dorothee (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, *438*(7064):117–121. ISSN 0028-0836. doi:10.1038/nature04105.

Elber, Ron (2005). Long-timescale simulation methods. *Current Opinion in Structural Biology*, *15*(2):151–156. ISSN 0959-440X. doi:DOI:10.1016/j.sbi.2005.02.004. Theory and simulation/Macromolecular assemblages.

Essmann, Ulrich; Perera, Lalith; Berkowitz, Max L.; Darden, Tom; Lee, Hsing; and Pedersen, Lee G. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, *103*(19):8577–8593. doi:10.1063/1.470117.

Gibrat, Jean-François and Gõ, Nobuhiro (1990). Normal mode analysis of human lysozyme: Study of the relative motion of the two domains and characterization of the harmonic motion. *Proteins: Structure, Function, and Genetics*, *8*(3):258–279. ISSN 1097-0134. doi:10.1002/prot.340080308.

Hammes, Gordon G. (2002). Multiple conformational changes in enzyme catalysis. *Biochemistry*, *41*(26):8221–8228. doi:10.1021/bi0260839. PMID: 12081470.

Hammes-Schiffer, Sharon (2002). Impact of enzyme motion on activity. *Biochemistry*, *41*(45):13335–13343. doi:10.1021/bi0267137. PMID: 12416977.

Harris, Sarah A. and Laughton, Charles A. (2007). A simple physical description of DNA dynamics: quasi-harmonic analysis as a route to the configurational entropy. *Journal of Physics: Condensed Matter*, *19*(7):076103. doi:10.1088/0953-8984/19/7/076103.

Harris, Sarah A.; Sands, Zara A.; and Laughton, Charles A. (2005). Molecular dynamics simulations of duplex stretching reveal the importance of entropy in determining the biomechanical properties of DNA. *Biophysical Journal*, *88*(3):1684–1691. ISSN 0006-3495.

Harrison, Robert W. (1984). Variational calculation of the normal modes of a large macromolecule: Methods and some initial results. *Biopolymers*, *23*(12):2943–2949. ISSN 1097-0282. doi:10.1002/bip.360231216.

Hayward, S and Gõ, N (1995). Collective variable description of native protein dynamics. *Annual Review of Physical Chemistry*, *46*(1):223–250. doi:10.1146/annurev.pc.46.100195.001255.

Henzler-Wildman, Katherine A.; Lei, Ming; Thai, Vu; Kerns, S. Jordan; Karplus, Martin; and Kern, Dorothee (2007). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, *450*(7171):913–916. ISSN 0028-0836. doi:10.1038/nature06407.

Herrick, James B. (1910). Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. *Arch Intern Med*, *VI*(5):517–521. doi:10.1001/archinte.1910.00050330050003.

Humphrey, William; Dalke, Andrew; and Schulten, Klaus (1996). VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, *14*:33–38.

Ishima, Rieko and Torchia, Dennis A. (2000). Protein dynamics from NMR. *Nature Structural & Molecular Biology*, *7*(9):740–743. ISSN 1072-8368. doi:10.1038/78963.

Israelachvili, Jacob and Wennerstrom, Hakan (1996). Role of hydration and water structure in biological and colloidal interactions. *Nature*, *379*(6562):219–225. doi:10.1038/379219a0.

Jensen, Morten O.; Mouritsen, Ole G.; and Peters, Gunther H. (2004). The hydrophobic effect: Molecular dynamics simulations of water confined between extended hydrophobic and hydrophilic surfaces. *The Journal of Chemical Physics*, *120*(20):9729–9744. doi:10.1063/1.1697379.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *32*(5):922–923. doi:10.1107/S0567739476001873.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *34*(5):827–828. doi:10.1107/S0567739478001680.

Karplus, Martin and Kushick, Joseph N. (1981). Method for estimating the configurational entropy of macromolecules. *Macromolecules*, *14*(2):325–332. doi:10.1021/ma50003a019.

Kavanaugh, Jeffrey S.; Moo-Penn, Winston F.; and Arnone, Arthur (1993). Accommodation of insertions in helixes: The mutation in hemoglobin catonsville (Pro 37α-Glu-Thr 38α) generates a $3_{10} \to \alpha$ bulge. *Biochemistry*, *32*(10):2509–2513. ISSN 0006-2960. doi:10.1021/bi00061a007.

Kundu, Sibsankar; Melton, Julia S.; Sorensen, Dan C.; and Phillips, George N. (2002). Dynamics of proteins in crystals: Comparison of experiment with simple models. *Biophysical Journal*, *83*(2):723–732. ISSN 0006-3495.

Lange, Oliver F. and Grubmüller, Helmut (2006). Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, *62*(4):1053–1061. doi:10.1002/prot.20784.

Lum, Ka; Chandler, David; and Weeks, John D. (1999). Hydrophobicity at small and large length scales. *The Journal of Physical Chemistry B*, *103*(22):4570–4577. doi: 10.1021/jp984327m.

Ma, Jianpeng (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, *13*(3):373–380. ISSN 0969-2126.

Marques, Osni and Sanejouand, Yves-Henri (1995). Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins: Structure, Function, and Genetics*, *23*(4):557–560. ISSN 1097-0134. doi:10.1002/prot.340230410.

Matsumura, M.; Wozniak, J. A.; Sun, D. P.; and Matthews, B. W. (1989). Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. *Journal of Biological Chemistry*, *264*(27):16059–16066.

Noid, D. W.; Fukui, K.; Sumpter, B. G.; Yang, C.; and Tuzun, R. E. (2000). Time-averaged normal coordinate analysis of polymer particles and crystals. *Chemical Physics Letters*, *316*(3-4):285–296. ISSN 0009-2614. doi:DOI:10.1016/S0009-2614(99)01152-5.

Noy, Agnes; Pérez, Alberto; Laughton, Charles A.; and Orozco, Modesto (2007). Theoretical study of large conformational transitions in DNA: the B-A conformational change in water and ethanol/water. *Nucleic Acids Research*, *35*(10):3330–3338. doi:10.1093/nar/gkl1135.

Pacheco, Peter S (1997). *Parallel Programming with MPI*. Morgan Kaufmann, San Francisco, CA.

Perahia, David and Mouawad, Liliane (1995). Computation of low-frequency normal modes in macromolecules: Improvements to the method of diagonalization in a mixed basis and application to hemoglobin. *Computers & Chemistry*, *19*(3):241–246. ISSN 0097-8485. doi:DOI:10.1016/0097-8485(95)00011-G. Third Conference on Computers in Chemistry.

Phillips, James C.; Braun, Rosemary; Wang, Wei; Gumbart, James; Tajkhorshid, Emad; Villa, Elizabeth; Chipot, Christophe; Skeel, Robert D.; Kalé, Laxmikant; and Schulten, Klaus (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, *26*(16):1781–1802. ISSN 1096-987X. doi:10.1002/jcc.20289.

Ramanathan, Arvind and Agarwal, Pratul K. (2009). Computational identification of slow conformational fluctuations in proteins. *The Journal of Physical Chemistry B*, *113*(52):16669–16680. ISSN 1520-6106. doi:10.1021/jp9077213.

Rod, Thomas H.; Radkiewicz, Jennifer L.; and Brooks, Charles L. (2003). Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(12):6980–6985. doi:10.1073/pnas.1230801100.

Rossmann, Michael G.; Morais, Marc C.; Leiman, Petr G.; and Zhang, Wei (2005). Combining X-Ray crystallography and electron microscopy. *Structure*, *13*(3):355–362. ISSN 0969-2126.

Ryckaert, J.; Ciccotti, G.; and Berendsen, H. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, *23*(3):327–341. ISSN 00219991. doi:10.1016/0021-9991(77)90098-5.

Saibil, Helen R. (2000). Conformational changes studied by cryo-electron microscopy. *Nature Structural & Molecular Biology*, *7*(9):711–714. ISSN 1072-8368. doi:10.1038/78923.

Scatena, L. F.; Brown, M. G.; and Richmond, G. L. (2001). Water at Hydrophobic Surfaces: Weak Hydrogen Bonding and Strong Orientation Effects. *Science*, *292*(5518):908–912. doi:10.1126/science.1059514.

Schlick, Tamar; Barth, Eric; and Mandziuk, Margaret (1997). Biomolecular dynamics at long timesteps: Bridging the timescale gap between simulation and experimentation. *Annual Review of Biophysics and Biomolecular Structure*, *26*(1):181–222. doi:10.1146/annurev.biophys.26.1.181.

Shaw, David E.; Dror, Ron O.; Salmon, John K.; Grossman, J. P.; Mackenzie, Kenneth M.; Bank, Joseph A.; Young, Cliff; Deneroff, Martin M.; Batson, Brannon; Bowers, Kevin J.; Chow, Edmond; Eastwood, Michael P.; Ierardi, Douglas J.; Klepeis, John L.; Kuskin, Jeffrey S.; Larson, Richard H.; Lindorff-Larsen, Kresten; Maragakis, Paul; Moraes, Mark A.; Piana, Stefano; Shan, Yibing; and Towles, Brian (2009). Millisecond-scale molecular dynamics simulations on anton. In *SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, pages 1–11. ACM, New York, NY, USA. ISBN 978-1-60558-744-8. doi:http://doi.acm.org/10.1145/1654059.1654099.

Snir, Marc; Otto, Steve; Huss-Lederman, Steven; Walker, David; and Dongarra, Jack (1998). *MPI: The Complete Reference*. MIT Press, Cambridge.

Tama, Florence; Gadea, Florent Xavier; Marques, Osni; and Sanejouand, Yves-Henri (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Structure, Function, and Genetics*, *41*(1):1–7. ISSN 1097-0134. doi:10.1002/1097-0134(20001001)41:1<1::AID-PROT10>3.0.CO;2-P.

Tanford, Charles (1980). *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*. Wiley, New York.

Teeter, Martha M. and Case, David A. (1990). Harmonic and quasiharmonic descriptions of crambin. *The Journal of Physical Chemistry*, *94*(21):8091–8097. doi:10.1021/j100384a021.

Tirion, Monique M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, *77*(9):1905–1908. doi:10.1103/PhysRevLett.77.1905.

Tournier, Alexander L. and Smith, Jeremy C. (2003). Principal components of the protein dynamical transition. *Phys. Rev. Lett.*, *91*(20):208106. doi:10.1103/PhysRevLett.91.208106.

Waight, Andrew B.; Love, James; and Wang, Da-Neng (2010). Structure and mechanism of a pentameric formate channel. *Nat Struct Mol Biol*, *17*(1):31–37. ISSN 1545-9993. doi:10.1038/nsmb.1740.

Wallqvist, Anders; Gallicchio, Emilio; and Levy, Ronald M. (2001). A model for studying drying at hydrophobic interfaces: Structural and thermodynamic properties. *The Journal of Physical Chemistry B*, *105*(28):6745–6753. doi:10.1021/jp010945i.

Wang, Xiaodong (2008). *Fundamentals of Fluid-Solid Interactions-Analytical and Computational Approaches*. Elsevier.

Weiner, J. H. (1983). *Statistical Mechanics of Elasticity*. CRC Press, New York. ISBN 0486422607.

Yang, Mingjun; Zhang, Xin; and Han, Keli (2010). Molecular dynamics simulation of SRP GTPases: Towards an understanding of the complex formation from equilibrium fluctuations. *Proteins: Structure, Function, and Bioinformatics*, *78*(10):2222–2237. ISSN 1097-0134. doi:10.1002/prot.22734.