IOWA STATE UNIVERSITY
Digital Repository

2016

Normal mode computations and applications

Hyuntae Na
Iowa State University

Normal mode computations and applications

by

Hyuntae Na

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Science (Computational Molecular Biology)

Program of Study Committee:
Guang Song, Major Professor
Xiaoqiu Huang
Robert L. Jernigan
Yan-Bin Jia
Zhijun Wu

Iowa State University

Ames, Iowa

2016

DEDICATION

To my family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to take this opportunity to express my gratitude to those who have helped me in various ways during my years as a Ph.D. student.

First and foremost, I am very grateful to Dr. Guang Song for his invaluable guidance throughout my research and the writing of this thesis. This work would not have been achieved without his constant support and kind encouragement.

I would also like to thank my committee members and collaborators for their valuable time, inputs, and comments: Dr. Daniel ben-Avraham in Clarkson University, Dr. Xiaoqiu Huang, Dr. Robert L. Jernigan, Dr. Yan-Bin Jia, and Dr. Zhijun Wu in Iowa State University.

I am thankful to members in our lab, Dr. Tu-Liang Lin and Dr. Santhosh K. Vammi for their feedback and discussions. I am thankful also to the other members in our lab and my friends for their friendship and support: Dr. Taekyung Lee, Dr. Huan Lin, Dr. Yuheng Long, Dr. Haihua Xie, Ce Zhang, Hailu Yang, and Jaekyun Song. I am thankful to members in Dr. Jernigan's lab for their inputs and discussions: Kannan Sankar, Kejue Jia, Sambit Mishra, and Yuan Wang.

I am extremely grateful to my parents and brother for supporting me to pursue this program. Especially, I could not finish this study if my parents have not raised me patient and strong and have not been supportive for me when I am in difficulty.

ABSTRACT

Proteins are essential structural and functional units in cells. Proteins form stable and yet somewhat flexible 3-D structures and often function by interacting with other molecules. Their functional behaviors are determined by their 3-D structures as well as their dynamics. Protein dynamics studies are thus very important.

One of the most powerful computational methods for studying protein dynamics is normal mode analysis (NMA). The low frequency modes especially are of great interest for many protein dynamics studies. Although it provides analytical solutions to a protein's collective motions, classical NMA is cumbersome to use and may become even prohibitive when the system being studies is too large. Many simplified NMA models have been developed, which use extremely simplified structural models and/or coarse-grained potentials. However, the dynamics given by such models may not always be fully realistic.

In this dissertation, I have alleviated these problems by addressing the following sequence of questions: (1) what is the contribution of inter-residue (inter-atom) forces to protein normal modes; (2) how to remove the cumbersome energy minimization step in NMA while preserving most of the accuracy of the model; (3) how to efficiently construct coarse-grained structural models from all-atom models while maintaining the accuracy in dynamics. Additionally, using my new models as well as the classical NMA, I have closely examined the vibrational spectrum of globular proteins in the whole frequency range, and have found a connection with experimental observations. Finally, as an application of normal modes, the last part of this thesis presents a novel approach in which normal modes are used to identify what breathing motions of myoglobin dynamically open ligand

migration channels. The results are in an excellent agreement with molecular dynamics simulation results and experimentally determined ligand entry rates.

CHAPTER 1. OVERVIEW AND OBJECTIVES

1.1 Introduction

Proteins are fundamental functional units in cells. As ubiquitous and versatile macro-molecules in living organisms, proteins have many different roles, varying from maintaining structures, binding ligands, catalyzing reactions, to sending signals to other systems, etc. Proteins form stable and yet somewhat flexible 3D structures and often function by interacting with other molecules. Their functional behaviors are determined by their 3-D structures as well as their flexibilities. It is fascinating to see how proteins exercise their precise controls in realizing many different functions, even though the exact mechanisms of many such processes are not fully known.

The importance of computational studies of protein dynamics and functions has long been recognized, most notably by the recent Nobel Prize in Chemistry awarded to Karplus, Levitt, and Warshel for "the development of multiscale models for complex chemical systems." [136] Molecular dynamics simulation is one of the most popular tools for studying the dynamics of proteins and other biological molecules. Another powerful tool for studying protein dynamics is normal mode analysis (NMA). While molecular dynamics simulation is stochastic by nature, normal mode analysis provides complete analytical solutions of protein dynamics locally at and around a specific conformation, usually an energetically minimized structure.

However, there are some major challenges in applying normal mode analysis. The classical normal mode analysis uses atomic structure modes and detailed force field po-

tentials that have hundreds and even thousands of parameters. Energy minimization is always required before computing the normal modes. The whole process is cumbersome and time consuming and may become even prohibitive for large systems. To overcome these hindrances, simplified NMA models such as elastic network models have been developed. However, the accuracy of these simplified models is questionable and there is a lack of tight connections between the classical NMA and the simplified NMAs.

In this dissertation, I have addressed some of these problems in normal mode computations.

1.2   Thesis Organization

The thesis is organized as follows:

Chapter 1: Overview and Objectives.   This chapter gives a general introduction to the thesis: presenting the research goals and the overall structure of this thesis.

Chapter 2: A Natural Unification of GNM and ANM and the Role of Inter-Residue Forces.   The Gaussian network model (GNM) and anisotropic network model (ANM) are two elastic network models that have been widely used to study protein fluctuation dynamics. Both models have strengths and weaknesses. Attempts have been made in the past to unify the two models but they had severe limitations. This work presents a novel theoretical result that shows how GNM and ANM can be unified through taking into account the effect of inter-residue forces. The unification reveals also the role of inter-residue forces in protein fluctuation dynamics. This new understanding of ANM triggered a follow-up study reported in Chapter 3.

Chapter 3: Bridging Between NMA and Elastic Network Models.   In this chapter, through steps of simplification that starts with NMA and ends with elastic

network model (ENM), we have built a tight connection between NMA and ENM. In the process of bridging the two, we have also discovered several high-quality simplified models. Our best simplified model has a mean correlation with the original NMA that is as high as 0.88. In addition, the model is force-field independent and does not require energy minimization, and thus can be applied directly to experimental structures. Another benefit of drawing the connection is a clearer understanding why ENMs work well and how it can be further improved. We have discovered that ANM can be greatly enhanced by including an additional torsional term and a geometry term.

Chapter 4: Bridging between NMA and Elastic Network Models: Preserving All-atom Accuracy in Coarse-grained Models. For large protein complexes, obtaining fine-grained all-atom descriptions of normal mode motions can be computationally prohibitive because of the limitation of available computational resources. For this reason, coarse-grained models have been used widely. However, most existing coarse-grained models use extremely simple potentials to represent the interactions within the coarse-grained structures and as a result, the dynamics obtained for the coarse-grained structures may not always be fully realistic. There is a gap between the quality of the dynamics of the coarse-grained structures given by all-atom models and that by coarse-grained models. In this chapter, we resolve an important question in protein dynamics computationshow can we efficiently construct coarse-grained models whose description of the dynamics of the coarse-grained structures remains as accurate as that given by all-atom models? Our method takes advantage of the sparseness of the Hessian matrix and achieves a high efficiency with a novel iterative matrix projection approach.

Chapter 5: Universality of Vibrational Spectra of Globular Proteins. In 1993, ben-Avraham found that the vibrational spectra of five proteins, when properly normalized, seemed to have one "universal" curve in the torsional space. [10] In this work, I have extended ben-Avraham's work to confirm the universality of vibrational

spectrum of globular proteins in both torsional and Cartesian spaces. Peaks in the universal spectrum curve are thus not protein specific but force field specific. This significant result implies that experimental spectra of proteins could be used to guide the fine tuning of theoretical empirical potentials, and the various features and peaks observed in theoretical studies could spur experimental confirmation.

Chapter 6: Quantitative Delineation of How Breathing Motions Open Ligand Migration Channels in Myoglobin and Its Mutants. Ligand migration and binding are central to the biological functions of many proteins and it is widely thought that protein breathing motions open up ligand channels dynamically. In this chapter, I present a novel normal mode-based method that quantitatively delineates what and how breathing motions open ligand migration channels. Results of applying the method to myoglobin wild-type and its several mutants are in excellent agreement with MD simulation results and experimentally determined ligand entry rates.

Chapter 7: Summary and Conclusion

This final chapter presents a summary and concludes the thesis.

CHAPTER 2.   A NATURAL UNIFICATION OF GNM AND ANM AND THE ROLE OF INTER-RESIDUE FORCES

Hyuntae Na[2][3] and Guang Song[2][3][4]

Abstract

Gaussian network model (GNM) and Anisotropic network model (ANM) are two of the elastic network models that have been widely used to study protein fluctuation dynamics. Both models have strengths and weaknesses. Attempts were made in the past to unify the two models but had severe limitations. This work presents a novel theoretical result that shows how GNM and ANM can be unified through taking into account the effect of inter-residue forces. The unified model, called Force Spring Model, or FSM, is reduced to ANM when all the inter-residue forces are set to be zero. Moreover, the unification reveals the role of inter-residue forces in protein fluctuation dynamics. Specifically, the effect of inter-residue forces is closely examined by studying the changes in mean-square fluctuations when the inter-residue forces are present.

---

[1]This chapter is reprinted with permission of *Phys. Biol.* 2014, 11(3), 036002.

[2]Graduate student and Associate Professor, respectively, Department of Computer Science, Iowa State University.

[3]Primary researchers and authors.

[4]Author for correspondence.

2.1 Introduction

Gaussian Network Model (GNM) [8] and Anisotropic Network Model (ANM) [2] have been widely used to study protein fluctuation dynamics and conformation changes. In 2008, Zheng [159] proposed a simple approach to unify the two methods by introducing an anisotropic parameter f_{anm}. The new method, named Generalized Anisotropic Network Model, or GANM, is a linear combination of the Anisotropic Network Model (ANM) and the Gaussian Network Model (GNM) with a contribution ratio f_{anm} that satisfies $0 \leq f_{anm} \leq 1$:

$$H^{\mathrm{GANM}} = H^{\mathrm{ANM}} - f_{anm} \cdot (H^{\mathrm{ANM}} - \Gamma^{\mathrm{GNM}} \otimes I_3), \tag{2.1}$$

where H^{ANM} is the ANM Hessian matrix, Γ^{GNM} is the GNM Kirchhoff matrix, I_3 is the 3×3 identity matrix, and $\otimes$ is the operator of the Kronecker product. The author then showed that GANM could outperform GNM and ANM in describing thermal fluctuations (B-factors) and conformation changes by tuning the parameter f_{anm}.

Zheng's work represented one of the first attempts to unify the two popular network models. However, there are a couple of drawbacks in this unification. First, the physical meaning of the parameter f_{anm} is not clear. Second, which is a more severe problem, GANM, as a linear combination of GNM and ANM, is not rotationally invariant [77, 137]. This is evident from the fact that it has three, instead of six, zero eigenvalues.

In this work, we present a Spring Force Model (SFM) in which a unification of GNM and ANM naturally arrives. The derivation uncovers how ANM and GNM should be unified and what is the physical meaning of the parameters that appear in the unification. Moreover, because the model's Hessian matrix is obtained by taking the second derivatives of a physically realistic potential, the model is intrinsically rotationally invariant.

In ANM model [2], each residue is represented by a single bead, usually the C_α atoms. As a significant improvement over NMA, ANM uses experimental structures

directly as input without the need for energy minimization, and assumes the spring mass system formed by the C_α atoms is at equilibrium. Inter-residue forces were considered and formulations for balancing the inter-residue forces to achieve the system equilibrium were given.

Specifically, for the equilibrium condition, ANM model requires only that the sum of all the inter-residue forces acting on one bead is zero. How these forces should be balanced was clearly presented in the original ANM paper (see equations (10) to (15) in [2]). These forces rebalance themselves when there are external forces exerted on the protein that cause the structure to deform. As a result, the protein will arrive at a new equilibrated state, where the net force at each node is zero but the inter-residue forces may not be zero, the details of which have been clearly worked out by the same authors in [3, 4, 36, 157].

However, when used to compute normal modes or B-factors for a given structure, ANM assumes the input structure is at equilibrium in the stricter sense that the all the springs are relaxed. That is, it assumes $s_{ij} = s_{ij}^0$ (where s_{ij} and s_{ij}^0 are instantaneous and initial separation distances between residues i and j), as stated in the text between equations (18) and (19) in the original ANM paper [2].

Our model is similar to ANM and is a simple extension of it: both are a spring mass model and assume the input structure is at equilibrium; both have the same requirement for the equilibrium condition, i.e., the sum of inter-residue forces on a given node is zero. The major difference between our model and ANM is that in ANM, probably for the sake of simplicity, all the springs are initially set to be relaxed (i.e., $s_{ij} = s_{ij}^0$) and thus all the inter-residue forces are initially set to be zero, while in our model, inter-residues forces are not zero initially. Indeed, this is a more realistic condition, since at equilibrated or energy-minimized structures, the net forces at all the atoms are zero but the inter-atom forces are usually not zero.

In the following Method section, we will derive the Hessian matrix for our model. Since it is based on both a spring network and non-zero initial inter-residue forces, we name it Force Spring Model, or FSM in short. As a spring-mass model, FSM has a similar potential function to that of ANM. The derivation of FSM Hessian matrix, however, is different from ANM in that the inter-residue forces, the first-derivatives of the potential are no longer assumed to be zero.

2.2   Methods

2.2.1   GNM and ANM

Gaussian Network Model (GNM) was first introduced in [8] under the assumption that the fluctuations Δr_{ij} between the ith and jth C_α atoms in the folded protein is Gaussianly distributed. The model use the Kirchhoff matrix Γ^{GNM} to describe the connectivity among the C_α atoms:

$$\Gamma_{ij}^{\text{GNM}} = \begin{cases} -1 & if \quad i \neq j \text{ and } r_{ij} \leq r_c \\ 0 & if \quad i \neq j \text{ and } r_{ij} > r_c \\ -\sum_{i, i \neq j} \Gamma_{ij} & if \quad i = j, \end{cases} \tag{2.2}$$

where i and j are the indices of the residues and r_c is the cutoff distance, which is usually set to be 7-8 Å in GNM. Conveniently, the expected value of residue fluctuations, $\langle \Delta r_i{}^2 \rangle$, and correlations, $\langle \Delta r_i \cdot \Delta r_j \rangle$, can be obtained from the inverse of the Kirchhoff matrix.

However, GNM can be used to compute only the magnitudes of protein fluctuations. Anisotropic Network Model (ANM) [2] was introduced to obtain the directions of protein fluctuations. ANM uses a harmonic Hookean potential to define the interactions among the atoms, i.e.,

$$V = (1/2) \sum_{i,j} k_{i,j} (r_{i,j} - r_{i,j}^0)^2. \tag{2.3}$$

For ANM and ANMr, a variant of ANM whose spring constants are inversely proportional to the squared separation distances [153], the spring constants $k_{i,j}$ between atoms i and j are defined as follows:

$$k_{i,j}^{\text{ANM}} = \begin{cases} 1, & \text{if } r_{i,j} < r_c \\ 0, & \text{otherwise}, \end{cases} \tag{2.4}$$

$$k_{i,j}^{\text{ANMr2}} = 1/r_{i,j}^2, \tag{2.5}$$

where $r_{i,j}$ is the Euclidean distance between atoms i and j and r_c is the cutoff distance, which is usually set to be 13 Å in ANM [2]. Other values for the cutoff distance were also tried later on. For example, Zheng et al. [161] found a cutoff distance between 10 and 15 Å performed equally while Riccardi et al. [106] found 10 Å to be optimal in reproducing crystallographic B-factors.

The ANM Hessian matrix can be easily obtained by taking the second derivatives of potential V (see equation (2.3)). Particularly, for atoms i and j that are in contact, the 3 by 3 block element $H_{i,j}^{\text{ANM}}$ of the ANM Hessian matrix is:

$$H_{i,j}^{\text{ANM}} = \frac{-1}{r_{i,j}^2} \begin{pmatrix} x_j - x_i \\ y_j - y_i \\ z_j - z_i \end{pmatrix} \begin{pmatrix} x_j - x_i & y_j - y_i & z_j - z_i \end{pmatrix}. \tag{2.6}$$

2.2.2 Unifying GNM and ANM: the Effect of Inter-Atom Forces

Given a protein with n residues and that each residue is represented by a single bead, say its C_α atom, one simple way to model the interactions within the system is to let the beads interact through Hookean springs. The potential energy of the system is:

$$V = (1/2) \sum_{i,j} k_{i,j} (r_{i,j} - r_{i,j}^0)^2, \tag{2.7}$$

where $r_{i,j}$ and $r_{i,j}^0$ are the instantaneous and equilibrium distances between interacting residues i and j.

The Hessian matrix of the system is a $3n \times 3n$ positive semi-definite matrix whose elements are the second derivatives of the potential energy V with respect to x_i, y_i, and

z_i coordinates of each atom $i \in \{1, ..., n\}$. It can be written as a $n \times n$ block matrix as follows:

$$H = \begin{pmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,N} \\ H_{2,1} & H_{2,2} & \cdots & H_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N,1} & H_{N,2} & \cdots & H_{N,N} \end{pmatrix}, \tag{2.8}$$

with each block element $H_{i,j}$ being a 3×3 matrix of the second derivatives:

$$H_{i,j} = \begin{pmatrix} \frac{\partial^2 V}{\partial x_i \partial x_j} & \frac{\partial^2 V}{\partial x_i \partial y_j} & \frac{\partial^2 V}{\partial x_i \partial z_j} \\ \frac{\partial^2 V}{\partial y_i \partial x_j} & \frac{\partial^2 V}{\partial y_i \partial y_j} & \frac{\partial^2 V}{\partial y_i \partial z_j} \\ \frac{\partial^2 V}{\partial z_i \partial x_j} & \frac{\partial^2 V}{\partial z_i \partial y_j} & \frac{\partial^2 V}{\partial z_i \partial z_j} \end{pmatrix}, \tag{2.9}$$

if $i \neq j$, and

$$H_{i,i} = -\sum_{j \neq i} H_{i,j} . \tag{2.10}$$

The component $\frac{\partial^2 V}{\partial x_i \partial x_j}$ in equation (2.9) can be rewritten as:

$$\frac{\partial^2 V}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i} \left(\frac{\partial V}{\partial r_{i,j}} \cdot \frac{\partial r_{i,j}}{\partial x_j} \right) \tag{2.11}$$

$$= \frac{\partial^2 V}{\partial r_{i,j}^2} \cdot \frac{\partial r_{i,j}}{\partial x_i} \cdot \frac{\partial r_{i,j}}{\partial x_j} + \frac{\partial V}{\partial r_{i,j}} \cdot \frac{\partial^2 r_{i,j}}{\partial x_i \, \partial x_j} \tag{2.12}$$

$$= k_{i,j} \cdot \frac{\partial r_{i,j}}{\partial x_i} \frac{\partial r_{i,j}}{\partial x_j} - f_{i,j} \cdot \frac{\partial^2 r_{i,j}}{\partial x_i \, \partial x_j} . \tag{2.13}$$

In the above equation, $f_{i,j} = -\frac{\partial V}{\partial r_{i,j}} = -k_{i,j}(r_{i,j} - r_{i,j}^0)$ is the inter-atom force between atoms i and j, and $k_{i,j} = \frac{\partial^2 V}{\partial r_{i,j}^2}$ is the spring constant of the spring connecting the two atoms. In ANM, it is assumed that pairwise distances $r_{i,j}$ of the input structure are the same as the equilibrium distances $r_{i,j}^0$, and consequently all $f_{i,j}$'s are zero. Here, we loosen this requirement and require instead only the net force at each atom is zero. As a result, $f_{i,j}$ is not necessarily zero, but $\vec{f}_j^{net} = \sum_i f_{i,j} \vec{u}_{i,j} = 0$ for all j's, where $\vec{u}_{i,j}$ is the unit vector pointing from atom i to atom j. In summary, we still assume the input structure is at equilibrium so that there is no need for energy minimization, but no longer do we assume the inter-atom forces are all zero.

Note that since $\frac{\partial r_{i,j}}{\partial x_i} = \frac{x_i - x_j}{r_{i,j}}$, the second term of the second addend in equation (2.13) can be rewritten as:

$$
\begin{aligned}
\frac{\partial^2 r_{i,j}}{\partial x_i \partial x_j} &= \frac{\partial}{\partial x_i}\left(\frac{\partial r_{i,j}}{\partial x_j}\right) = -\frac{1}{r_{i,j}} + \frac{(x_i - x_j)^2}{r_{i,j}^3} \\
&= -\frac{1}{r_{i,j}}\left(1 + \frac{\partial r_{i,j}}{\partial x_i}\cdot\frac{\partial r_{i,j}}{\partial x_j}\right).
\end{aligned}
\tag{2.14}
$$

Using $f_{i,j}$, $k_{i,j}$, and equation (2.14), the partial derivative in equation (2.11) can be rewritten as:

$$
\frac{\partial^2 V}{\partial x_i \partial x_j} = \left(k_{i,j} + \frac{f_{i,j}}{r_{i,j}}\right)\cdot\left(\frac{\partial r_{i,j}}{\partial x_i}\cdot\frac{\partial r_{i,j}}{\partial x_j}\right) + \frac{f_{i,j}}{r_{i,j}}.
\tag{2.15}
$$

In a similar manner, term $\frac{\partial^2 V}{\partial x_i \partial y_j}$ in equation (2.9) can be written as:

$$
\frac{\partial^2 V}{\partial x_i \partial y_j} = \left(k_{i,j} + \frac{f_{i,j}}{r_{ij}}\right)\cdot\left(\frac{\partial r_{i,j}}{\partial x_i}\cdot\frac{\partial r_{i,j}}{\partial y_j}\right).
\tag{2.16}
$$

Likewise, the rest of the matrix in equation (2.9) can be written out. Note that $\frac{\partial r_{i,j}}{\partial x_i}\cdot\frac{\partial r_{i,j}}{\partial x_j}$ and $\frac{\partial r_{i,j}}{\partial x_i}\cdot\frac{\partial r_{i,j}}{\partial y_j}$ in the above two equations are the same as the block element of the ANM Hessian matrix when the spring constant is 1 (see equation (2.6)). Therefore, putting equations (2.15) and (2.16) together, the block Hessian matrix in equation (2.9) can be rewritten as:

$$
\begin{aligned}
H_{i,j} &= \left(k_{i,j} + \frac{f_{i,j}}{r_{i,j}}\right)\cdot H_{i,j}^{\text{ANM}} + \frac{f_{i,j}}{r_{i,j}} \otimes I_3 \tag{2.17} \\
&= k_{i,j}\cdot H_{i,j}^{\text{ANM}} + \frac{f_{i,j}}{r_{i,j}}\left(H_{i,j}^{\text{ANM}} - \Gamma_{i,j}^{\text{GNM}} \otimes I_3\right), \tag{2.18}
\end{aligned}
$$

where $H_{i,j}^{\text{ANM}}$ and $\Gamma_{i,j}^{\text{GNM}}$ are the i,j elements of the ANM block Hessian matrix and the GNM matrix, respectively. And they are:

$$
H_{i,j}^{\text{ANM}} = \frac{-1}{r_{i,j}^2}\begin{pmatrix} x_j - x_i \\ y_j - y_i \\ z_j - z_i \end{pmatrix}\begin{pmatrix} x_j - x_i & y_j - y_i & z_j - z_i \end{pmatrix},
\tag{2.19}
$$

$$
\Gamma_{i,j}^{\text{GNM}} = -1.
\tag{2.20}
$$

Remarkably, equation (2.18) naturally unifies the two widely used models, ANM and GNM. It reveals how they should be combined in order to maintain the rotational invariance and the role of inter-atom forces in this unification. It shows that the parameter f_{anm} used in GANM by Zheng [159] to linearly combine ANM and GNM actually should not take arbitrary values, but have to satisfy the constraint that $\sum_i f_{i,j} \vec{u}_{i,j} = 0$, i.e., the net force at each atom has to be zero. The term $\frac{f_{i,j}}{r_{i,j}}$ clearly has the right units. As a counterpart of $k_{i,j}$, it represents a special kind of "spring", whose effect is proportional to the magnitude of inter-atom forces and inversely proportional to the inter-atom distances. If we require all inter-atom forces $f_{i,j}$ to be zero, as is in ANM, this Hessian matrix reduces to ANM Hessian matrix as expected.

It is also possible for the Hessian matrix in equation (2.18) to be reduced to that of GNM by setting parameters $f_{i,j}$, the inter-atom forces, to be proportional to inter-atom distances so that $\frac{f_{i,j}}{r_{i,j}} = -1$. In doing so, however, the net force at each atom may no longer remain zero, rendering the system to be out of equilibrium. Indeed, since the Hessian matrix in equation (2.18) is rotationally invariant, there does not exist a physically correct force assignment that reduces it to the rotationally-variant GNM, while maintaining the equilibrium.

Equation (2.18) shows also the way in which the inter-atom forces contribute to the Hessian matrix and in turn, to protein dynamics. In the following section, we will examine closely the effect of inter-atom forces on protein fluctuation dynamics.

2.2.3 Generating Random Forces

In this section, we present an algorithm for generating random inter-atom forces that satisfy the stationary point condition. At the stationary point condition, all the atoms have a net force of zero, i.e., $\vec{f}_j = \sum_{i \neq j} f_{i,j} \vec{u}_{i,j} = \vec{0}$, where $f_{i,j}$ is the force exerted by atom i on j. A positive/negative $f_{i,j}$ means i exerts a repulsive/attractive force on j. $\vec{u}_{i,j}$ is the unit vector from i to j in the Cartesian coordinate, and $\vec{0}$ is 3×1 zero vector.

Let $F = \{f_{i,j} \mid \forall i, j\}$ be the initial set of pairwise forces $f_{i,j}$ that are randomly generated. By default, F does not satisfy the stationary point condition. To modify F so that it does satisfy, we take the following steps:

- Compute the net force $\vec{f_i}$ at each atom.

- Multiply the Hessian matrix inverse H^{-1} with $\vec{f_i}$ to obtain the instantaneous displacement $\vec{\delta_i}$ for each atom.

- Now imagine we make a displacement of $-\vec{\delta_i}$ for each atom i. This displacement will create forces $g_{i,j}$ among the atom pairs, specifically, $g_{i,j} = H_{i,j}(\vec{\delta_j} - \vec{\delta_i})$.

- Reset $f_{i,j}$ to be $f_{i,j} + g_{i,j}$. F now satisfies the stationary point condition.

Algorithm 1 describes the procedure in mathematical details. In the algorithm, $P = \{\vec{p_1}, \vec{p_2}, ..., \vec{p_n}\}$ denotes atom coordinates, while H the $3n \times 3n$ Hessian matrix.

Lastly, it is worth pointing out without proof that there exist an infinite number of sets of inter-atom forces that satisfy the stationary point condition.

Algorithm 1 Stationary(H, P, F)

1: $\vec{u}_{i,j} \leftarrow (\vec{p_j} - \vec{p_i})/\|\vec{p_j} - \vec{p_i}\|$
2: $\vec{f_i} \leftarrow \sum_{j \neq i} \vec{u}_{i,j} f_{i,j}$
3: $\vec{\delta_i} \leftarrow \sum_j (H^{-1})_{i,j} \vec{f_i}$
4: $g_{i,j} \leftarrow \langle H_{i,j}(\vec{\delta_j} - \vec{\delta_i}), \vec{u}_{i,j} \rangle$
5: $F = \{f_{i,j} \mid f_{i,j} = f_{i,j} + g_{i,j}, \forall i, j\}$
6: return F

2.3 Results

In this section, we examine what impact inter-atom forces have on the dynamic behaviors of proteins. Specifically, the effect of forces is evaluated by observing the changes in mean-square fluctuations when the inter-atom forces are switched from being "absent" to "present", and when different sets of inter-atom forces are applied. To

compute mean square fluctuations, we will use ANMr model. ANMr is a special kind of ANM whose only difference from ANM is that its spring constant $k_{i,j}$ is not uniform but is inversely proportional to the squared distance between a pair of atoms. We choose to use ANMr to compute mean-square fluctuations because ANMr has been shown to perform significantly better than ANM [153]. Since ANMr is a special kind of ANM, the following general statements about ANM apply to ANMr as well.

Recall that, in ANM, a residue is usually approximated by one bead, its C_α atom, and neighboring C_α's interact via Hookean springs that are set to be at their equilibrium at the input structure. This conveniently sets the whole input structure at equilibrium without any energy minimization. The inter-residue forces, among the C_α atoms, also are zero.

In reality, even at equilibrium, where the net force at each atom is zero, a protein's inter-atom forces are not necessarily zero. These strains that exist inside a protein will persist even if the system is viewed at the residue level. So for coarse-grained model like ANM, these strains should take the form of inter-residue forces[1]. Therefore, if we consider a coarse-grained model as a mean-field average of an all-atom model, it would be more accurate to take into account the effect of inter-residue forces. And that is our focus here.

To study the effect of inter-residue forces, there is, however, another challenge. While many force fields exist for all-atom models and are used to define precisely how atoms should interact, it is not clear how residues should interact with one another. Statistical potentials, especially distance-dependent ones [9, 53, 119, 162] define residue-residue contact potentials and in theory can be used to compute inter-residue forces. However, these potentials are the statistical averages over a large number of proteins. When applied to a specific protein, the inter-residues forces they assign often lack accuracy

---

[1] More precisely speaking, these strains may take also the form of inter-residue torques that involve three body or even four-body interactions, but multi-body interactions are not considered here for the sake of simplicity.

and specificity. Particularly, chances are that it would not set the input structure at equilibrium.

For this reason, in this work, we do not address the problem of how to compute inter-residues forces and but leave it as an open question for future research. Instead, we focus on the effect of these forces if they are ever present. For this purpose, we generate inter-residues forces randomly. The only constraint we set on these inter-residues forces is that the net force at each residue is zero. This will guarantee that the input structure is at stationary point condition.

2.3.1 Generating Sets of Inter-Residue Forces That Satisfy Stationary Point Condition

Given a protein conformation, which in coarse-grained models such as ANM is often represented by the C_α coordinates, to study the effect of the inter-residue forces, we randomly generate sets of inter-residue (or inter-C_α) forces that satisfy the stationary point condition. Since we are using ANMr model whose spring constants are inversely proportional to the squared distances between pairs of atoms, we require that the random inter-residue forces also be roughly proportional to the inverse of the squared distances. To achieve that, we do the following:

1. Assign to each atom i a random "charge" $c_i \in \{-1, 0, 1\}$, with a probability of 20% being 1 or -1, and 60% being 0. Note that these "charges" are given not to represent the actual electric charges of the residues, but as a convenient way to initialize the inter-atom forces.

2. Initialize the pair-wise forces as: $f_{i,j} = \frac{c_i c_j}{\|\vec{p}_i - \vec{p}_j\|^2}$, where $\vec{p}_i$ is the coordinate of residue i.

3. Update $f_{i,j}$ by applying Algorithm 1 (see Methods section), which guarantees that the new $f_{i,j}$'s satisfy the stationary point condition.

2.3.2 The Effect of Forces on Mean-Square Fluctuations

Now let F be such a set of randomly generated forces. Since F satisfies the stationary point condition, the corresponding Hessian matrix in equation (2.18) is guaranteed to have six zero eigenvalues. However, F may represent a saddle point, causing the Hessian matrix to have negative eigenvalues. In such a case F will be regenerated until the eigenvalues are all positive (except for the six zero eigenvalues).

Denote $\mathbf{b}^{(cF)}$ as the mean-square fluctuations of a protein whose inter-atom forces are cF, where c is a constant scaling factor. Note that $\mathbf{b}^{(0)}$ reduces back to the mean-square fluctuations of the original $\mathbf{b}^{\text{ANMr2}}$, where forces are not considered. To study the effect of inter-atom forces on mean-square fluctuations, $\mathbf{b}^{(cF)}$ are computed for 1000 different sets of random F's and two different scaling factors $c \in \{1, 2\}$.

Figure 2.1 shows the distributions of the mean-square fluctuations when the inter-residue forces are applied. The periplasmic copper/silver-binding protein CusF of *E. coli* is used for the experiment. The structure reported in the PDB (id: 2QCP) has 80 residues and composed of 5 beta strands, as shown in figure 2.1(a) in a cartoon image. The cartoon is colored according to the crystallographic B-factor values. It shows that loops (orange) and tails (red) are more flexible than the beta strands (blue). In (b) and (c), the distributions of mean-square fluctuations $\mathbf{b}^{(F)}$ and $\mathbf{b}^{(2F)}$ are plotted, respectively. In both figures, the black line represents the median of the 1000 computed B-factors at each residue, while the gray band represents the range of B-factors that are between 25 and 75 percentiles (of the 1000 computed B-factors at each residue), and two outer gray lines mark the boundaries of 5 and 95 percentiles, respectively. As a reference, the B-factor $\mathbf{b}^{\text{ANMr2}}$ without forces, or $\mathbf{b}^{(0)}$, is plotted as the red line. $\mathbf{b}^{(0)}$ is nearly the same as the medians and consequently the two lines are mostly indistinguishable in the two figures.

Making the flexible regions more flexible. From figure 2.1 it is seen that the gray band that represents the range of B-factors between 25 and 75 percentiles is quite

(a)

(b)

(c)
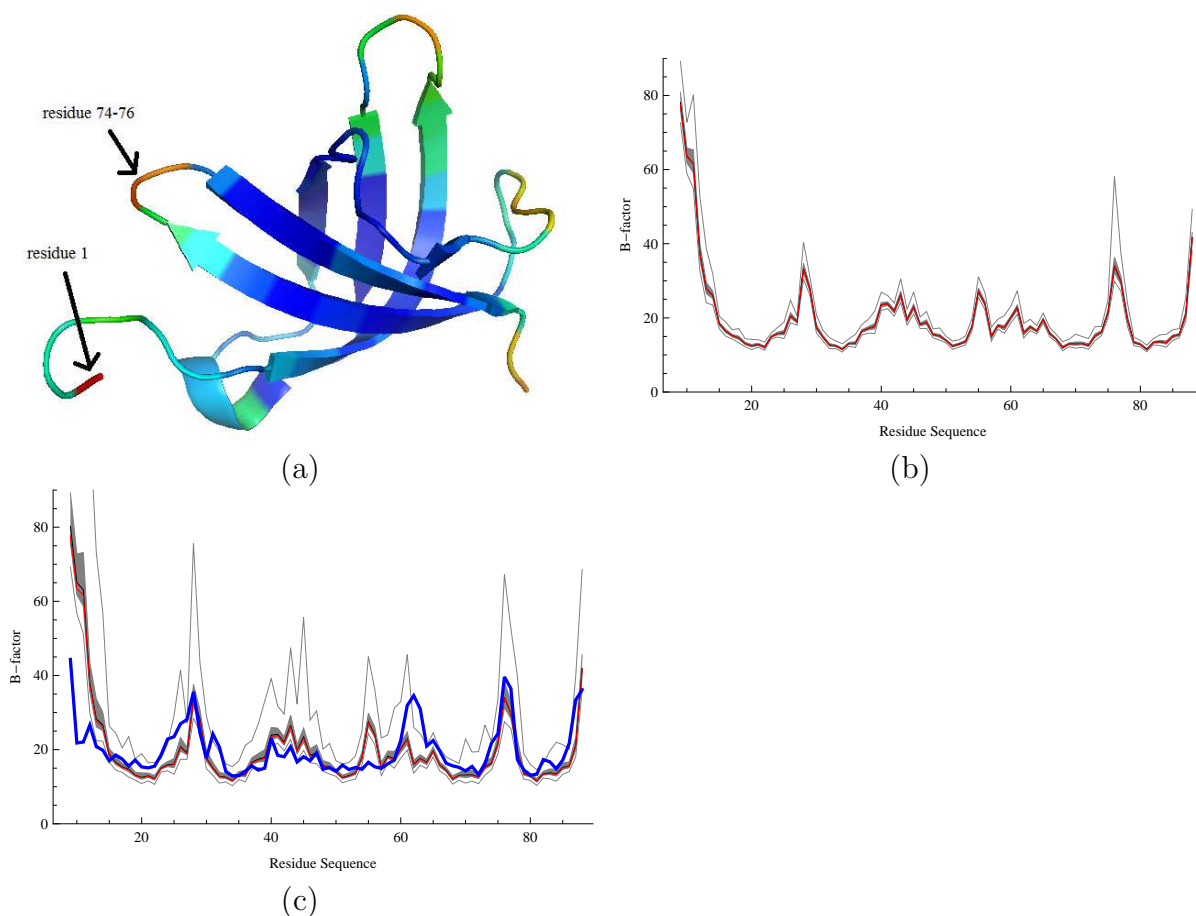
Figure 2.1 **The effect of inter-atom forces on a protein's mean-square fluctuations.** 1000 sets of inter-atom forces are applied and the variations these forces cause on each residue's mean square fluctuation are collected and statistically analyzed. (a) shows a cartoon image of the protein, the periplasmic copper/silver-binding protein CusF of E. coli (pdbid: 2QCP, resolution: 1 Å), colored according to the crystallographic B-factors of the C_α atoms. (b) the distributions of mean-square fluctuations as affected by inter-atom forces. the black line represents the median of the 1000 computed B-factors at each residue, while the gray band being the range of B-factors that are between 25 and 75 percentiles (of the 1000 computed B-factors at each residue), the two outer gray lines being the boundaries of 5 and 95 percentiles, and the red line being $\mathbf{b}^{(0)}$, or the B-factors when forces are all zero. (c) is the same as (b) except the magnitudes of all the forces are scaled by a factor of 2. The blue line in (c) represents the experimental B-factors.

narrow, closely surrounding the median line, suggesting most of the times the effect of forces on a residue's fluctuations is small. However, the 95 percentile line (the top line) shows that the inter-residue forces are capable to greatly increase a residue's fluctuation magnitude, especially for residues that are already flexible, i.e., those at or near the peaks. On the other hand, the 5 percentile line (the bottom line) indicates the inter-residue forces' effect in reducing a residue's fluctuation magnitude is relatively much smaller.

Figure 2.1(c) shows the same results as (b), except that the magnitudes of the inter-residue forces are doubled (using $2F$ instead of F). Interestingly, the increase in mean-square fluctuations (i.e., the difference between the top line and the red line) at most residues is more than doubled as a result of doubling the forces.

In summary, we have seen that, i) the forces affect the flexible regions more than others, and ii) the forces have more of an effect in increasing a residue's magnitude of fluctuation than decreasing it.

The effect of inter-residue forces on improving correlations with experimental B-factors. Also shown in the figure 2.1(c) is a thick blue line, which represents the crystallographic B-factors $\mathbf{b}^{\mathrm{PDB}}$. In the figures, $\mathbf{b}^{\mathrm{PDB}}$ is scaled by s and translated by t to minimize its superposition error with $\mathbf{b}^{\mathrm{ANMr2}}$ (the red line):

$$e(s, t, \mathbf{b}^{\mathrm{PDB}}, \mathbf{b}^{\mathrm{ANMr2}}) = \sum_{i=1}^{n} \left((\mathbf{b}^{\mathrm{PDB}})_i \cdot s + t - (\mathbf{b}^{\mathrm{ANMr2}})_i \right)^2, \tag{2.21}$$

where $(\mathbf{b})_i$ is ith element of vector $\mathbf{b}$, and n is the number of residues.

From figure 2.1(c) it is seen that the values of $\mathbf{b}^{\mathrm{PDB}}$ (the blue line) fall mostly within the 5 percentile and the 95 percentile range of $\mathbf{b}^{(2F)}$, except for residues 61-65. This implies that there may exist, for this particular protein, assignments of inter-atom forces that can greatly improve its correlations with the experimental B-factors.

2.3.3 Application to Other Proteins

We repeat the same analysis as above also for a dataset of eight other proteins that have less than 50% sequence similarity to one another and whose resolutions are better than 0.8 Å and whose lengths are from 64 to 158 amino acids, as used in [117]. These proteins are: Type III antifreeze protein rd1 (pdb id: 1ucs) (64 residues), syntenin Pdz2 domain (1r6j) (82 residues), high-potential iron-sulfur protein (1iua) (83 residues), carbohydrate Binding Domain Cbm36 (1w0n) (121 residues), Lys-49 phospholipase A2 homologue (lysine 49 PLA2) (1mc2) (122 residues), cobratoxin (1v6p) (62 residues each, two chains), bacterial photoreceptor pyp (1nwz) (125 residues), and E. Coli pyrophos-phokinase HPPK (1f9y) (158 residues).

Figure 2.2 shows the results, from which it is seen that, common to all the proteins, the experimental B-factors mostly fall within the boundaries of the 5 and 95 percentile of the 1000 computed B-factors when inter-residue forces are present. This suggests that inter-residue forces may be the reason, or part of the reason for some of the observed difference between experimental B-factors and theoretical B-factors as computed by ANMr-like models that do not consider the effect of inter-atom forces. On the other hand, we do not expect that inter-residue forces alone could account for all the differences between experimental B-factors and computed B-factors. For some of the residues shown in figure 2.2, the difference is too large to come from inter-residue forces alone. In addition, the difference may also be due to the coarse-grained nature of ANMr-like models. The uncertainties in experimental B-factors are another factor. Experimental B-factors are not an exact representation of the mean-square fluctuations of residues since they are subject to the influence of lattice disorder, crystal packing, etc. Indeed, a number of studies that took into account the effect of crystal packing found that including the effect of crystal packing significantly improved the fittings between experimental and calculated B-factors [41, 45, 63, 106, 120].
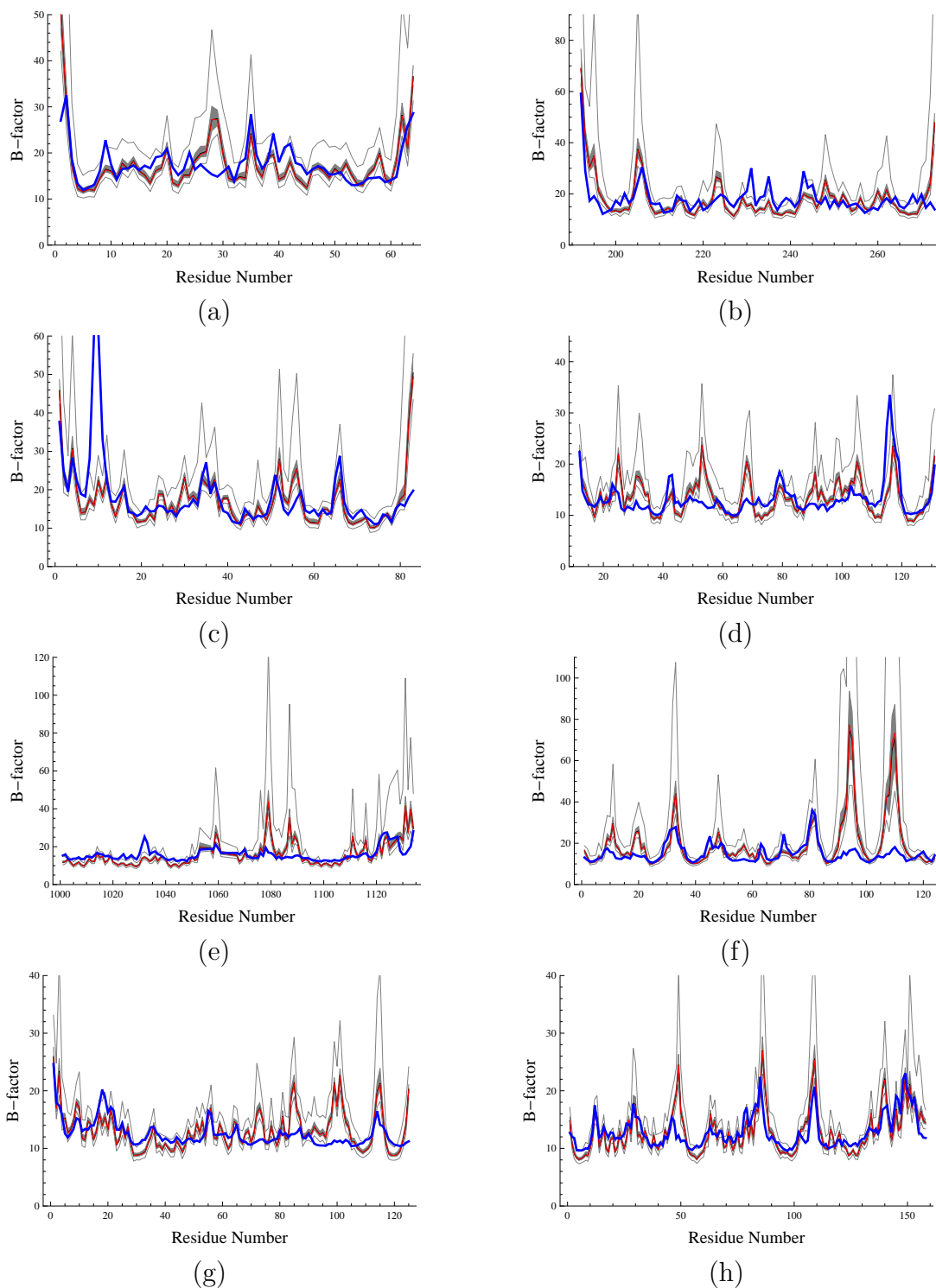
Figure 2.2 **The effect of inter-atom forces on the mean-square fluctuations of eight other proteins.** This figure is the same as figure 2.1(c) except that the proteins are different. The pdb-id for the proteins are: (a) 1UCS, (b) 1R6J, (c) 1IUA, (d) 1W0N, (e) 1MC2, (f) 1V6P, (g) 1NWZ, and (h) 1F9Y.

2.4   Discussions

This paper presents a new theoretical result that shows how GNM and ANM can be unified through taking into account the effect of inter-residue forces. The unification reveals also the role of inter-residue forces in protein fluctuation dynamics. Depending on the magnitudes and directions of the inter-residue forces, the unification presents a continuous spectrum of models that are, to various extents, a mix of GNM and ANM. Perceivably there may exist an optimal mix that is able to combine the strengths of the two models in depicting protein dynamics. When no forces are assigned, the model is reduced to ANM. An open question is whether or not there exist a way to assign the forces systematically and realistically in such coarse-grained models.

Many elastic network models have been developed to study how a protein system responds to binding or external forces. Yilmaz and Atilgan [157] discussed how one could insert minimal concerted fluctuations in a set of non-bonded contacts to regulate the motion of residues at sensory positions, and how the induced changes in inter-residue separations could be adaptively annihilated. Ming and Wall [87] modeled the interaction between a ligand and a protein using ENM by introducing a larger spring constant and a larger cutoff distance and discovered that binding at the native binding sites caused a large change in protein dynamics. Zheng et al. [160, 161] introduced perturbations in the contact spring constants to show that some residues, when their interactions with neighboring residues were perturbed, affected more strongly the frequencies of the functional modes than the other residues. They showed that these dynamically important residues were conserved. Sacquin-Mora and Lavery [110] and Eyal and Bahar [30] computed effective spring constants between pairs of residues by applying an external force to mechanically unfold a protein and found them to be in good correlation with experimental unfolding forces. Atilgan et al [3, 4] developed a perturbation scanning method that computes conformation changes in response to external forces. These models demon-

strate how perturbations in spring constants and/or contact distances affect the normal modes or how external forces induce conformation changes. In several of these models, inter-residue forces were explicitly included in the formulation when computing how a protein system responds to the perturbations. However, in all of these models, at the initial equilibrium state inter-residue forces have always been set to zero. That is to say, in all of these models, the state before any perturbation takes place contains only relaxed springs and no inter-residue forces. The uniqueness of our work is the presence of inter-residue forces even at the initial equilibrium state and we have studied how inter-residue forces affect protein fluctuation dynamics at the equilibrium state.

It would be highly interesting to extend our work to see how having a different initial equilibrium state, in which explicit inter-residues forces are present, may alter the outcomes of the aforementioned force-related ENM studies in which inter-residue forces were absent at the initial equilibrium state. For example, if one explicitly considers the inter-residue forces at the equilibrium state, will they still be able to identify the same set of dynamically conserved residues as found in [160, 161]? and how would it affect the simulation results of mechanical unfolding presented in [30, 110]? etc.

This work thus opens a new way to study the effect of external forces on protein fluctuations and dynamics using elastic network models. Binding is one of the most fundamental processes in protein functions. Ligand binding exerts external forces on a host protein, causing it to rebalance itself, in which process the host protein often makes noticeable conformation changes, moving from the initial open form (apo) to the final "closed" form. The closed form may display different motion patterns from the open form. Such changes in the motion patterns are often attributed to the conformation differences between the ligand-free state and the ligand-bound state. However, external forces exert their influence not only by inducing conformation changes, but also by creating strains, or inter-atom forces and/or torques, within the host protein. As our results demonstrate here, such strains also are able to affect directly the host protein's fluctuation dynamics.

One possible benefit of considering the effect of inter-atom forces is that it may provide insight into situations where there is much dynamics change but little structure change, such as when a ligand's binding causes little change in the conformation of the host protein but much change in its fluctuation patterns. For example, experimental studies on hen egg lysozyme showed the binding of several antibodies causes little conformation change in the host protein but significant changes in the hydrogen exchange protection factors that are closely related to the magnitudes of local fluctuations [34, 54]. This is puzzling and difficult to understand using elastic network models whose output depends solely on the shape and geometry of the input structures, since with such models it is generally expected that the same conformation produces the same dynamics and little change in the conformations means little change in the fluctuations. However, incorporating the effect of inter-residue forces into elastic network models as done here changes this general picture. As seen in figure 2.2, the same conformations can have different fluctuation dynamics when different inter-residue forces are applied. Under this new paradigm, it becomes possible to interpret the aforementioned puzzle regarding antibody binding using elastic network models: even though there is little conformation change taking place to interpret the significant changes in fluctuations, it is possible that the observed changes in fluctuations are caused by the strains created by the ligand binding. Other feasible interpretations of this phenomenon also exist [34].

It is worth noting that FSM as of now is not yet readily applicable to dynamics studies of real protein systems, due to the fact that a systematical way for assigning inter-residue forces is still lacking. The purpose of our experiment with artificial random forces is not to establish FSM as a new mature model. Instead, the experiment serves only to illustrate to what extent inter-residue forces may *potentially* affect B-factor computations. The next challenge is thus to find a way to assign inter-residue forces systematically and meaningfully and correctly for a given structure. This is an important open question and is beyond the scope of this work. Once this problem is solved, FSM may then be applied to tackle the aforementioned problems.

Inter-residue forces may have a significant role in better understanding protein conformation changes and other function related motions. We plan to investigate this in the future after we have found a way to systematically estimate inter-residue forces.

Acknowledgement

CHAPTER 3.   BRIDGING BETWEEN NMA AND ELASTIC NETWORK MODELS

Hyuntae Na[2][3] and Guang Song[2][3][4]

Abstract

Normal mode analysis (NMA) has been a powerful tool for studying protein dynamics. Elastic network models (ENM), through their simplicity, have made normal mode computations accessible to a much broader research community and for many more biomolecular systems. The drawback of ENMs, however, is that they are less accurate than NMA. In this work, through steps of simplification that starts with NMA and ends with elastic network models we build a tight connection between NMA and elastic network models. In the process of bridging between the two, we have also discovered several high-quality simplified models. Our best simplified model has a mean correlation with the original NMA that is as high as 0.88. In addition, the model is force-field independent and does not require energy minimization, and thus can be applied directly to experimental structures. Another benefit of drawing the connection is a clearer under-

---

[1]This chapter is reprinted with permission of *Proteins* 2014, 82(9), 2157–2168.

[2]Graduate student and Associate Professor, respectively, Department of Computer Science, Iowa State University.

[3]Primary researchers and authors.

[4]Author for correspondence.

standing why elastic network models work well and how it can be further improved. We discover that ANM can be greatly enhanced by including an additional torsional term and a geometry term.

3.1 Introduction

Protein dynamics has long been recognized to be critical to protein functions. It also has been recognized that function related motions occur mostly along a small number of collective coordinates, the normal modes. Though protein conformation changes are often anharmonic in nature, it has been shown that conformation changes mostly take place along one or a few harmonic normal modes. Normal mode analysis (NMA) thus has been a popular and powerful tool for studying protein motions and dynamics since the 80's. [17, 38, 72]

To apply NMA to compute normal modes, a given structure has to be first energetically minimized. This minimization process uses a complex semi-empirical potential and takes a significant amount of computer time and memory, especially for large systems. Moreover, the minimized structure usually deviates from the original structure, partly due to the fact that the minimization is mostly done in vacuo, [81] and partly due to the imperfection in the force field potentials.

In mid 90's, a seminal work by Tirion [139] showed that a much simpler potential, a single parameter potential, was sufficient to reproduce the slow dynamics in a quality comparable to that by a detailed, complex potential. Her work thus greatly simplified the process of computing normal modes and opened the way for a flurry of research activities in normal mode computations, analysis, and applications. [2, 8, 25, 43, 47, 58, 63, 66, 70, 76, 77, 94, 108, 121, 127, 129, 130, 149, 153, 154, 159, 161, 163] Many models have been developed to compute normal modes and applied to study protein dynamics.

There is no doubt that one of the main contributors to the popularization of normal mode computations and their success is simplicity. The simplicity, achieved by using a much simpler potential that requires no energy minimization, certainly presents a great advantage over the original NMA. However, with the simplification also comes a loss of accuracy. While there is significant gain from exploiting the simplicity of these simplified models and from what they can do, there are few works that examine these models' accuracy in comparison to the original NMA. [60, 163]

The validity of Tirion's simplification of NMA's full potential with a single-parameter potential, as well as that of many others, was justified a posteriori, for example, by showing that the slow dynamics computed by the simplified model matches with that by NMA using a full potential. [139] It is often not obvious why such simplified models should work, though some insightful comments and plausible explanations have been given. [137]

In this work, we bridge NMA and elastic network models and in so doing have developed a different approach to derive simplified NMA models. Instead of creating from the beginning a new model and justifying it *a posteriori*, we start with the original NMA, identify what is essential to its accuracy, and then take reasonable steps of simplification. The advantages of proceeding in this way are several. First, in so doing we are able to simplify NMA while preserving its essential components and keeping track of its accuracy. Second, in this process of starting with NMA, simplifying it in steps, and finally arriving at the commonly used elastic network models, a bridge is being built connecting all-atom NMA and elastic network models. This bridge is an important way to show the tight connection between usual atomic models and the elastic network models. Thirdly, the connection clearly reveals the reasons why elastic network models work well and in what ways they can be improved to have a better agreement with the original NMA.

3.2   Methods

In this section, we will first give an overview of NMA, and then describe in details the steps we take to simplify it.

3.2.1   Overview of NMA

NMA uses a complex all-atom force-field potential, such as Amber, [145] CHARMM, [83] etc., that contains many interaction terms and can be divided into two-body, three-body, or four-body interactions. Two-body interactions include bond stretching, van der Waals interactions, and electrostatic interactions. Three-body interactions include the bond angle interactions, while the four-body interactions are those via dihedral angles. Before applying NMA to a protein system to study its normal mode motions, a given input structure has to be first energetically minimized. This minimization process takes a significant amount of computer time and memory, especially for large systems. Moreover, the minimized structure usually deviates from the original structure.

It is important to realize that at the minimized structure, even though the whole system is at equilibrium and the net force at each atom is brought to be zero, the interatomic forces are not necessarily zero. For example, there are atom pairs that interact only through non-bonded interactions, such as the electrostatic interactions, and the forces they exert on each other persist even at the equilibrium state.

Once the input structure is energetically minimized, the Hessian matrix can be written out, from which normal modes can be obtained by solving for its eigenvalues and eigenvectors. As the second derivative of the force-field potential, Hessian matrix depends directly on force field parameters, many of which are the spring constants of various kinds, such as bond stretching spring constants, bond angle spring constants, torsional spring constants, and improper angle spring constants, etc.

A key realization in our simplification of NMA is that Hessian matrix depends not only on these spring constants, but also on the inter-atomic forces or torques. Whether it is of a two-body potential, or a three-body or four-body one, the following derivation shows that the Hessian matrix, as a second derivative of the potentials, can always be written as a summation of a spring-based term and a force-based term. The spring-based term specifies the contributions from the force-field spring constants, while the force-based term specifies contributions from the inter-atomic forces or torques, which as we reasoned above, are not zero even at the equilibrium structure.

3.2.2 NMA Hessian Matrix as a Summation of Spring-Based Terms and Force-Based Terms

As aforementioned, the first key realization in simplifying NMA is that NMA Hessian matrix, as a second derivative of the potential, consists of the two kinds of contributions. One is related to force field spring constants while the other the inter-atomic forces or torques.

First, let us consider the three-body potential, specifically that of the bond angle interactions. Let $\theta = \angle ijk$ be the instantaneous angle formed by three atoms i, j, and k. The bond angle potential of atoms i, j, and k is defined as $V_\theta = \frac{1}{2}k_\theta(\theta - \theta_0)^2$, where k_θ is the bond angle spring constant, and θ_0 is the equilibrium angle. The block Hessian matrix H_θ for the bond angle interaction is a 9×9 second derivative matrix of V_θ with respect to x, y, and z coordinates of atoms i, j, and k. Write one component $\frac{\partial V_\theta}{\partial X_i \partial Y_k}$ of H_θ as follows:

$$
\begin{aligned}
\frac{\partial V_\theta}{\partial X_i \partial Y_k} &= \frac{\partial}{\partial Y_k}\left(\frac{\partial V_\theta}{\partial \theta}\frac{\partial \theta}{\partial X_i}\right) \\
&= \frac{\partial^2 V_\theta}{\partial \theta^2}\frac{\partial \theta}{\partial X_i}\frac{\partial \theta}{\partial Y_k} + \frac{\partial V_\theta}{\partial \theta}\frac{\partial^2 \theta}{\partial X_i \partial Y_k} \\
&= k_\theta \cdot \frac{\partial \theta}{\partial X_i}\frac{\partial \theta}{\partial Y_k} - f_\theta \cdot \frac{\partial^2 \theta}{\partial X_i \partial Y_k},
\end{aligned} \tag{3.1}
$$

where $f_\theta = -\frac{\partial V_\theta}{\partial \theta}$ is the bending force (which is actually a torque). Notice that Eq. (3.1)

is a combination of the physical terms (k_θ and f_θ) and geometric terms (the partial derivatives), which represent the projection of physical interactions into a particular coordinate system. In a similar fashion, the rest of the elements of the block Hessian matrix H_θ can be written out using k_θ and f_θ. Finally, the block Hessian matrix H_θ can be rewritten as a summation of two terms:

$$H_\theta = k_\theta \cdot H_{\theta|k_\theta} - f_\theta \cdot H_{\theta|f_\theta}, \tag{3.2}$$

where $H_{\theta|k_\theta}$ and $H_{\theta|f_\theta}$ are 9×9 matrices that are fully determined by protein geometry and atom coordinates, k_θ is a force field parameter, and $f_\theta = -k_\theta(\theta - \theta_0)$ is the torque acting on the bond angle.

Now for the four-body interactions, let H_ϕ be the 12×12 block Hessian matrix for the torsional interaction among four atoms i, j, k, and l. Let $k_\phi = \frac{\partial^2 V}{\partial \phi^2}$ and $f_\phi = -\frac{\partial V}{\partial \phi}$ be the torsional spring constant and the torsional bending force (torque), respectively. Similar to Eq. (3.2), the Hessian matrix H_ϕ can be written as a function of k_ϕ and f_ϕ:

$$H_\phi = k_\phi \cdot H_{\phi|k_\phi} - f_\phi \cdot H_{\phi|f_\phi}. \tag{3.3}$$

Since $V(\phi) = K_\phi(1 - \cos(n(\phi - \phi_0)))$ in most force fields, where K_ϕ and ϕ_0 are force field parameters, and n is the multiplicity, $k_\phi = \frac{\partial^2 V}{\partial \phi^2} = n^2 K_\phi \cos(n(\phi - \phi_0))$.

Likewise, the Hessian matrix H_l for two-body interactions between a pair of atoms i and j can be determined:

$$H_l = k_l \cdot H_{l|k_l} - f_l \cdot H_{l|f_l}. \tag{3.4}$$

There are usually three types of two-body interactions in an all-atom potential, i.e., bond stretching, van der Waals interactions, and electrostatic interactions, and thus three different k_l's. For the bond stretching potential V_{bond}, which is usually expressed as $V_{\mathrm{bond}} = K_{\mathrm{bond}}(r - r_0)^2$, we have,

$$k_l(\mathrm{bond}) = \frac{\partial^2 V_{\mathrm{bond}}}{\partial r^2} = 2K_{\mathrm{bond}}. \tag{3.5}$$

For the van der Waals term, since $V_{\text{vdW}} = \epsilon((\frac{r_0}{r})^{12} - 2(\frac{r_0}{r})^6)$, we have,

$$k_l(\text{vdW}) = \frac{\partial^2 V_{\text{vdW}}}{\partial r^2} = \frac{12\epsilon}{r^2}\left(13\left(\frac{r_0}{r}\right)^{12} - 7\left(\frac{r_0}{r}\right)^6\right). \tag{3.6}$$

Lastly, for the electrostatic term, since $V_{\text{elec}} = \frac{332q_i \cdot q_j}{rD}$, where q_i is partial charge of atom i, and D is the dielectric constant and is set to be 1, k_l is thus:

$$k_l(\text{elec}) = \frac{\partial^2 V_{\text{elec}}}{\partial r^2} = \frac{2 \cdot 332 q_i \cdot q_j}{r^3} = \frac{664 \cdot q_i \cdot q_j}{r^3}. \tag{3.7}$$

Finally, given n the number of atoms, the $3n \times 3n$ full Hessian matrix H^{NMA} for the whole system can be written as a summation of a spring constant based term $H^{\text{NMA}}_{\text{spr}}$ and a force/torque based term $H^{\text{NMA}}_{\text{frc}}$:

$$H^{\text{NMA}} = H^{\text{NMA}}_{\text{spr}} + H^{\text{NMA}}_{\text{frc}}, \tag{3.8}$$

where

$$H^{\text{NMA}}_{\text{spr}} = \sum_{\theta \in \Theta} k_\theta H_{\theta|k_\theta} + \sum_{\phi \in \Phi} k_\phi H_{\phi|k_\phi} + \sum_{l \in L} k_l H_{l|k_l},$$

$$H^{\text{NMA}}_{\text{frc}} = -\left(\sum_{\theta \in \Theta} f_\theta H_{\theta|f_\theta} + \sum_{\phi \in \Phi} f_\phi H_{\phi|f_\phi} + \sum_{l \in L} f_l H_{l|f_l}\right),$$

where Θ, Φ, and L are the sets of angular, dihedral, and pairwise interactions, respectively.

3.2.3 First Step of Simplification: the Spring-Only NMA

A second key realization in simplifying NMA is that force/torque-based terms contribute much less than their corresponding spring-based terms to protein fluctuation dynamics.[1] Consequently, omitting them incurs only a small deterioration to a model's

---

[1] In (Na and Song, 2013, A Natural Unification of GNM and ANM, *under review*), we show that the effect of any inter-atomic force $f_{i,j}$ between atoms i and j is about the same as adding an additional spring whose constant is $k^{\text{eff}}_{i,j} = \frac{f_{i,j}}{r_{i,j}}$, where $r_{i,j}$ is the distance between atoms i and j. This effective spring constant $k^{\text{eff}}_{i,j}$ is usually much weaker than the actual spring $k_{i,j}$ between atoms i and j. Take the bond stretching term for example, $f_{i,j} = k_{i,j} \cdot \Delta r_{i,j}$. Therefore, the ratio $\frac{k^{\text{eff}}_{i,j}}{k_{i,j}} = \frac{\Delta r_{i,j}}{r_{i,j}} \ll 1$. Similar arguments can be made for other force terms.

accuracy. On the other hand, one huge gain in omitting the force/torque-based terms is that it becomes much easier to write down the Hessian matrix, since forces are more difficult to be estimated correctly than the spring constants. The absence of forces in a model also makes energy minimization unnecessary.

Therefore, our first step of simplification is to use a spring-only NMA. We name this model sbNMA, or spring-based NMA.

sbNMA assumes that all forces and torques are zero. Therefore,

$$H^{\text{sbNMA}} = H^{\text{NMA}}_{\text{spr}} = \sum_{\theta \in \Theta} k_\theta H_{\theta|k_\theta} + \sum_{\phi \in \Phi} k_\phi H_{\phi|k_\phi} + \sum_{l \in L} k_l H_{l|k_l}, \tag{3.9}$$

where k_l represents all the two-body spring constants and includes bond stretching spring constants $k_l(\text{bond})$ (see Eq. (3.5)), spring constants due to van der Waals interactions $k_l(\text{vdw})$ (see Eq. (3.6)), and spring constants due to electrostatic interactions $k_l(\text{elec})$ (see Eq. (3.7)).

Like elastic network models, sbNMA is a fully spring-based model. However, in order to apply sbNMA to compute normal modes and mean-square fluctuations, another step of approximation is needed. This is because sbNMA as of now contains springs with negative spring constants. Negative spring constants can cause the input protein structure to become unstable by setting it at a saddle point. Mathematically, the Hessian matrix will have negative eigenvalues.

To set an input structure at equilibrium and to avoid having negative eigenvalues in the sbNMA Hessian matrix, the following approximations are made.

First, sbNMA assumes that the input structure has the equilibrium values for all its torsional angles, i.e., $\phi = \phi_0$, as normally done in Go-like potential. [125] Therefore,

$$k_\phi = n^2 K_\phi \cos(n(\phi - \phi_0)) = n^2 K_\phi. \tag{3.10}$$

This will guarantee that k_ϕ are always positive. It also removes the dependence on the force-field parameter ϕ_0.

Secondly, it is possible that spring constants due to van der Waals interactions, $k_l(\text{vdw})$ (see Eq. (3.6)), and spring constants due to electrostatic interactions, $k_l(\text{elec})$ (see Eq. (3.7)), may be negative.

Figure 3.1(A) shows the distribution of the spring constants for electrostatic interactions, $k_l(\text{elec})$, between all pairs of atoms. Each dot in the figure represents one spring constant $k_l(\text{elec})$ between one pair of atoms, computed using Eq. (3.7), with the coordinates taken from the minimized structure of one of the proteins in the dataset (pdb-id: 2XRH.pdb). The minimization has been done using the Tinker program with the CHARMM22 force field. The partial charges, q_i and q_j in Eq. (3.7), of all the atoms are taken also from CHARMM22 force field. Similarly, Figure 3.1(B) shows the distribution of the spring constants for van der Waals interactions between all pairs of atoms, computed using Eq. (3.6) and the same minimized structure. All van der Waals parameters, which are atom-specific, are taken from CHARMM22 force field.
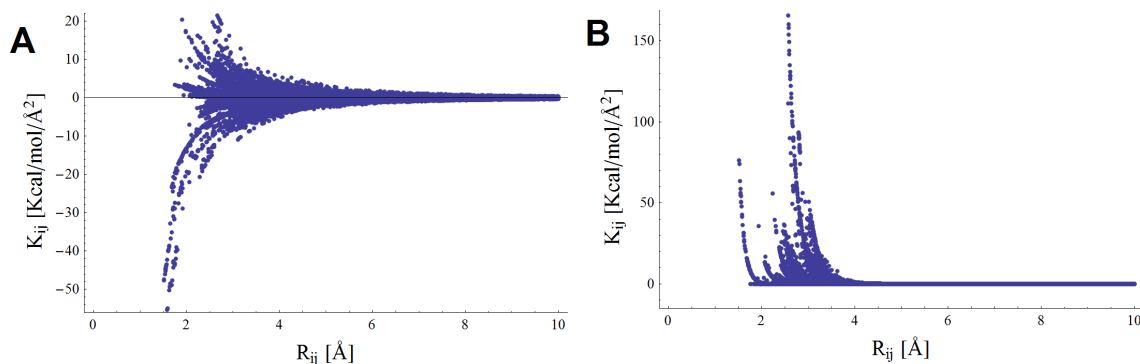
Figure 3.1   **The distributions of (A) electrostatic spring constants $k_l(\text{elec})$ and (B) van der Waals spring constants $k_l(\text{vdW})$ over r_{ij}.** The distributions are based on protein 2XRH.pdb. Other proteins have a similar distribution. Most of the contributions from van der Waals interactions come from the range where $r_{i,j}$ is around or less than 4 Å, where $k_l(\text{vdW})$ is large. $k_l(\text{elec})$ is nearly symmetric, having nearly equal numbers of positive and negative spring constants.

From Figure 3.1 it is seen that the magnitudes of electrostatic interactions-based spring constants $k_l(\text{elec})$ appear to be fairly large, especially at the short range, but are

a few times smaller than those of van der Waals interactions. On the other hand, $k_l(\text{elec})$ decreases more slowly as the separation distance increases and has a longer interaction range than $k_l(\text{vdW})$, as expected. Another contrast between the two is that, while the van der Waals based spring constants are mostly positive, electrostatic interactions-based spring constants have nearly equal number of positive and negative values and their distributions appear to be nearly symmetric along the abscissa (see Fig. 3.1(A)). The net effect of these quite large positive and negative spring constants on protein fluctuations however are smaller than they appear as they mostly cancel out each other. This may not be obvious at atomic level, but if we zoom out a little and look at a protein at the residue level, we see most residues as single units have a net charge of zero. Therefore, the strength of the electrostatic interactions among most of the residues is on the order of dipole-dipole interactions or even multi-pole multi-pole interactions, which are much weaker. Consequently, the contribution of the electrostatic interactions to the Hessian matrix and to a protein's fluctuations is much smaller than what the magnitudes of atomic-level spring constants would suggest. Secondly, at short ranges (i.e., small R_{ij}) where their contributions are large, they are dominated by the even larger spring constants from the van der Waals (Fig. 3.1(B)). For these reasons and to avoid negative spring constants, we choose not to include the electrostatic term in sbNMA. Later experimental results where a strong correlation between sbNMA and NMA is found (see the first row in Table 3.1) further confirm that this is a reasonable approximation.

Notice that the van der Waals based spring constants $k_l(\text{vdW})$ (see Fig. 3.1(B)) also may become negative. However, most of this happens when r is large, where the magnitude of $k_{i,j}$ is extremely small. Most of the contributions from van der Waals interactions come from when r is around or less than r_0, the equilibrium distance between a pair of atoms and which is usually near or below 4 Å. Thus, we set $k_l(\text{vdW}) = \max(k_l(\text{vdW}), 0)$, to ensure that $k_l(\text{vdW})$ is non-negative.

Table 3.1   The effect of different modelings of protein geometry on the fluctuation dynamics

Case	H_{geom}	H_ϕ	H_{nbond}	corr. with NMA
0 (sbNMA)	ff[a]	ff	ff	0.878
1	inf[b]	ff	ff	0.867
2	$100\times$const[c]	ff	ff	0.871
3	$0.01\times$const	ff	ff	0.816
4	const	ff	ff	0.878

[a]ff stands for force field, where rigidity of bond stretching, bond angle terms, etc. are modeled according the force fields;
[b]inf stands for infinity, where all protein geometry related spring constants are set to be infinity, i.e., are set to be totally rigid;
[c]const means a force-field independent, atom-type independent uniform constant values are used. For example, 340 Kcal/mol/$\mathring{A}^2$ is used for all the bond stretching terms.

With above approximations to remove negative spring constants, H^{sbNMA} becomes,

$$H^{\text{sbNMA}} = k_\theta \cdot H_{\theta|k_\theta} + \left(n^2 K_\phi + K_{\text{improper}}\right) \cdot H_{\phi|k_\phi} + (k_l(\text{bond}) + k_l(\text{UB}) + k_l(\text{vdW})) \cdot H_{l|k_l},$$

(3.11)

where the improper angle and the Urey-Bradley terms are also considered and K_{improper} and $k_l(\text{UB})$ represents their spring constants, respectively.

Our work in [77] details how $H_{\theta|k_\theta}$, $H_{\phi|k_\phi}$, and $H_{l|k_l}$ can be computed. To compute H^{NMA}, one may use software packages such as Amber, [111] Charmm, [18] Tinker, [102] or Gromacs, [103] etc.

3.2.4   Further Simplification of NMA: Approximating the Force Field Parameters with a Small Set of Constants

sbNMA as formulated above is a fully spring-based model like elastic network models. However, like NMA, it still uses extensive parameters from a force field, such as bond stretching spring constants, bond angle spring constants, parameters for van der Waals interactions, etc.

Our goal in this section is to identify what are the essential ingredients of sbNMA and what are not, and then to simplify the model while keeping all of its essential ingredients. The process of doing this (see the Results section) leads us to approximate the extensive force field parameters with a small set of force-field independent constants. This is important in making the tight connection between NMA and elastic network models.

Our hypothesis is that, i) bond stretching, bond angle, and the improper angle terms have large spring constants already and variations in their values should have only minor effect on the fluctuation dynamics; ii) torsional interactions, on the other hand, have much smaller spring constants and thus much of a protein flexibility should come from the torsional degrees of freedom; iii) the effect of non-bonded interactions on protein fluctuations is mainly contributed by the van der Waals term, while the contribution from electrostatic is smaller. Non-bonded interactions further reduce and modify a protein's flexibility that originates mostly from the torsional degrees of freedom. A proper modeling of the non-bonded interactions should be important.

Specifically, we make the following simplifications:

1. use one and the same K_{bond}, K_θ, K_ϕ, K_{improper}, or K_{UB} for all proteins;

2. use a single generic set of van der Waals radii, such as the Bondi radii, [12] for van der Waals interactions.

A significant advantage of making this simplification is that the model is now force-field independent. We name this further simplified model ssNMA, or simplified spring-based NMA, to distinguish it from sbNMA. ssNMA as of now is highly similar to other elastic network models, and as with elastic network models, one can write down the ssNMA Hessian matrix without resorting to a force-field. ssNMA and sbNMA represent intermediate models that bridge NMA and elastic network models and show how they are connected. As will become clear in the Results section, the connection between two

sheds insights on their relationship and allows one to see how elastic network models such as ANM can be further improved so that it matches better with NMA.

3.2.5 ANM Model

One of the most widely used elastic network models is Anisotropic Network Model, [2] or ANM. In ANM, the spring constant between atoms i and j is defined as follows:

$$k_{i,j}^{\text{ANM}} \quad = \quad \begin{cases} 1, & \text{if } r_{i,j} < r_{\text{cutoff}} \\ 0, & \text{otherwise,} \end{cases} \qquad (3.12)$$

where $r_{i,j}$ is the Euclidean distance between atoms i and j, and r_{cutoff} is the cutoff distance, which is a parameter in the model. In this work, to draw the tight connection between NMA and elastic network models, we use the fine-grained ANM, where all the atoms are included and are mass-weighted.

3.3 Results and Discussions

3.3.1 The Close Match between sbNMA and NMA

We first apply the sbNMA model to a large number of proteins and show that the fluctuation dynamics produced by sbNMA matches closely with that of NMA. To exclude the potential bias created by crystal packing or lattice disorder on protein fluctuations, the atomic fluctuations computed from NMA and sbNMA are compared with each other and not with the experimental B-factors. However, interested readers may refer to Supplemental Materials for the performance of the models developed in this work in their correlations with *experimental B-factors*. Such correlations must be interpreted with caution as experimental B-factors are strongly affected by crystal packing, lattice order, etc., which are not considered in these models.

To compute the fluctuations, all the structures are first energetically minimized using the Tinker program [102] with the CHARMM22 force field. [83] The minimized structures

are then used by NMA and sbNMA, and later on, ANM model, to compute the mean-square fluctuations. Force field parameters in CHARMM22 are used in computing the sbNMA Hessian matrix.

Let M be the $n \times n$ diagonal mass matrix, I be the 3×3 identity matrix, and $\otimes$ be the operator of the Kronecker product. Denote $\mathbf{b}^{\text{NMA}}$ and $\mathbf{b}^{\text{sbNMA}}$ as the mean-square fluctuations by NMA and sbNMA, respectively. The following procedure is used to compute them:

1. Use Tinker to run the energy minimization and determine the minimized structure $\mathcal{C}$, whose potential energy as defined by CHARMM22 is locally minimized;

2. Compute the mean-square fluctuations $\mathbf{b}^{\text{NMA}}$ using the Hessian matrix (provided by Tinker) of the minimized structure;

3. Compute the sbNMA Hessian matrix H^{sbNMA} of $\mathcal{C}$, whose parameters are from CHARMM22;

4. Determine frequencies f_i and modes $\mathbf{m}_i$ of H^{sbNMA} in the mass-weighted Cartesian coordinate as follows, where $i = 7, 8, ..., 3n$:

    (a) $\tilde{H}^{\text{sbNMA}} \leftarrow (M^{1/2} \otimes I_3)^{-1} H^{\text{sbNMA}} (M^{1/2} \otimes I_3)^{-1}$;

    (b) $\langle f_i, \tilde{\mathbf{m}}_i \rangle \leftarrow i$th eigenvalue and eigenvector of $\tilde{H}^{\text{sbNMA}}$;

    (c) $\mathbf{m}_i \leftarrow (M^{1/2} \otimes I_3)^{-1} \tilde{\mathbf{m}}_i$;

5. Compute the mean-square fluctuations $\mathbf{b}^{\text{sbNMA}}$ using f_i and $\mathbf{m}_i$;

6. Compute the correlation between $\mathbf{b}^{\text{NMA}}$ and $\mathbf{b}^{\text{sbNMA}}$.

The procedure is repeated on a dataset of 177 proteins that have less than 30% sequence similarity. All these 177 proteins are high-resolution crystal structures, containing the ANISOU entries for anisotropic B-factors, and whose sizes are greater or equal to 60 residues but less than 150, due to the computational costs of running NMA. Structures

that fail to pass Tinker's energy minimization or Hessian matrix computation procedure are excluded. The pdb-id's of the whole list of the proteins are given in the Supporting Information. Figure 3.2 shows the histogram of the root mean square deviations from the crystal structures after the energy minimization. It is seen that for most proteins, the structure deviation falls within 2-3 Å, but some are further away.
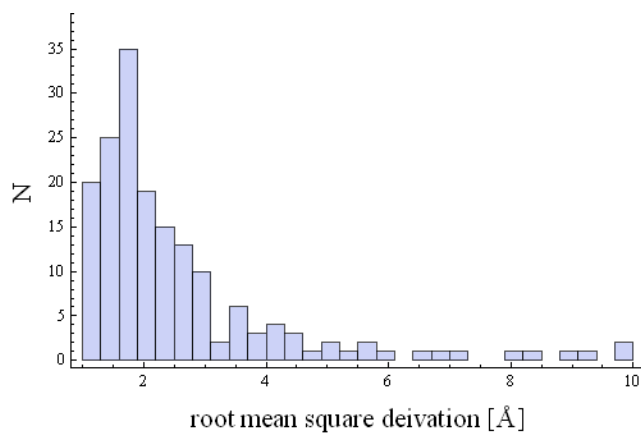
Figure 3.2 **The histogram of the C_α root mean square deviations from the crystal structures after the energy minimization for the 177 proteins used in this study.**

The high correlations between $\mathbf{b}^{\text{NMA}}$ and $\mathbf{b}^{\text{sbNMA}}$ are shown in Fig. 3.3 as a histogram. The mean correlation value is as high as 0.88, indicating that sbNMA represents a high-quality approximation to NMA. The results also confirm that the spring-based terms in the NMA Hessian matrix indeed have a much bigger contribution than the force-based terms.

3.3.2 Identifying the Essential Components of sbNMA and Further Simplification

In a regular force field for proteins, there are hundreds and perhaps even thousands of parameters. For example, there are over a hundred finer atom types in most force fields. The fine distinctions among atom types and the interactions that depend on them are considered necessary for accurate molecular dynamics simulations and studies.
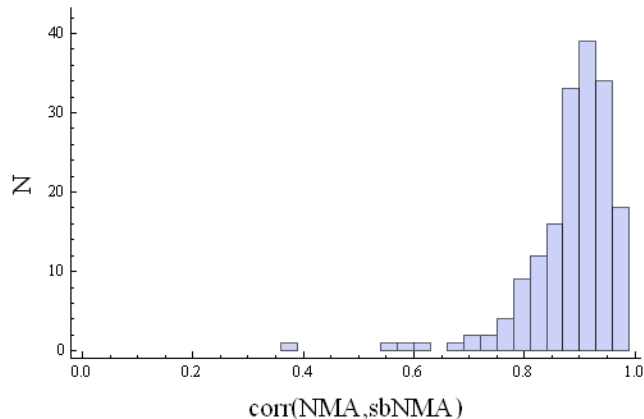
Figure 3.3 **The histogram of the correlations between the mean-square fluc-
tuations computed by NMA and by sbNMA.** The mean correlation
is 0.88. 177 proteins are used.

However, such fine distinctions may not be necessary in our model where the main aim is
to produce high-quality normal modes that resemble closely those of NMA. Wherever the
fine distinctions are unnecessary for this purpose, our model can be simplified without
sacrificing any or much accuracy. Indeed, the model will be easier to use if only a small
number of parameters are needed and the preparation process for the computation of
the normal modes is simplified. This simplification is also important in making the tight
connections between NMA and elastic network models.

To this end, we first divide all the terms that contribute to sbNMA Hessian matrix
into three groups. Let H_ϕ denote the part of Hessian matrix contributed by the torsional
term, H_{nbond} that by the non-bonded terms, specifically the van der Waals term. Finally
let H_{geom} denote the rest of terms, which include the bond stretching, bond angle, Urey-
Bradley, and the improper terms. The full Hessian matrix is the sum of all these three
terms, i.e.,

$$H = H_{\mathrm{geom}} + H_\phi + H_{\mathrm{nbond}}. \tag{3.13}$$

Note that the terms in H_{geom} are mainly for maintaining protein geometry. They spec-
ify interactions within parts of a protein that are mostly rigid and thus large spring
constants are used. The inclusion of this term is important for representing parts of a

protein as nearly rigid while rendering the rest as flexible. Indeed, A protein's flexibility comes mainly from the remaining degrees of freedom, the torsional rotations. Torsional rotations are not totally free. H_ϕ specifies the effect of angular springs that constrain torsional rotations. As will be seen later, a proper modeling of H_ϕ is essential for accurately reproducing protein flexibility. Unfortunately, most contact-based protein models, such as most elastic network models, do not have this term. Lastly, H_{nbond} represents the contributions of non-bonded interactions. Non-bonded interactions further constrain a protein's fluctuations by reducing their scales and modifying the fluctuation patterns. Without them a protein's fluctuations would be totally characterized by a sum of many independent torsional rotations/fluctuations. Non-bonded interactions cause these otherwise independent torsional rotations to become entangled and the final fluctuation patterns of the system to become more complex. A proper modeling of non-bonded interactions is thus important, since whatever non-bonded interactions are employed specify how the torsional rotation modes are to be mixed. Most elastic network models use uniform cutoff distance and uniform springs to define non-bonded interactions. We will show below that using proper van der Waals radii to define non-bonded interactions can bring significantly better fluctuation dynamics.

In the following, we will show the quantitative effect of these three terms on the fluctuation dynamics. This will allow us to identify the most essential ingredients of a good model. Intending to simplify the model, we will examine whether or not these three terms can be well approximated by force field independent constant values. If such an approximation can be done, it will greatly simplify NMA and connect it with elastic network models.

Tables 3.1-3.2 give the comparisons among the cases where the three different terms are approximated with constant values. Table 3.1 examines the effect of different modelings of H_{geom}, the protein geometry related terms, on the fluctuation dynamics. The first row, where all the terms take force field values (ff), represents the un-simplified

sbNMA. The rows below list different approximations of the H_{geom} term. "const" means that a single force-field independent, atom-type independent uniform spring constant is used. For example, a single spring constant 340 Kcal/mol/Å^2 is used for all the bond stretching terms, 45 Kcal/mol/rad^2 for all bond angles, 70 Kcal/mol/rad^2 for all the impropers, and 10 Kcal/mol/Å^2 for Urey-Bradley. These values are estimated averages over the ranges of values that are used in the CHARMM force field [83] for K_{bond}, K_{θ}, K_{improper}, and K_{UB}. The results presented in this paper are quite insensitive to these parameters. Other force fields, such as Amber [145], having force field parameters in similar ranges, would render similar estimated averages. That is why we say our parameters are force-field independent. It is clearly seen from Table 3.1 that approximation using constant values as these gives nearly the same results as the otherwise force field parameters. Using much stronger springs (100 times more or even infinity, i.e., fully rigid) and much weaker springs worsens the result.

Table 3.3 explores the effect of different modelings of the torsional term on the fluctuation dynamics. Here the approximation "const" means using a force-field independent, atom-type independent uniform constant values for k_{ϕ}, which is 1 Kcal/mol/rad^2. The multiplicity n is set to be 1. Similar to the geometry-related term H_{geom}, approximating all the torsional terms with a single parameter produces nearly the same results as using the force field parameters. The gain is great simplification. A torsional spring constrains the torsional motions and its presence is highly important for accurately reproducing protein dynamics. Its importance tends to be under-estimated, as many contact-based protein models do not include a torsional term. Its importance is clearly demonstrated in Table 3.3, where a significant deterioration is seen when the torsional term is weakened or even ignored.

Table 3.2 examines the effect of different modelings of the non-bonded term on the fluctuation dynamics. Again "ff" stands for force field, which uses many fine atom types (can be over 100) and thus many different van der Waals radii, one for each atom type.

Table 3.2   The effect of different modelings of the non-bonded term on fluctuation dynamics

Case	H_{geom}	H_ϕ	H_{nbond}	corr with NMA
6	const	const	ff[a]	0.860
7	**const**	**const**	**const vdW[b]**	**0.881**
8	const	const	vdW contacts[c]	0.863
9[d]	$\sim$const[e]	const	ANM[f]	0.806

[a]ff stands for force field, which uses many fine atom types and thus many different van der Waals radii, one for each atom type;
[b]const vdW means a single van der Waals radius for each major type of atoms, namely O, H, N, C, etc., using the widely-used Bondi radii;
[c]vdW contacts means the same Bondi radii are used only to define interacting pairs, or contacts, while the interaction strength is set to be a constant value of 1;
[d]Case 9 represents a standard ANM plus an explicit geometry term and a torsional term;
[e]$\sim$const means *almost* the same as const. This is because ANM itself has a weak (much weaker than const) geometry term, since it implicitly considers 1-2 and 1-3 bonded interactions but their spring constants are only 1;
[f]ANM means the non-bonded interactions, including 1-4 interactions, are specified by ANM model, which uses a cutoff distance of 4.5 Å and a spring constant of 1 Kcal/mol/Å^2.

"const vdW" means approximating the non-bonded interactions with a single set of van der Waals radii, one for each major atom type, namely O, H, N, C, etc.. Bondi radii [12] are used. "vdW contacts" uses the same set of van der Waals radii as "const vdW", but only to define interacting pairs, or contacts, while the interaction strength is set to be constant, which is 1. "ANM" in case 9 (see the last row of Table 3.2) means that the non-bonded interactions are specified by ANM model. Case 9 represents a standard ANM plus an explicit geometry term (H_{geom}) and a torsional term (H_ϕ). Though the standard ANM has an implicit geometry term since it considers 1-2 and 1-3 interactions, their springs are much weaker than those in "const" and their contributions are thus are negligible. Therefore, the difference between case 9 and the rest of the cases in Table 3.2 comes primarily from their difference in non-bonded interactions.

Table 3.3   The effect of different modelings of the torsional term on fluctuation dynamics

Case	H_{geom}	H_ϕ	H_{nbond}	corr with NMA
4	const	ff[a]	ff	0.878
5	const	$0.01\times$const[b]	ff	0.435
6	const	const	ff	0.860

[a]ff stands for force field, where the torsional interactions are modeled according to the force fields, which use many different torsional spring constant values;
[b]const means a single force-field independent, atom-type independent spring constant is used, which is 1 Kcal/mol/rad^2. The multiplicity n is set to be 1.

The first observation from Table 3.2 is that non-bonded interactions specified by the force field can be well approximated by a single set of van der Waals radii, and a single ϵ, which is set to be -0.1 Kcal/mol. Another observation is that non-bonded interactions modeled by ANM are less accurate. Lastly, the effect of van der Waals interactions is mostly captured by the interaction pairs they define. Interestingly, the interaction strength k_{ij}, as shown in Fig. 3.1(B), when approximated by a single uniform value of 1 (as in "vdW contacts") did not deteriorate the correlations much.

Models such as ANM are purely contact-based models. In ANM, the bonded terms are treated in the same way as non-bonded, both of which use a uniform spring constant of 1. To investigate what effects an explicit protein geometry term (H_{geom}) and/or a torsional term (H_ϕ) may have on such models, and to make the connection between ANM and NMA, we add these two terms to ANM and compute the changes in the fluctuation dynamics. The results are given in Table 3.4, from which it is seen that adding a simple torsional term greatly improves the ANM model. Having an explicit protein geometry term helps too, but to a smaller extent.

In summary, we conclude that the two most important ingredients in a good model of protein fluctuation dynamics are the torsional term and the non-bonded van der Waals term. This is not surprising since most of a protein's flexibility comes from the torsional degrees of freedom. The non-bonded interactions modify protein motions that otherwise

Table 3.4 The effects of a strong explicit bonded term and/or a torsional term on ANM model.

Case	H_{geom}	H_ϕ	H_{nbond}	corr with NMA
9^a	$\sim$const^b	const	ANM	0.806
10	$\sim$const	0	ANM	0.504
11	inf^c	const	ANM	0.799
12	inf	0	ANM	0.659
13	ANM^d	const	ANM	0.789
14 (ANM)	ANM^d	0^e	ANM	0.465

^{a}Case 9 represents a standard ANM plus an explicit geometry term and a torsional term;
$^b\sim$const means *almost* the same as const. This is because ANM itself has a weak (much weaker than const) geometry term, since it implicitly considers 1-2 and 1-3 bonded interactions but their spring constants are only 1;
^{c}inf stands for infinity, where all protein geometry related spring constants are set to be infinity, i.e., are set to be totally rigid;
^{d}ANM's geometry term. It includes 1-2 and 1-3 interactions whose spring constants are 1;
^{e}ANM does not have an explicit torsional term. It has 1-4 interactions but they are taken into account in the H_{nbond} term along with the rest of non-bonded interactions.

would be contributed solely by the torsional degrees of freedom. Different models of non-bonded interactions represent different ways in which the fluctuations are modified. A proper modeling of the non-bonded interactions is thus critical for properly reproducing a protein's fluctuation dynamics. The rest of the terms, which serve mostly to maintain protein geometry, are nearly rigid and are often treated as fully rigid in many models.

3.3.3 The Best Simplified Model

The results from Tables 3.1-3.4 clearly show that the best simplified model is the one that uses a single set of constant van der Waals radii for the non-bonded interactions (H_{nbond}) and constant values for the geometry term and the torsional term, as highlighted in bold in the second row of Table 3.2. We denote this model as ssNMA, or simplified spring-based NMA. Table 3.5 lists all the parameters used in ssNMA, divided into three categories, i.e., protein geometry related, torsional, and non-bonded.

Table 3.5   A summary of all the parameters used in the simplified ssNMA model[a]

H_{geom}	H_ϕ	H_{nbond}
$K_b = 340,$	$K_\phi = 1,$	$\epsilon = -0.1,\ r_H = 1.2,$
$K_\theta = 45,$	$n = 1$	$r_N = 1.85\ (1.55^b),$
$K_{\text{improper}} = 70,$		$r_O = 1.70\ (1.40^b),$
$K_{\text{UB}} = 10$		$r_C = 1.90,\ r_S = 2.0$

[a]The units are Kcal/mol/Å^2 for K_b and K_{UB}, Kcal/mol/rad^2 for K_θ and K_ϕ and K_{improper}, Kcal/mol for ϵ, and Å for all radii;
[b]The value in parentheses is for 1-4 interactions.

Connecting Elastic Network Models with NMA. The contributions of these three categories to the quality of a model are illustrated in Fig. 3.4, where ANM, an elastic network model, is used as the base model to show how the model can be enhanced to approach NMA as more terms are added or refined. Specifically, adding a proper geometry term to ANM marginally improves it. Adding both a geometry term and a torsional term significantly improves ANM. We name this model eANM, or enhanced ANM. A further improvement over eANM is achieved by replacing its ANM-based non-bonded term with a more accurate van der Waals based non-bonded term, which is the ssNMA model. Having a correlation value with NMA that is nearly 0.9, ssNMA is our best simplified model and uses only a few parameters. Like other elastic network models, ssNMA is force-field independent and does not require energy minimization. ssNMA has a similar performance to sbNMA, which is the same as ssNMA except that it is force-field dependent and uses extensive force field parameters. Lastly, on the last column of the figure, when adding the force-based terms back to sbNMA, we have the original NMA. Because of the force-based terms, energy minimization becomes necessary in order to bring an input protein system to equilibrium. Based on the requirement for a force field and/or energy minimization, the models listed in Fig. 3.4 are divided into three classes. Class I is the elastic network models, which require neither a force field nor energy minimization. Class II includes models such as sbNMA, which do not require

energy minimization but a force-field. NMA belongs to class III, which requires both a force field and energy minimization.
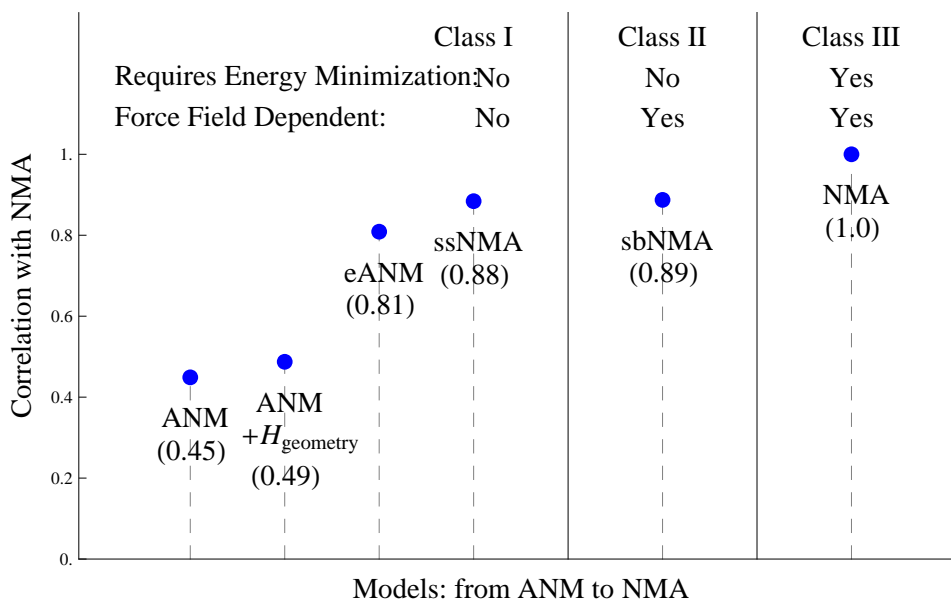
Figure 3.4 **From ANM to NMA:** the roles of three major terms (namely protein geometry, torsional, and non-bonded) to protein fluctuations and the extent of their contributions in improving ANM-like models to become NMA-like models. All evaluations are done by comparing with NMA, specifically correlations in mean-square fluctuations.

3.4   Conclusions

Normal mode analysis (NMA) is one of the few powerful tools for studying protein dynamics. However, computing normal modes has been greatly hindered by the cumbersome energy minimization process needed to bring a structure to equilibrium before NMA can be applied. Moreover, the minimized structure is often a few angstroms away from the input structure.

Elastic network models, due to their simplicity, have made normal mode computations much more accessible, to a much broader community, and for many more bio-molecular systems, even for large systems such as ribosome, [146] nuclear pore complex, [75] etc. Compared to NMA, the weakness of elastic network models is that they are less accurate.

The objective of this work is to bridge NMA and elastic network models and to make the strong connection between the usual atomic models that use full force fields and the elastic network models that use simplified potentials. The connection between NMA and elastic network models is made in the following way. Starting with NMA, we first identify what is essential to its accuracy, and then take steps of simplification, with the goal of simplification being to reach elastic network models. There are a few key realizations that have helped in this process of simplification. The first realization is that NMA Hessian matrix, as a second derivative of the potentials, consists of two types of contributions. One is related to force field spring constants while the other the inter-atomic forces or torques. The second key realization is that the contribution from the force-based terms is small and the full NMA can be well approximated by the spring-based NMA, or sbNMA. The third and last key realization is that the extensive force field parameters used in NMA and sbNMA, numbering in hundreds or even thousands, can be well approximated by a very small set of force-field independent constants.

The simplification process that starts with NMA and reaches ANM presents itself also a new way to derive high-quality elastic network models. Indeed, in drawing the connection between NMA and elastic network models, we have discovered several new elastic network models that, i) closely resemble the accuracy of full-scale NMA, and yet, ii) are simple, easy to use, without the complexity of energy minimization that NMA requires. Particularly, we have identified ssNMA as one of the best simplified models. ssNMA has the simplicity of elastic network models while maintaining a high correlation with NMA.

Since ssNMA requires only an input structure and a few force field independent parameters, it can be applied directly to experimental structures without the need for energy minimization. This is highly significant. Since formerly with NMA, even if we want to believe that an input structure, say a crystal structure, represents an equilibrated native state, chances are that we would not be able to find a force field that agrees with

us that the structure is energetically minimized. Thus, we would not be able to apply NMA without first "minimizing" and thus deforming the input structure that we have believed to be native. With ssNMA, it is different. ssNMA can be directly applied to the structure. The minimization step is skippable in ssNMA - this is highly beneficial, especially when the minimization step is unnecessary and can do more harm than good to an input structure.

In the process of connecting NMA and elastic network models, we also gain a clearer understanding why elastic network models work well and how it can be further improved. We discovered that ANM can be greatly enhanced by including an additional torsional term and a geometry term. The new model, eANM, or enhanced ANM, has a much higher correlation with NMA than ANM does.

Compared with other modified elastic network models existing in the literature [25, 77, 153, 154, 159, 163] that modify the original elastic network models in various ways and show improvement to various extents, our approach is unique in that it is not based on a heuristic model whose validity is justified *a posteriori*, e.g., by showing that a model is able to produce experimental B-factors well. Our derivation of models is "first-principle" based. It starts with NMA, and arrives at elastic network models through steps of reasonable simplification. The advantage of doing in this way is that it maintains a tight link to the original NMA model and gives us a deeper understanding of what makes a good model. The disadvantage is, as is true with most first-principle based derivations, there is a limitation on what simplifications can be made, and consequently, what models can be reached by these steps of simplifications.

It is worth noting that the whole process of bridging NMA and elastic network models and all the discussions so far apply to atomistic models only. In other words, this work shows how to bridge between the classical NMA, which is atomistic by nature, and *atomistic* elastic network models. This work has discovered some high-quality *atomistic* elastic network models such as ssNMA and eANM. These models are highly valuable

especially when computing normal modes for proteins where energy minimization is undesired and an atomistic model is needed. In such situations, ssNMA or eANM is an ideal alternative to NMA, for which energy minimization is required.

ssNMA or eANM is intrinsically fine-grained models and does not have direct coarse-grained counterparts as ANM does. To apply them to coarse-grained systems, one can project their fine-grained Hessian matrices into the C_α space to get a coarse-grained ssNMA or eANM, by doing (as in Eqs. (1) and (2) of [163]):

$$H_{\text{all}} = \begin{vmatrix} H_{C_\alpha} & H_{\text{int}} \\ H_{\text{int}}^\top & H_{\text{other}} \end{vmatrix}; \tag{3.14}$$

$$H_{C_\alpha}^* = H_{C_\alpha} - H_{\text{int}} \times H_{\text{other}}^{-1} \times H_{\text{int}}^\top; \tag{3.15}$$

where H_{C_α}, H_{other}, and H_{int} are submatrices of the C_α atoms, the other (non-C_α) atoms, and their interactions, respectively. $H_{C_\alpha}^*$ represents the reduced coarse-grained Hessian matrix. By projecting ssNMA or eANM to the coarse-grained level, the computational cost is reduced. However, the inversion of matrix H_{other} can still be costly. Such a projection can be very useful when a higher-quality coarse-grained model (and higher quality normal modes) is desired and computational time is not an issue.

We have applied ssNMA and eANM directly, without energy minimization, to a number of sizable proteins and our results show that eANM and ssNMA give better correlations with experimental B-factors than the coarse-grained ANM, see Table S2 in Supplemental Materials. However, as we pointed out earlier, cautions must to be taken to interpret the results, since experimental B-factors are strongly influenced by crystal packing, lattice order, etc., which are not considered in these models.

Because ssNMA or eANM is intrinsically fine-grained and have to be projected (through matrix inversion of a sub-matrix) in order to be used at the coarse-grained level, their applicability to protein systems is more limited than those models that are naturally coarse-grained, such as ANM. Indeed, as shown in Table S3 in Supplemental Materials, eANM takes longer time to compute than ANM, and for even larger proteins,

eANM may become even inapplicable. ssNMA is in a similar situation. Therefore, though the present work provides a nice bridge between NMA and fine-grained elastic network models, future work is still needed to extend the bridge to coarse-grained elastic network models, particularly regarding to developing higher-quality coarse-grained models that are also computationally as efficient as or comparable to existing coarse-grained models. Perhaps what has been learned here from how fine-grained ANM can be enhanced (i.e., the eANM model) may inspire the development of such kind of enhanced coarse-grained elastic network models.

In summary, in this work through steps of simplification we have built a bridge between NMA and elastic network models. In the process of bridging the two, we have also discovered several high-quality simplified models. Being all-atom and using simplified potentials, these models help make the tight connection between the usual atomic models and the elastic network models. They also have the advantage of being able to incorporate a higher level of cooperativity through including more springs up to a longer distance and through multi-body interactions.

Acknowledgment

CHAPTER 4.   BRIDGING BETWEEN NMA AND ELASTIC NETWORK MODELS: PRESERVING ALL-ATOM ACCURACY IN COARSE-GRAINED MODELS

Hyuntae Na[24], Robert L. Jernigan[34], and Guang Song[245]

Abstract

Dynamics can provide deep insights into the functional mechanisms of proteins and protein complexes. For large protein complexes such as GroEL/GroES with more than 8,000 residues, obtaining a fine-grained all-atom description of its normal mode motions can be computationally prohibitive and is often unnecessary. For this reason, coarse-grained models have been successfully used. However, most existing coarse-grained models use extremely simple potentials to represent the interactions within the coarse-grained structure and as a result, the dynamics obtained for the coarse-grained structure may not always be fully realistic. There is a gap between the quality of the dynamics of the coarse-grained structure given by all-atom models and that by coarse-grained models. In this work, we resolve an important question in protein dynamics computations – how

---

[2]Graduate student and Associate Professor, respectively, Department of Computer Science, Iowa State University.

[3]Professor, Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University.

[4]Primary researchers and authors.

[5]Author for correspondence.

can we efficiently construct coarse-grained models whose description of the dynamics of the coarse-grained structure remains as accurate as that given by all-atom models? Our method takes advantage of the sparseness of the Hessian matrix and achieves a high efficiency with a novel iterative matrix projection approach. The result is highly significant since it can provide descriptions of normal mode motions at an all-atom level of accuracy even for the largest biomolecular complexes. The application of our method to GroEL/GroES offers new insights into mechanism of this biologically important chaperonin, such as the conformational transitions of this protein in its functional cycle are even more strongly connected to only first few lowest frequency modes than with other coarse-grained models.

Author Summary

Proteins and other biomolecules are not static but are constantly in motion. Moreover, they possess intrinsic collective motion patterns that are tightly linked to their functions, and resemble mechanical systems. Thus, an accurate and detailed description of their motions can provide deep insights into their functional mechanisms. For large protein complexes with hundreds of thousands of atoms or more, an atomic level description of the motions can be computationally prohibitive, and so coarse-grained models with fewer structural details are often used instead. However, there can be a big gap between the quality of motions derived from atomic models and those from existing coarse-grained models. In this work, we solve an important problem in protein dynamics studies: how to preserve the atomic-level accuracy in describing molecular motions while using coarse-grained models? We accomplish this by developing a novel iterative matrix projection method that dramatically speeds up the computations. This method is significant since it promises more accurate descriptions of protein motions approaching an all-atom level even for the largest biomolecular complexes. Results shown here for a large molecular chaperonin demonstrate how this can provide new insights into functional process.

4.1 Introduction

Protein dynamics plays a key role in describing the function of most proteins and protein complexes. The importance of protein dynamics studies has been increasingly recognized alongside with the importance of the structures themselves. Experimentally, protein dynamics can be studied using nuclear magnetic resonance (NMR) [42, 88], time-resolved crystallography [134], fluorescence resonance energy transfer (FRET) [6] and other single-molecule techniques [61], etc. Computationally, the study of protein dynamics most commonly takes relies upon molecular dynamics (MD) simulations [56, 74, 85]. Normal mode analysis (NMA) is another popular and powerful tool for studying protein motions and dynamics that was first applied to proteins in the early 80's [17, 38, 72]. The advantage of using normal modes over MD is that these can most efficiently describe protein motions near the native state. To apply NMA, a structure is first energetically minimized. The minimized structure is then used to construct the Hessian matrix, from which normal modes can be obtained from its eigenvectors and eigen-frequencies. This method poses a huge demand on computational resources, especially memory, since applications to large supramolecules may have hundreds of thousands of atoms. The time spent on computing the eigenvalues/eigenvectors also is large, of the order of the cube of the number of atoms. Consequently, its applications are limited to smaller systems.

For this reason, many simplified models [2, 8, 43, 47, 58, 59, 66, 75–77, 79, 80, 84, 89, 126, 130, 139, 146, 149, 150, 155, 160] have been developed for efficient normal mode computations. These models use simplified structural models or simplified force fields or commonly, both. One commonly applied type of coarse-grained models are the elastic network models [2, 8], which usually treats each residue as one node, and residue-residue interactions as Hookean springs. It has been demonstrated for a large number of cases that these extremely simple models can still capture quite well the slow dynamics of a protein [139]. And because of their high level of simplicity, they have been successfully

applied to study the normal mode motions of the largest structural complexes such as GroEL/GroES [55, 58, 82, 132, 133, 156], ribosome [64, 67, 68, 146], nuclear pore complex [75], etc.

However, along with the significant gains from this simplicity comes also some loss of accuracy, particularly in the accuracy of the normal modes [91, 142]. The validity of most simplified models was justified *a posteriori*, by comparing with experimental B-factors or sets of multiple experimental structures for example. How well they preserve the accuracy of the original NMA has rarely been assessed directly [89]. To overcome this problem of accuracy, we are building a strong connection between NMA and elastic network models (ENMs) through a series of steps of simplification that begin with NMA and end with ENMs, and propose a new way to derive accurate elastic network models in a top-down manner (by gradually simplifying NMA) [89]. Our derivation on the realization that the Hessian matrix of the original NMA can be written as a summation of two main terms, the spring-based terms and the force/torque-based terms, with the former contributing significantly more than the latter. By ignoring the latter term, we obtained at a new model, sbNMA (or spring-based NMA), that has high accuracy and closely resembles the original NMA and requires no energy minimization. sbNMA, like the original NMA, is force-field dependent and uses many parameters. By further simplifying it, we arrived at two force-field independent elastic network models, ssNMA (simplified spring-based NMA) and eANM (enhanced ANM), both of which use many fewer parameters and yet still preserve most of its accuracy [89]. For example, the mean square fluctuations predicted by ssNMA for a set of small to medium proteins have an average correlation of nearly 0.9 with those predicted with the original NMA [89]. It was shown [91] also that ssNMA modes are more accurate than those from other elastic network models. However, this bridging, as detailed in Ref. [89], connected NMA only with all-atom elastic network models but not with coarse-grained ones. Both ssNMA and eANM, though strongly resembling NMA, are by nature all-atom models and cannot be directly applied to coarse-grained structures.

There is little doubt that for very large biomolecular systems, coarse-grained structure representations are needed, since all-atom normal mode analyses for such systems are computationally often out of reach. Now the aim here is to extend the idea of bridging between NMA and elastic network models to coarse-grained models while preserving sufficient accuracy to obtain accurate protein dynamics even for very large systems. Is it possible to efficiently construct coarse-grained models whose description of the dynamics of a coarse-grained structure remains as accurate as that given by all-atom models? Coarse-grained models, such as C_α-based models, obviously do not have all the structural details of all-atom models. But, is it possible that the dynamics of the C_α atoms can be given by the coarse-grained models is as accurately as with all-atom models? Is it possible to have *both* the simplicity of coarse-grained structure and the accuracy of all-atom interactions? These questions are the focus of this work. And we demonstrate affirmative answers to these question by employing a novel iterative matrix projection technique. While our earlier work [89] bridged between NMA and all-atom elastic network modes and represents a force-field simplification of NMA while maintaining most of its accuracy, the present work presents the additional structural simplification from all-atom elastic network models to coarse-grained elastic network models. Once this full bridging is completed, it should reveal deep insights for how to develop coarse-grained elastic network models that preserve most of the accuracy of NMA.

4.2   Methods

A coarse-grained model has two key components: i) a coarse-grained structure representation, and ii) an interaction model for the coarse-grained structure. The challenge that one normally faces in developing coarse-grained models is that there is no prescription for how to represent the interactions among the coarse-grained structure

precisely [51]. Most semi-empirical force field potentials are for atomic models. There are now a few coarse-grianed potentials for use in dynamics. Highly simplified Hookean springs are commonly used to model residue-residue interactions. These clearly provide only a very rough approximation to the atomic models. Other studies that have also been linked atomic and coarse-grained models have applied force-matching [51] or required frequency spectra to have similar distributions [65]. A statistical mechanical foundation was developed by the same research group [95] to show that many-body potentials of mean force that govern the motions of the coarse-grained sites can be generated. Regarding coarse-grain structure representation, C_α atoms are normally used to represent residues, although other coarse-grained representations have also been investigated [158].

In this work, to extend the accurate all-atom models to coarse-grained models without losing accuracy in the dynamics, we take two steps. First, we show that it is possible to define a precise interaction model for the coarse-grained structure so that its dynamics are the same as that of its all-atom counterpart. Second, we show that the construction of such a precise interaction model can be performed efficiently and straightforwardly.

4.2.1 How to Construct a Precise Interaction Model for a Coarse-Grained Structure?

It is useful first to perform an operation that separates out the atoms used for the coarse-graining from the remainder of the atoms. Mathematically, it is possible to define a *precise* interaction model (in the form of a Hessian matrix) for the coarse-grained structure by rearranging the original Hessian matrix $\mathbf{H}_{all}$ into parts for the coarse-grained atoms and the remainder of the atoms in separate subspaces, as was done by Eom et. al. [28] and Zhou and Siegelbaum [163]:

$$\mathbf{H}_{all} = \begin{pmatrix} \mathbf{H}_{cc} & \mathbf{H}_{cr} \\ \mathbf{H}_{cr}^\top & \mathbf{H}_{rr} \end{pmatrix}, \tag{4.1}$$

$$\tilde{\mathbf{H}}_{cc} = \mathbf{H}_{cc} - \mathbf{H}_{cr}\mathbf{H}_{rr}^{-1}\mathbf{H}_{rc}^{\top}, \tag{4.2}$$

where c stands for the atoms used for the coarse-graining, r stands for the remainder of the structure, and $\top$ represents the matrix transpose. It can be shown mathematically [11, 147] that $\tilde{\mathbf{H}}_{cc}$ maintains the same description of the mean-square fluctuations and cross-correlations of the coarse-grained structure as the original Hessian matrix. All elements in $\tilde{\mathbf{H}}_{cc}^{-1}$ are the same as their corresponding elements in $\mathbf{H}_{all}^{-1}$. A similar idea of using matrix projection to obtain the motions for subsystems was previously used also by Brooks and Zheng and their co-workers [40, 148] to develop their VSA (vibration subsystem analysis) model.

However, this mathematical rearrangement in Eq. (4.2) requires the inversion of $\mathbf{H}_{rr}$, which appears to be nearly as difficult as computing the inverse of the original all-atom Hessian matrix, assuming the number of atoms in the coarse-grained structure is much smaller than that of the original all-atom model. Therefore, unless $\tilde{\mathbf{H}}_{cc}$ can be computed in an efficient way, the precise interaction model defined in Eq. (4.2) will be computationally too expensive to apply for very large systems and thus of little practical utility.

In the next section, we present a novel way for computing $\tilde{\mathbf{H}}_{cc}$ efficiently, without directly inverting $\mathbf{H}_{all}$ or $\mathbf{H}_{rr}$. As a result, this permits the construction of coarse-grained models that can represent the dynamics of the coarse-grained structure as accurately as for all-atom models.

4.2.2 Efficiently Construct the Coarse-Grained Hessian Matrix through Iterative Projection

To efficiently obtain the Hessian matrix $\tilde{\mathbf{H}}_{cc}$ from Eq. (4.2) but without having to directly invert $\mathbf{H}_{rr}$, we take advantage of the fact that for the Hessian matrix $\mathbf{H}_{all}$ of the the second derivatives of the potential, can be highly sparse for some all-atom models.

$\mathbf{H}_{all}$ is not so sparse for the conventional NMA, due to the persistence of electrostatic interactions to long distances. However, it is sparse for ssNMA, an accurate all-atom model that closely resembles NMA as mentioned above.

The potential for ssNMA includes most of the same interaction terms as for NMA, except for the electrostatic interactions [89]. As a simplified model of spring-based NMA (or sbNMA), ssNMA uses one single uniform spring constant for all bond stretching terms, one uniform spring constant for all the bond-bending terms, and one for the torsional terms. Its non-bonded van der Waals interactions are truncated near the equilibrium distance to avoid negative spring constants in the Hessian matrix [89]. A single set of van der Waals radii are used for all van der Waals interactions. All the equilibrium values such as bond lengths, bond angles, and torsional angles are taken from the reference structure. Consequently, most of the off-diagonal elements in the ssNMA Hessian matrix are zero.

In the following, we use ssNMA to construct the all-atom Hessian matrix $\mathbf{H}_{all}$ and show how a precise interaction model $\tilde{\mathbf{H}}_{cc}$ can be efficiently constructed through an iterative matrix projection procedure. We call this model coarse-grained ssNMA, or CG-ssNMA, which will preserve the same accuracy as the all-atom ssNMA in its description of the dynamics of the coarse-grained structure.

The procedure, as detailed before, takes full advantage of the sparseness of the Hessian matrix. Given a protein that has n atoms, one can iteratively reduce its size (or coarse-grain it) by removing one atom, or a group of r atoms, at a time without losing accuracy in depicting the motions of the remaining atoms. This can be done by adding a correction term to the interactions among the remaining atoms. Define by $\mathbf{H}$ the Hessian matrix with n atoms as follows:

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{kk} & \mathbf{H}_{kr} \\ \mathbf{H}_{kr}^{\top} & \mathbf{H}_{rr} \end{pmatrix}, \tag{4.3}$$

where $\mathbf{H}_{kk}$ is the block matrix of $\mathbf{H}$ for the remaining $n - r$ atoms, $\mathbf{H}_{rr}$ the block

matrix for r atoms to be removed, and $\mathbf{H}_{kr}$ represents the interactions between the group of atoms to be removed and the remaining atoms. The effective Hessian matrix $\tilde{\mathbf{H}}_{kk}$ of the remaining atoms after taking into account the correction term can be written as [28, 91, 163]:

$$\tilde{\mathbf{H}}_{kk} = \mathbf{H}_{kk} - \mathbf{H}_{kr}\mathbf{H}_{rr}^{-1}\mathbf{H}_{kr}^{\top}, \tag{4.4}$$

with the term $\mathbf{H}_{kr}\mathbf{H}_{rr}^{-1}\mathbf{H}_{kr}^{\top}$ being the correction term.

It can be shown the motions of the remaining atoms as described by $\tilde{\mathbf{H}}_{kk}$ is the same as those by the original Hessian matrix $\mathbf{H}$. This numerical preservation is crucial when an all-atom Hessian matrix is gradually coarse-grained by repeatedly removing non-C_α atoms, since it guarantees that the quality of the description of the C_α atoms remains the same while the size of the Hessian matrix is being reduced.

Note that each atom interacts only with a few, say m on average, atoms due the sparseness of the Hessian matrix. As a result, $\mathbf{H}_{kr}$ has only a small number (rm) of non-zero elements, representing the interactions between the group of atoms to be removed and the rest of the atoms. Therefore, the term $\mathbf{H}_{kr}\mathbf{H}_{rr}^{-1}\mathbf{H}_{kr}^{\top}$ in (4.4) can be computed in $O(r^3 + r^2m^2)$ time. Coarse-graining the whole protein structure takes roughly n/r iterations and thus requires in total $O((r^2 + rm^2)n)$ time, which is *linear* in the protein size n.

To further reduce the running time, matrix elements that are near zero (weak interactions) are set to zero if their absolute values are less than a predetermined threshold value ξ. A properly chosen ξ can further improve computation speed while preserving the accuracy, by effectively reducing the number of interactions, especially those between the atoms being removed and retained C_α atoms. Different ξ values were tested, as detailed in the next section.

Fig. 4.1 illustrates how the sparseness of the Hessian matrix is maintained throughout the iterative matrix projection procedure. At the initial step, atoms are shuffled so that C_α atoms are grouped together and placed on the left-most side of the Hessian matrix, as

shown in Fig. 4.1(A), where the grouped C_α and non-C_α atoms are separately represented by dark and light gray blocks, respectively. Blue dots represent the non-zero elements of the Hessian matrix. The non-C_α atoms can then be rearranged, for example, using the Cuthill-McKee algorithm [24], so that the atoms that interact with one another are placed close together in the matrix. As a result, the non-zero elements are relocated near the diagonal of the matrix (see Fig. 4.1(B)). In such a sparse matrix, Fig. 4.1(C) shows the effect of applying one matrix projection using Eq.(4.4), where the red dots represent the elements of the matrix whose values are modified. Note that the sparseness of the non-C_α region is mostly unaffected by the projection. The sparseness of the white region (interactions with C_α atoms) can be maintained by using an appropriate threshold value ξ mentioned earlier.

Algorithm 2 below lists the steps that iteratively reduce the all-atom Hessian matrix to a coarse-grained one. The algorithm takes as input the all-atom Hessian matrix $\mathbf{H}$, a set of C_α atom indices $\{k_1, ..., k_n\}$, and a threshold value ξ. All matrix elements whose absolute values are less than ξ are set to 0. In practice, it turns out that lines 4-11 run more efficiently if each iteration of the coarse-graining process removes not single atoms but a group of atoms (R_i as in line 2). Removing a group of adjacent atoms reduces the average number of interactions (m in the above Big-O notation) with the remaining atoms. These groups of atoms are determined by spatially partitioning the whole structure (3-D) into cubic blocks (18 Å for each dimension, about 500 atoms in each block). These blocks represent initial groups of atoms. The reason why atoms are partitioned in this way is to minimize the number of interactions among the different groups. Blocks are then sorted by their sizes (i.e., the number of atoms) in descending order. Next, starting with the smallest one, blocks on the "small" end (usually blocks on the outsides of a structure) are iteratively merged together with the next smallest block as long as the size of the merged group does not exceed the size limit (which is about taken to be around 500 atoms per group, the number of atoms in a regular cubic

(A)

Ca atoms

non-Ca atoms

(B)

Ca atoms

non-Ca atoms
reshuffled by
reverse CM
algorithm

H_{kk} H_{kr}

H_{rk} H_{rr}

(C)

updated by
$H_{kr}H_{rr}^{-1}H_{rk}$
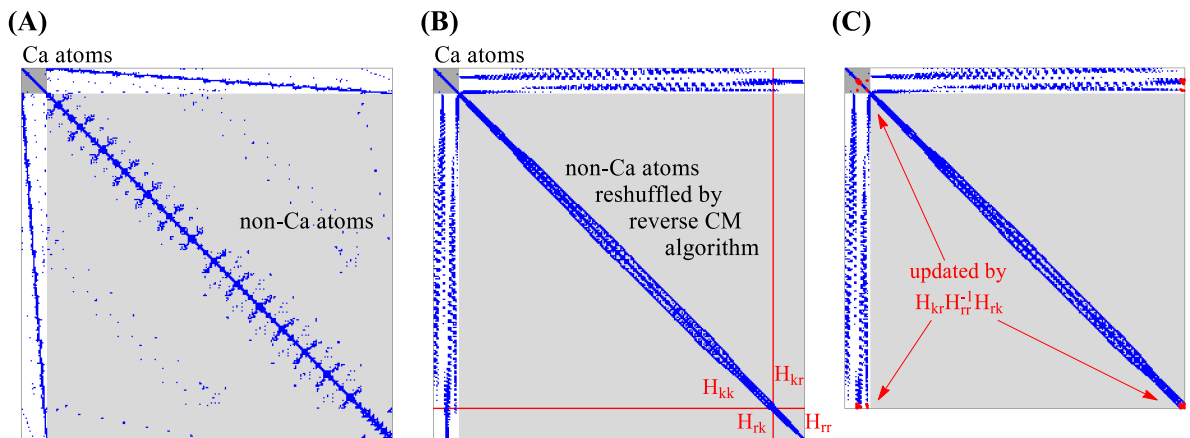
Figure 4.1   **Illustration of how the sparseness of the Hessian matrix can be maintained throughout the iterative matrix projection procedure, when coarse-graining is performed by selecting the C_α atoms for retention.**   (A) In a first step the original Hessian matrix is shuffled so that C_α atoms (in dark gray at the top-left corner) are separated from the non-C_α atoms (in light gray). Blue dots represent non-zero elements. (B) In a second step the non-C_α atoms are rearranged again so that those interacting with one another are placed close together in the matrix using, for example, the Cuthill-McKee algorithm [24]. As a result, most non-zero elements are placed near the diagonal. (C) Matrix after performing one projection to remove atoms in group r. The red dots represent the blocks modified by the projection. The sparseness of the non-C_α region is mostly unaffected. The sparseness of the white region (interactions with C_α atoms) can be maintained by using an appropriate threshold value ξ, see text.

block). The merging process stops when there are no small blocks left to be merged. In lines 7 and 9, sparse($\mathbf{A}, b$) returns a sparse matrix of $\mathbf{A}$ by setting to zero $\mathbf{A}$'s elements that satisfy $|\mathbf{A}_{i,j}| < b$, where $|\mathbf{A}_{i,j}|$ is the absolute value of $\mathbf{A}_{i,j}$. Threshold ξ/m is used in line 9 since the addition (or subtraction) in line 10 is accumulated m times. Line 9 prevents very small values from being added to $\mathbf{H}$ in line 10 and then removed in line 7 at the next iteration.

Algorithm 2 CoarseGrain($\mathbf{H}, \{k_1, ..., k_n\}, \xi$)

1: $K \leftarrow \{k_1, ..., k_n\}$
2: $R \leftarrow \{R_1, R_2, ..., R_m\}$
3: $\mathbf{H} \leftarrow$ Hessian matrix of $\mathbf{H}$ reshaped in the order of $K, R_1, R_2, ..., R_m$
4: **for** $i = m, m - 1, ..., 1$ **do**
5:      $k \leftarrow |K| + \sum_{j=1}^{i-1} |R_j|$
6:      $r \leftarrow k + |R_i|$
7:      $\mathbf{B} \leftarrow \text{sparse}(\mathbf{H}_{1..k, k+1..r}, \xi)$
8:      $\mathbf{D} \leftarrow \mathbf{H}_{k+1..r, k+1..r}$
9:      $\mathbf{E} \leftarrow \text{sparse}(\mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\top, \xi/m)$
10:     $\mathbf{H}_{1..k, 1..k} \leftarrow \mathbf{H}_{1..k, 1..k} - E$
11: **end for**
12: $\mathbf{H} \leftarrow \text{sparse}(\mathbf{H}_{1..|K|, 1..|K|}, \xi)$
13: **return H**

4.3   Results

4.3.1   Validation of Model Accuracy and Efficiency

In this section, we first verify computationally that the coarse-grained ssNMA model constructed according to the proposed procedure indeed not only preserves the accuracy of all-atom models in its description of the motions of the coarse-grained structure but also is computationally efficient. To this end, we first show, by applying it to a dataset of 177 small to medium proteins, that with a properly chosen threshold value ξ, the coarse-grained ssNMA preserves full accuracy. We then extend the same coarse-graining procedure, using the same ξ value, to construct coarse-grained ssNMA Hessian matrices for 80 large superamolecules of different sizes and show that the construction of these ssNMA Hessian matrices requires only a nearly linear time and can thus be carried out quickly, even for large systems.

4.3.2   The Iterative Coarse-Graining Procedure Preserves Accuracy

To validate the accuracy of the method, Algorithm 2 is applied to 177 small-to-medium proteins whose sizes are greater or equal to 60 residues but less than 150. This is the same set of proteins that was used in our earlier work [89]. Only small to medium

sized proteins are used at this stage due to the high computational costs of running all-atom models, which have also been computed here for comparison purposes.

Each protein structure is first energetically minimized. From the all-atom ssNMA Hessian matrix, two coarse-grained Hessian matrices, $\mathbf{H}$ and $\hat{\mathbf{H}}$, are computed. $\mathbf{H}$ is computed by direct matrix projection (as in Eq. (4.2)), which is an exact but very expensive computation, while $\hat{\mathbf{H}}$ is computed with the proposed iterative projections as in Algorithm 2. To show that $\hat{\mathbf{H}}$ preserves the same full accuracy as $\mathbf{H}$, we compute the correlations between mean square fluctuations (MSF) computed with $\mathbf{H}$ and those from $\hat{\mathbf{H}}$, and the eigenvalue-weighted overlaps between modes by $\mathbf{H}$ and those by $\hat{\mathbf{H}}$. The eigenvalue-weighted mode overlap is defined as:

$$\sum_{i=7}^{3n} \frac{w_i}{w} |\mathbf{m}_i \cdot \hat{\mathbf{m}}_i|, \tag{4.5}$$

where n is the number of atoms, $\mathbf{m}_i$ (and $\hat{\mathbf{m}}_i$) is the ith mode of $\mathbf{H}$ (and $\hat{\mathbf{H}}$), $w_i = 1/\lambda_i$ is the relative weight and is set to be the inverse of the ith eigenvalue of $\mathbf{H}$, and $w = \sum_{i=7}^{3n} w_i$ is the normalization factor. The reason why we use the modes with the same indices ($\mathbf{m}_i$ and $\hat{\mathbf{m}}_i$) instead of the best matching modes when computing the weighted-overlap is to measure also how well the order of the modes is preserved. (Lower frequency modes are given higher weights in this weighted overlap measure. The intuition behind this weighted mode scheme is that it represents how similar the modes (including their orders) are between the two models.

Table 4.1 shows the levels of accuracy that can be achieved when different threshold values ξ are applied to ssNMA [89]. It is seen that ssNMA preserves the full accuracy (1.0 in correlations and overlaps) in mean square fluctuations and modes when a threshold value (ξ) as large as 0.01 is used. Similar results are also seen for the enhanced ANM model (eANM) [89], another all-atom model that closely resembles NMA. Using a large threshold value allows the sparseness of the Hessian matrix to be maintained during the iterative matrix projection process and consequently the construction of the coarse-grained ssNMA Hessian matrix to be carried out quickly.

Table 4.1   **The accuracy of models at different threshold values ξ.**

ξ^a	NMA (0.0^b)		ssNMA (0.99^b)		eANM (0.98^b)	
	$corr^c$	w-ovlp^d	corr	w-ovlp	corr	w-ovlp
0.0001	0.99	0.96	1.00	1.00	1.00	1.00
0.001	0.85	0.62	1.00	1.00	1.00	1.00
0.01	0.82	0.69	1.00	1.00	1.00	0.99
0.1	0.56	0.53	0.99	0.92	0.98	0.83

The accuracy of ssNMA, in both mean-square fluctuations and mode details, is fully preserved at $\xi = 0.01$. The initial sparseness of the Hessian matrix, in parentheses, is 0.0, 0.99, 0.98 for NMA, ssNMA, and eANM, respectively.
$^a\xi$: the threshold value used to set to zero the smallest elements in the Hessian matrix;
^{b}initial sparseness of the Hessian matrix;
^{c}corr: mean-square fluctuation correlation;
^{d}w-ovlp: eigenvalue-weighted mode overlap as defined in Eq. (4.5).

For conventional NMA, however, the iterative coarse-graining approach as described above does not work nearly as well (see Table 4.1). This is due to the slowly-decreasing, long-range electrostatic interactions.

4.3.3   The Iterative Coarse-Graining Procedure Is Efficient

Secondly, we look at the efficiency, i.e., how much time does this iterative coarse-graining procedure require? To this end, we apply the same iterative coarse-graining procedure to construct coarse-grained ssNMA Hessian matrices for a number of large proteins and protein complexes. The same threshold value, $\xi = 0.01$, is used, which has been shown in the previous section to preserve the full accuracy.

Fig. 4.2 shows the efficiency (computational time) of the proposed method as a function of the system size. In the figure, each blue and red point represents respectively, for a protein of that size, the coarse-graining time, i.e., the time required to construct the coarse-grained ssNMA Hessian matrix (with $\xi = 0.01$), and the diagonalization time of that coarse-grained Hessian matrix. The dashed lines show the growth rates of the time

cost as a function of the system size. The curves are obtained from the least squares fitting to a non-linear function $f(x) = ax^b$. As shown in the figure, the diagonalization time (red curve) grows approximately as the cube, while the coarse-graining time grows approximately linearly. Especially for large complexes, the time needed for coarse-graining the all-atom Hessian matrix using Algorithm 2 becomes increasingly smaller relative to the diagonalization time. As a result, the total time for computing the normal modes for such large protein complexes using the coarse-grained ssNMA Hessian matrices is about the same as for other coarse-grained elastic network models such as ANM.
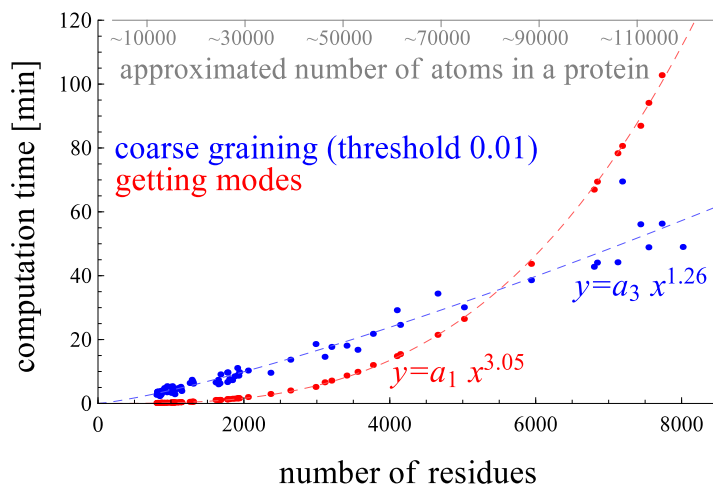
Figure 4.2   **Comparison of the proposed coarse-graining time and the diagonalization time of the coarse-grained Hessian matrix.**

In summary, the results in this section demonstrate that the proposed iterative coarse-graining procedure not only preserves the accuracy in depicting the motions of the coarse-grained structure but also is computationally efficient in the time it takes to construct coarse-grained Hessian matrices, being negligible comparing to the time needed for computing normal modes for large protein complexes.

This result is significant since it means that we can construct coarse-grained models that preserve all-atom accuracy even for very large protein complexes, which was not previously possible. Next, as an application, we apply the proposed procedure to compute and analyze the dynamics of the GroEL/GroES complex.

4.3.4 Application to GroEL/GroES complex

The GroEL/GroES complex [151] is a molecular chaperone that assists the unfolding of partially folded or misfolded proteins, by providing them with the chance to refold. GroEL consists of *cis* and *trans* rings, each of which has 7 subunits. Each subunit is about 547 residues. GroES also has 7 chains and each chain contains about 97 residues. The GroEL *cis*-ring and GroES form a capped chamber that can hold proteins and facilitate protein unfolding partly through their intrinsic collective motions, such as compressing, stretching, twisting, shearing, and relaxing. Fig. 4.3 shows the GroEL/GroES structure (pdbid: 1AON) in top and front views. In Fig. 4.3(A), the three domains of *cis* and *trans* rings are distinguished with different colors: equatorial (green), intermediate (yellow), and apical (blue) domains.
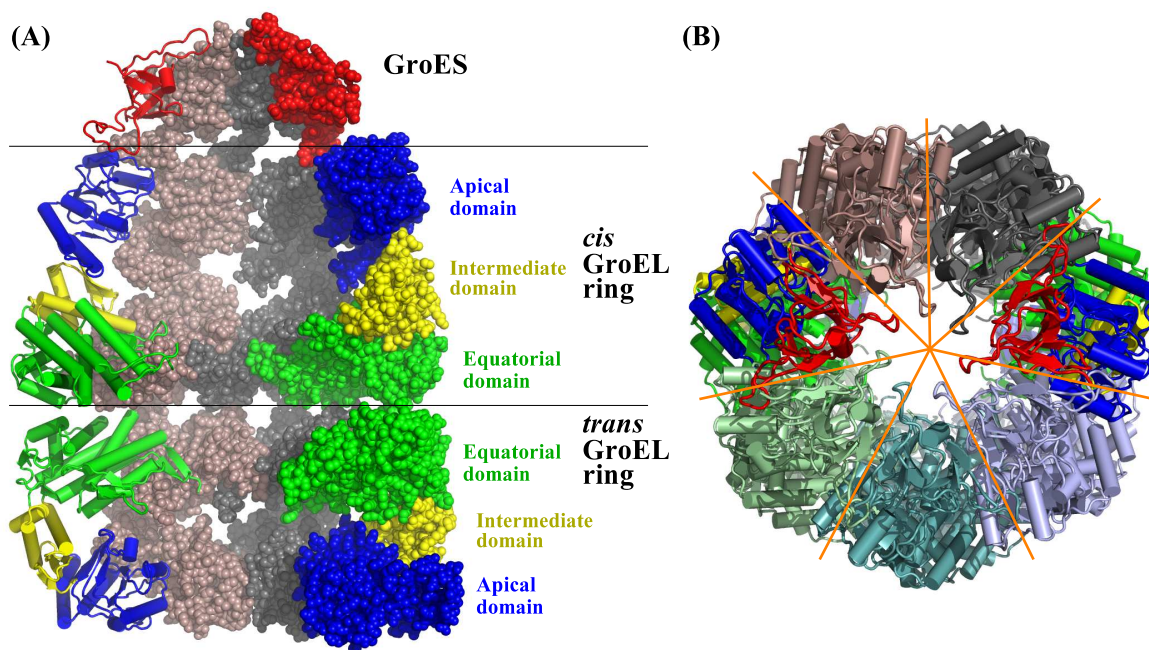
Figure 4.3 **Structure of the GroEL/GroES complex in (A) front and (B) top views.** For subunits of the GroEL, the equatorial, intermediate, and apical domains of *cis* and *trans* rings are colored green, yellow, and blue, respectively. The GroES cap is displayed in red.

To understand its functional mechanisms, it is informative to obtain the intrinsic motions of this complex. However, for large protein complexes such as GroEL/GroES that has over 8,000 residues, standard all-atom NMA will take a prohibitively large memory and a long time to run. Consequently, past normal mode studies on this complex were limited to coarse-grained models [58, 132], or all-atom models of single subunits [82]. Though a more accurate description of its normal modes is highly desirable and may provide deeper insights into the functional mechanism of the complex, it was lacking due to computational constraints.

Here, we apply the proposed iterative procedure to obtained a coarse-grained ssNMA Hessian matrix for the entire GroEL/GroES complex. This coarse-grained ssNMA (or CG-ssNMA) model preserves the all-atom accuracy in its description of the motions of the coarse-grained structure as in the original ssNMA.

4.3.5   Mean-Square Fluctuations

First, we apply CG-ssNMA to compute mean-square fluctuations. To this end, we use the GroEL-GroES-$(ADP)_7$ complex (pdbid: 1AON) [151] as the initial structure. This structure is composed of the co-chaperone GroES, the *cis*-ring whose subunits are bound with 7 ADPs, and the *trans*-ring.

Structure Preparation.   The residues whose side-chains are not present in the PDB structure (1AON) are effectively treated as alanines (no side chains have been added). Since the crystal structure contains only heavy atoms, hydrogen atoms are added using the psfgen program from VMD [50] and energetically minimized. Lastly, the Hessian matrix of all-atom ssNMA [89] is determined, and is coarse-grained using the proposed procedure as detailed in Algorithm 2.

Fig. 4.4 shows the mean-square fluctuations (MSFs) determined by CG-ssNMA (in red) and by the coarse-grained C_α-based ANM (in gray), and the experimental B-factors

(in black). In (A), all 8015 residues' MSFs and B-factors are shown for three separate parts: the *cis*-ring with a white background, the *trans*-ring with a light gray background, and the GroES cap with a white background. In (B), the first subunits of the three parts (*cis* and *trans* rings, and GroES) are re-plotted to show the MSF in more detail. In the figure, the mean-square fluctuations by ssNMA and ANM are computed using all the modes (including all the high-frequency modes) and scaled to minimize the root-mean-square deviation from the experimental B-factors. The correlation between experimental and predicted B-factors is 0.69 for ssNMA, and 0.52 for ANM. Note that there are a few high peaks in ssNMA MSFs.
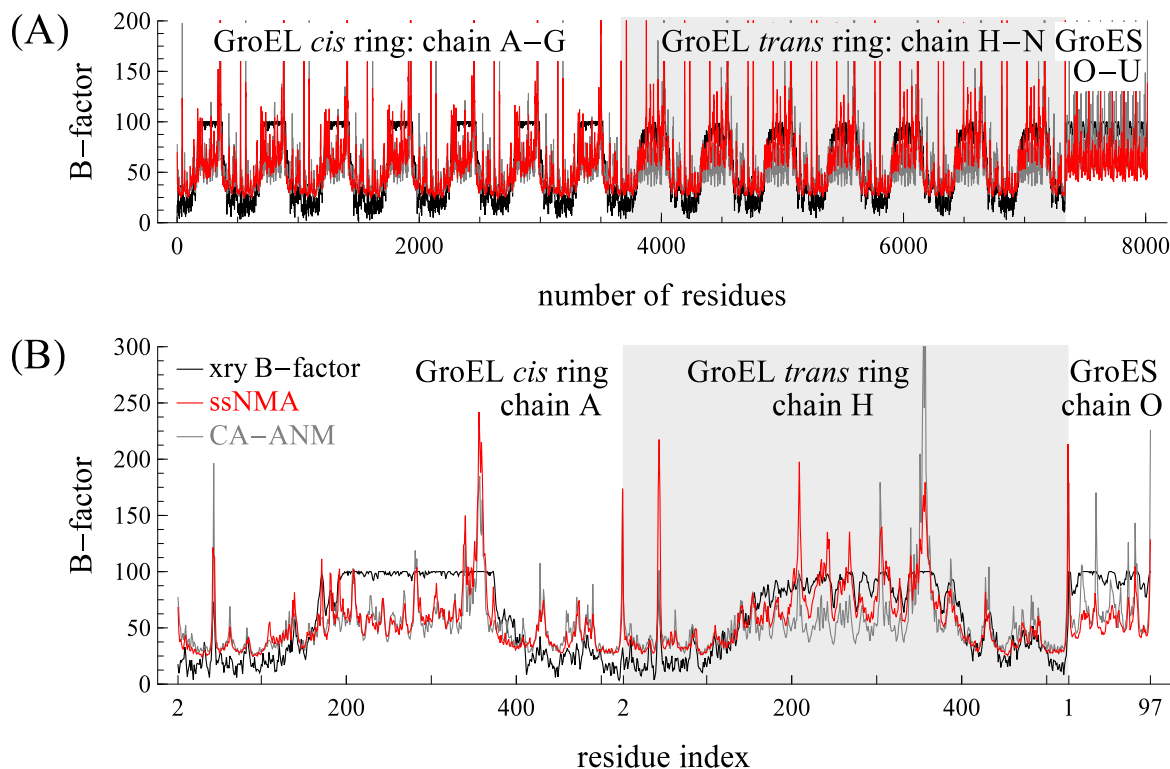
Figure 4.4 **Comparisons of the experimental B-factors with the mean-square fluctuations (MSFs) computed with the new coarse-grained ss-NMA and by ANM, for (A) all residues and (B) only the first subunit in each ring.** The middle gray region is the *trans*-ring of GroEL, and the left and right white regions are the *cis*-ring of GroEL and the GroES cap, respectively.

4.3.6 Motion Correlations and Cooperativity

The motion correlation (or cooperativity) $C_{i,j}$ between the i-th and j-th residues can be expressed as follows:

$$C_{i,j} = \frac{\langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle}{(\langle \mathbf{r}_i \cdot \mathbf{r}_i \rangle \langle \mathbf{r}_j \cdot \mathbf{r}_j \rangle)^{1/2}}, \tag{4.6}$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the displacement vectors for the i-th and j-th residues in a given mode, respectively, $\mathbf{a} \cdot \mathbf{b}$ is the dot product of two vectors $\mathbf{a}$ and $\mathbf{b}$, and $\langle a \rangle$ is the average value a within the first k lowest frequency modes. Fig. 4.5 shows the cooperativity of residue motions within each subunit and across the whole protein complex. The cooperativity plot is generated from the first 15 dominant (i.e., lowest frequency) modes given by the coarse-grained ssNMA.
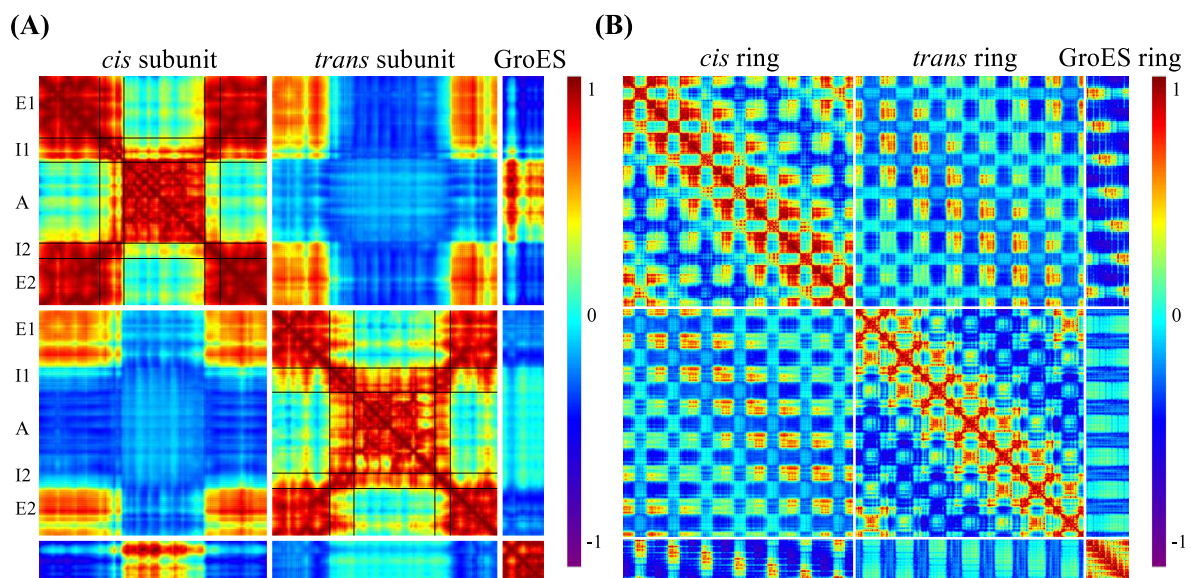
Figure 4.5 **Cooperativity of residue motions using the first 15 lowest frequency modes of the coarse-grained ssNMA model.** (A) The cooperativity within a single set of subunits: chain A from the *cis* ring, chain H from the *trans* ring, and chain O from GroES. (B) The cooperativity among all residue pairs in the GroEL/GroES complex.

Fig. 4.5(A) shows the cooperativity among residue pairs within a single set of subunits: one subunit from the *cis* ring (chain A of 1AON), one from the *trans* ring (chain

N), and one from GroES (chain O). The cooperativity of residue pairs is color coded: red for strong correlated motions ($C_{i,j} = 1$), cyan for uncorrelated ($C_{i,j} = 0$), and purple/blue for anti-correlated ($C_{i,j} = -1$). The most noticeable difference between the *cis* and *trans* rings is the involvement of the intermediate domain in the motions of the apical or equatorial domain. In the *cis* ring, the red regions indicate that the motions of intermediate domain (I1 and I2) are strongly correlated with those of the equatorial domain (E1 and E2), while the motions of the apical domain (A) are largely independent of them. In the *trans* ring, however, the motions of intermediate domains (I1' and I2') are more correlated with those of the apical domain (A') than with the equatorial domain (E1' and E2'). A similar cooperativity plot for the ANM model is given in Supporting information (4.5). Overall, the two methods give similar correlation patterns. The main noticeable difference is that the relative motions between equatorial (E1' and E2') and apical (A) domains of the trans-ring subunit are more clearly shown as anti-correlated (i.e., the region appears to be bluer) in Fig. 4.5 (given by ssNMA) than what is found with ANM shown in 4.5.

One general role of the intermediate domain is connecting the apical and equatorial domains and facilitating the communication between them. The results in Fig. 4.5 imply that the dynamics or motion partner of the intermediate domain depends on the structural form of the GroEL ring: *cis* or *trans*. Considering the structure transitions of *cis*→*trans* and *trans*→*cis* that take place during the GroEL/GroES functional cycle, it is not surprising that the transition path in the former case may be different from a simple reverse of the latter. Additionally, Fig. 4.5(A) shows that the motions of GroES and the apical domain (A) of the *cis* ring also are highly correlated.

The cooperativity of all the residues in the complex is presented in Fig. 4.5(B), along the off-diagonal where there are four dark blue mesh bands, implying that the apical domains of subunits that sit on opposite sides across the rings, such as chain C/D and chain A, are strongly anti-correlated. Another interesting observation is that the motions of GroES are strongly anti-correlated to the equatorial domain of the *cis* ring.

4.3.7 The Characteristics and Quality of the ssNMA Modes

The ssNMA model presented in this work, though coarse-grained in structure, maintains an all-atom level accuracy in its description of the interactions and consequently an all-atom level accuracy in its description of the normal mode motions of the coarse-grained structure. Such an accurate description of the normal mode motions is highly desirable and has not been performed before for such a large protein complexes as GroEL/GroES with over 8,000 residues. In the following, we will examine closely the first few lowest frequency modes of ssNMA and characterize their motions. The quality of these modes is then assessed. A comparison with C_α-based ANM modes is made at the end.

Fig. 4.6 characterizes the slow dynamics of GroEL/GroES in individual modes or pairs of modes. The first lowest frequency mode portrays a rotational motion around the cylindrical axis of the complex. This mode matches with the first mode of ANM nearly perfectly, with a high overlap of 0.97. The third mode is about opening the gate of the *trans* ring to receive substrates into its chamber, by moving its apical domains to conform its structure to resemble that of the *cis* ring. The second and fourth modes are mainly about a swing motion of the *trans* ring. This motion also helps to open the chamber gate of the *trans* ring. In ssNMA, this gate opening motion in the *trans* ring is clearly captured by these three distinct modes, especially the third mode, whose importance is manifested also in the conformation transitions during the GroEL/GroES functional cycle that will be described in the next section. In ANM, there is not a single mode that closely matches the third mode of ssNMA. The gating opening motion seems to spread into several modes in ANM and be mingled with other motions. The 5th–6th modes are shearing motions of the GroES cap and the apical domains of the *cis* ring. This motion causes them to shift significantly relative to the equatorial domains. This motion (in the 5th/6th modes) is similar, to some extent, to that in the second and third modes of ANM, which in turn have some resemblance also to the second/fourth modes of ssNMA.

The 7th–10th modes display alternating motions of compression and extension of the whole complex. The 11th mode is mainly about stretching/compressing the chamber of the *cis*-ring. To some extent, this motion (of the 11th mode) changes the structure of the *cis* ring towards the shape of the *trans* ring. The 12th–13th modes are mainly about tilting the *cis/trans* rings and the GroES cap.
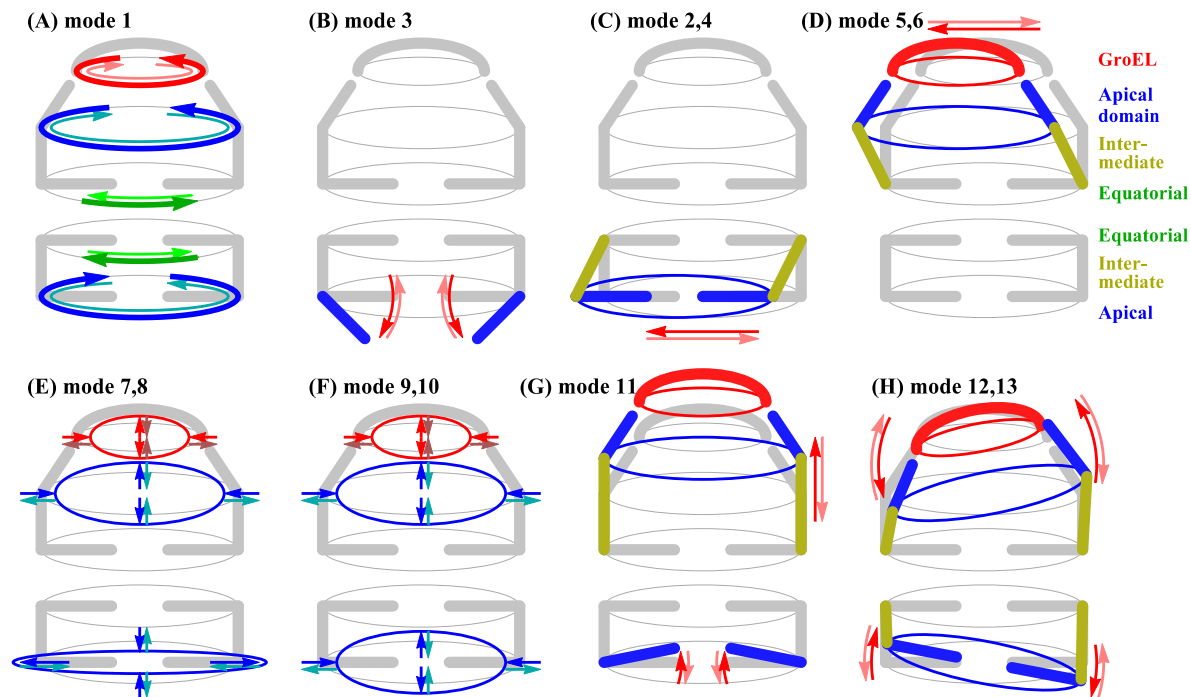
Figure 4.6 **Descriptions of the first 13 lowest frequency modes of GroEL/GroES, determined by the coarse-grained ssNMA.**

The animations of the top 13 dominant modes (lowest frequency) of ssNMA (and ANM) are made available at http://www.cs.iastate.edu/~gsong/CSB/coarse.

Next, we compare more quantitatively the modes of ssNMA and ANM.

Quantitative comparisons of the normal modes of ssNMA and ANM. Table 4.2 summarizes the overlaps between the lowest frequency modes of coarse-grained ssNMA and ANM. Note that the first ssNMA mode matches nearly perfectly with the 1st ANM mode with a high overlap value of 0.97, while other modes match only moder-

ately well. The order of modes between the two models also seems to be scrambled. The fairly low overlap values indicate that only the lowest frequency mode is well preserved in ANM, but significantly less so for other modes. This is consistent with our previous observations [91, 92]. The third ssNMA mode is mainly about opening the gate of the *trans* ring by moving its apical domains apart so that its structure becomes more similar to the *cis* ring. This mode is functionally important as it describes a key protein transition (see the next section). However, in ANM, the closest resemblance of this motion is to the 20th mode that describes a mixed motion of expanding/compressing of both GroEL chambers.

Table 4.2   ssNMA modes and their corresponding best matching modes in ANM.

ssNMA	ANM	overlap	ssNMA	ANM	overlap
1	1	0.97	8	7	0.83
2	4	0.65	9	9	0.64
3	20	0.62	10	10	0.66
4	5	0.66	11	8	0.78
5	2	0.68	12	11	0.60
6	3	0.72	13	12	0.62
7	6	0.77			

The table contains ssNMA modes and their corresponding best matching modes in ANM with which they have the largest overlaps. Results shown are for the first 13 lowest frequency modes, the same modes whose motion characteristics are presented in Fig. 4.6.

Fig. 4.7 shows how well the quality of the secondary structures are preserved as the protein complex moves in the directions of the modes of ssNMA or ANM. In this study, for each mode, the protein structure is deformed along the mode direction until its RMSD changes 1 Å from the initial structure. The RMSDs of individual secondary structures (alpha-helices or beta-sheets) are determined independently, and the average RMSDs of these secondary structures are then computed. This procedure is repeated for the first 100 lowest frequency modes of both coarse-grained ssNMA and ANM. In the figure, the

solid red (black) line represents the secondary structure deviations by the coarse-grained ssNMA (or ANM), and the dashed lines are the least-square fits to the solid lines. The plot shows that secondary structures are preserved about twice as well with ssNMA as for ANM.
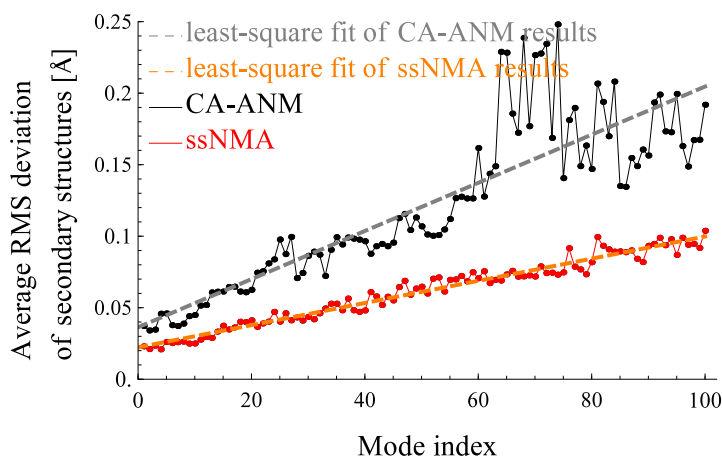
Figure 4.7 **Preservation of secondary structures in mode motions.** The solid red (or black) line represents the average structure deviations of all secondary structures of the GroEL/GroES complex when it moves along a normal mode of ssNMA (or ANM). The dashed lines are the least-square fits to the solid lines.

In summary, there are two major quality improvements in ssNMA modes over ANM modes, both of which can be attributed to the all-atom accuracy that is maintained in ssNMA. First, the secondary structures are better preserved in ssNMA modes than in ANM. The modes determined by coarse-grained ssNMA appear to be more accurate and realistic. This is consistent with the the more realistic potential that ssNMA employs. ssNMA has several terms in its potential function that enforce covalent geometry while the ANM model treats the whole system with uniform elastic springs. Second, which is related to the first, the modes by coarse-grained ssNMA seem to characterize the different collective motion patterns of the protein complex better. So, interestingly there is some significant amount of cohesion that is lost in the coarse-graining with ANM, which is retained in the ssNMA.

4.3.8 Normal Models Facilitate the Functional Conformation Transitions

In this section, we apply CG-ssNMA to interpret the conformation transitions in the functional cycle of GroEL/GroES. Our hypothesis is that the intrinsic normal mode motions of the complex should facilitate its conformation transitions. To measure how well the modes are related to the conformation transitions, we compute the overlaps between normal modes and a given transition. We then repeat the computations and analysis using ANM and compare the results with those from CG-ssNMA.

In total there are six conformation transitions among five known conformation states of the complex (see Table 4.3) considered: T→R, T→R‴, R''_{nocap}→$R''_{nocap,flipped}$, R''→$R''_{flipped}$, R''→S, and S→R'', where "nocap" stands for the absence of the GroES cap. Table 4.4 summarizes, for these transitions, the top 3 largest overlaps found using CG-ssNMA and ANM. The indices of the modes that give the largest overlaps also are given. The first two cases represent transitions from the apo form to ATP/GroES bound forms. The transitions R''→$R''_{flipped}$ and R''_{nocap}→$R''_{nocap,flipped}$ were thought to take place during the functional cycle of GroEL/GroES [104], in which the two GroEL units alternate as a functional chaperone. However, recent work [31] suggested that in *vivo* the GroEL/GroES complex assumes a football shape in the functional process and that both GroEL's might work simultaneously as protein unfolding chaperones. For this reason, we consider also the functional transitions between states R'' and S. Table 4.4 lists the results.

T→R and T→R''_* : Transitions T→R and R''_* in Table 4.4 show that these is mostly achieved with a torsional motion along the vertical axis of the structure. Both the CG-ssNMA and ANM models capture this torsional motion, but their mode indices are different. It is the fourth mode in CG-ssNMA that gives the largest overlap while it is the first in ANM. The results clearly show that the motion to R (as induced by ATP binding) is along the path to R''_*, as observed by Roseman et al. [107] from low resolution cryo-EM images.

Table 4.3   The five conformations of the GroEL/GroES complex used in this work.

conformation	pdb-id	description
T state	1GR5	The tense state.
R	2C7E	The relaxed state, 7 ATP bound
R″	1AON	Bullet-shaped structure, 7 ADP bound, GroES bound
R″$_*$	1GRU	Bound with 7 ATP and 7 ADP, GroES bound
S	4PKO	S is obtained by removing one GroES ring from the football-shaped complex 4PKO that is bound with two GroES rings and 14 ADP.

R″→R″$_\mathbf{flipped}$: Ranson et al. [104] suggested that the functional process of GroEL/GroES involves alternations to the two GroEL rings as functional units and the complex is bullet-shaped [151] in *vivo*.

Here we consider the transition between a bullet-shaped complex (R″) to its flipped counterpart. In this transition, one of the GroEL rings goes from the *trans* form to the *cis* form, while the other ring changes from *cis* to *trans*. Results in Table 4.4 show that the coarse-grained ssNMA captures well the transition from *trans* to *cis* using its fourth mode, which has the second largest overlap, while the 17th mode has the best overlap and characterizes mostly the transition from *cis* to *trans* ring, as well as a partial transition from *trans* to *cis*. ANM, on the other hand, describes the transition of *trans→cis* and *cis→trans* using the 17th and 18th modes, each of which is the mixture of both *cis*-ring and *trans*-ring deformations.

It is thought that after the binding of the ATPs to the trans ring, the GroES cap is removed and the substrate protein is released. Then the two GroEL rings go through *trans→cis* and *cis→trans* transitions, respectively, and another GroES will bind the opposite ring, completing a cycle. The GroES cap stabilizes the *cis* ring in its conformation and prevents its transition to a *trans* conformation. However, after the ATP binding at the opposite ring, the GroES cap is removed, which makes the transition from a *cis* to a *trans* conformation easier. The larger overlap seen in this transition without the

Table 4.4 Top three overlaps between structure displacements and normal modes.

transition models	T→R		T→R''$_*$		R''→R''$_\text{flipped}$		R''$_\text{nocap}$→ R''$_\text{nocap,flipped}$		R''→S		S→R''	
	ovlp	mode	ovlp	mode	ovlp	mode	ovlp	mode	ovlp	mode	ovlp	mode
coarse-grained ssNMA	0.56	4	0.64	4	0.40	11	0.49	17	0.53	1	0.41	2
	0.24	46	0.27	60	0.36	15	0.29	4	0.39	3	0.26	3
	0.24	16	0.25	17	0.26	3	0.29	13	0.27	15	0.24	14
CA-ANM	0.57	1	0.55	1	0.46	13	0.51	18	0.52	1	0.47	3
	0.33	51	0.40	11	0.36	32	0.43	17	0.35	20	0.39	14
	0.33	10	0.23	96	0.33	20	0.27	116	0.26	68	0.24	24

Structure S is obtained by removing one GroES from the football-shaped structure (pdbid: 4PKO). For the transition from R'' to its flipped counterpart, the normal modes are computed either with or without the GroES cap, in both of these cases only the two GroEL rings are used to computed the conformation displacement. The values in each table entry are the overlaps between the given conformation transition and a mode, with the mode index also given.

GroES cap (see Table 4.4) provides evidence that GroES is probably removed first before the *cis*↔*trans* conformation transitions take place rather than occurring simultaneously. This agrees with the idea that structures facilitate functional transitions.

R''→S (opening the *trans* ring gate): Recent work by Fei et al. [31] suggested that the GroEL/GroES complex in *vivo* should have a football shape. The formation of a football-shaped GroEL/GroES complex was thought to be promoted by substrate protein (SP), and that "SP shifts the equilibrium between the footballs and bullets in favor of the former, consequently making them the predominant species." [31]

Here, we examine the transitions between a football-shaped complex and a bullet-shaped complex. Transition R''→S opens the gate of the *trans* ring to receive a substrate protein (unfolded or misfolded) in its chamber. This is accomplished by conforming the structure of its apical domain to that of a *cis* ring (see the third mode in Fig. 4.6 and in 4.5). 4.5 highlights the conformation change that takes place within a trans-ring

monomer in this transition. The overlaps between the transition and normal modes reveal a large contribution by the torsional rotation along the vertical axis (mode 1), as the *trans* ring of S is rotated about 8 degree counter-clockwise from that of R″ [31]. Secondly, this transition is captured by the third ssNMA mode that mainly depicts a chamber-opening motion. In contrast, CA-ANM provides this transition mainly using its 20th mode, which is a mixture of the chamber opening motion and some other deformation of the *cis* ring and the GroES cap.

S→R″ (closing the *cis* ring gate): Transition S→R″ closes the gate of the *cis* ring to conform its structure to that of a *trans* ring. Similar to the transition R″→S, this transition requires torsional rotations and gate-closing motions. The coarse-grained ssNMA captures this transition using the second and third low frequency modes. CA-ANM captures the torsional rotation properly using the third mode, but has to rely on higher-frequency modes to capture the gate-closing transition (See Table 4.4, last column).

Summary. For all the above conformation transitions, CG-ssNMA's interpretation of them involves more of the first few lowest frequency modes than for ANM. This is consistent with the observation made earlier that ssNMA modes tend to preserve the secondary structures better and thus likely are of better quality. Indeed, it is expected that the all-atom accuracy that CG-ssNMA maintains should render a more accurate description of protein motions.

4.4   Conclusions and Discussions

Normal mode analysis (NMA) is an indispensable tool for obtaining the patterns of intrinsic collective dynamics of biomolecular systems around their native states. Such dynamics studies and computations are important since dynamics is tightly linked to

functional mechanisms and can reveal insights that studies based on static structures alone cannot provide. For very large complexes and eventually even a cell, all-atom descriptions of the dynamics of the system are neither feasible nor necessary. A coarse-grained structure representation is often sufficient. But what about the dynamics for a coarse-grained structure? Even though the structure representation is coarse-grained, we still would like to have an accurate description of its dynamics, ideally as close in accuracy to an all-atom model as possible.

It was by the use of coarse-grained models that past normal mode studies of very large biomolecular systems were carried out and remarkable insights were gained in these studies [58, 64, 67, 75, 132, 146]. There is no doubt that the levels of coarse-graining chosen for studying these large systems were appropriate. However, what was not previously assessed was the quality of the dynamics that was provided by the coarse-grained structure representations, by comparing against atomic results. Since most coarse-grained models use extremely simple potentials to model the interactions within the coarse-grained structure, the dynamics they render are likely to have some deficiencies.

In this work, we have successfully bridged this gap and have presented a new method that can be used to efficiently construct a coarse-grained model whose for which the dynamics of the coarse-grained structure remains as accurate as that for by all-atom model. The method takes advantage of the sparseness of the Hessian matrix and iteratively reduces its size through projection until it is reduced to that of the desired coarse-grained structure. Since the projections maintain the accuracy of the interactions, the final Hessian matrix represents the precise interactions within the coarse-grained structure. Compared with the RTB (rotation-translation block) method [128] or BNM (block normal modes) [76], which assumes rigidity and ignores flexibility within each block, our method provides a more accurate description of the motions of coarse-grained systems. Compared with the VSA model (vibration subsystem analysis) [40, 148], the advantage of our method is that it is computationally significantly more efficient.

Results presented in this work are highly significant since they promise to provide descriptions of normal mode motions at the all-atom level of accuracy even for the largest biomolecule complexes. While preserving all-atom accuracy through matrix projection is not new and has been done previously [11, 40, 148, 163], one of our key contributions here is developing a new algorithm that can carry out this matrix projection highly efficiently and therefore make it applicable to very large structure complexes, which has not been done previously. Such accurate descriptions of the intrinsic dynamics may help reveal new insights into the functional mechanisms of many biomolecular systems. It should be noted that because we are able to efficiently obtain a precise interaction model (the Hessian matrix) for the coarse-grained system, we can solve it not only for a few low frequency modes, but for all the modes. If only a few low frequency modes are needed, then there are some alternative methods that may be more efficient.

Our application of the method to GroEL/GroES reveals some new insights into the functional process of this biologically important chaperonin. For example, our results show that the conformational transitions of this protein complex in its functional cycle are even more closely linked to relatively few of its lowest frequency modes than was previously observed using other coarse-grained models.

This work is a continuation of our previous work that aimed to bridge NMA with elastic network models [89]. While the previous work bridged between NMA and all-atom elastic network models, this work represents the second half of developing this bridge, namely between *all-atom* elastic network models and *coarse-grained* elastic network models. Combined together, the two works demonstrate how one can bridge between the conventional NMA that uses an all-atom model with a full force-field and coarse-grained elastic network models that are nowadays the preferred choice for normal mode computations due to their simplicity. This bridging reveals novel insights on how one may develop coarse-grained models that are not only simple to use, but also maintain most of the accuracy of the original NMA.

Limitations and Future Work. Although the proposed iterative coarse-graining procedure can be used to efficiently construct coarse-grained models whose description of dynamics of the coarse-grained structure preserves all-atom accuracy, it is limited in that it can be applied only to some of the models, such as ssNMA or eANM or sbNMA (see 4.5). It cannot be applied to the original NMA. This is because the potential of NMA contains electrostatic interactions that decay rather slowly and consequently the NMA Hessian matrix is not sparse; however, there remain some uncertainties about how to best compute the electrostatics.

A possible partial solution is to add a switch function to the non-bonded interactions of NMA and make it decay to zero at some cutoff distance, as is commonly done in MD simulations. This will make the Hessian matrix much sparser and make it possible to apply the proposed iterative procedure to NMA. We have shown this to be the case (see results in 4.5). However, this is only a partial solution since it recovers only the short range part of the electrostatics. The long range electrostatic interactions, which may have a pronounced contribution to long-range collective motions and cooperativity, are still missing. Additionally, the cumbersome energy minimization (which ssNMA does not require) becomes necessary, which can be a challenge when working with large biomolecular complexes.

One possible future work is to study the effects of electrostatic interactions on normal modes, specifically the extents of contributions by short-range and long-range electrostatic interactions. If the short-range component of the electrostatic interactions dominates the long range component in contributing to normal modes, then the aforementioned partial solution will provide an excellent approximation.

4.5   Supporting Information

S1 Video.

The important gate-opening mode (mode 3) in R″ →S transition. The video shows the motions of the GroEL/GroES complex along this important gate-opening mode (link). More animations for transitions listed in Table 4.4 are available at `http://www.cs.iastate.edu/~gsong/CSB/coarse`.

S1 Table.

Table 4.5   **Accuracy of screened NMA and sbNMA at different threshold values ξ.**

ξ^a	screened-NMA^b (0.87^c)		sbNMA (0.99^c)	
	corr^d	w-ovlp^e	corr	w-ovlp
0.0001	1.00	1.00	1.00	1.00
0.001	1.00	1.00	1.00	1.00
0.01	1.00	0.98	1.00	1.00
0.1	0.95	0.62	0.99	0.92

This table is an extension of Table 1 and contains results for two more models: screened-NMA and sbNMA, whose accuracy also is (nearly) fully preserved at $\xi = 0.01$. The initial sparseness of the Hessian matrix, in parentheses, is 0.87 and 0.99 for screened-NMA and sbNMA, respectively.
(All the following remarks except ^{b} are the same as those in Table 1.);
$^a\xi$: the threshold value used to set to zero the smallest elements in the Hessian matrix;
^{b}screened-NMA: same as NMA except that its non-bonded interactions (electrostatics and van der Waals) are tapered to zero at 9.0 Å;
^{c}initial sparseness of the Hessian matrix;
^{d}corr: mean-square fluctuation correlation;
^{e}w-ovlp: eigenvalue-weighted mode overlap as defined in Eq. (4.5).

S1 Fig.

(A)

(B)
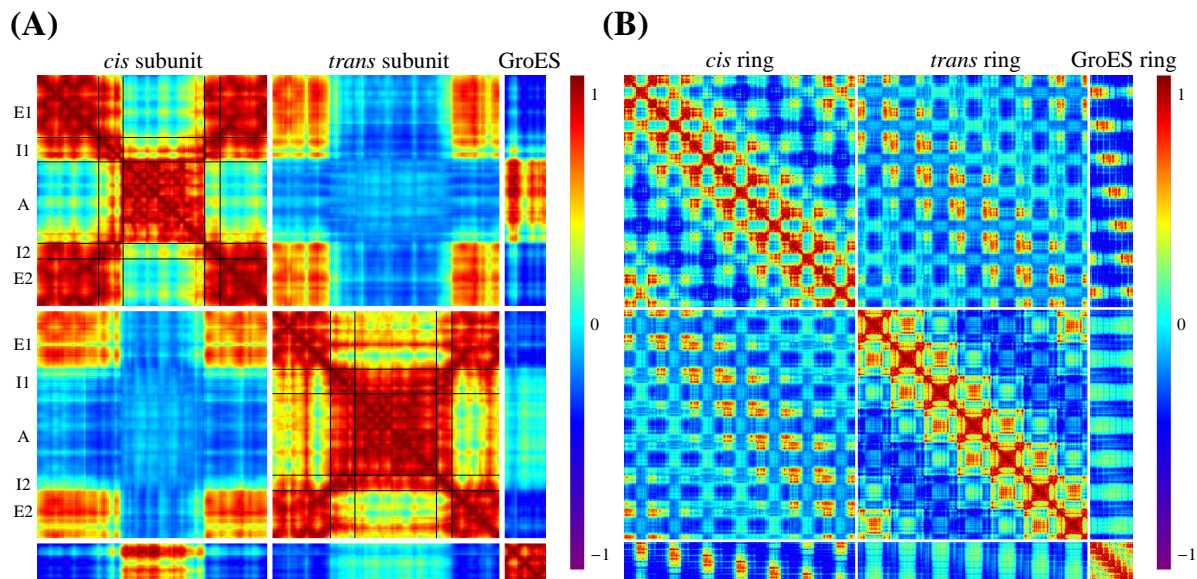
Figure 4.8 **Cooperativity of residue motions using the first 15 lowest frequency modes of the CA-ANM model.** (A) The cooperativity within a single set of subunits: chain A from the *cis* ring, chain H from the *trans* ring, and chain O from GroES. (B) The cooperativity among all residue pairs in the GroEL/GroES complex.

S2 Fig.
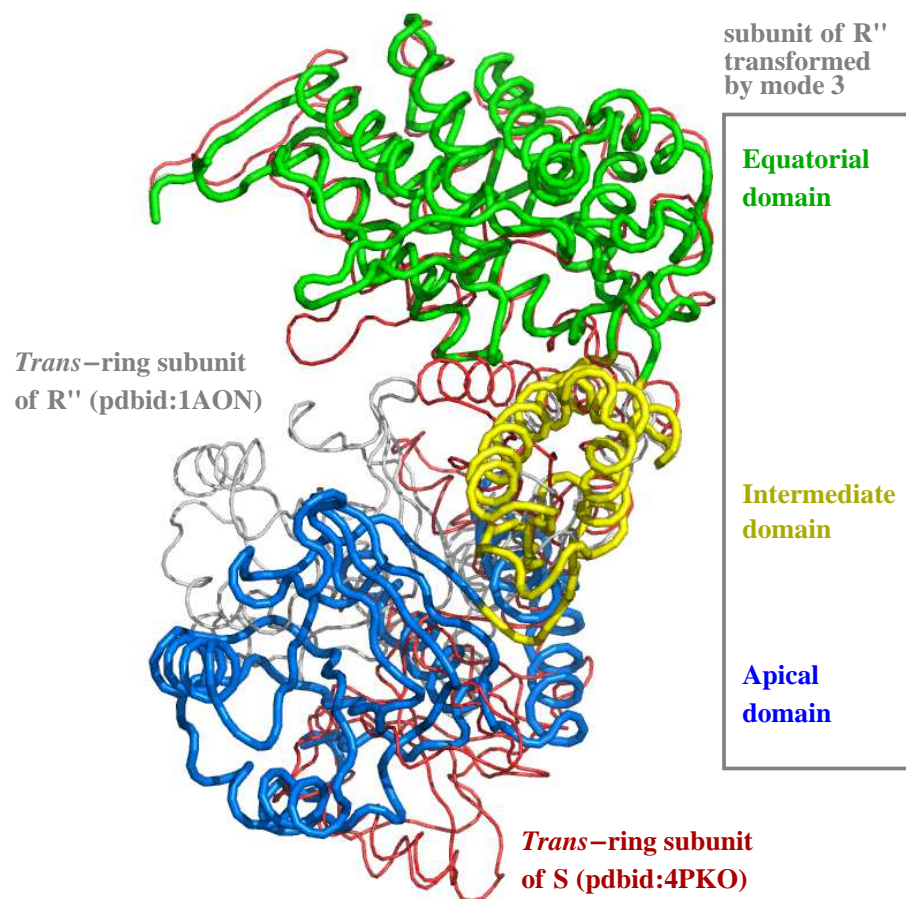
Figure 4.9   **The conformation changes within a trans-ring subunit in R″ →S transition.** The *trans*-ring subunit of conformation R″ is represented by the thin gray line, while that of conformation S by the thin red line. The thick curve (in blue, yellow, and green) displays, for this R″ →S transition, the conformation change contributed by the third mode (of the ssNMA model) alone. This figure shows that a large conformation change takes place within the subunits in this conformation transition and is well captured by the third mode of ssNMA. The three conformations shown are aligned by the equatorial domain (in green).

CHAPTER 5.   UNIVERSALITY OF VIBRATIONAL SPECTRA OF GLOBULAR PROTEINS

Hyuntae Na[2][4], Guang Song[2][4][5], and Daniel ben-Avraham[3][4]

Abstract

It is shown that the density of modes of the vibrational spectrum of globular proteins is universal, i.e., regardless of the protein in question, it closely follows one universal curve. The present study, including 135 proteins analyzed with a full atomic empirical potential (CHARMM22) and using the full complement of all atoms Cartesian degrees of freedom, goes far beyond previous claims of universality, confirming that universality holds even in the frequency range that is well above $100\,\mathrm{cm}^{-1}$ ($300 - 4000\,\mathrm{cm}^{-1}$), where peaks and turns in the density of states are faithfully reproduced from one protein to the next. We also characterize fluctuations of the spectral density from the average, paving the way to a meaningful discussion of rare, unusual spectra and the structural reasons for the deviations in such "outlier" proteins. Since the method used for the derivation of the vibrational modes (potential energy formulation, set of degrees of freedom employed,

---

[1]This chapter is reprinted with permission of *Phys. Biol.* 2016, 13(1), 016008.

[2]Graduate student and Associate Professor, respectively, Department of Computer Science, Iowa State University.

[3]Professor, Department of Physics, Clarkson University.

[4]Primary researchers and authors.

[5]Author for correspondence.

etc.) has a dramatic effect on the spectral density, another significant implication of our findings is that the universality can provide an exquisite tool for assessing and improving the quality of potential functions and the quality of various models used for NMA computations. Finally, we show that the input configuration too affects the density of modes, thus emphasizing the importance of simplified potential energy formulations that are minimized at the outset. In summary, our findings call for a serious two-way dialogue between theory and experiment: Experimental spectra of proteins could now guide the fine tuning of theoretical empirical potentials, and the various features and peaks observed in theoretical studies – being universal, and hence now rising in importance – would hopefully spur experimental confirmation.

5.1 Introduction

The atomic structures of thousands of proteins have been elucidated and display recurring patterns of folding, such as the common globin and Greek key folds, and the β-barrel folds. These repeating structural motifs obtain distinct flexibility signatures. Identifying these intrinsic deformability characteristics is required to ascertain and better understand protein functionality. Historically, the characterization of any object's internal deformabilities under small perturbations has been achieved by a normal mode analysis of its internal degrees of freedom. While a normal mode analysis is a well-defined and straightforward computation, the identification of a suitable set of internal degrees of freedom and an appropriate potential energy formulation to quantify the effects of deformations, remains more of an art. Here we examine the spectrum of vibrations obtained for a large number of proteins, using several of the more traditional approaches for normal mode analysis. Within a given approach, the density of the spectrum of vibrations is universal, despite the many significant differences among individual proteins.

Normal modes of proteins have been studied since the early 1980's [17, 38, 72, 131]. A normal mode calculation requires as input an empirical potential function for the various forces between the protein's atoms: the more detailed the atomic potential function, the more reliable the results, but on expense of more cumbersome computations. Starting with the seminal work of Tirion [139], various simpler alternatives to a detailed potential have been explored [2, 8, 9, 43, 47, 76, 77, 80, 89, 91, 130, 142, 149, 153, 160]. In addition to simplifications in the potential energy formulation, reduced sets of internal degrees of freedom (dofs) have been explored. Two common choices include the restricted set of dihedral angles degrees of freedom, or *torsional* dofs, for short (the rationale being that changes in bond lengths and angles require far larger energy investment than dihedral or torsional changes) [38, 72, 141] and Cartesian dofs for the reduced set of only the C_α atoms [8, 9].

Early normal mode analyses examined the density of the modes by frequency range, $g(\omega)$, and deliberated the meaning of the various features in the curves found for each protein. However, it was soon found out that, when properly normalized, the $g(\omega)$ of different proteins seem to collapse onto one universal curve, characteristic of globular proteins in general [10, 141]. This initial finding was based on merely 5 proteins, and on a normal mode analysis with only torsional dofs. Most later studies of the distribution of normal modes did little to confirm the universality of $g(\omega)$, as they focused on properties of the spectrum only at the low frequency range (up to $\sim 20\,\mathrm{cm}^{-1}$) and tended to rely on simplified potential functions. An exception is the recent analyses of Hinsen et al., [44, 46] of crambin, lysozyme, and myoglobin, using the AMBER potential, that suggested that the universality of the density of the modes extends to all frequencies.

In this paper, we re-examine the hypothesis that $g(\omega)$ is universal. Advances in computer technology in recent years allow us to consider 135 globular protein structures whose resolutions are better than 2.5 Å and whose sequence identity is less than 30%, and obtain spectra of normal modes with a detailed atomistic empirical potential and

the full complement of Cartesian degrees of freedom. (Some of our results are presented for torsional dofs only and/or for simplified potentials.) This wealth of information lets us do much more than simply confirm the putative universality of $g(\omega)$: (1) Our main result is that the density of the spectrum of vibrations, $g(\omega)$, is universal also for the full complement of Cartesian dofs, down to the seemingly idiosyncratic peaks and details in the high-frequency range. This is a big surprise: in the low-frequency range universality is expected on the grounds that slow modes involve long-wavelength coherent motion of large domains of a protein, and therefore the many interactions involved (at the surfaces between domains) average out in the same fashion, regardless of details. In contrast, high-frequency oscillations involve small coherence lengths and motions of *small* groups of atoms relative to one another, so here universality is unexpected. (2) Our large data set allows us to characterize not only a reliable average for $g(\omega)$, but also the typical fluctuations from that average. Specific features in the $g(\omega)$ of a protein are unusual only in comparison to these fluctuations, so the old notion of identifying and discussing the meaning of salient features of $g(\omega)$ of a protein finally becomes possible. For example, our data allows us to identify subtle, yet meaningful differences in the spectra of proteins of different folds. Some of these observations are echoed in experimental findings. (3) The universal curve for $g(\omega)$ depends on the specific empirical potential one uses, its parameters, etc., whether the potential is detailed or simplified, as well as on the set of degrees of freedom (e.g., Cartesian or torsional). We show that the comparison of the $g(\omega)$'s arising in each case is a very sensitive way to assess the accuracy and success of the various approximations and approaches. (4) Working with an atomistic detailed potential, the first step in an NMA involves minimizing the potential function, thereby altering the input PDB structure. In other simplified approaches, one posits a potential that is minimized at the given configuration (PDB, or other) at the outset. We show that energy-minimized starting configurations obtain significantly different spectra $g(\omega)$ than the original PDB starting configurations, and we discuss the implications of this finding.

The remainder of this paper is organized as follows. In Section 5.2 we describe our protein dataset and briefly review the theoretical technique of normal modes analysis and the various approaches (full and simplified potential functions, choices of dofs, etc.) considered in this work. Our results are presented and analyzed in Section 5.3. Final conclusions and promising open problems are discussed in Section 5.4.

5.2   Materials and Methods

5.2.1   The Protein Dataset

The protein dataset used in this study is the same as the one used in a previous work by Na and Song [92]. The dataset includes 135 proteins resolved to better than $2.5\,\text{Å}$ and following minimization none of the proteins undergoes more than a $6.0\,\text{Å}$ RMSD change. The proteins are quite evenly divided between different fold classes, including 42 all-α proteins, 37 all-β proteins, and 56 α/β-proteins. Their sizes range from 61 residues (pdb-ids: 1I2T, 1I0M, 2J5Y, 3MP9) to 149 residues (pdb-ids: 1GU1, 2Y9F, 3AXC); the distribution of the proteins by size is illustrated in figure 5.1(A). Only small to medium sized proteins are used here due to the large computational cost of running NMA. The protein structures are energetically minimized using the Tinker program [102] with the CHARMM22 force field [83]. The amount of structure deviations due to energy minimization is given in figure 5.1(B). No cutoff distance is specified in the process. As a result, the program does not taper the electrostatic or the van der Waals potential with any smoothing function but considers all pair-wise non-bonded interactions. The minimized structures and the original PDB structures are available at http://www.cs.iastate.edu/~gsong/CSB/NMAdb/135.html.
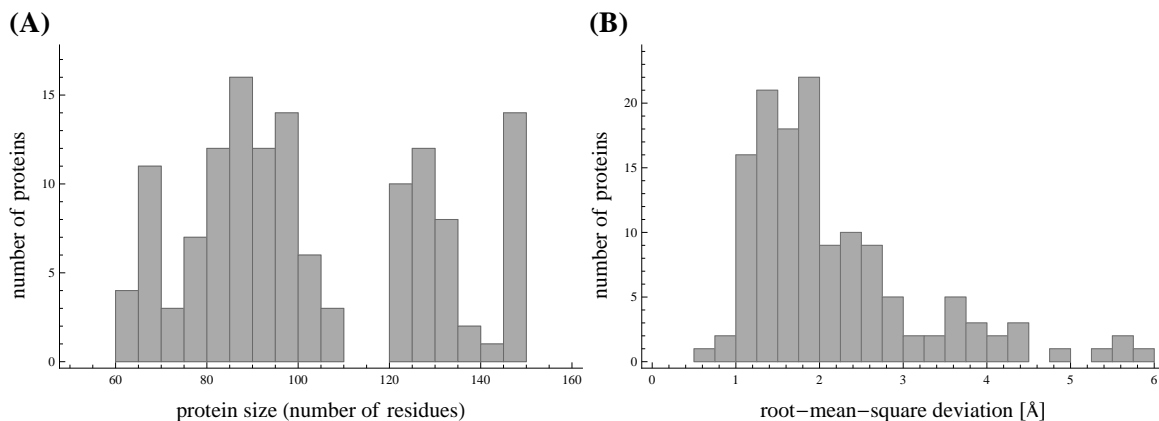
Figure 5.1   **(A) The size distribution of the 135 proteins used in this work; (B) the extent of structure deviations caused by energy minimization among the same 135 proteins.**

5.2.2   Normal Modes Analysis

Normal modes analysis (NMA) was first applied to proteins in the early 80's. [17, 38, 72, 131]. Conventional NMA proceeds from a detailed atomic potential function, V, for the interactions within the protein system. Generally, the input structure (mostly a PDB structure) is not at an energy minimum according to V. As required by NMA, the potential function has to first be minimized — a computationally expensive operation that also distorts the starting configuration by as many as several angstroms (RMSD).

Using the minimized structure, one constructs the Hessian matrix **H**, which is the second derivative of the potential energy with respect to the protein's degrees of freedom $\{q_i\}_{i=1}^{N}$;

$$\mathrm{H}_{ij} = \frac{\partial^2 V}{\partial q_i \partial q_j} \, , \tag{5.1}$$

as well as the mass, or inertia matrix **M**;

$$\mathrm{M}_{ij} = \sum_k m_k \frac{\partial \mathbf{r}_k}{\partial q_i} \cdot \frac{\partial \mathbf{r}_k}{\partial q_j} \, , \tag{5.2}$$

where the sum runs over all atoms k of the protein, and m_k and $\mathbf{r}_k$ are the k-atom's mass and location, respectively. One then solves the generalized eigenvalue problem:

$$\mathbf{H}\mathbf{v}_i = \lambda_i \mathbf{M}\mathbf{v}_i. \tag{5.3}$$

Here $\mathbf{v}_i$ is the i-th vibrational eigenmode, and the eigenvalue $\lambda_i = \omega_i^2$ encodes its (angular) frequency; $\mathbf{v}_i(t) = \mathbf{v}_i(0)\cos(\omega_i t)$. For comparison with experimental work, it is customary to express ω_i in terms of the corresponding inverse wavelength of electromagnetic radiation (measured in cm^{-1}). This is achieved by dividing its value (in radians/sec) by $2\pi c$, where c is the speed of light, $c = 2.997925 \times 10^{10}\,\mathrm{cm/sec}$.

Finally, we compute the density of vibrational modes $g(\omega)$, the focus of this work, in the following way. Subdivide the frequency range into bins of width $\Delta\omega$ and count the number of modes n_j that have frequency ω within the j-th bin, i.e., modes with $\omega_j - \frac{1}{2}\Delta\omega < \omega < \omega_j + \frac{1}{2}\Delta\omega$. Then $g(\omega_j) = n_j/(N\Delta\omega)$, where N is the total number of dofs for the protein. (Typically, $\Delta\omega = 5$ or $10\,\mathrm{cm}^{-1}$.) This procedure is done for each of the 135 proteins in our dataset and their $g(\omega)$'s are processed as needed (averaged, compared to one another, etc.)

There exist several choices for the $\{q_i\}_{i=1}^N$ degrees of freedom. The simplest choice, conceptually, is the full complement of Cartesian degrees of freedom; the (x, y, z) coordinates for each of the atoms in the protein. This yields a diagonal $\mathbf{M}$ matrix and minimization is conceptually simpler than with generalized dofs, but the number of dofs N can become prohibitively large. A common alternative in early work was using the quite smaller set of torsional and dihedral angle dofs; bond lengths and bond angles are frozen in this method (approximating the fact that these are much stiffer than the torsional and dihedral dofs). A drawback of torsional dofs is that the $\mathbf{M}$ matrix is more complicated (though smaller), and the minimization algorithm is trickier. In the present study our torsional modes are obtained by projecting a Cartesian Hessian onto the torsional space, as done in the Torsional Network Model [86].

5.2.3 Simplified Normal Mode Analyses

Because conventional NMA is cumbersome to use, due to its complicated all-atom potential and energy minimization process, in 1996 Tirion proposed a simplified potential that required no minimization [139]. Tirion's approach uses torsional dofs (freezing bond lengths and angles) and postulates a universal Hookean potential between non-bonded atom pairs:

$$V = \sum_{\langle ij \rangle} \frac{1}{2} C (r_{ij} - r_{ij}^0)^2 \,, \tag{5.4}$$

where r_{ij} and r_{ij}^0 are the current and the starting-configuration distance between atoms i and j, respectively, and the sum runs over all $\langle ij \rangle$ non-bonded atom pairs that are sufficiently close to one another: $r_{ij}^0 < r_{VdW}^i + r_{VdW}^j + r_c$ (r_{VdW}^a is the Van der Waals radius of atom a and r_c is a cutoff distance, typically a few angstroms). The big advantage of the Tirion potential is that it requires no minimization, it is minimized at the outset at the starting configuration $\{\mathbf{r}_i^0\}$. Other potentials and approaches that require no minimization (known generally as *elastic network models*) have been developed since Tirion's seminal work. We now review the main ingredients of the simplified approaches discussed in this paper.

ANM. The ANM, or Anisotropic Network Model, was developed by Atilgan et al., [2] in 2001. It is mainly a coarse-grained version of the Tirion potential, with each residue represented only by its C_α atom. It has been used also as an all-atom model, though to a much lesser extent. Simplifying further still, ANM employs the easier to use Cartesian dofs. However, on obliterating the constraints of bond lengths and angles it washes out the distinction between bonded and non-bonded interactions and further loses in accuracy. Because of its easy implementation ANM has been widely used in many normal mode-based studies and analyses.

sbNMA. In 2014, Na and Song [89, 90] developed a new way for deriving simplified models for normal mode computations. They employed a top-down approach and derived

several high-quality elastic network models (i.e., require no minimization) by gradually simplifying the conventional NMA. The most accurate of these approaches is the spring-based NMA (sbNMA). Structurally, sbNMA is the same as the conventional NMA and is an all-atom model. The interaction model of sbNMA, on the other hand, is different from the NMA force field from which it is derived. While the Hessian in general consists of spring-constant-based terms $\mathbf{H}^{\mathrm{spr}}$ (or spring-based, for short) — terms that are proportional to the spring constants — and force/torque-based terms $\mathbf{H}^{\mathrm{frc}}$ (terms that are proportional to the inter-atomic forces or torques) [89, 90], i.e., $\mathbf{H}^{\mathrm{NMA}} = \mathbf{H}^{\mathrm{spr}} + \mathbf{H}^{\mathrm{frc}}$, sbNMA keeps only the spring-based terms. The rationale was that the force/torque terms contribute significantly less to the overall dynamics than the spring-based terms [89]. To ensure stability, regions where the spring constants become negative are excluded. For example, electrostatic interactions (which were shown to contribute much less than van der Waals interactions [89]), are not included, as attractive forces give rise to negative spring constants. The spring-based NMA (or sbNMA) preserves much of the complexity of the original NMA, and the neglect of the force/torque terms has minimal impact. As a result, sbNMA yields very high-quality vibrational modes and closely resembles NMA [89].

A very similar approach to sbNMA, dubbed ATMAN (for Atomic Torsional Mode Analysis), was developed independently by Tirion and ben-Avraham [142]. ATMAN too keeps only spring-based terms, derived from a detailed atomic potential, and only wherever these are positive. The main difference to sbNMA is that ATMAN allows for "stretching" the range of positive spring constants, to compensate for the loss of range where the spring constants are negative. This, however, adds tunable parameters.

ssNMA. A further simplification beyond sbNMA is achieved by the simplified spring-based NMA, or ssNMA. It combines many of the different constants in sbNMA into one single parameter, thus requiring a much smaller set, of 17 parameters in total. For example, it uses a single bond-stretching spring constant for all bonded pairs of atoms,

regardless of their types. This of course results in some loss of accuracy, compared to sbNMA. Below, we limit our study to NMA (with the CHARMM22 force field), the sbNMA and ssNMA derived from it, and ANM. Note that only the original NMA requires minimization, while all of the simplified approaches can start from any given protein configuration.

5.2.4 Computing the Contribution from Various Interaction Types

The CHARMM22 potential energy function, which we use for NMA and sbNMA, consists of several types of terms: (a) bond stretching, (b) bond-angle bending, (c) improper angle distortions, (d) torsional and dihedral rotations, and (e) non-bonded interactions, including Van der Waals and electrostatic forces. We group Urey-Bradley interactions along with the bond-angle bending terms. One can compute the relative individual contribution from each type of interaction as follows. Since $V = V_{\text{bond}} + V_{\text{angle}} + \cdots + V_{\text{nonbonded}}$, the Hessian decomposes into mutually exclusive matrices;

$$\mathbf{H} = \mathbf{H}_{\text{bond}} + \mathbf{H}_{\text{angle}} + \mathbf{H}_{\text{improper}} + \mathbf{H}_{\text{torsional}} + \mathbf{H}_{\text{nonbonded}} . \tag{5.5}$$

Then, the relative contribution c_{ij} of interaction type
$j \in \{\text{bond}, \text{angle}, \text{improper}, \text{torsional}, \text{nonbonded}\}$ to the ith mode $\mathbf{v}_i$, is

$$c_{ij} = \mathbf{v}_i^\top \mathbf{H}_j \mathbf{v}_i . \tag{5.6}$$

Note that the c_{ij} are properly normalized, $\sum_j c_{ij} = 1$, because of equation (5.5) and the fact that our eigenvectors are $\mathbf{H}$-normalized; $\mathbf{v}_i^\top \mathbf{H} \mathbf{v}_i = 1$. c_{ij} is guaranteed to be greater or equal to 0 with approaches like sbNMA and ATMAN, where the various $\mathbf{H}_j$ are positive semi-definite. Intuitively, c_{ij} reveals the extent to which interaction type j constrains the vibration along mode i.

5.3   Results

5.3.1   Universality of the Density of Vibrational Modes

Universality in the full complement of Cartesian dofs.   Our main result is presented in figure 5.2, which demonstrates that the density of vibrational modes for each of the proteins in our dataset is very nearly the same. In other words, *the density of the spectrum of vibrations is universal.* To obtain this figure, we have conducted a full NMA on each of the proteins in the dataset, employing the CHARMM22 atomic potential function, and using all of the Cartesian dofs of each protein's atoms, and obtained their $g(\omega)$'s. In the plot, we show the average of $g(\omega)$ over all 135 proteins (black curve); fluctuations from the average are indicated by colored bands that include proteins within different percentile ranges: 25–75 percentile (orange), 5–95 percentile (red), 0–100 percentile (gray). An animation that displays the vibrational spectra of the proteins one by one and illustrates how they all share a common spectrum pattern is given in the Supporting Information.

Surprisingly, even accounting for extreme fluctuations (the 0–100 percentile includes *all* of the proteins in the dataset), the various main features of $g(\omega)$ — seemingly idiosyncratic turns and peaks — are faithfully reproduced throughout the whole frequency range. These peaks must thus correspond to some physical characteristics of the structure of globular proteins in general, and to physical interactions within them that are independent of the details of each individual protein structure.

The universality of $g(\omega)$ would seem to exclude any possibility of gleaning particular knowledge of a protein from its specific vibrational spectrum. This is not necessarily so: If the density of modes of a protein deviates *significantly* from the average $g(\omega)$, perhaps the deviations can tell us something about the structure of the specific protein in question. Whether a deviation is significant, could be decided from the fluctuation bands in figure 5.2. There could be, however, a different cause for deviations, besides anomalous
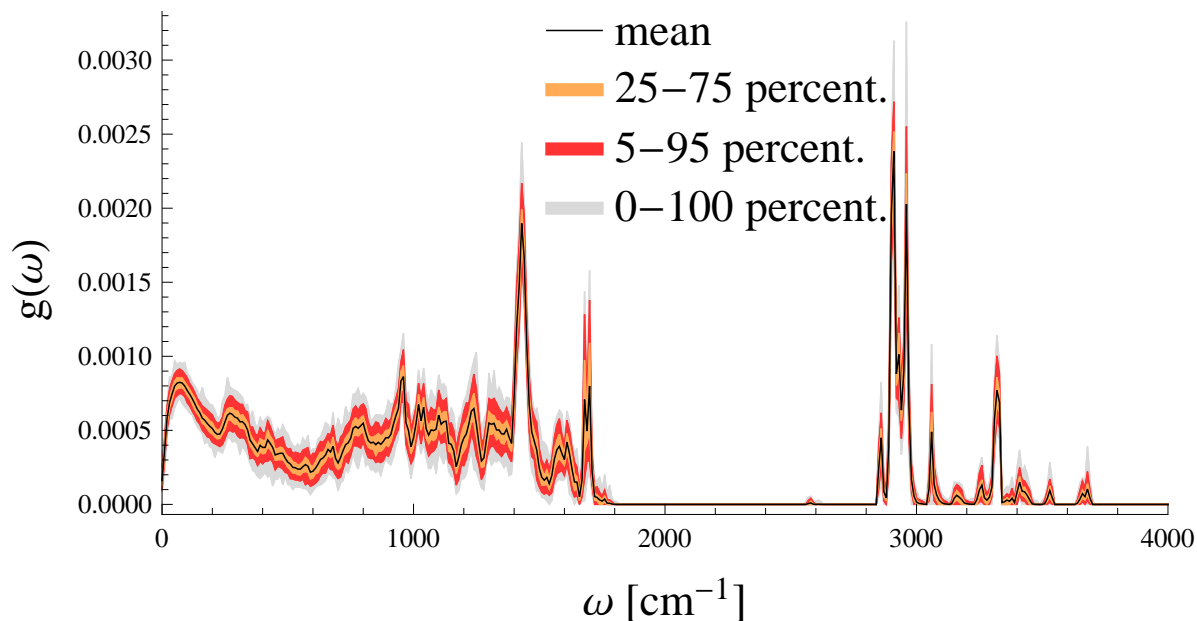
Figure 5.2 **Universality of the density of vibrational modes of globular proteins.** The black line shows the average of the 135 proteins in the dataset. The fluctuations from the average are represented by the color bands, demarcating the fraction of proteins that are included within various ranges: 25–75 percentile (orange), 5–95 percentile (red), 0–100 percentile (gray). The bin size $\Delta\omega$ used here for computing the density of modes is $10\,\mathrm{cm}^{-1}$.

structure: Relative fluctuations of a random variable tend to decrease inversely proportional to the square root of the system's size (the protein's size), so smaller proteins would exhibit larger fluctuations, even if their structure is not anomalous. With that in mind, we have examined the proteins that fall outside of the 5–95 percentile, but found no correlation with size: The distribution of protein sizes in that range is very similar to that of the dataset at large (figure 5.1(A)). Based on these preliminary results, we conclude that rare fluctuations are more likely due to structural anomalies. A more detailed study of the dependence (or independence) of fluctuations on size, and of rare outlier proteins is left for future work.

We stress that ours is the first study of universality of $g(\omega)$ that employs the full complement of Cartesian degrees of freedom. Universality was first discovered by study-

ing the spectrum from the restricted set of torsional dofs, in the 0–$300\,\mathrm{cm}^{-1}$ frequency range [10, 141]. For low frequencies, universality is expected because slow modes involve motion of large domains of a protein, and the many interactions at the surfaces between domains average out uniformly, by the law of large numbers, regardless of details. High-frequency oscillations, on the other hand, involve small coherence lengths and only a few atoms, so the law of large numbers cannot be as easily invoked. We were therefore unprepared for the striking results of figure 5.2. A possible explanation is that, because the high-frequency modes represent the oscillations of only a few atoms relative to the rest of the protein, that has a much bigger mass, the size and structural details of the rest of the protein is mostly irrelevant and consequently the characteristics of the oscillations depend only on the structural composition of the few oscillating atoms and are largely protein independent. We also note that the scope of the present study, of 135 proteins, far surpasses the scope of previous works on the subject, proving universality beyond all doubt, at least from a theoretical point of view. Experimental study on four proteins that have different mixtures of secondary-structures showed that their vibrational spectra have "a common appearance" [37]. This experimental result in the low-frequency range thus seems to confirm universality as well [37] (see also Section 5.3.3).

Cartesian vs. torsional degrees of freedom. We now examine the density of vibrations for the restricted set of torsional dofs. Recall that changes in torsional angles are a lot easier to effect than changes in bond lengths and angles, so most of the protein's motion under thermal excitation could be accounted by torsional changes alone. Moreover, because the reduced set of torsional angles is much smaller than the full complement of Cartesian dofs, it allows analysis of much larger proteins and protein systems, a fact that accounts for much of their popularity.

In figure 5.3 we compare the vibrational modes in Cartesian dofs (dashed black lines) to those in torsional dofs (solid black lines) for four randomly selected proteins in our dataset: 3NBC (a), 3RHB (b), 2QCP (c), 3MP9 (d). Since the total number of dofs is

different for Cartesian and torsional coordinates, we show the actual count of modes in each $5\,\mathrm{cm}^{-1}$-bin, instead of the usual density of modes. In this way one can see that the additional bond-length and angle bending dofs included in the Cartesian complement add vibrational modes in each bin. Nevertheless, there is a clear relationship between the Cartesian and torsional curves: Both the *main peak*, at about $80\,\mathrm{cm}^{-1}$, and the *secondary peak*, at about $300\,\mathrm{cm}^{-1}$, are in nice agreement in the two representations (albeit with more modes present in the Cartesian dofs). As the frequency increases, torsional modes die out and a comparison becomes irrelevant. The first two peaks are particularly important, as they encompass the low frequencies that account for most of the protein's thermal motions (B-factors, etc.).

Same as for the Cartesian case, the density of torsional modes for the proteins in the dataset clusters around a single universal curve (the solid red curve in figure 5.3, which is the averaged density of modes over all the proteins in the dataset). Knowing that $g(\omega)$ is universal in either choice of dofs, Cartesian or torsional, and that the two representations agree on the first two low-frequency peaks, allows one to use whichever set of dofs is convenient for the question at hand. Indeed, in what follows, we shift back and forth between these two choices.

The origin of the peaks. We now address the question of what gives rise to the various detailed features in the spectrum density $g(\omega)$. For each particular eigen-frequency, ω_i, we compute, using the sbNMA model, the relative contribution c_{ij} ($j = a, b, \ldots, e$) of each of five interaction types: (a) bond stretching, (b) bond-angle bending, (c) improper angle distortions, (d) torsional and dihedral rotations, and (e) non-bonded interactions, as explained in Section 5.2.4. The contribution of interactions of type j to the spectrum density is then given by $c_{ij}g(\omega_i)$. The relative contributions of the five interaction types are shown in figure 5.4.

Of the five types of interaction only (d) and (e) are accessible with torsional dofs, since torsional and dihedral angle changes can affect neither bond lengths nor bond angles.
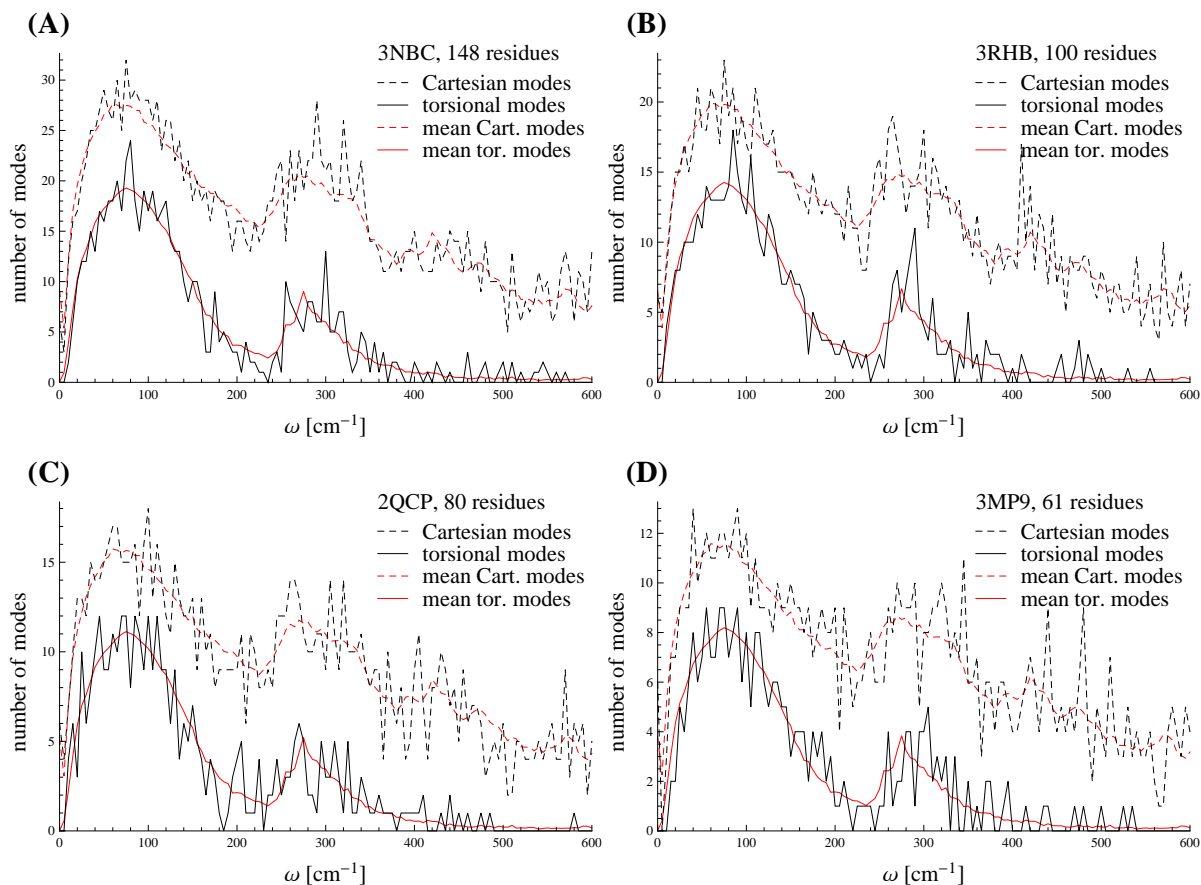
Figure 5.3 **Spectrum of vibrations for Cartesian vs. torsional dofs for four example proteins**: 3NBC (**A**), 3RHB (**B**), 2QCP (**C**), and 3MP9 (**D**). The averaged density of modes (over all the proteins in the dataset) for Cartesian and torsional dofs are represented respectively by the dashed and solid red curves.

Thus one expects those two types of interaction to account for all of the spectrum density observed with the restricted set of torsional dofs. Indeed, the dark gray- and light gray-shaded regions in figure 5.4, corresponding to the two types (d) and (e) respectively, bear a remarkable resemblance to the torsional spectra of figure 5.3.

Both the main and secondary peaks originate largely from torsional motions. The first peak is mainly due to torsional motions along the backbone, the ϕ or ψ rotations, since they involve the motions of larger masses and hence tend to oscillate at lower frequencies, while the second peak is probably more influenced by the torsional motions
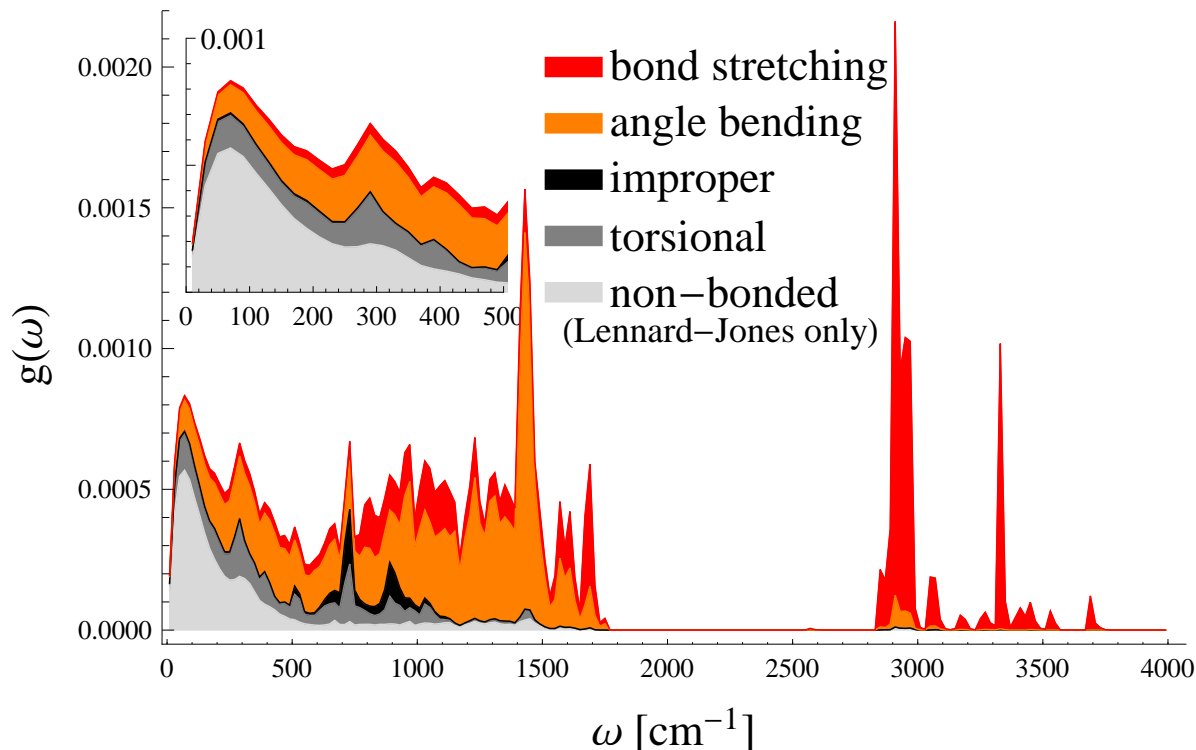
Figure 5.4 **Relative contribution of the various interaction terms to the vibrational spectrum.** Inset: The main and secondary peaks shown in more details.

of the side-chain rotamers, which is of higher frequency since the mass of the rotating component is smaller. This is in agreement with a previous study that suggested that the low frequency modes are contributed mainly by rigid body motions of the entire residues and side chain rotations [46].

As the frequency increases beyond about $500\,\mathrm{cm^{-1}}$ the bond stretching and bond angle interactions account for the lion's share of the spectrum density. The peaks at about $1{,}000 - 1{,}700\,\mathrm{cm^{-1}}$ are largely dominated by angle bending interactions, while those in the $3{,}000 - 4{,}000\,\mathrm{cm^{-1}}$ region arise mostly from bond stretching. Indeed, experimental results [83, 100] involving gas-phase infrared and Raman spectroscopy, though limited to small molecules such as N-methylacetamide (CH_3-CO-NH-CH_3) or alanine dipeptide, confirm the vibrations of bonds and angles at frequencies similar to the peaks we find, and have been even used for adjusting force-field parameters. The additional (universal)

information provided by the spectrum density $g(\omega)$ of globular proteins, combined with such experiments, could help better fine-tune the existing empirical potential energy functions (see, also, Section 5.3.3).

One can confirm the origin of the main and secondary peaks in yet another, perhaps more direct way. In figure 5.5 we plot the torsional dofs spectrum density of sbNMA under three different conditions: (i) The original sbNMA, in gray (NMA is shown in black, for reference), (ii) without the torsional and dihedral interaction terms (blue), and (iii) with those terms included, but without the non-bonded interaction terms (red). The elimination of the torsional interactions results indeed in the obliteration of the secondary peak, and the elimination of non-bonded interactions results in a big distortion of the main peak, in agreement with the conclusions of the foregoing analysis. Notice also that the correspondence between main peak and non-bonded interactions, and secondary peak and torsional interactions is not quite perfect: Elimination of torsional terms has some effect on the shape of the main peak as well, and elimination of the non-bonded interactions seems to have quite a dramatic effect not only on the main peak but also on the secondary peak. This agrees nicely with figure 5.4), where the shading indicates that both non-bonded interactions and torsional terms contribute to the main and secondary peaks, though to a different extent (the relative torsional contribution is larger for the secondary peak).

5.3.2 Vibrational Spectra for Different Protein Folds

Are there different vibrational spectra for proteins belonging to different classes of fold? It has long been known [73] that secondary elements such as α-helices, β-sheets, and turns exhibit different typical vibrational frequencies in the range below $\sim 100\,\mathrm{cm}^{-1}$. These differences should in principle show in the vibrational spectra of proteins belonging to different fold classes, though the typical fluctuations from one protein to the next (figure 5.2) would seem to impose a formidable obstacle to observing this phenomenon.
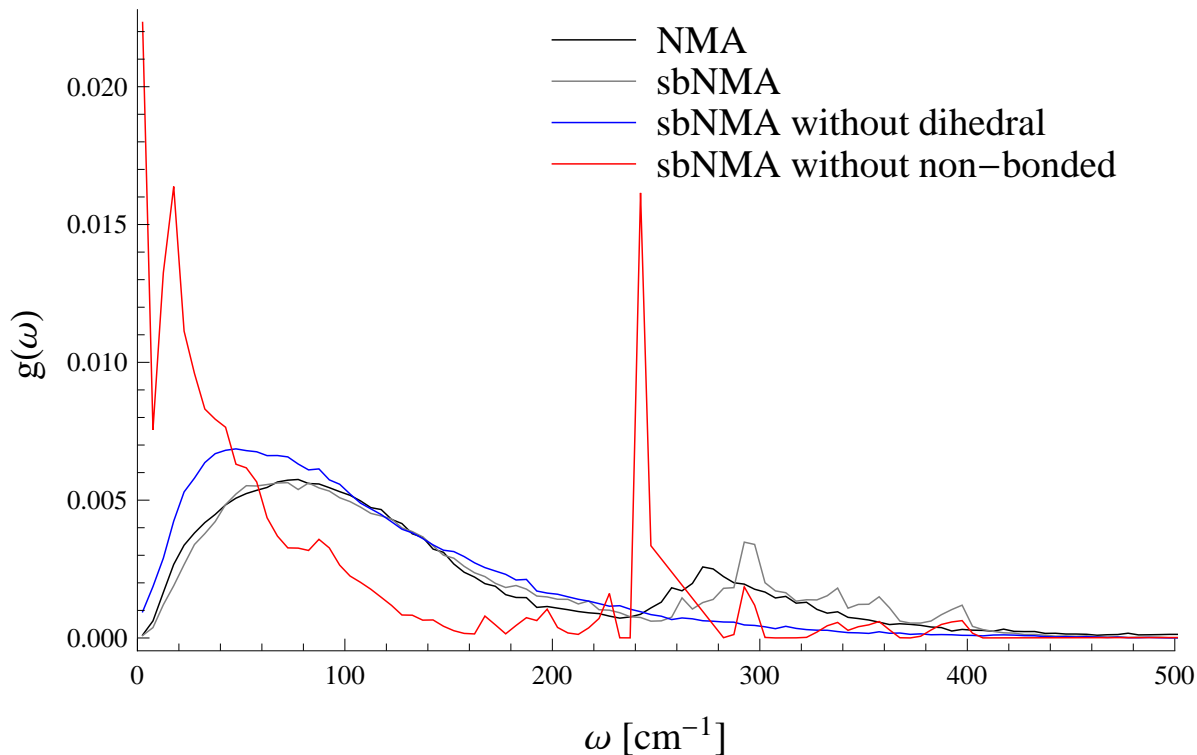
Figure 5.5 **The torsional-dofs spectrum of vibrations with and without various interaction terms.** The black line is the spectrum of NMA while the gray line that of sbNMA, a model that closely resembles NMA. When the torsional interaction term is removed from sbNMA's potential, the second peak disappears. The first peak disappears (or reshapes significantly) when non-bonded interaction term is removed.

Indeed, the spectra of any two proteins in our data set belonging to a different fold, an all-α protein and an all-β protein, say, are clearly within the typical fluctuations range. This result is in accord with experimental findings of Giraud et al., [37] who failed to observe any significant differences between the spectra of α-rich and β-rich proteins in their studies with ultrafast OHD-RIKES spectroscopy. With some care, however, one can tease out small but significant differences from our theoretical simulations.

Since the differences, if any, are expected in the low-frequency range, we carry our analysis in only torsional dofs, which capture the features of $g(\omega)$ in this range most cleanly. We have used the CATH catalog of protein structure classification [118] for identifying the fold type of most of the proteins in our dataset; the remainder of the

proteins were sorted out by manual inspection. We thus identified 42 alpha-proteins (27 by CATH and 15 manually), 37 beta-proteins (27 by CATH and 10 manually), and 56 α/β-proteins (38 by CATH and 18 manually). We then averaged the torsional vibrational frequency distributions $g(\omega)$ within each group. The three resulting curves are plotted in different colors in figure 5.6a (all-α in red, all-β in blue, and α/β in gray). Although the three curves are very similar to one another, systematic deviations can be clearly seen upon closer inspection: The main peak shifts progressively to the left, from all-α to α/β- to all-β proteins, and the opposite trend occurs at the farther slope of the secondary peak. That these shifts are systematic can be most clearly seen by randomly reassigning the proteins in the dataset to three groups of corresponding sizes, and recomputing the averages in the random groups: The three curves now intertwine seemingly at random and the systematic deviations disappear (figure 5.6a; inset).
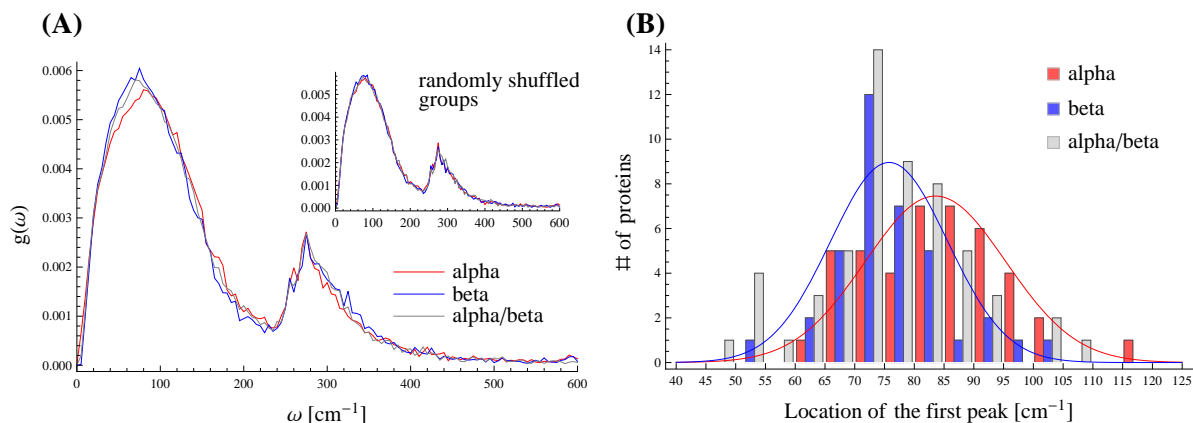
Figure 5.6 **Vibrational spectra and statistics of the main peak location for different protein folds.** (**A**) Vibrational spectra for different protein folds: 42 all-α proteins (red), 37 all-β proteins (blue), and 56 α/β-proteins (gray). Notice the systematic shifts in the main peak and the far slope of the secondary peaks. Inset: Plots of $g(\omega)$ when equivalent number of proteins are assigned *randomly* to the three groups as in the main plot. The systematic deviations disappear. (**B**) Statistics of the location of the main peak for the three groups of protein (color coding same as (**A**)) roughly fits a Gaussian distribution (curves shown for all-α and all-β proteins only) and clearly demonstrate the differences between the various types.

A better view of the systematic deviations is provided by the statistics of the location of the main peak, where the differences are most pronounced. In figure 5.6b we show histograms for the locations of the main peak in the three groups (same color coding as before), along with best gaussian fits to the histograms for the two extreme groups (all-α and all-β). Despite the large overlap the shifts are quite apparent: The peak for all-α proteins is at about $85\,\mathrm{cm}^{-1}$, while that of beta-proteins is at $70 - 75\,\mathrm{cm}^{-1}$. The fact that α-helices are more compactly structured than β-sheets may account for the stiffer (higher frequency) peak for all-α proteins.

In the high-frequency range (and using all dofs), differences between α- and β-rich proteins are observed near the range of amide vibration frequencies that have been extensively used to distinguish between α-helix and β-sheet in protein infrared (IR) spectroscopy [20, 21, 35, 39, 62, 93, 124, 152]. Figure 5.7(A) shows the vibrational spectra in Cartesian dofs of all-α proteins in a red curve, all-β proteins in a blue curve, α/β-proteins in a gray curve, and three amide regions (I, II, and III) in light gray bands. Figure 5.7(B) shows an enlarged view of the three amide regions with additional arrows that point at peaks in the frequency curves. Though the overall vibrational spectrum is universal for all globular proteins, Figure 5.7 shows that there are small, yet noticeable local differences in the three amide regions between all-α and all-β proteins. Remarkably, in each amide region, not only the general locations of the peaks but also the magnitudes of the shifts between all-α and all-β proteins as predicted by our method match well with results from infrared spectroscopy [20, 21, 35, 39, 62, 93, 124, 152].

5.3.3 Using the Vibrational Spectrum to Assess and Improve Theoretical Approaches

While $g(\omega)$ is universal for a given atomic potential function, one would expect to see different curves for different formulations and parameterizations of the potential. There is, however, only one reality and the "true" shape of $g(\omega)$ can only be decided
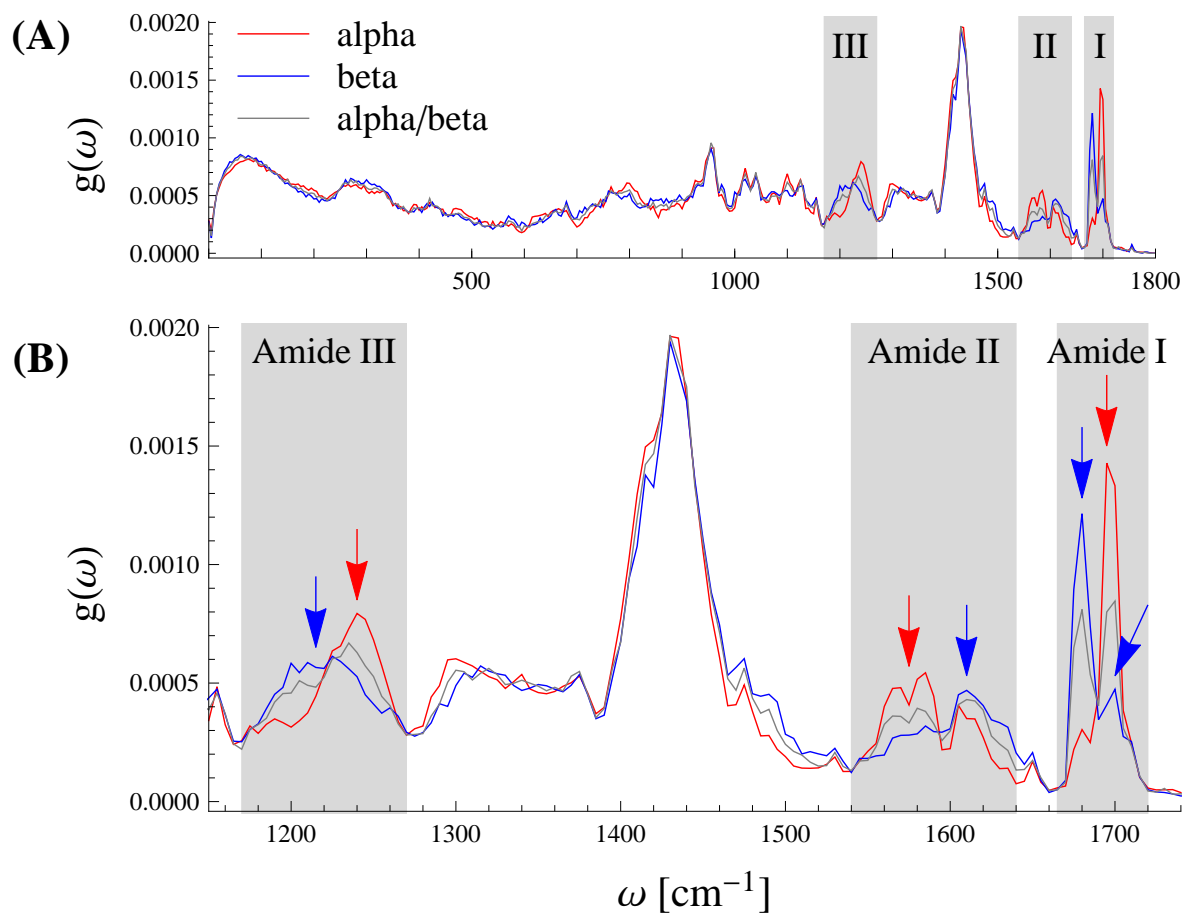
Figure 5.7 **Vibrational spectra of amide groups for different protein folds.** (**A**) shows the vibrational spectra for different protein folds in the range of amide vibration frequencies: 42 all-α proteins (red), 37 all-β proteins (blue), and 56 α/β-proteins (gray). The frequency range of amide I, II, and III are highlighted in light gray bands. (**B**) zooms in on the three amide regions. Arrows point out peaks of frequency curves of all-α and all-β proteins in the amide regions.

by experiment. We show below how this effect can help one choose between different potential formulations. The vibrational spectrum is also quite sensitive to different levels of approximation (simplified models/potentials, restricted dofs) and this can be exploited to assess the accuracy of various simplified models and help fine-tune their parameters.

Sensitivity to different empirical potentials. In order to demonstrate the sensitivity of $g(\omega)$ to the potential function one uses for the analysis, we compute the vibrational spectrum in two different ways: (a) With the atomic detailed CHARMM22 potential [83] (Figure 5.8, in solid line), and (b) with the same potential but where the Van der Waals radii of the various atoms is *replaced* with the values from the L79 potential function [71, 73], and with a uniform torsional spring constant, $K_\phi = 0.1\,\mathrm{Kcal/mol/rad}^2$ (in dashed line). The shift of the main peak to smaller frequencies, in the latter case, and the overall shape of the curve is quite in agreement with the spectra of proteins obtained with the L79 potential from the start [142]. (For simplicity, we performed the analysis in the restricted set of torsional dofs, which suffices for our purposes.) This demonstrates quite cleanly the influence of different parameter values in different (or the same) potential function(s). It also lends further support to our claim that the location of the main peak has to do largely with non-bonded interactions (see Section 5.3.1).

Which potential function gives a better parameterization of the Van der Waals radii? This, and similar issues, can only be decided by experiment. Experimental spectra are hard to obtain, since their extraction often requires various uncontrolled assumptions and approximations, and they remain a challenge. One recent experimental study, employing ultrafast OHD-RIKES spectroscopy [37] in the low-frequency spectrum, finds a main peak at about $80\,\mathrm{cm}^{-1}$ and a broader, diffuse peak at around $300\,\mathrm{cm}^{-1}$ (locations marked in the figure by vertical dotted lines). This seems to favor the CHARM22 potential formulation over that of L79. Agreement between theory and experiment remains an elusive goal, though: We hope that theoretically obtained $g(\omega)$'s would spur experimental interest in
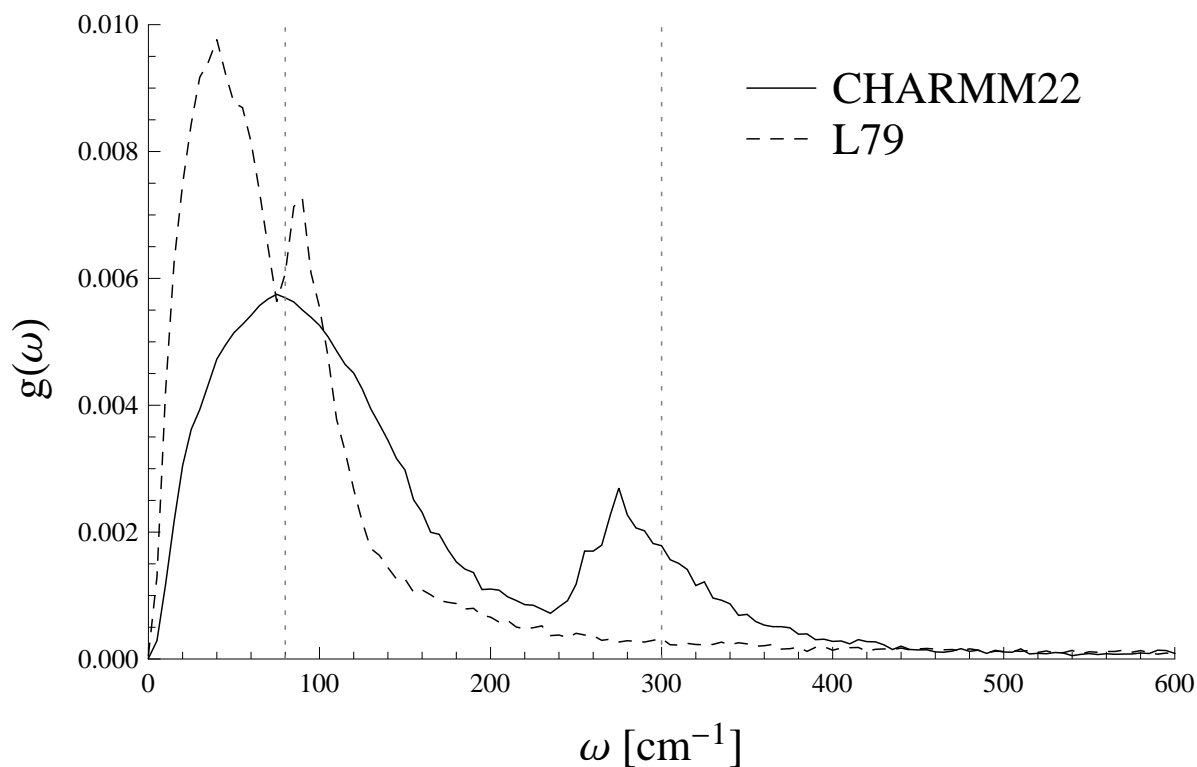
Figure 5.8 **Vibrational spectra obtained with the CHARMM22 potential (solid) and the approximated L79 potential (dashed).** The two vertical dotted lines mark the locations of the first and second peaks observed experimentally [37], which favor the CHARMM22 potential formulation over that of L79.

searching to confirm the various peaks, and conversely, that the various peaks observed by experiments would help fine-tune the theoretical potential functions, thus bringing greater understanding.

Sensitivity to different levels of approximation. The spectrum distribution $g(\omega)$ is also sensitive to various levels of approximation that are used in simplified NMA models. We have already seen one obvious example of this in Section 5.3.1, in the different spectra one obtains with the restricted set of torsional dofs vs. the full complement of Cartesian dofs. We now examine the effect of other common simplifications. For simplicity we use once again torsional dofs only, since they suffice to capture the differences.

In figure 5.9 we plot $g(\omega)$ as obtained from NMA with the full atomic CHARMM22 potential (black line), along with decreasing levels of approximation: sbNMA (in blue), ssNMA (red), and ANM (gray). Since NMA requires minimization of the CHARMM22 potential function we use the minimized structures also in all the approximate techniques, so as to obtain a fair comparison. (The effect of the starting configuration is discussed separately, in the next section.) The quality of the approximate approaches is clearly reflected by the general trend: the better the approximation the better the fit of its $g(\omega)$ to that of NMA.
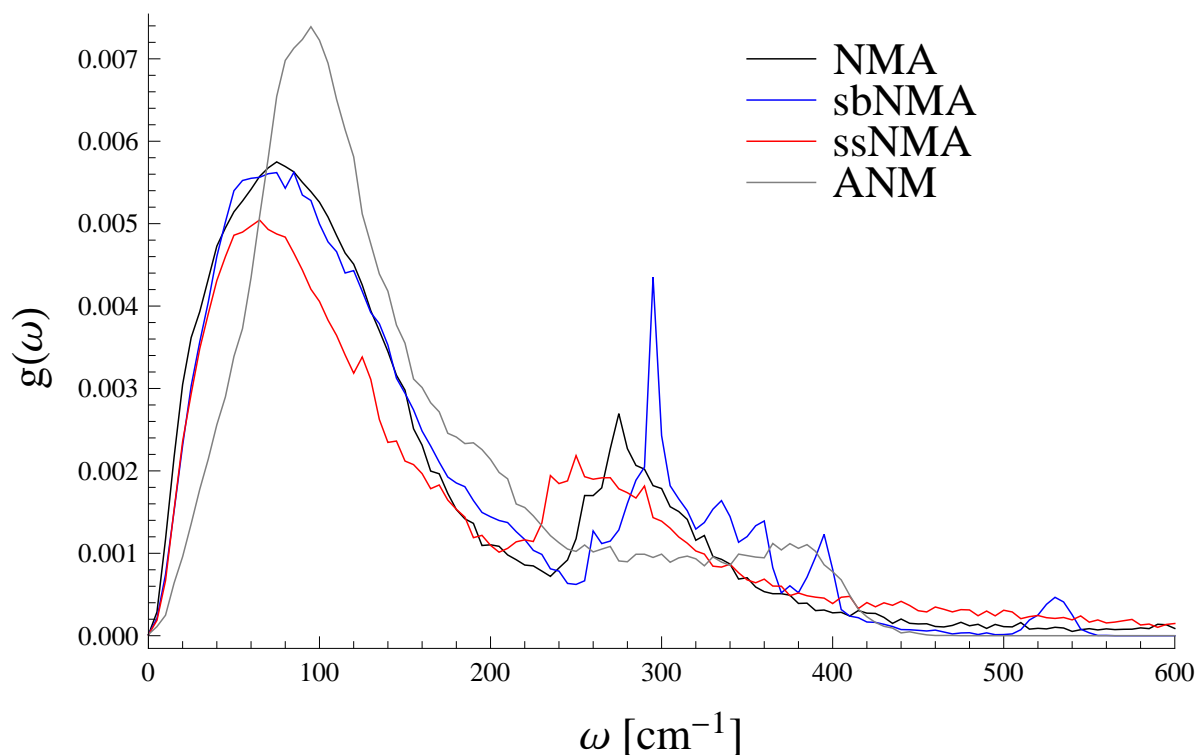
Figure 5.9 **The vibrational spectra obtained by the original NMA and various simplified models.** Vibrational spectra provide a critical assessment of the quality of the simplified models.

The best approximation, sbNMA, also yields the best fit: The main peak is reproduced very faithfully and significant differences show only in the secondary peak, which seems shifted somewhat towards higher frequencies. This makes sense, in view of the

fact that the torsional terms in the potential (responsible for the secondary peak) are more heavily approximated in this technique than the non-bonded interaction terms (that shape the main peak): The spring constant for the torsional terms is fixed at a one value, regardless of the amount of rotation, as opposed to the non-bonded spring constants whose values are a function of the distance between the interacting atoms. The near perfect reproduction of the first peak confirms that electrostatic interactions, which are neglected in sbNMA, indeed contribute significantly less to the normal mode motions than the van der Waals interactions [89]. The ssNMA, that uses a smaller set of parameters than sbNMA, results in further deterioration of $g(\omega)$. Finally, the coarsest approximation, ANM, with only one universal spring constant for all interactions, yields the worst fit to the $g(\omega)$ of the original NMA.

One can use the $g(\omega)$ to improve the various approximations. As an example, the maximum of the secondary peak of sbNMA, at about $290\,\mathrm{cm}^{-1}$, could be shifted to the NMA's maximum at $270\,\mathrm{cm}^{-1}$ by softening the torsional spring constants by a factor of $\sqrt{290/270} \approx 0.87$ (since ω is proportional to the square root of the spring constant, and the secondary peak is mostly determined by torsional energy terms). Indeed, this ploy succeeds in effecting the desired shift (results not shown), though the quality of the main peak somewhat deteriorates. One could in principle fine-tune the various sbNMA parameters to achieve an *optimal* fit to $g(\omega)$ of the original NMA.

It is less clear how to improve the fit of the main peak with ssNMA. Juxtaposing the results of sbNMA and ssNMA, it becomes apparent that detailed spring constants for non-bonded interactions, dependent on atom type and distance, as in sbNMA [89] (and ATMAN [142]), are crucial to a faithful reproduction of the main peak. Likewise, the torsional spring constants are essential to a successful reproduction of the secondary peak. For the simplest approximation, ANM, varying the single available spring constant would only result in an overall scaling of the frequencies, or the ω's. In other words, the whole curve would compress or dilate uniformly, as the spring constant is softened or

strengthened. Thus, softening the spring constant might achieve a better fit of the maximum location of the main peak, that would then shift to the left, however, this would also result in further narrowing of the peak (which seems already too narrow in comparison to NMA). In short, it seems rather impossible to achieve a satisfying fit with such a simple approximation.

In closing this section, we note that the quality of the various approximations has formerly been assessed by comparing individual modes, rather than the distribution of their frequencies, as suggested here. For example, modes comparison has been used by Na and Song [91] to conclude that a good quality approximation, such as sbNMA or ssNMA, requires geometric terms that maintain proper bond lengths and bond angles, distance-dependent van der Waals based spring constants for non-bonded interactions, plus torsional spring constants, as a minimal set of parameters. Interestingly, Tirion and ben-Avraham [142] reached the same conclusion in their development of the closely related ATMAN approach, but using $g(\omega)$ as a guide. Clearly the two techniques, comparing individual modes and comparing frequency distributions, have their own problems and merits and are complementary to one another, enriching our chest of theoretical tools.

5.3.4 How Input Structures Affect the Vibrational Spectrum

A great advantage of simplified potential functions of the type introduced by Tirion [139] in 1996, such as ANM, sbNMA, ATMAN, etc., is that they require no minimization: The potential function is at a minimum at the outset, regardless of the protein's given configuration. Thus, using such potentials one can obtain the normal modes of a protein for any number of different *starting configurations*, or *input structures*. It is well known that as long as the input structures do not differ by a large amount, the first few slowest modes remain quite unchanged (for a recent detailed, quantitative study, see Na and Song [92]). We here show, however, that the overall distribution of mode frequencies, $g(\omega)$, *is* affected by different input structures. The question then arises of what is the *proper* input structure for a normal mode analysis.

To demonstrate the effect, in figure 5.10 we show the spectra obtained with sbNMA with two different input structures: (i) The PDB configurations of the proteins, and (ii) the configurations obtained by minimizing the CHARMM22 potential energy. Recall that the sbNMA parameters are determined from the CHARMM22 potential, but being a Tirion-type potential it allows us to obtain spectra with the two different input sets. (In contrast, a full detailed potential such as CHARMM22 must first be minimized and is limited to only the minimized structures.) For the minimized structures, sbNMA and NMA using CHARMM22 obtain very similar spectra, as we have already demonstrated in Section 5.3.3 (the curve for CHARMM22 is included in the figure, as a reminder). The point of this plot is the significant differences between the spectra of (i) and (ii): For example, the maximum of the main peak, located at about $80\,\mathrm{cm}^{-1}$ for the minimized structures, shifts to about $50\,\mathrm{cm}^{-1}$ for the (non-minimized) PDB inputs. On one hand, universality as shown in figure 5.2 indicates the spectrum is not protein-specific, arising from structural properties common to globular proteins in general. On the other hand, results from figure 5.10 imply that the spectrum depends quite strongly on whether the structures are minimized or not. (The spectra obtained for the same 135 structures but without energy minimization are universal as well. See Figure S1 in Supplemental Information). The two are not contradictory to each other. Their difference, specifically the shift of the first peak, can be understood as follows. First, recall that the spectrum at the low frequency end is contributed mainly by non-bonded interactions, especially the van der Waals interactions, which are sensitive to inter-atomic distances. Second, energy minimization causes the structures to relax according to the force field. The structure change inevitably alters the inter-atomic distances and consequently the van der Waals terms in the potential function. The change in the latter in turn mostly determines the shift in the location of the first peak. At the high-frequency end, however, there is little difference between spectra of minimized structures and non-minimized structures (see Figure S2 in Supplemental Information).
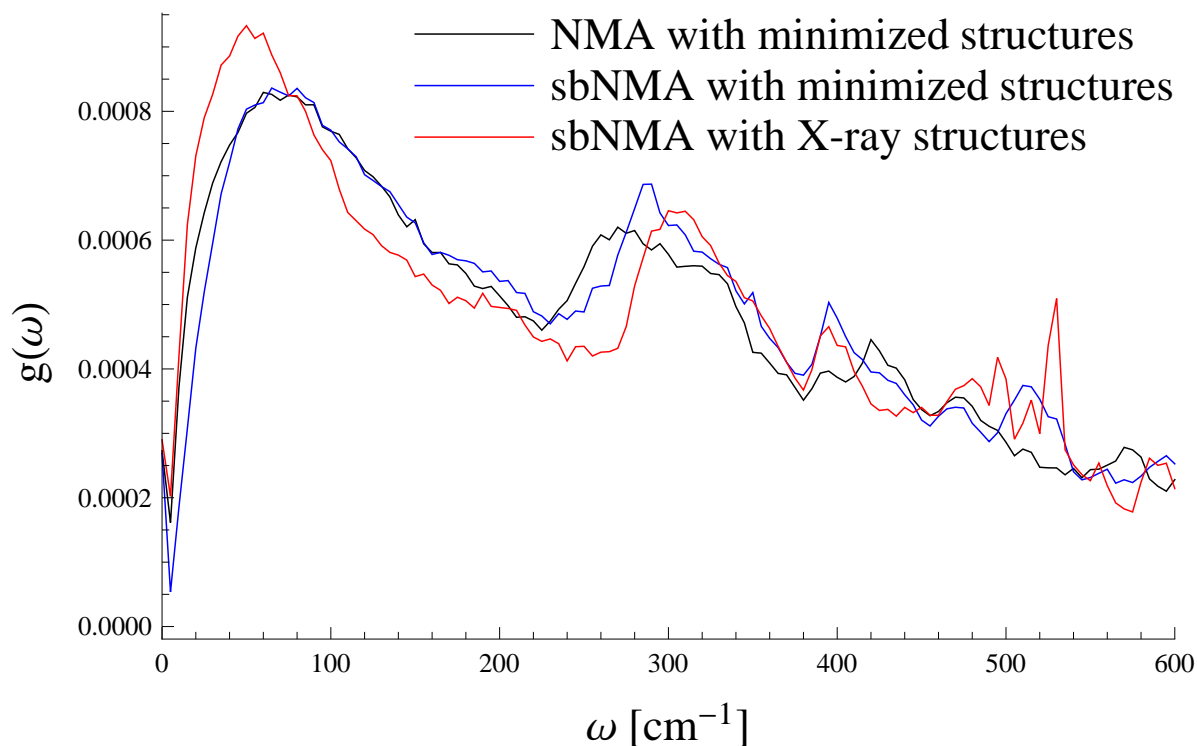
Figure 5.10 **Dependence of the vibrational spectrum on input structures.** The spectrum of sbNMA varies significantly (for example, in the location of the main peak) when different input structures are used. The spectrum of NMA is shown in the background as a reference.

So, what is the proper starting configuration for a normal mode analysis? Ideally, one would like to use the equilibrium configuration of a *single* protein in its natural state, but the PDB configurations are obtained from *crystal* structures and it is not quite clear whether these two are the same. Taking into account that the $g(\omega)$ from the minimized structures agree better with the experimental results of Giraud et al., [37] (main peak at about $80\,\mathrm{cm}^{-1}$), there are two possibilities: If the CHARM22 potential is to be trusted then this suggests that the PDB crystal structure is different from a single protein's equilibrium structure; conversely, if the PDB crystal structure is the same as the protein's equilibrium structure, this suggest that CHARMM22 potential is not quite right and its various parameters need to be adjusted. On the one hand, there is enough reason

to suspect that the crystal packing distorts equilibrium configurations. For example, the configuration of G-actin, as obtained from its crystal form, needs to be distorted significantly in order to fit with the structure derived in the different packing of F-actin filaments [143]. On the other hand, experimental spectra are difficult to interpret and the results are far from uniform. It might very well be that the PDB structures ought to be trusted more than the CHARMM22 parameters. If the latter is the case, normal modes analysis must proceed from undistorted PDB structures (using a Tirion-type potential such as sbNMA, or ATMAN), as argued by Na and Song [92]. The actual answer seems important, either way.

5.4   Conclusion and Discussion

In this work, we have shown that the density of modes in the vibrational spectrum of globular proteins is universal: The density of modes of different globular proteins, when properly normalized, tend to aggregate around one universal curve. We find this universality to be true not only for the low frequency range and for the restricted set of torsional dofs, as observed in earlier studies [10, 141], but for the whole frequency spectrum and for the full complement of dofs available to the proteins' atoms. This surprising result is highly significant, in that it implies that the universal patterns of the spectrum, its turns and peaks, are not protein-specific but rather force-field specific, arising from structural properties and inter-atomic interactions common to globular proteins in general.

The universality of the spectrum density and the fact that the actual $g(\omega)$ curve depends on the empirical potential used for the normal modes analysis calls for a serious two-way dialogue between theory and experiment: Experimental spectra of proteins could now guide the fine tuning of theoretical empirical potentials, and the various features and peaks observed in theoretical studies — being universal, and hence now rising in importance — would hopefully spur experimental confirmation.

The characterization of the typical *fluctuations* from the average $g(\omega)$ paves the way to the interpretation of *salient features* in the spectra of *individual* proteins, thus promising to fulfill a decades-old goal of continued work on normal mode analysis. That this is possible, in principle, was clearly demonstrated by the discernible differences in the spectra of proteins of different fold families (see Section 5.3.2).

The universality of $g(\omega)$ also provides us with an exquisite tool for the assessment of various approximation approaches. We have thus seen that in order to obtain a faithful resemblance of the NMA spectrum, an approximate technique must include, at the very least, spring constants for the non-bonded interactions that are atom-type dependent and distance-dependent, and include energy terms for changes of torsional and dihedral dofs, as done in sbNMA [89], or ATMAN [142]. This level of accuracy is indispensable for highly sensitive tasks such as finding the different vibrations of closely related crystal isoforms of a protein [140] and can improve, in general, the results of many normal mode-based studies, such as identifying folding cores [7] or hot-spot residues [99]. In the opposite extreme, the simplest ANM approximation, while useful for predicting the general shape of the slowest modes, cannot simultaneously account for both the location and the width of even the main peak of $g(\omega)$.

What is the source of the universality we observe in $g(\omega)$? For low frequency modes, it has been argued that the coherent motions of large domains of a protein involve mainly interactions between atoms of adjacent domain surfaces. Those interactions average out in the same way, for all proteins, simply because the number of interacting pairs is large and one can then invoke the central limit theorem to describe their combined effect. This argument does not work, however, for higher frequencies, where the coherence length of the modes is tiny and the moving components involve but a few atoms. One possibility is that, due to the very different stiffnesses of torsional changes, angle bending, and bond stretching, the three elements dominate different parts of the spectrum: torsional terms in the low frequency range of $0 - 500\,\mathrm{cm}^{-1}$, angle bending in the intermediate range

of $500 - 2000\,\mathrm{cm}^{-1}$, and bond stretching in the high-frequency range, above $2000\,\mathrm{cm}^{-1}$ (see figure 5.4). Different proteins may have similar percentages of the various angle and bond types, explaining the universality in the mid- and high-frequency range. Only future work could unravel the full causes for the universality of $g(\omega)$, and whether the distinct stiffness magnitude for the three types of interactions play a decisive role in it.

Among the many other interesting open questions left by this study we mention the precise relation between Cartesian and torsional dofs. The densities of the two spectra look very similar, but there is an excess of modes with Cartesian dofs (see figure 5.3). Quantitatively speaking, it seems like that there are about 30% more modes in Cartesian space, at the low frequency end. Why? A possible explanation is that some Cartesian modes represent a mix of torsional and non-torsional motions (such as bond bending motions), as indicted by figure 5.4, but still oscillate at low frequencies. Whether this is the case, as well as a detailed comparison of the slow modes themselves in Cartesian and torsional dofs, is left for future work.

Is the universality of $g(\omega)$ true only for globular proteins or does it encompass other types of proteins? It would be interesting if the same spectrum density, or features of it, resulted also for non-globular proteins. How would various ligands affect the spectrum density? In a recent work by Wynne and co-workers [144], the spectrum of vibrations was found to be modified after ligand binding (and the range of affected modes was postulated to play an important role in the ligand binding process as well). Since proteins of different fold families exhibit discernible variations in the spectrum density, it is plausible that other types of proteins, and ligands, would also affect the spectrum. Is protein size more relevant than suggested by our results? In the present study, we were limited to small- to medium-size proteins, for the ease of all-atom NMA computations. We expect an inverse correlation between protein size and deviations from universality — the smaller the protein, the larger the deviation — but we were unable to see that in our data. A larger study is needed to prove or disprove this notion. Ultimately, a full characterization

of the fluctuations is important in order to correctly identify outlier proteins. It would be interesting to study a few outlier proteins whose spectra truly differs from the average, beyond the expected fluctuations.

The question of the proper starting conformation for a normal mode analysis remains a genuine puzzle (Section 5.3.4). Given an experimentally determined structure, should we first minimize it before performing NMA? We have seen that the location of the main peak shifts significantly, from $80\,\mathrm{cm}^{-1}$ for potential-minimized structures, to about $50\,\mathrm{cm}^{-1}$ for the original crystal structures given by the PDB files. We have postulated that part of the effect is due to the conformational distortions undergone by the proteins in the crystal packing. Future spectral studies based on nuclear magnetic resonance (NMR) structures that are not influenced by crystal packing might shed light on this issue. Whatever the answer, the fact that the input configuration makes a big difference in the outcome emphasizes the need for further development of good Tirion-type potential energy formulations, such as sbNMA and ATMAN, that are minimized at the outset.

An issue that we have left untouched in this study is the question of the precise nature of the spectrum of vibrations in the low-frequency range. In early work [10, 26] it was suggested that the low-frequency spectrum has an anomalous *spectral* dimension of $d_s \leq 2$ (instead of 3, as expected for a three-dimensional crystal): This implies that the low-frequency spectrum behaves in power-law fashion, $g(\omega) \sim \omega^{d_s-1}$. Later studies [19, 105] found a weak dependence of the spectral (and fractal) dimension with protein size. The power-law (and anomalous dimensions) interpretation has been contested by Etchegoin and Nöllmann [29, 96], who maintained that the low-frequency spectrum rather fits a log-normal distribution and is better explained by the analogous behavior in glasses (their analysis, however, relied on spectra obtained with only torsional dofs). We have not attempted to delve into this argument, mostly due to the limited range of sizes of our proteins. A future study, involving a larger dataset and heavier proteins, and using the NMA method with an all-atom potential and the full complement of dofs, as done in the present work, will shed much needed light on this interesting problem.

Acknowledgement

CHAPTER 6.   QUANTITATIVE DELINEATION OF HOW BREATHING MOTIONS OPEN LIGAND MIGRATION CHANNELS IN MYOGLOBIN AND ITS MUTANTS

Hyuntae Na[23] and Guang Song[234]

Abstract

Ligand migration and binding are central to the biological functions of many proteins such as myoglobin (Mb) and it is widely thought that protein breathing motions open up ligand channels dynamically. However, how a protein exerts its control over the opening and closing of these channels through its intrinsic dynamics is not fully understood. Specifically, a quantitative delineation of the breathing motions that are needed to open ligand channels is lacking. In this work, we present and apply a novel normal mode-based method to quantitatively delineate what and how breathing motions open ligand migration channels in myoglobin and its mutants. The motivation behind this work springs from the observation that normal mode motions are closely linked to the breathing motions that are thought to open ligand migration channels. In addition, the method

---

[1]This chapter is reprinted with permission of *Proteins* 2015, 83(4), 757–770.

[2]Graduate student and Associate Professor, respectively, Department of Computer Science, Iowa State University.

[3]Primary researchers and authors.

[4]Author for correspondence.

provides a direct and detailed depiction of the motions of each and every residue that lines a channel and can identify key residues that play a dominating role in regulating the channel. The all-atom model and the full force-field employed in the method provide a realistic energetics on the work cost required to open a channel, and as a result, the method can be used to efficiently study the effects of mutations on ligand migration channels and on ligand entry rates. Our results on myoglobin and its mutants are in excellent agreement with MD simulation results and experimentally determined ligand entry rates.

6.1   Introduction

Proteins are one of the fundamental functional units in cells. It is fascinating to see how proteins exercise precise controls in different functions. Among these, a particular feat is seen in how a protein regulates the ins and outs of small ligands through its matrix. This process is of paramount importance in the proper function of many proteins, such as many enzymes whose efficient catalysis relies directly on the uptake of O_2 or other gaseous molecules. However, the ligand update mechanism employed by these proteins is poorly understood. As there are usually no open channels for ligands to enter into or exit from the interior of a host protein at the static structure, protein dynamics has been often thought to open the channels dynamically but it is not fully understood how it does so in many proteins.

Experimentally, flash photolysis and mutagenesis studies were often employed to study the recombination kinetics in heme proteins and to identify ligand migration channels. [5, 33, 48, 98, 115, 116] For example, site-directed mutagenesis of 27 residues was used to map out the ligand pathways. [116] Random mutagenesis studies conducted by Huang and Boxer [48] revealed that single mutations of several clusters of residues far away from the pathways profoundly affected the ligand-binding kinetics. Time-resolved

X-ray crystallography [122, 123] literally allowed one to track a photo-disassociated ligand as it migrated through the protein, as well as structure relaxation, over a broad range of timescales, from a few nanoseconds to as long as a few milliseconds. [15, 16, 113, 114] It provided direct insight into the gating role of the correlated motions between the backbone and side-chains in ligand migration. [114]

Computationally, molecular dynamics (MD) has been extensively applied to study ligand migration since late 70's. [1, 13, 14, 22, 27, 49, 97] A recent work by Ruscio et al. [109] obtained a cumulative 7-μs simulation on myoglobin and identified many different trajectories and entry/exit portals on the protein surface. The advantage of using MD is that one can observe actual events of ligand passing in and out of the protein matrix. Its drawback is that it takes extensive time to run the simulations and the process is stochastic. As a result, the less frequently traveled channels are difficult to identify. Implicit ligand sampling (ILS), [23] an innovative approach developed recently, on the other hand, computes the potential of mean force (PMF) corresponding to the placement of a ligand everywhere inside a protein. ILS provides a complete three-dimension map that identifies the potential cavity sites and the pathways connecting them, some of which are in regions that are difficult to probe experimentally.

Proteins have intrinsic dynamical behaviors that contribute directly to their functions. [42] Most of these dynamical behaviors are captured by its normal mode motions. [17, 38, 72] For many proteins such as myoglobin, these motions, often called breathing motions, open the channels. However, a quantitative delineation of exactly what and how breathing motions open a given channel is lacking. In this work we present a novel method that is able to determine exactly what combinations of the intrinsic normal modes may be used to open a channel. Given a structure of the protein to be studied, the method has two key steps. The first step is to apply Voronoi diagram to estimate where the putative ligand migration channels are. This efficient step ($O(n \log n)$ time) can quickly identify putative channels. [78] The second step is to gradually stretch

open each channel by identifying and applying the best combination of normal modes (see Methods section for details). The product of this stretching process is a sequence of conformation changes that eventually lead to the full opening of the channel.

The strengths and weaknesses of our method are summarized in Table 6.1 in comparison with two commonly used methods for channel mapping and identification: MD and ILS. [23] Compared to ILS, our method is superior in that i) it provides a quantitative description of what combinations of normal modes are needed to open a channel, and ii) it identifies the key residues whose motions contribute the most in opening the channels. Compared to MD, the advantage of our method is that i) it does not require simulations, and ii) the conformation changes needed to open a channel are fully about the channel's opening process and are not tangled with the background thermal fluctuations of the protein. This separation may be critical in identifying key motions and key residues that regulate the channels. Lastly, ILS and MD are both simulation-based and suffer an intrinsic limitation that simulation-based approaches share: the narrow range of sampling. Thus, "the effects of slow conformational and allosteric changes will not be observed during the course of the simulation. Therefore, there is no guarantee that all biologically relevant pathways will be discovered through simulation." [23] Our method is normal-mode based and can identify channels that open rarely and require slow conformation changes. The weakness of our method is that it does not consider the interaction between the ligand and the protein. It is thus more suitable for studying the migration of small ligands.

Our proposed method provides a direct and detailed depiction of the motions of each and every residue that lines a channel and thus allows one to identify key residues that play a dominating role in regulating the channel. The all-atom model and the full force-field employed in the method provide a realistic energetics on the work cost required to open a channel, and as a result, the method presents itself as an efficient computational tool for studying the effects of mutations on ligand migration channels and on ligand

Table 6.1 Comparison between our method and two other well-known computational methods.

	our method	ILS[a]	MD[b]
Method basis (require simulations?)	normal modes (no)	MD (yes)	MD (yes)
Completeness in mapping ligand channels	close to complete	close to complete	not complete
Transition pathways that open the channels	Yes Yes	No No	Yes Yes
Identify normal modes contribute the most	Yes Yes	No No	No No
Running time	short to medium	short to medium	long
Ligand size	small	small	small or large
Channel prediction quality	estimate	estimate	more realistic
Ligand-protein interactions	not considered	not considered	considered

The table shows the comparison between our method and the two other well-known computational methods for ligand channel mapping and identification.
[a]implicit ligand sampling (ILS) [23];
[b]molecular dynamics (MD).

entry rates. Our results on myoglobin and its mutants are in excellent agreement with MD simulation results and experimentally determined ligand entry rates.

6.2 Methods

To quantitatively determine the most favorable breathing motions that gradually open up a ligand channel, we first define the constraints on the breathing motions. These constraints serve to guarantee that the radius of the channel continually increases by a small amount at each iteration until the channel is fully opened. We then define the criterion for selecting the optimal breathing motion among those that satisfy the constraints.

6.2.1 Constraints Needed for Breathing Motions that Gradually Open a Channel

6.2.1.1 Definitions

First let us define a channel and the radius of a channel in mathematical terms. As in [78], given an input protein structure, we first compute its Voronoi diagram. Protein cavities are denoted by Voronoi vertices that are inside the protein and have large enough clearance. If we think a channel as a sequence of cylindrical pipes connecting together, the axis of the channel is represented by consecutive Voronoi edges, or line segments, that connect an internal cavity with the solvent. A channel may have different clearances at different segments of the channel. According to Voronoi diagram computation, each line segment, or Voronoi edge, matches to three atoms that have equidistance to the edge. This distance represents the minimum clearance at this segment of the channel. As a channel is composed of a sequence of channel segments, the channel radius or channel clearance is defined as the smallest clearance of all channel segments. The location where the clearance is the smallest is the bottleneck of the channel. Let $l_1, ..., l_k$ be the axes of the consecutive segments of a channel, a_i, b_i, c_i be the three atoms corresponding to Voronoi edge l_i, and r_i be the circumradius of triangle $\triangle a_i b_i c_i$, $1 \leq i \leq k$ (see Figure 6.1). The radius of the channel, denoted by a capital R, is the smallest radius of all the circumradii, i.e., $R = \min(r_1, ..., r_k)$.

6.2.1.2 The Derivative of Channel Clearance with Respect to Normal Modes.

Denote by a, b, and c the three atoms of channel segment l, and by $\mathbf{p}_i$ the Cartesian coordinate of atom i. The circumradius $r_{a,b,c}$ of triangle $\triangle abc$ is determined using the trigonometry:

$$r_{a,b,c} = \frac{\|\mathbf{p}_{a,b}\| \cdot \|\mathbf{p}_{b,c}\| \cdot \|\mathbf{p}_{c,a}\|}{2\|\mathbf{p}_{a,b} \times \mathbf{p}_{b,c}\|}, \tag{6.1}$$
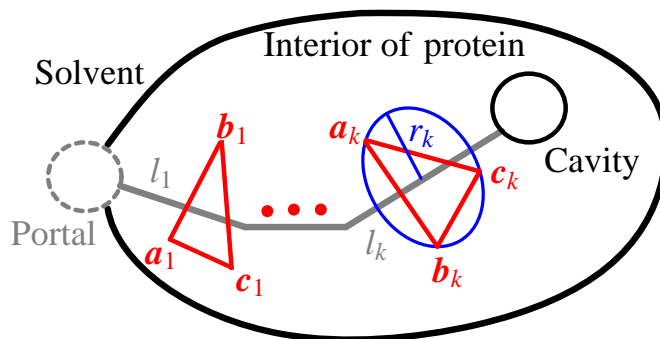
Figure 6.1 **Illustration of a channel.** The channel between the solvent and an internal cavity is represented by a series of line segments, $l_1, ..., l_k$, or Voronoi edges. Each channel/line segment (or Voronoi edge) l_i matches to three atoms, a_i, b_i, c_i, where $1 \leq i \leq k$. The atoms represent the lining atoms of the channel and define the channel clearance, which is the circumradius of the triangle formed by the three atoms.

where $\mathbf{p}_{a,b} = \mathbf{p}_a - \mathbf{p}_b$, $\|\mathbf{p}\|$ is the norm of vector $\mathbf{p}$, and $\mathbf{a} \times \mathbf{b}$ is the cross product of vectors $\mathbf{a}$ and $\mathbf{b}$.

Let $\mathbf{w}_j$ be the jth normal mode of the protein, and $\mathbf{w}_{ja}$ its component for atom a. For an instantaneous movement of $t_j \mathbf{w}_j$, where t_j is a small scalar value, the new position of the atom a can be written as $\mathbf{p}_a + t_j \mathbf{w}_{ja}$. In a similar manner, the new positions of atoms b and c can be obtained. Now, let $\mathrm{r}(t_j)$ be the circumradius (see Eq. (6.1)) of triangle $\triangle abc$, as a function of t_j. The first derivative of radius $\mathrm{r}(t_j)$ with respect to mode $\mathbf{w}_j$ can be written as:

$$
\begin{aligned}
\dot{r}_{\mathbf{w}_j} &= \frac{\partial \mathrm{r}(t_j)}{\partial t_j} = \frac{\partial \mathrm{r}(t_j)}{\partial \mathrm{r}(t_j)^2} \cdot \frac{\partial \mathrm{r}(t_j)^2}{\partial t_j} \\
&= \frac{1}{4 r_{a,b,c}} \cdot \frac{f(a,b,c) + f(c,a,b) + f(b,c,a)}{\|\mathbf{p}_{ab} \times \mathbf{p}_{bc}\|^2} \\
&\quad - r_{a,b,c} \cdot \frac{(\mathbf{p}_{ab} \times \mathbf{p}_{bc})^\top (\mathbf{p}_{ab} \times \mathbf{w}_{jb,jc} + \mathbf{w}_{ja,jb} \times \mathbf{p}_{b,c})}{\|\mathbf{p}_{ab} \times \mathbf{p}_{bc}\|^2},
\end{aligned} \tag{6.2}
$$

where $\mathbf{a}^\top$ is the transpose of a vector $\mathbf{a}$, $f(a,b,c) = ((\mathbf{w}_{ja} - \mathbf{w}_{jb})^\top \mathbf{p}_{a,b})\|\mathbf{p}_{a,c}\|^2 \|\mathbf{p}_{b,c}\|^2$, and $\mathbf{w}_{ja,jb} = \mathbf{w}_{ja} - \mathbf{w}_{jb}$. $\dot{r}_{\mathbf{w}_j}$ represents the rate of change in radius $\mathrm{r}(t_j)$ with respect to the movement along mode $\mathbf{w}_j$. A large (small) $\dot{r}_{\mathbf{w}_j}$ for a particular mode means the channel radius quickly (slowly) increases or decreases as the protein makes a breathing

motion along that mode. As the protein fluctuates, the circumradius $\mathrm{r}(t_j)$ of $\triangle abc$ is approximated as:

$$\mathrm{r}(t_j; a, b, c) = r_{a,b,c} + \dot{r}_{\mathbf{w}_j} \cdot t_j. \tag{6.3}$$

Now consider the effect of a movement that involves a combination of m modes, i.e., $\sum_{i=1}^{m} t_j \mathbf{w}_j$. The circumradius r should become:

$$\mathrm{r}(t_1, ..., t_m; a_i, b_i, c_i) = r_{a_i, b_i, c_i} + \sum_{i=1}^{m} \dot{r}_{\mathbf{w}_j|i} \cdot t_j, \tag{6.4}$$

where r_{a_i, b_i, c_i} is the circumradius of the ith triangle $\triangle a_i b_i c_i$, and $\dot{r}_{\mathbf{w}_j|i}$ represents $\dot{r}_{\mathbf{w}_j}$ for $\triangle a_i b_i c_i$.

6.2.1.3   The Needed Constraints on Normal Modes that Gradually Open a Channel

Now we determine what combination of normal modes are needed to gradually open an initially closed channel. Equation (6.4) specifies how the circumradius changes as a result of mode motions. Our goal is to find the right combination of t_js such that the channel radius is guaranteed to increase by a certain small amount at each iteration step.

Assume that a channel is composed of k consecutive channel segments, and that each channel segment i is lined by three atoms a_i, b_i, and c_i, where $1 \leq i \leq k$. Recall that the channel radius R is the smallest radius of all channel segments:

$$R = \min_{1 \leq i \leq k} r_{a_i, b_i, c_i}. \tag{6.5}$$

After a small movement of $\sum_{i=1}^{m} t_i \mathbf{w}_i$, the channel radius becomes:

$$R_{\text{after}} = \min_{1 \leq i \leq k} \mathrm{r}(t_1, ..., t_m; a_i, b_i, c_i). \tag{6.6}$$

Now we require that after each iteration, the channel radius increases by a small amount of s, which is a model parameter, i.e.,

$$R_{\text{after}} - R \geq s. \tag{6.7}$$

Let $\mathbf{t} = (t_1, ..., t_m)^\top$ be the vector form of mode motion, and $\dot{\mathbf{r}}_i = (\dot{r}_{\mathbf{w}_1|i}, \dot{r}_{\mathbf{w}_2|i}, ..., \dot{r}_{\mathbf{w}_m|i})^\top$. The vector $\mathbf{t}$ that increases the channel radius by s satisfies the following constraints:

$$\dot{\mathbf{r}}_i^\top \mathbf{t} \geq u_i \qquad \forall 1 \leq i \leq k, \tag{6.8}$$

where

$$u_i = \min(r_{a_1,b_1,c_1}, ..., r_{a_k,b_k,c_k}) + s - r_{a_i,b_i,c_i}. \tag{6.9}$$

6.2.2 Selecting the Best Combination of Normal Modes

The constraints in Eq. (6.8) specify what breathing motions can gradually open a channel. There exist many combinations of modes satisfying these constraints. Which one should we choose? Which criterion should we follow to identify the most plausible breathing motion? One apparent choice is to minimize the amount of work required to open a channel. The optimal breathing motion that can open a channel should be the one that takes the least effort. Another consideration, which is less obvious but very necessary, is to realize that the derivation of Eq. (6.8) requires that the channel opening process should take place via many small, ideally infinitesimal, steps. Therefore, in the process of searching for the optimal breathing motion, we have also to require that the magnitude of the motion at each iteration step, which is $\|\mathbf{t}\|$, be small.

Denote by E, $\mathbf{f}$, and H the potential energy, the force, and the Hessian matrix at the current conformation, respectively. Denote by W the matrix form of the modes: $W = (\mathbf{w}_1, ..., \mathbf{w}_m)$. A protein movement can be written as $W\mathbf{t}$, where $\mathbf{t}$ is the vector form of the modes' contributions and is $\mathbf{t} = (t_1, ..., t_m)^\top$. The potential energy change δE due to movement $W\mathbf{t}$ can be approximated to the second order as:

$$\delta E = -\mathbf{f}^\top W\mathbf{t} + \frac{1}{2}\mathbf{t}^\top W^\top HW\mathbf{t}. \tag{6.10}$$

Let $\widetilde{H} = W^\top HW$. In this work, we use all the torsional modes solved by TNM [86] for W. H is the full Hessian matrix written in the Cartesian space. In general $\mathbf{w}_i$ is

not an eigenvector of H. Let Λ and V be the eigenvalues and eigenvectors of $\widetilde{H}$ in the matrix form, respectively, i.e., $\widetilde{H} = V\Lambda V^\top$.

Now define,

$$\mathbf{t}^* \;=\; |\Lambda|^{1/2}V^\top\mathbf{t}; \tag{6.11}$$

$$\mathbf{f}^* \;=\; |\Lambda|^{-1/2}V^\top W^\top\mathbf{f}, \tag{6.12}$$

where $|\Lambda|$ is a matrix whose elements take the absolute values of the corresponding elements in Λ, and $|\Lambda|^{1/2}$ is the square root of matrix $|\Lambda|$. Note Λ is a diagonal matrix. Let $\text{sign}(\Lambda)$ be a matrix whose elements take the signs of the elements of Λ. We have, $\Lambda = |\Lambda|^{1/2}\,\text{sign}(\Lambda)\,|\Lambda|^{1/2}$.

Using transformed variables $\mathbf{t}^*$ and $\mathbf{f}^*$, Eq. (6.10) becomes:

$$\delta E = -\mathbf{f}^{*\top}\mathbf{t}^* + \frac{1}{2}\mathbf{t}^{*\top}\,\text{sign}(\Lambda)\,\mathbf{t}^*. \tag{6.13}$$

Note that $\mathbf{t}^*$ in Eq. (6.11) is scaled by the square roots of the eigenvalues (i.e., $|\Lambda|^{1/2}$), and $\mathbf{f}^*$ is the generalized force and is the negative gradient of the potential δE with respect to $\mathbf{t}^*$. The advantage of using $\mathbf{t}^*$ over $\mathbf{t}$ is that the preference for lower frequency modes is naturally taken into account and the components of $\mathbf{t}^*$ can now be treated equally when finding the optimal $\mathbf{t}^*$. Λ as a diagonal matrix contains the eigenvalues of the Hessian matrix H in the mode space defined by W. $\text{sign}(\Lambda)$ represents the signs of these eigenvalues. When the conformation is at a local energy minimum, all the eigenvalues are positive and $\text{sign}(\Lambda)$ becomes an identity matrix. At other places, such as at a saddle point, some eigenvalues may be negative and $\text{sign}(\Lambda)$ may contain some -1's along the diagonal.

Recall our discussion in the beginning of this section that in finding the most plausible breathing motion that can open a channel, we have two considerations. One is to find a motion that takes the least work, and the other is to have a small $\|\mathbf{t}\|$ or $\|\mathbf{t}^*\|$ at each iteration. Only minimizing the work without requiring the magnitude of $\mathbf{t}$ to be small

may result in large, unrealistic moves. To allow searching in all directions in the mode space and yet to bias the search towards the direction of the potential energy's gradient descent, we set the search space to be of the shape of an ellipsoid, whose one focus is the origin of the mode space for $\mathbf{t}^*$ and whose other focus is in the direction of the force $\mathbf{f}^*$, i.e., the direction of the steepest descent of the potential energy. The eccentricity of the ellipsoid controls the degree of the biasedness towards the direction of the steepest descent. An eccentricity of 0 means no bias, while an eccentricity of 1 means the search is fully along the direction of the steepest descent. Eccentricity is a model parameter in this work. Experiments show that an eccentricity of 0.7 works well. The function of such an ellipsoid is:

$$\|\mathbf{t}^*\|(1 - e \cdot \cos(\theta)) = const, \tag{6.14}$$

where e is the eccentricity, θ is angle between the direction $\mathbf{t}^*$ and the direction of the steepest descent of the potential energy, i.e.,

$$\cos(\theta) = -\frac{\mathbf{f}^{*\top}\mathbf{t}^*}{\|\mathbf{f}^*\| \cdot \|\mathbf{t}^*\|}. \tag{6.15}$$

Now the search for the most plausible breathing motion $\mathbf{t}^*$ that opens a channel becomes an optimization problem:

$$\operatorname{argmin} \mathbf{t}^* \, \|\mathbf{t}^*\|^2 (1 - e \cdot \cos(\theta))^2. \tag{6.16}$$

The above optimization is subject to the constraints given in Eq. (6.8), which can be rewritten, using the new variable $\mathbf{t}^*$, as follows:

$$(\mathbf{r}_i^*)^\top \mathbf{t}^* \geq u_i \qquad \forall 1 \leq i \leq k, \tag{6.17}$$

where

$$\mathbf{r}_i^* = |\Lambda|^{-1/2} V^\top \dot{\mathbf{r}}_i. \tag{6.18}$$

Algorithm 3 lists the steps for finding the most plausible breathing motion that increases a channel's radius by a pre-specified amount of s.

Algorithm 3 ConstraintGuidedMotion($\mathbf{f}, H, W, \{\dot{\mathbf{r}}_1, ..., \dot{\mathbf{r}}_k\}, \{u_1, ..., u_k\}$)

1: $V \leftarrow$ eigenvectors of $W^\top H W$
2: $\Lambda \leftarrow$ eigenvalues of $W^\top H W$
3: $\mathbf{f}^* \leftarrow |\Lambda|^{-1/2} V^\top W^\top \mathbf{f}$
4: $\mathbf{r}_i^* = |\Lambda|^{-1/2} V^\top \dot{\mathbf{r}}_i \qquad \forall 1 \le i \le k$
5: $\mathbf{t}^* \leftarrow \operatorname{argmin} \mathbf{t}^* \|\mathbf{t}^*\|^2 (1 - e \cdot \cos(\theta))^2$
$\qquad \underset{1 \le i \le k}{\operatorname{subjto}} (\mathbf{r}_i^*)^\top \mathbf{t}^* \ge u_i$
6: $\mathbf{t} \leftarrow V |\Lambda|^{-1/2} \mathbf{t}^*$

6.2.3 The Iterative Procedure for Opening up a Channel

In this section, we summarize the steps to open up a channel. Given a protein structure, it is first energetically minimized using the CHARMM22 force field. [83] From the minimized structure, Voronoi diagram is computed and cavities and putative ligand migration channels are identified.

Algorithm 4 lists the rest of the steps that follow. The algorithm receives as inputs the initial protein conformation $\mathbf{p}_0$ (i.e., the minimized structure), a putative channel path L_0, the target channel radius h that is required for an open channel, and the radius increase s at each iteration. In the algorithm, channel L_0 is gradually opened by applying at each iteration the best combination of normal modes determined by Algorithm 3, until the channel radius reaches h. In each iteration, the mode matrix W is determined by using the torsional network model (TNM). [86] The optimal value $\mathbf{t}$ determined by Algorithm 3 is then used to update the conformation. There are two major advantages in using torsional modes: i) it reduces the number of degrees of freedom; ii) it avoids sharp potential energy increases caused by distorted protein geometry.

6.3 Results

In this section, we apply the proposed method to study ligand migration in myoglobin and its mutants.

Algorithm 4 ExpandChannel($\mathbf{p}_0, L_0, h, s$)

1: $\mathbf{p} \leftarrow \mathbf{p}_0, \quad L \leftarrow L_0$
2: $M \leftarrow \{\langle a_i, b_i, c_i \rangle \mid \text{triangle atoms } a_i, b_i, c_i \text{ of channel segment } l_i \in L\}$
3: **while** $\min(r_{a_1,b_1,c_1}, r_{a_2,b_2,c_2}, ...) < h$ **do**
4:     Compute force $\mathbf{f}$ and Hessian matrix H of conformation $\mathbf{p}$
5:     Compute modes $W = (\mathbf{w}_1, ..., \mathbf{w}_m)$ using TNM. [86]
6:     Determine constraints $\dot{\mathbf{r}}_i^\top \mathbf{t} \geq u_i$ according to (6.8), for all $\langle a_i, b_i, c_i \rangle \in M$
7:     $\mathbf{t} \leftarrow \text{ConstraintGuidedMotion}(\mathbf{f}, H, W, \{\dot{\mathbf{r}}_1, ...\}, \{u_1, ...\})$
8:     Update conformation: $\mathbf{p} \leftarrow \mathbf{p} + W\mathbf{t}$
9:     Update channel segments of L using the new $\mathbf{p}$
10:     Update M.
11: **end while**

6.3.1   General Experimental Procedure

Given a protein structure, it is first energetically minimized in vacuum using the Tinker [102] software and the CHARMM22 force field. [83] Hydrogen atoms are added to the structure during this process. From the minimized structure, we first compute the Voronoi diagram.

In this work, we use the all the atoms in the myoglobin structure, including all the hydrogen atoms, to construct the Voronoi diagram. Each atom in myoglobin is represented by a point, with one exception. To take into account the size difference between different atom types, especially between hydrogen atoms and heavy atoms, each heavy atom within a short distance from the channel being studied is represented by a set of points (12 points in total). These points are uniformly distributed, as in a regular icosahedron, around the surface of the heavy atoms, each of which is given a radius that is the difference between the van der Waals radii of the heavy atom and a hydrogen atom. The reason why only the heavy atoms close to the channel are treated in this way is for computational efficiency.

Next, putative ligand channels L_0 are determined using the Voronoi diagram as a guide. [78] Specifically, any path in the Voronoi graph that is between a cavity (which is represented by a Voronoi vertex, see next section) and the solvent (which is represented by any Voronoi vertex that is outside the protein and has a large clearance) and whose clearance is greater than a given threshold is considered as a putative channel.

Lastly, each putative channel is iteratively expanded using Algorithm 4 until it is fully open. A channel is considered to be fully open if its clearance surpasses a preset threshold h. In this work, the threshold is set to be 1.8 Å. The algorithm identifies at each step the best combination of normal modes that are able to open up the channel gradually. The sequence of steps from the initial conformation to the final conformation where the channel is fully open form the transition pathway needed to open the channel. Consequently, protein motions along this pathway provide a quantitative description of what breathing motions are needed and how they open or close the channel. The amount of work needed to open each channel is also computed and is used to estimate the relative likelihood that a channel may be used by a ligand to enter into or exit from the protein matrix. This amount of work corresponds to the enthalpy change between the initial and the final conformations:

$$\Delta H = W. \tag{6.19}$$

The free energy change between the two states, i.e.

$$\Delta F = \Delta H - T \cdot \Delta S, \tag{6.20}$$

is an even more desirable measure in predicting the likelihood for a channel to open. However, the computation of free energy or entroy is much more complicated. In the following, we will show the enthalpy change alone (ΔH) already presents itself as a good measure for predicting ligand migration channels.

6.3.2 Cavities in Myoglobin

Myoglobin is known to have seven large internal cavities. Besides its four Xenon biding sites (Xe1–Xe4) [112, 138] and distal pocket (DP), [135] two additional cavities Ph1 and Ph2 (renamed here as S1 and S2) were identified in an MD simulation [14] (See Figure 6.2). The coordinates of the centers of these seven cavities in reference to a crystal structure (1A6G.pdb) are given in Table S1 in Supplemental Information.
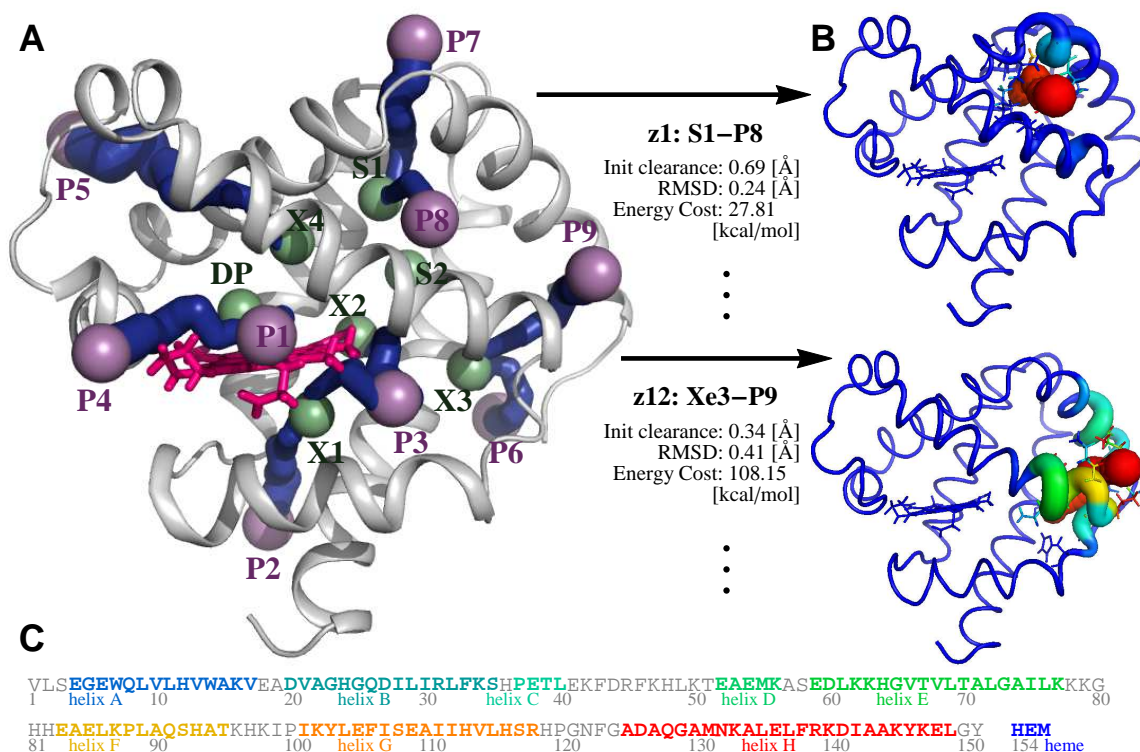
Figure 6.2 **Ligand migration channels in myoglobin.** (A) A cartoon image of myoglobin (pdbid: 1A6G) overlaid with cavities (green spheres), portals (purple spheres with labels that start with P), and nine channels (rugged blue tubes) identified by both our method (see the full list in Table 6.2) and MD. (B) Different conformation changes needed to open these channels as determined by our method. The thickness along the backbone trace shows the magnitude of the motions of the residues as each channel opens up. The RMSD deviations between the final conformations and the initial minimized structure and their potential energy differences are also given. (C) The primary sequence of myoglobin and its secondary structures (helices A to H).

To obtain the coordinates of the cavities in a new myoglobin structure (wild type or mutants), we first align the structure to the reference crystal structure. Once aligned, the coordinates of the cavities in the reference structure are copied and used as "estimated cavity centers" for the new structure.

As aforementioned, each cavity is represented by one Voronoi vertex in the Voronoi diagram. Here is how the vertex is picked. Since for each cavity the coordinates of the "estimated cavity center" are known, the Voronoi vertex that is the closest to the the estimated cavity center and whose clearance surpasses a given threshold is selected as the representative vertex for that cavity.

6.3.3 Ligand Migration Channels in Myoglobin

In this study, 1A6G (pdb-id) is used as the initial structure. In the energy minimization process, in order to prevent the heme from being bended, the positions of its heavy atoms are fixed. After energy minimization, Voronoi diagram is constructed from the minimized structure and used as a guide [78] to generate putative channels. As a result, 41 putative channels are determined that connect the solvent with the seven internal cavities. Next, all 41 putative channels are tried to be stretched open using Algorithm 4 and the work costs to open them are recorded.

6.3.3.1 Putative Channels and the Work Cost to Open Them

Table 6.2 summarizes the results for all 41 putative channels. The channels are sorted by the energy cost to open them, and compared with the channels determined in a previous MD simulation (which are labeled by portal numbers: P1 to P9). [109] A channel from our work is assigned to one of the numbered portals (i.e., P1 to P9) that were previously identified in an MD simulation [109] if at least two residues lining its opening to the solvent are the same as the residues reported to line the numbered portal. [109] Remarkably, most of the channels that require the lowest energy cost to open find a

match with one of the channels identified in MD simulations. [109] This implies that the amount of entropy changes may be similar for most of the channels or proportional to the amount of enthalpy changes. And as a result, the change of enthalpy, or the work cost, alone presents itself a very good indicator in predicting ligand channels. However, there are some exceptions. There exist a few channels that are easy to open according to our method (i.e., small ΔH) but have not been observed during MD simulations, such as z4, z7, z10, and z11. One plausible explanation for this is that these channels actually do open up dynamically, but they are overlooked in MD. This is quite possible since MD simulations are stochastic by nature and are incomplete in conformation sampling. The opening of these channels may also represent a more rare event and thus present a higher entropy cost than other channels and consequently, they may be less favored by the ligand kinetically.

It is helpful also to realize that results in Table 6.2 are predictions of ligand migration channels sorted by the amount of change in enthalpy (or the work cost), not by the amount of change in free energy. It is foreseeable that thermal fluctuations, which are quite large in macromolecules like proteins (see chapter 12.5 in [52]), are able to surpass these barriers and open these channels. Results in Table 6.2 do not tell us how often these channels may open, since that depends also on the entropy cost (or gain).

Figure 6.3 plots the relationship between the energy cost in opening a channel and its initial clearance. The general trend is that as initial clearance decreases, the cost in opening a channel increases, though there are many exceptions. There are some channels (such as z37 and z41) that are more difficult to open though they have relative larger initial clearances, while other channels (such as z5, z8 and z10), which have smaller initial clearances, are easier to open. This implies that initial channel clearance is only a fair indicator of a putative channel's openability.

Table 6.2   Prediction results on ligand migration channels.

idx[a]	cvty[b]	clrs[c]	rmsd[d]	cost[e]	portal[f]	idx	cvty	clrs	rmsd	cost	portal
z1	S1	0.69	0.24	27.81	P8	z22	Xe2	0.22	0.40	134.44	
z2	Xe1	0.62	0.25	44.42	P2	z23	Xe2	0.38	0.46	134.74	
z3	Xe2	0.46	0.29	47.09	P3	z24	Xe1	0.21	0.48	137.97	
z4	Xe1	0.26	0.38	55.83		z25	S2	0.32	0.47	142.69	
z5	Xe1	0.25	0.36	56.89	P3	z26	DP	0.25	0.46	160.75	P4
z6	S1	0.55	0.31	58.09	P7	z27	Xe4	0.27	0.46	175.84	
z7	Xe1	0.43	0.35	73.85		z28	Xe4	0.24	0.51	182.89	P5
z8	Xe3	0.23	0.30	74.82	P6	z29	S2	0.28	0.54	191.27	
z9	DP	0.47	0.36	76.11	P1	z30	S2	0.33	0.50	191.96	
z10	Xe3	0.14	0.29	85.93		z31	Xe4	0.29	0.48	195.40	
z11	Xe1	0.32	0.40	99.58		z32	Xe1	0.31	0.50	196.32	
z12	Xe3	0.34	0.41	108.15	P9	z33	S2	0.17	0.52	199.17	
z13	Xe3	0.40	0.42	109.18		z34	DP	0.19	0.53	203.14	
z14	S1	0.11	0.38	111.42		z35	Xe1	0.26	0.50	204.24	
z15	Xe3	0.17	0.45	113.47		z36	Xe4	0.28	0.53	244.69	
z16	S2	0.42	0.39	116.53		z37	Xe4	0.32	0.51	256.18	
z17	Xe3	0.42	0.41	120.40		z38	Xe4	0.22	0.57	288.13	
z18	Xe4	0.49	0.42	127.60		z39	Xe4	0.27	0.55	289.11	
z19	S2	0.22	0.42	129.97		z40	Xe1	0.19	0.53	295.08	
z20	Xe1	0.30	0.49	132.59		z41	Xe4	0.37	0.57	306.50	
z21	S1	0.18	0.44	132.73							

[a]putative channel index;
[b]cavity to which the channel is connected;
[c]initial clearance of the channel [Å];
[d]RMSD between the initial and final conformations [Å];
[e]total energy cost ΔH to expand the channel [kcal/mol];
[f]corresponding portal as determined in MD. [109]
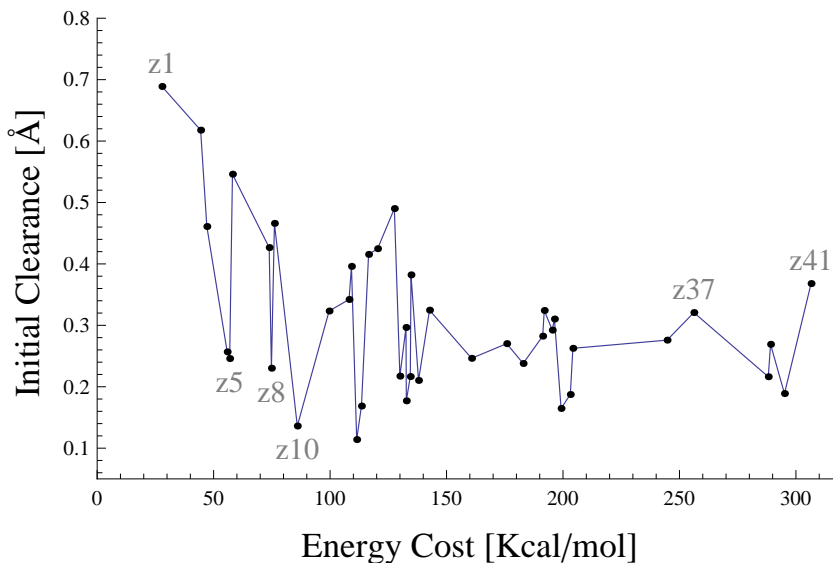
Figure 6.3 **The relationship between the energy cost required to open a channel and the channel's initial clearance.**

6.3.3.2 Quantitative Delineation of the Protein Motion Trajectories that Open the Channels

Our channel stretching procedure, when applied to the putative channels listed in Table 6.2, not only predicts which channels are more likely to open than others in terms of the work cost, but also determines the breathing motions that open these ligand migration channels. It identifies at each iteration step the best combination of normal modes that is able to gradually open up the channel. This sequence of steps from the initial conformation to the final conformation where the channel is fully open represent the transition pathway needed to open the channel. The rocking motion along this pathway provides a quantitative description of what breathing motions are needed and how they open and close the channel. One advantage of our approach over MD simulations is that one has full control in selecting which channel to study and thus focusing on that channel only, while for MD-based simulations, one has to rely on pure chance and wait for the event of an ligand entering through the channel of interest to happen, due to the stochastic nature of the simulation process. Thus the rarer a channel is used by a

ligand, the more difficult it becomes to computationally study and analyze that channel using MD. Another potential benefit is that the approach examines how a channel may be opened based on enthalpy alone and it separates out the thermal fluctuations that are involved in the actual process. Since thermal motions are filtered out, it may provide a clearer and more insightful understanding of how residue-residue interactions open and close a channel.

All the final conformations (in PDB format) at which the channels are open, the transition pathways (in PDB format) needed to open each and every of these channels, as well as movies that display the breathing motions needed to open all these channels are available at http://www.cs.iastate.edu/~gsong/CSB/channels/.

6.3.3.3 Identify Key Normal Mode Motions and Key Residues that Regulate the Channels

The quantitative delineation of the breathing motions that open a channel allows us to closely examine the motions and identify key normal mode motions or key residues that regulate the channel. Since the modes selected at each iteration are torsional modes, the combination of these modes represent a displacement in the torsional space. The residues whose torsional angles (which could be either backbone ϕ/ψ angles or side-chain χ angles) display the largest change are those that contribute the most to opening the channels. On the other hand, the residues that have the most strain are those that block the channel from opening. We define the strain of a residue as the potential energy change within the residue between its initial and final conformations.

In the following, we present the results of a few closely examined channels. For each channel, we present in a plot (see Figure 6.4) the residues that line the channel and their distances to the cavity, the initial clearances along the channel, the amount of backbone motions and side chain motions required of the residues to open the channels, and the magnitudes of motions in Cartesian space as measured by root mean square distance

(RMSD), and lastly, the strain that incurs on each residue as a result of opening the channel. Movies that show the opening processes of these channels and the movements of the residues that line the channels are available at http://www.cs.iastate.edu/~gsong/CSB/channels/.
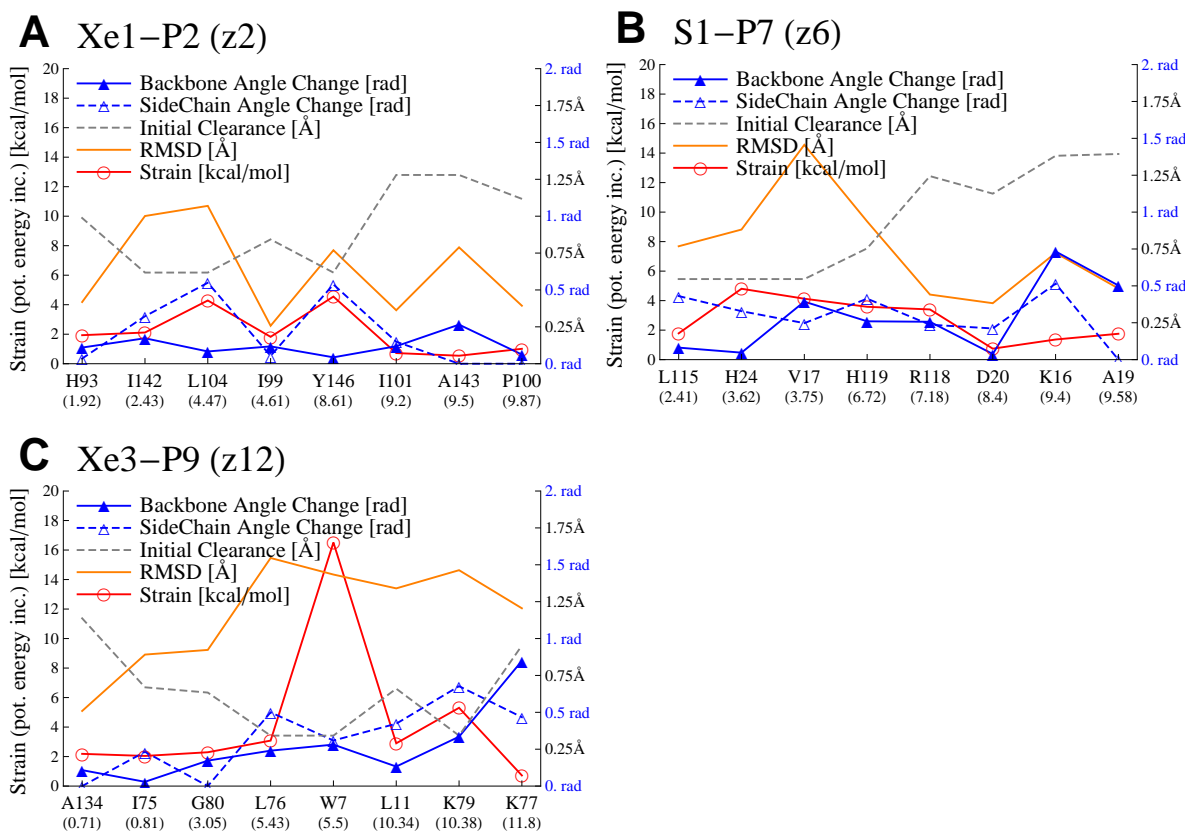
Figure 6.4 **The initial clearances, conformation changes required of the channel lining residues, and the strains incurred on them as a result of opening some of the channels**: (A) the channel from Xe1 to portal 2, (B) the channel from S1 to portal 7, and (C) the channel from Xe4 to portal 9.

Channel Xe1 to P2 (z2): For this channel, side-chain motions play a dominating role, especially those of Leu104 and Tyr146. Figure 6.4(A) clearly shows that these two residues have the largest side-chain rotations and experience the most strain. The channel is opened by the side-chain swing motions of Leu104 and Tyr146.

Channel S1 to P7 (z6): The S1 to P7 channel, on the other hand, is more controlled by the motions of backbone torsional angles of several residues, namely Lys16, Val17, and Ala19, as is evident from Figure 6.4(B). The combined backbone motions of these residues cause a large Cartesian space displacement for Val17 and bring it away from the center of the channel and thus increase the channel radius.

Channel Xe3 to P9 (z12): This is a case where one residue (Trp7) plays a dominating role in blocking the channel path while the other residues are freer to move. In Figure 6.4(C), it is seen that Trp7, compared to other residues lining the channel, clearly takes most of the strain (potential energy increase). Trp7 itself does not undergo much side-chain or backbone rotations in the channel opening process. It resists the motions and as a result, a large amount of internal strain is created in the residue. Replacing Trp7 with a smaller residue such as ALA should greatly lower the energy cost to open this channel.

6.3.3.4 Can the Channels Be Opened by Backbone Motions or Side-Chain Motions Alone?

The opening and closing of each channel are often controlled by the interplay of backbone motions and side-chain motions. However, the necessity of having side-chain motions or backbone motions in opening a channel has not been systematically assessed before. One advantage of our proposed method is that it allows one to assess independently the contributions of side-chain motions and backbone motions. In the following, we apply our method to investigate if the channels can be opened by either motions alone, and if so, what extra work is needed.

Figure 6.5 compares the energy cost to open some of the channels when only backbone motions (blue dot-dashed) or only side-chain motions (orange dashed) or both (black solid line) are allowed. We denoted by the infinity sign, ∞, the case that a channel cannot be opened.
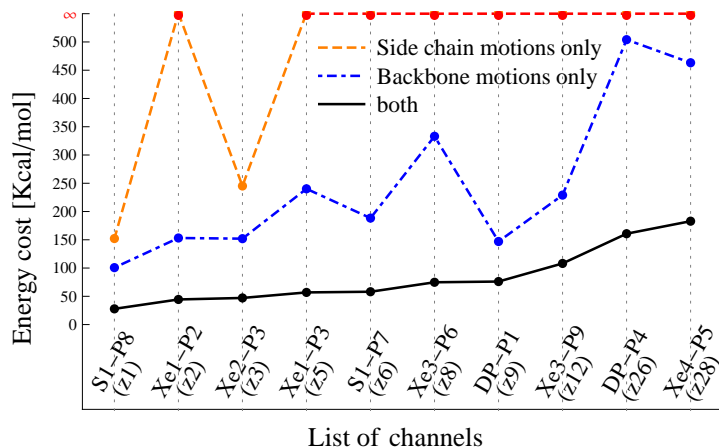
Figure 6.5 **Comparisons of energy costs required to open a channel when only backbone motions (blue dot-dashed) or only side-chain motions (orange dashed) or both (black solid) are allowed.** Channels are sorted by the opening energy cost when both motions are allowed. ∞ means that the channel cannot be opened.

Results in Figure 6.5 show that many channels cannot be opened by side-chain motions alone (orange dashed). The backbone motions are capable to open each channel even though it requires much more work. Interestingly, the well-known HIS channel (DP to P1) cannot be fully opened by side-chain motions alone, though it is generally accepted that His64 plays a gating role in opening the channel. A plausible explanation is that backbone breathing motions are needed to create enough space for the side-chain of His64 to swing open.

6.3.4 Myoglobin Mutants: How Mutations Affect the Histidine Channel

Another advantage of our proposed method is that it can be easily extended to study the effects of mutations on ligand migration channels. Mutagenesis studies can provide deep insights into the behaviors of proteins. Deep mutational scanning as reported by Fowler and Fields [32] promises that large-scale mutagenesis data might become available experimentally. Our proposed method provides a convenient computational approach for studying the effects of mutations. Such computational studies can be combined with

experimental mutagenesis studies to gain deeper insights into the functional behaviors of proteins.

Myoglobin has been long studied since its 3-D structure was first revealed by X-ray crystallography. [57] His64 was soon proposed to play a gating role of the channel between the heme pocket and solvent for ligand entry and exit. [101] This was confirmed by many mutagenesis studies that followed, some of which studied how mutations affects ligand entry and exit rates. For example, Scott et al. [116] reported biomolecular rate constants for ligand entry/escape of the wild type myoglobin and its several mutants. These rates represent how fast a ligand enters into or escapes from the protein. It has been well established experimentally that most of the ligand entry is through the HIS channel. [98]

To computationally study the effects of mutations, we apply our method to study the HIS channel of myoglobin wild type (WT) and four mutants (F46A, F46W, H64A, H64W). For starting conformations, structures 2MGK and 3OGB are used for WT and mutant H64W, respectively. The starting conformations of F46A, F46W, and H64A are manually generated using the psfgen program from VMD. [50]

6.3.4.1   Mutations at His64

It is expected that replacing His64 with a Tryptophan makes it more difficult for the HIS channel to open. Table 6.3 lists the strain energy of each residue or the strain between an adjacent pair of residues. Strain is defined as the potential energy increase of a residue (or between a pair of residues) as a result of changes in structure. The results show that the H64W mutation greatly increases the strain in residue 64 (now a Trp) and in the heme. This is probably because the larger side-chain of Trp reduces the available room in the channel and consequently, stretching open the channel creates more strain on the lining residues, especially on Trp64 and the heme. Moreover, Our results indicate that Trp64 rotation alone cannot fully open the channel. The motions of Thr67 contribute the other half. Figure 6.6(A) shows that in WT, His64 can rotate enough

to open the channel from initial (in gray) to final (in color) conformations. However, in Figure 6.6(B), rotation of Trp64 in mutant H64W is hindered by part of the heme (highlighted in red), and consequently Thr67 has to be moved away to fully open the channel.

Table 6.3   Energy costs and strains of opening HIS channel of Mb wild type and its 4 mutants.

wild type or mutant	clear -ance[a]	opening cost[b]	entry rate[c]	strain of residue i					strain between residues i-j					
				Arg 45	Phe 46	His 64	Thr 67	Heme	60 -45	60 -46	64 -45	64 -46	Heme -64	Heme -67
WT	0.57	76.71	34±7	5.7	0.3	8.6	2.0	10.6	4.3	-0.2	5.9	0.1	0.4	2.3
F46A	0.57	66.24	110	2.6	0.3	6.8	1.9	7.4	3.1	0.0	2.4	0.1	-1.0	2.1
F46W	0.40	71.20	35	7.6	0.1	11.0	4.3	8.6	6.9	1.1	9.0	-0.7	-2.7	3.8
H64A	0.89	39.33	410	0.9	0.9	1.5	0.2	6.0	3.1	0.1	-0.1	0.3	0.4	1.3
H64W	0.25	88.67	8.6	0.7	0.1	13.0	3.7	17.6	5.2	-0.2	-4.3	0.3	5.2	3.8

The table shows the energy costs and strains created in opening the HIS channel of myoglobin wild type and its four mutants.
[a]initial clearance [Å];
[b]energy cost to open the HIS channel [kcal/mol];
[c]ligand entry rate as measured by Scott et al. [116] [$\mu M^{-1} s^{-1}$]

By replacing His64 with Alanine, many channel-opening energy barriers are removed. It was thought [98] that this drastic effect was mainly due to the enlarged void created by the mutation. Our computational results confirm this postulation. The strain of on residue 64 (now an Alanine) is almost gone after the mutation. The strain on Arg45 also drops significantly since the formal pressing between Arg45 and His64 due to steric constraints is gone.

6.3.4.2   Mutations at Phe46

H46A and H46W were studied experimentally by Lai et al. [69] to test whether or not Phe46 sterically restricts the swing rotations of His64. Our results in table 6.3 and Figure 6.6(C)/(D) suggest that it is actually Arg45 that directly constrains His64's
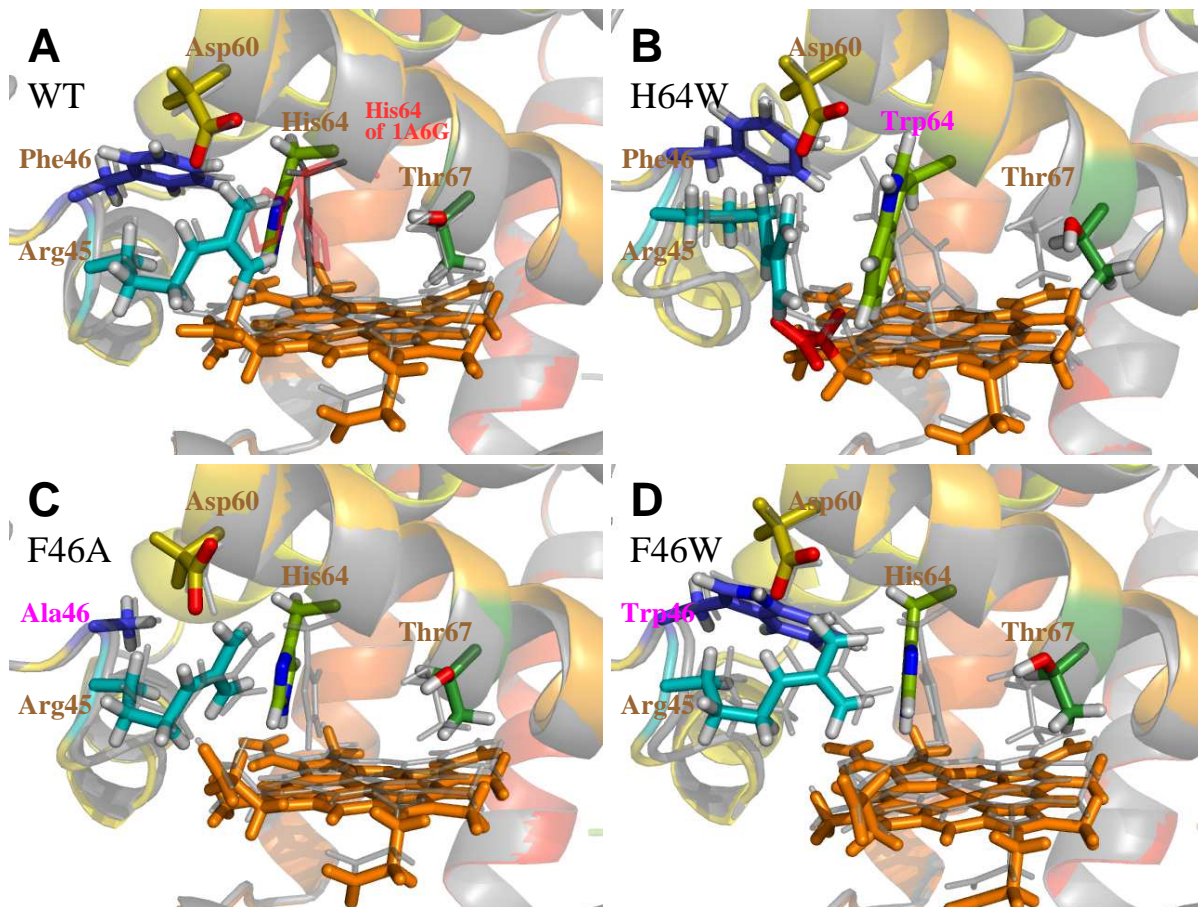
Figure 6.6 **The interplay of residues in opening the HIS channel of myoglobin (A) wild type, and (B)-(D) its three mutants.** The initial conformations (gray transparent) and the final conformations (in color) of the key residues (labelled) are shown. In (A), both the open and closed conformations of His64 from crystal structure 1A6G are shown in transparent red.

rotations. Arg45 comes between His64 and Phe46. Phe46 comes into play indirectly by constraining the motions of Arg45. In mutant H46A, in which Phe46 is replaced with Alanine, Arg45 becomes less constrained and can move more freely. As a result, His64 can rotate without difficulty. However, in H46W, Trp46 and Asp60 greatly constrain the motions of Arg45, which in turn blocks His64's rotations. In the end, since His64 cannot open the channel large enough, Thr67 has to budge to create a large enough hole for a ligand to pass through.

6.3.4.3 Understanding How Mutations Affect Ligand Entry Rates

Our proposed method computes the work cost to open a channel. When applied to both the wild type and the mutants of the same protein, it can be used to study the effects of mutations computationally.

From our computations, we find that, among the wild type myoglobin and its mutants, the difficulty to open the HIS channel is at its highest in the H64W mutant, followed by WT, mutants F46W, F46A, and H64A. The actual energy costs to open the HIS channel in these structures are summarized in Table 6.3. Notice that the order of the energy costs (3rd column) matches well with that of the entry rates (4th column) reported experimentally: [116] the larger the entry rate, the smaller the energy cost.

Now the energy cost (ΔH) to open a channel and the channel opening frequency (and therefore the ligand entry rate k) can be related in the following way:

$$k \propto \exp\left(\frac{-\Delta H}{k_{\mathrm{B}}T}\right). \tag{6.21}$$

That is, there is a linear relationship between the energy cost and the logarithm of the inverse of the rate:

$$\Delta H = a \cdot \log(1/k) + b, \tag{6.22}$$

where a (positive) and b are two constants. Ideally, we should use the free energy change ΔF in the above equation. However, if the amount of entropy change ΔS involved in opening the HIS channel is assumed to be the same for the WT and the mutants, the above equation is valid for ΔH as well.

In Figure 6.7, we plot the energy costs as predicted by our method to open the HIS channel versus the logarithms of ligand entry rates as measured experimentally, [116] for myoglobin wild type and it mutants. Remarkably, the data points in the figure clearly show that there is a strong linear relationship between our predicted energy costs and the logarithms of the experimental rate constants. Together with the channel prediction

results in Table 6.2, results here strongly demonstrate that our NMA-based method is effective in predicting ligand migration channels and the effects of mutations.
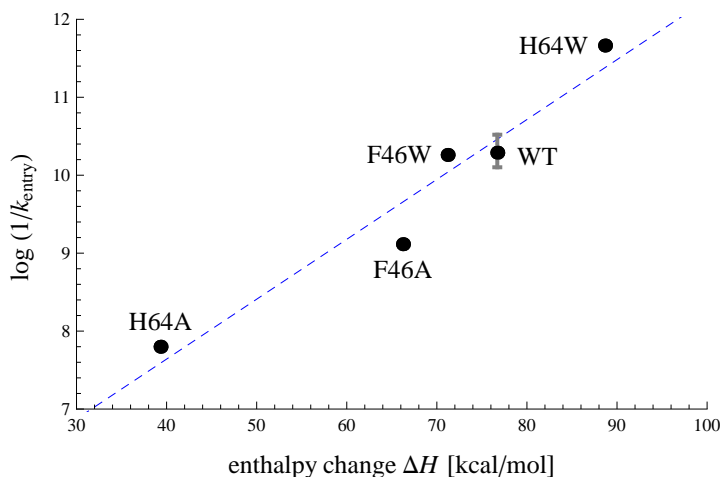
Figure 6.7 **Linear relationship between the amount of change in enthalpy and the logarithm of ligand entry rate.**

6.4 Summary and Discussions

Molecular dynamics (MD) and normal mode analysis (NMA) are two widely used computational methods for studying protein dynamics. Both tools are powerful and they complement each other in important ways. The advantage of NMA is that it has a closed-form analytical solution and as a result it is able to cover much more efficiently the conformation space near the native state. For the problem of ligand migration pathways and channels, most of existing computational studies have been done using MD. In comparison, little work has been done that uses a NMA-based method.

In this work, we have presented and applied a novel normal mode-based method to efficiently predict ligand migration channels of myoglobin and its mutants. The motivation behind this work springs from the observation that normal mode motions are closely linked to breathing motions that are often thought to be the cause that opens ligand migration channels. Our results in this work are important for several reasons. First, since protein normal modes are closely linked to protein breathing motions, our normal

mode-based method allows us to quantitatively delineate what breathing motions open a channel. Second, the method allows us to gain a direct and detailed depiction of the motions of each and every residue that lines a channel and thus allows us to identify key residues that play a dominating role in regulating the channel, either through backbone motions or side-chain motions or a combination of both. Third, the all-atom based model and the full force-field employed in our method allow us to gain a realistic energetics related to the work cost required to open a channel. Fourth, our method provides an efficient computational tool for studying the effects of mutations on ligand migration channels and on ligand entry rates. Lastly, the features summarized above mean that our method has strong predictive power over ligand migration channels in other proteins of known structures, over the effects of mutations, and in recognizing key residues.

Our results on myoglobin and its mutants are in excellent agreement with MD simulation results and experimentally determined ligand entry rates. Most of the channels that require the least amount of work to open as predicted by our method match with channels identified in MD simulations. When applied to myoglobin mutants related to the HIS channel, the work costs predicted by our method, or the enthalpy changes required to open the HIS channel, are found to match closely with experimentally measured ligand entry rates. In addition, the method has predicted key residues and their roles in opening a channel and these predictions also are in agreement with MD simulation results and mutagenesis studies. Lastly, our method provides atomic-scale transition pathways (conformation changes of the protein) needed to open each and every channel. Transition pathways (in PDB format) and movies that display how breathing motions open these channels are available at http://www.cs.iastate.edu/~gsong/CSB/channels/. Such atomic-scale transition pathways are not available from other channel mapping methods such as ILS, [23] and it would take a much longer time to obtain them using MD simulations.

There are three notable contributions in the method development. The first is to compute the derivative of a channel clearance with respect to any given mode. This

derivative describes how rapidly (or slowly) a breathing motion along a particular normal mode increases or decreases the clearance of a given channel. It is foreseeable that such derivatives with respect to the normal modes may be used to study other structural or functional properties of proteins that depend on protein dynamics. To the best of our knowledge, this has not been done before. Secondly, finding the best combination of normal modes that gradually open a channel is formulated as an optimization problem. Thirdly, the optimization function we used is innovative. When defining what is the best combination of normal modes that slowly open a channel, there are two attributes that we desire to minimize at each iteration: one is the work required and the other is the magnitude of the conformation change. How to minimize one without neglecting the other? Though it is desired that the work cost to stretch open a channel should be minimized, however, if we focus only on minimizing the work cost, it may result in large unrealistic movements. Therefore, we need to somehow minimize the work cost while keeping the magnitude of the motions (i.e., $\|\mathbf{t}\|$) small at the same time. Our optimization function, which is of ellipsoidal form and has an adjustable eccentricity factor, provides a perfect balance between the two requirements.

The limitation of the current method is that it does not consider the potential interactions between the ligand and the host protein. While this is acceptable for small gaseous ligands, [33] care must be taken when applying this method to larger ligands or charged ligands (such as proton or metal ions), since in those cases van der Waals or electrostatic interactions between the ligand and the host protein may strongly affect the normal modes of the host protein.

Acknowledgments

CHAPTER 7.   SUMMARY AND CONCLUSION

In this dissertation, I have made several significant contributions to the field of computational biology, especially in the area of computational studies of protein dynamics that use *normal modes.* Specifically, I have developed a new approach that bridges classical normal mode analysis (NMA) with elastic network models and a series of novel schemes for deriving simplified NMA models that are both efficient and accurate. Since NMA is a widely used tool for studying protein dynamics, contributions made in this dissertation may have far-reaching impacts.

My first computational model, the force spring model (FSM), was developed to unify two popular elastic network models: GNM and ANM. The model was based on a key realization of how inter-residue forces or torques precisely influences normal mode computations. It was found that NMA Hessian matrix can always be written as a sum of spring-based terms and force-based terms and that the total contribution of the force-based terms (by inter-residue forces) is significantly smaller than that by the spring-based terms.

The first study then triggered the development of spring-based NMA (sbNMA) and simplified spring-based NMA (ssNMA). Both models keep only the dominant spring-based terms. In so doing, they remain nearly accurate as the classical NMA is and yet avoid the cumbersome energy minimization step that is required by classical NMA.

While the above two contributions are on the simplification of interaction models in normal mode computations, my third contribution is on the simplification of structural models. Coarse-grained structural models are often favored in normal mode compu-

tations for large proteins or protein complexes. However, many coarse-grained models are limited in their accuracy in representing the dynamics. I have developed a novel coarse-graining scheme that is able to preserve the atomic accuracy in dynamics while coarse-graining the structure. This is highly desirable and useful especially in dynamics studies of very large structure complexes. The method utilizes the sparseness of Hessian matrix to achieve efficient coarse-graining.

My fourth contribution is on the vibrational spectrum of globular proteins. I have found the vibrational spectrum of globular proteins is universal. That is, regardless of the protein in question, it closely follows one universal curve. The aforementioned accurate sbNMA was especially helpful in understanding what the different peaks in the vibrational spectrum represent. This work makes possible a potential two-way dialogue between theory and experiment regarding the vibrational spectrum: experimental spectra of proteins may be used to fine tune theoretical empirical potentials, and the various features and peaks observed in theoretical studies may be used to guide experimental studies.

My fifth contribution is on developing a new *normal-mode-based* computational method that predicts ligand migration channels and ligand entry rates in proteins and mutants. To this end I developed several new techniques, such as channel expansion ratios by normal modes, a novel ellipsoidal-shaped minimization function, and a constraint-guided motion planning approach, etc. The new computational method identifies and gradually opens a ligand channel while minimizing the potential energy cost. The method has been successfully applied to find ligand migration channels of myoglobin. The prediction results matched well with both molecular dynamics simulation results and experimentally-determined ligand entry rates.

My thesis work opens up a number of possible directions for future research. Future research may include but is not limited to: (1) classification of mutants by their functional motions, specifically how they open up ligand migration channels, as an extension

of my study on myoglobin mutants; (2) a deeper understanding of the degeneracy of protein normal modes, such as why some modes are more prone to degeneracy than others. (3) improving sbNMA by taking into account explicit solvent and studying the effect of solvation on protein complexes and on protein-protein interactions. (4) A deeper dynamics-based understanding of how antibiotics can hinder the proper function of bacterial ribosomes.

BIBLIOGRAPHY

[1] Anselmi, M., Di Nola, A., and Amadei, A. (2008). The kinetics of ligand migration in crystallized myoglobin as revealed by molecular dynamics simulations. *Biophys. J.*, 94(11):4277–81.

[2] Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80(1):505–515.

[3] Atilgan, C. and Atilgan, A. R. (2009). Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput. Biol.*, 5(10).

[4] Atilgan, C., Gerek, Z. N., Ozkan, S. B., and Atilgan, A. R. (2010). Manipulation of conformational change in proteins by single-residue perturbations. *Biophys. J.*, 99(3):933–943.

[5] Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H., and Gunsalus, I. C. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry*, 14(24):5355–73.

[6] Bae, W., Choi, M. G., Hyeon, C., Shin, Y. K., and Yoon, T. Y. (2013). Real-time observation of multiple-protein complex formation with single-molecule fret. *J. Am. Chem. Soc.*, 135(28):10254–10257.

[7] Bahar, I., Atilgan, A. R., Demirel, M. C., and Erman, B. (1998). Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Phys. Rev. Lett.*, 80:2733–2736.

[8] Bahar, I., Atilgan, A. R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design*, 2(3):173–181.

[9] Bahar, I. and Jernigan, R. (1997). Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.*, 266(1):195–214.

[10] ben-Avraham, D. (1993). Vibrational normal-mode spectrum of globular proteins. *Phys. Rev. B*, 47(21):14559–60.

[11] Bernstein, D. S. (2005). *Matrix mathematics: Theory, facts, and formulas with application to linear systems theory.* Princeton University Press.

[12] Bondi, A. (1964). Van der waals volumes and radii. *J. Phys. Chem.*, 68(3):441–51.

[13] Bossa, C., Amadei, A., Daidone, I., Anselmi, M., Vallone, B., Brunori, M., and Di Nola, A. (2005). Molecular dynamics simulation of sperm whale myoglobin: effects of mutations and trapped co on the structure and dynamics of cavities. *Biophys. J.*, 89(1):465–74.

[14] Bossa, C., Anselmi, M., Roccatano, D., Amadei, A., Vallone, B., Brunori, M., and Di Nola, A. (2004). Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin. *Biophys. J.*, 86(6):3855–62.

[15] Bourgeois, D., Vallone, B., Arcovito, A., Sciara, G., Schotte, F., Anfinrud, P. A., and Brunori, M. (2006). Extended subnanosecond structural dynamics of myoglobin revealed by laue crystallography. *Proc Natl Acad Sci U S A*, 103(13):4924–9.

[16] Bourgeois, D., Vallone, B., Schotte, F., Arcovito, A., Miele, A. E., Sciara, G., Wulff, M., Anfinrud, P., and Brunori, M. (2003). Complex landscape of protein structural

dynamics unveiled by nanosecond laue crystallography. *Proc Natl Acad Sci U S A*, 100(15):8704–9.

[17] Brooks, B. and Karplus, M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA*, 80(21):6571–6575.

[18] Brooks, B. R., Brooks, III, C. L., Mackerell, Jr., A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009). CHARMM: The biomolecular simulation program. *JOURNAL OF COMPUTATIONAL CHEMISTRY*, 30(10, Sp. Iss. SI):1545–1614.

[19] Burioni, R., Cassi, D., Cecconi, F., and Vulpiani, A. (2004). Topological thermal instability and length of proteins. *Proteins*, 55(3):529 – 35.

[20] Cai, S. and Singh, B. R. (1999). Identification of beta-turn and random coil amide iii infrared bands for secondary structure estimation of proteins. *Biophys. Chem.*, 80:7–20.

[21] Cai, S. and Singh, B. R. (2004). A distinct utility of the amide iii infrared band for secondary structure estimation of aqueous protein solutions using partial least squares methods. *Biochemistry*, 43:2541–2549.

[22] Case, D. A. and Karplus, M. (1979). Dynamics of ligand binding to heme proteins. *J Mol Biol*, 132(3):343–68.

[23] Cohen, J., Olsen, K., and Schulten, K. (2008). Finding gas migration pathways in proteins using implicit ligand sampling. *Methods Enzymol*, 437:439–457.

[24] Cuthill, E. and McKee, J. (1969). Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th National Conference*, ACM '69, pages 157–172, New York, NY, USA. ACM.

[25] Doruker, P., Jernigan, R. L., and Bahar, I. (2002). Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.*, 23(1):119–127.

[26] Elber, R. and Karplus, M. (1986). Low frequency modes in proteins: use of effective-medium approximation to interpret fractal dimension observed in electron-spin relaxation measurements. *Phys. Rev. Lett.*, 56:394 – 7.

[27] Elber, R. and Karplus, M. (1990). Enhanced sampling in molecular dynamics: use of the time-dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J Am Chem Soc*, 112:9161–9175.

[28] Eom, K., Baek, S., Ahn, J., and Na, S. (2007). Coarse-graining of protein structures for the normal mode studies. *J. Comput. Chem.*, 28:1400–10.

[29] Etchegoin, P. (1998). Glassylike low-frequency dynamics of globular proteins. *Phys. Rev. E*, 58(1):845 – 8.

[30] Eyal, E. and Bahar, I. (2008). Toward a molecular understanding of the anisotropic response of proteins to external forces: Insights from elastic network models. *Biophys. J.*, 94(9):3424–3435.

[31] Fei, X., Ye, X., LaRonde, N. A., and Lorimer, G. H. (2014). Formation and structures of GroEL:GroES$_2$ chaperonin footballs, the protein-folding functional form. *Proc. Natl. Acad. Sci. USA*, 111(35):12775–12780.

[32] Fowler, D. and Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, 11:801–807.

[33] Frauenfelder, H., McMahon, B. H., Austin, R. H., Chu, K., and Groves, J. T. (2001). The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proc Natl Acad Sci U S A*, 98(5):2370–4.

[34] Freire, E. (1999). The propagation of binding interactions to remote sites in proteins: Analysis of the binding of the monoclonal antibody D1.3 to lysozyme. *Proc. Natl. Acad. Sci. USA*, 96(18):10118–10122.

[35] Fu, F.-N., DeOliveira, D. B., Trumble, W. R., Sarkar, H. K., and Singh, B. R. (1994). Secondary structure estimation of proteins using the amide iii region of fourier transform infrared spectroscopy: Application to analyze calcium-binding-induced structural changes in calsequestrin. *Appl. Spectrosc.*, 48:1432–1441.

[36] Gerek, Z. and Ozkan, S. B. (2011). Change in allosteric network affects binding affinities of pdz domains: Analysis through perturbation response scanning. *PLoS Comput. Biol.*, 7(10):e1002154.

[37] Giraud, G., Karolin, J., and Wynne, K. (2003). Low-frequency modes of peptides and globular proteins in solution observed by ultrafast OHD-RIKES spectroscopy. *Biophys. J.*, 85(3):1903–13.

[38] Go, N., Noguti, T., and Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA*, 80(12):3696–3700.

[39] Goormaghtigh, E., Cabiaux, V., and Ruysschaert, J. M. (1990). Secondary structure and dosage of soluble and membrane proteins by attenuated total reflection fourier-transform infrared spectroscopy on hydrated films. *Eur. J. Biochem.*, 193:409–420.

[40] Hafner, J. and Zheng, W. (2009). Approximate normal mode analysis based on vibrational subsystem analysis with high accuracy and efficiency. *J. Chem. Phys.*, 130:194111.

[41] Hafner, J. and Zheng, W. (2010). Optimal modeling of atomic fluctuations in protein crystal structures for weak crystal contact interactions. *J. Chem. Phys.*, 132(1):014111.

[42] Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, 450(7172):964–972.

[43] Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33(3):417–429.

[44] Hinsen, K. (2006). Normal mode theory and harmonic potential approximations. In Cui, Q. and I. Bahar, ., editors, *Normal Mode Analysis*, chapter 1, pages 1–16. CRC Press.

[45] Hinsen, K. (2008). Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics*, 24(4):521–8.

[46] Hinsen, K. and Kneller, G. R. (2000). Projection methods for the analysis of complex motions in macromolecules. *Mol. Sim.*, 23:275–292.

[47] Hinsen, K., Thomas, A., and Field, M. J. (1999). Analysis of domain motions in large proteins. *Proteins*, 34(3):369–382.

[48] Huang, X. and Boxer, S. G. (1994). Discovery of new ligand binding pathways in myoglobin by random mutagenesis. *Nat Struct Biol*, 1(4):226–9.

[49] Hummer, G., Schotte, F., and Anfinrud, P. A. (2004). Unveiling functional protein motions with picosecond x-ray crystallography and molecular dynamics simulations. *Proc Natl Acad Sci U S A*, 101(43):15330–4.

[50] Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD – Visual Molecular Dynamics. *J Molec Graphics*, 14:33–38.

[51] Izvekov, S. and Voth, G. A. (2005). A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B*, 109:2469–2473.

[52] Jackson, M. B. (2006). *Molecular and Cellular Biophysics.* Cambridge University press.

[53] Jones, D. T., Taylort, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, 358:86–89.

[54] Jr, D. W., Benjamin, D., Poljak, R., and G.S., R. (1996). Global changes in amide hydrogen exchange rates for a protein antigen in complex with three different antibodies. *J Mol Biol.*, 257:866–76.

[55] Karplus, M., Gao, Y. Q., Ma, J., van der Vaart, A., and Yang, W. (2005). Protein structural transitions and their functional role. *Philos. Trans. A Math. Phys. Eng. Sci.*, 363(1827):331–356.

[56] Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9:646–652.

[57] Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., and Wyckoff, H. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*, 181:662–666.

[58] Keskin, O., Bahar, I., Flatow, D., Covell, D. G., and Jernigan, R. L. (2002). Molecular mechanisms of chaperonin GroEL-GroES function. *Biochemistry*, 41(2):491–501.

[59] Kim, M. H., Seo, S., Jeong, J. I., Kim, B. J., Liu, W. K., Lim, B. S., Choi, J. B., and Kim, M. K. (2013). A mass weighted chemical elastic network model elucidates closed form domain motions in proteins. *Protein Sci.*, 22:605–613.

[60] Kondrashov, D. A., Van Wynsberghe, A. W., Bannen, R. M., Cui, Q., and Phillips Jr., G. N. (2007). Protein structural variation in computational models and crystallographic data. *Structure*, 15(2):169–177.

[61] Kotamarthi, H. C., Sharma, R., Narayan, S., Ray, S., and Ainavarapu, S. R. K. (2013). Multiple unfolding pathways of leucine binding protein (lbp) probed by single-molecule force spectroscopy (smfs). *J. Am. Chem. Soc.*, 135(39):14768–14774.

[62] Krimm, S. and Bandekar, J. (1986). Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv. Protein Chem.*, 38:181–364.

[63] Kundu, S., Melton, J. S., Sorensen, D. C., and Phillips Jr., G. N. (2002). Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys. J.*, 83(2):723–732.

[64] Kurkcuoglu, O., Doruker, P., Sen, T. Z., Kloczkowski, A., and Jernigan, R. L. (2008). The ribosome structure controls and directs mrna entry, translocation and exit dynamics. *Phys. Biol.*, 5(4):046005.

[65] Kurkcuoglu, O., Jernigan, R. L., and Doruker, P. (2005). Collective dynamics of large proteins from mixed coarse-grained elastic network model. *QSAR Comb. Sci.*, 24:443–448.

[66] Kurkcuoglu, O., Jernigan, R. L., and Doruker, P. (2006). Loop motions of triosephosphate isomerase observed with elastic networks. *Biochemistry*, 45(4):1173–1182.

[67] Kurkcuoglu, O., Kurkcuoglu, Z., Doruker, P., and Jernigan, R. L. (2009a). Collective dynamics of the ribosomal tunnel revealed by elastic network modeling. *Proteins*, 75(4):837–845.

160

[68] Kurkcuoglu, O., Turgut, O. T., Cansu, S., Jernigan, R. L., and Doruker, P. (2009b). Focused functional dynamics of supramolecules by use of a mixed-resolution elastic network model. *Biophys. J.*, 97:1178–1187.

[69] Lai, H. H., Li, T., Lyons, D. S., Phillips Jr., G. N., Olson, J. S., and Gibson, Q. H. (1995). Phe-46(cd4) orients the distal histidine for hydrogen bonding to bound ligands in sperm whale myoglobin. *Proteins*, 22:322–39.

[70] Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., and Ortiz, A. R. (2005). An analysis of core deformations in protein superfamilies. *Biophys. J.*, 88(2):1291–1299.

[71] Levitt, M. (1980). *Protein Folding*, pages 17–39. Elsevier, North-Holland, Amsterdam.

[72] Levitt, M., Sander, C., and Stern, P. S. (1983). The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *Int. J. Quant. Chem.*, 10:181–199.

[73] Levitt, M., Sander, C., and Stern, P. S. (1985). Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, 181:423–447.

[74] Levitt, M. and Warshel, A. (1975). Computer simulation of protein folding. *Nature*, 253:694–698.

[75] Lezon, T. R., Sali, A., and Bahar, I. (2009). Global motions of the nuclear pore complex: insights from elastic network models. *PLoS Comput. Biol.*, 5(9).

[76] Li, G. and Cui, Q. (2002). A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca(2+)-ATPase. *Biophys. J.*, 83:2457–2474.

[77] Lin, T.-L. and Song, G. (2010). Generalized spring tensor models for protein fluctuation dynamics and conformation changes. *BMC Structural Biology*, 10(Suppl 1):S3+.

[78] Lin, T.-L. and Song, G. (2011). Efficient mapping of ligand migration channel networks in dynamic proteins. *Proteins*, 79(8):2475–2490.

[79] Lu, M. and Ma, J. (2005). The role of shape in determining molecular motions. *Biophys. J.*, 89(4):2395–2401.

[80] Ma, J. (2004). New advances in normal mode analysis of supermolecular complexes and applications to structural refinement. *Curr. Protein Pept. Sci.*, 5:119–123.

[81] Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13(3):373–380.

[82] Ma, J. and Karplus, M. (1998). The allosteric mechanism of the chaperonin GroEL: A dynamic analysis. *Proc. Natl. Acad. Sci. USA*, 95(15):8502–8507.

[83] MacKerell, A. D., Bashford, D., Bellott, Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616.

[84] Maragakis, P. and Karplus, M. (2005). Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.*, 352:807–822.

[85] McCammon, J. A., Gelin, B. R., and Karplus, M. (1977). Dynamics of folded proteins. *Nature*, 267:585–590.

[86] Mendez, R. and Bastolla, U. (2010). Torsional network model: Normal modes in torsion angle space better correlate with conformation changes in proteins. *Phys. Rev. Lett.*, 104(22):228103+.

[87] Ming, D. and Wall, M. E. (2006). Interactions in native binding sites cause a large change in protein dynamics. *J. Mol. Biol.*, 358(1):213–223.

[88] Mittermaier, A. and Kay, L. E. (2006). New tools provide new insights in nmr studies of protein dynamics. *Science*, 312(5771):224–228.

[89] Na, H. and Song, G. (2014a). Bridging between normal mode analysis and elastic network models. *Proteins*, 82:2157–2168.

[90] Na, H. and Song, G. (2014b). A natural unification of GNM and ANM and the role of inter-residue forces. *Phys. Biol.*, 11(3):036002.

[91] Na, H. and Song, G. (2015a). Conventional NMA as a better standard for evaluating elastic network models. *Proteins*, 83:259–267.

[92] Na, H. and Song, G. (2015b). The performance of fine-grained and coarse-grained elastic network models and its dependence on various factors. *Proteins*, 83:1273–1283.

[93] Nevskaya, N. A. and Chirgadze, Y. N. (1976). Infrared spectra and resonance interactions of amide-i and ii vibrations of α-helix. *Biopolymers*, 15:637–648.

[94] Ni, F., Poon, B. K., Wang, Q., and Ma, J. (2009). Application of normal-mode refinement to x-ray crystal structures at the lower resolution limit. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 65(7):633–643.

[95] Noid, W. G., Chu, J. W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., Das, A., and Andersen, H. C. (2008). The multiscale coarse-graining method. i. a rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.*, 128:244114.

[96] Nöllman, M. and Etchegoin, P. (1999). Universal low-frequency vibrations of proteins from a simple interaction potential. *Phys. Rev. E*, 60(4):4593 – 6.

[97] Nutt, D. R. and Meuwly, M. (2004). Co migration in native and mutant myoglobin: atomistic simulations for the understanding of protein function. *Proc Natl Acad Sci U S A*, 101(16):5998–6002.

[98] Olson, J. S., Soman, J., and Phillips Jr., G. N. (2007). Ligand pathways in myoglobin: a review of trp cavity mutations. *IUBMB Life*, 59:552–62.

[99] Ozbek, P., Soner, S., and Haliloglu, T. (2013). Hot spots in a network of functional sites. *PLoS ONE*, 8(9):e74320.

[100] Palmo, K., Mannfors, B., Mirkin, N. G., and Krimm, S. (2003). Potential energy functions: From consistent force fields to spectroscopically determined polarizable force fields. *Biopolymers*, 68:383–394.

[101] Perutz, M. and Mathews, F. (1966). An x-ray study of azide methaemoglobin. *J Mol Biol.*, 21:199–202.

[102] Ponder, J. W. and Richards, F. M. (1987). An efficient newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.*, 8:1016–1024.

[103] Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., and Lindahl, E. (2013). Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854.

[104] Ranson, N. A., Farr, G. W., Roseman, A. M., Gowen, B., Fenton, W. A., Horwich, A. L., and Saibil, H. R. (2001). ATP-bound states of GroEL captured by cryo-electron microscopy. *Cell*, 107:869–879.

[105] Reuveni, S., Granek, R., and Klafter, J. (2008). Proteins: Coexistence of stability and flexibility. *Phys. Rev. Lett.*, 100:208101.

[106] Riccardi, D., Cui, Q., and Phillips Jr., G. N. (2009). Application of elastic network models to proteins in the crystalline state. *Biophys. J.*, 96(2):464–75.

[107] Roseman, A. M., Chen, S., White, H., Braig, K., and Saibil, H. R. (1996). The chaperonin ATPase cycle: mechanism of allosteric switching and movements of substrate-binding domains in GroEL. *Cell*, 87:241–251.

[108] Rueda, M., Chacon, P., and Orozco, M. (2007). Thorough validation of protein normal mode analysis: A comparative study with essential dynamics. *Structure*, 15(5):565–575.

[109] Ruscio, J. Z., Kumar, D., Shukla, M., Prisant, M. G., Murali, T. M., and Onufriev, A. V. (2008). Atomic level computational identification of ligand migration pathways between solvent and binding site in myoglobin. *Proc. Natl. Acad. Sci. USA*, 105(27):9204–9209.

[110] Sacquin-Mora, S. and Lavery, R. (2009). Modeling the mechanical response of proteins to anisotropic deformation. *ChemPhysChem*, 10(1):115–118.

[111] Salomon-Ferrer, R., Case, D., and Walker, R. (2013). An overview of the amber biomolecular simulation package. *WIREs Comput. Mol. Sci.*, 3:198–210.

[112] Savino, C., Miele, A. E., Draghi, F., Johnson, K. A., Sciara, G., Brunori, M., and Vallone, B. (2009). Pattern of cavities in globins: The case of human hemoglobin. *Biopolymers*, 91:1097–1107.

[113] Schmidt, M., Nienhaus, K., Pahl, R., Krasselt, A., Anderson, S., Parak, F., Nienhaus, G. U., and Srajer, V. (2005). Ligand migration pathway and protein dynamics

in myoglobin: a time-resolved crystallographic study on l29w mbco. *Proc Natl Acad Sci U S A*, 102(33):11704–9.

[114] Schotte, F., Lim, M., Jackson, T. A., Smirnov, A. V., Soman, J., Olson, J. S., Phillips, G. N., J., Wulff, M., and Anfinrud, P. A. (2003). Watching a protein as it functions with 150-ps time-resolved x-ray crystallography. *Science*, 300(5627):1944–7.

[115] Scott, E. E. and Gibson, Q. H. (1997). Ligand migration in sperm whale myoglobin. *Biochemistry*, 36(39):11909–17.

[116] Scott, E. E., Gibson, Q. H., and Olson, J. S. (2001). Mapping the pathways for o2 entry into and exit from myoglobin. *J Biol Chem*, 276(7):5177–88.

[117] Sen, T. Z., Feng, Y., Garcia, J. V., Kloczkowski, A., and Jernigan, R. L. (2006). The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models. *J. Chem. Theory Comput.*, 2:696–704.

[118] Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J., and Orengo, C. A. (2015). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.*, 43:D376–D381.

[119] Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mena force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213(4):859–883.

[120] Soheilifard, R., Makarov, D. E., and Rodin, G. J. (2008). Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic b-factors. *Phys. Biol.*, 5(2):026008.

[121] Song, G. and Jernigan, R. L. (2007). vGNM: a better model for understanding the dynamics of proteins in crystals. *J. Mol. Biol.*, 369(3):880–893.

[122] Srajer, V., Ren, Z., Teng, T. Y., Schmidt, M., Ursby, T., Bourgeois, D., Prader-vand, C., Schildkamp, W., Wulff, M., and Moffat, K. (2001). Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved laue x-ray diffraction. *Biochemistry*, 40(46):13802–15.

[123] Srajer, V., Teng, T., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Schild-kamp, W., Bourgeois, D., Wulff, M., and Moffat, K. (1996). Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography. *Science*, 274(5293):1726–9.

[124] Susi, H. and Byler, D. M. (1986). Resolution-enhanced fourier transform infrared spectroscopy of enzymes. *Methods Enzymol.*, 130:290–311.

[125] Taketomi, H., Ueda, Y., and Gō, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. *International Journal of Peptide and Protein Research*, 7(6):445–459.

[126] Tama, F. and Brooks III, C. L. (2006). Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu. Rev. Biophys. Biomol. Struct.*, 35:115–133.

[127] Tama, F., Gadea, F. X., Marques, O., and Sanejouand, Y. H. (2000a). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, 41(1):1–7.

[128] Tama, F., Gadea, F. X., Marques, O., and Sanejouand, Y. H. (2000b). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, 41:1–7.

[129] Tama, F., Miyashita, O., and Brooks, C. L. I. (2004). Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J. Mol. Biol.*, 337(4):985–999.

[130] Tama, F. and Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, 14(1):1–6.

[131] Tasumi, M., Takeuchi, H., Ataka, S., Dwivedi, A. M., and Krimm, S. (1982). Normal vibrations of proteins: Glucagon. *Biopolymers*, 21:711–714.

[132] Tehver, R., Chen, J., and Thirumalai, D. (2009). Allostery wiring diagrams in the transitions that drive the GroEL reaction cycle. *J. Mol. Biol.*, 387:390–406.

[133] Tekpinar, M. and Zheng, W. (2010). Predicting order of conformational changes during protein conformational transitions using an interpolated elastic network model. *Proteins*, 78:2469–2481.

[134] Tenboer, J., Basu, S., Zatsepin, N., Pande, K., Milathianaki, D., Frank, M., Hunter, M., Boutet, S., Williams, G. J., Koglin, J. E., Oberthuer, D., Heymann, M., Kupitz, C., Conrad, C., Coe, J., Roy-Chowdhury, S., Weierstall, U., James, D., Wang, D., Grant, T., Barty, A., Yefanov, O., Scales, J., Gati, C., Seuring, C., Srajer, V., Henning, R., Schwander, P., Fromme, R., Ourmazd, A., Moffat, K., Van Thor, J. J., Spence, J. C., Fromme, P., Chapman, H. N., and Schmidt, M. (2014). Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science*, 346(6214):1242–1246.

[135] Teng, T. Y., Srajer, V., and Moffat, K. (1997). Initial trajectory of carbon monoxide after photodissociation from myoglobin at cryogenic temperatures. *Biochemistry*, 36(40):12087–100.

[136] The Nobel Prize in Chemistry 2013. http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/. [Online; used 19-April-2016; accessed 19-April-2016].

[137] Thorpe, M. F. (2007). Comment on elastic network models and proteins. *Phys. Biol.*, 4:60–3.

[138] Tilton, R.F. Jr, Kuntz, I.D. Jr, and Petsko, G.A. (1984). Cavities in proteins: structure of a metmyoglobin xenon complex solved to 1.9 Å. *Biochemistry*, 23(13):2849–2857. PMID: 6466620.

[139] Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905–1908.

[140] Tirion, M. M. (2015). On the sensitivity of protein data bank normal mode analysis: an application to GH10 xylanases. *Phys. Biol.*, 12(6):066013.

[141] Tirion, M. M. and ben-Avraham, D. (1993). Normal mode analysis of g-actin. *J. Mol. Biol.*, 230(1):186–95.

[142] Tirion, M. M. and ben-Avraham, D. (2015). Atomic torsional modal analysis for high-resolution proteins. *Phys. Rev. E*, 91(8):032712.

[143] Tirion, M. M., ben-Avraham, D., Lorenz, M., and Holmes, K. C. (1995). Normal modes as refinement parameters for the f-actin model. *Biophys. J.*, 68(1):5–12.

[144] Turton, D. A., Senn, H. M., Harwood, T., Lapthorn, A. J., Ellis, E. M., and Wynne, K. (2014). Terahertz underdamped vibrational motion governs protein-ligand binding in solution. *Nat. Commun.*, 5:3999.

[145] Wang, J., Cieplak, P., and Kollman, P. A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21(12):1049–1074.

[146] Wang, Y., Rader, A. J., Bahar, I., and Jernigan, R. L. (2004). Global ribosome motions revealed with elastic network model. *J. Struct. Biol.*, 147(3):302–314.

[147] Wikipedia: Invertible matrix. http://en.wikipedia.org/wiki/Invertible_matrix. [Online; used 17-August-2015; accessed 19-April-2016].

[148] Woodcock, H. L., Zheng, W., Ghysels, A., Shao, Y., Kong, J., and Brooks, B. R. (2008). Vibrational subsystem analysis: A method for probing free energies and correlations in the harmonic limit. *J. Chem. Phys.*, 129:214109.

[149] Wynsberghe, A. W. V. and Cui, Q. (2005). Comparison of mode analyses at different resolutions applied to nucleic acid systems. *Biophys. J.*, 89(5):2939–2949.

[150] Xu, C., Tobi, D., and Bahar, I. (2003). Allosteric changes in protein structure computed by a simple mechanical model: Hemoglobin t↔r2 transition. *J. Mol. Biol.*, 333(1):153–168.

[151] Xu, Z., Horwich, A. L., and Sigler, P. B. (1997). The crystal structure of the asymmetric GroEL-GroES-$(ADP)_7$ chaperonin complex. *Nature*, 388:741–750.

[152] Yang, H., Yang, S., Kong, J., Dong, A., and Yu, S. (2015). Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nat. Protoc.*, 10(3):382–396.

[153] Yang, L., Song, G., and Jernigan, R. L. (2009a). Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci. USA*, 106(30):12347–12352.

[154] Yang, Q. and Sharp, K. A. (2009). Building alternate protein structures using the elastic network model. *Proteins*, 74(3):682–700.

[155] Yang, Z., Májek, P., and Bahar, I. (2009b). Allosteric transitions of supramolecular systems explored by network models: Application to chaperonin GroEL. *PLoS Comput. Biol.*, 5(4):e1000360+.

[156] Yang, Z., Májek, P., and Bahar, I. (2009c). Allosteric transitions of supramolecular systems explored by network models: Application to chaperonin GroEL. *PLoS Comput. Biol.*, 5(4):e1000360.

[157] Yilmaz, L. S. and Atilgan, A. R. (2000). Identifying the adaptive mechanism in globular proteins: Fluctuations in densely packed regions manipulate flexible parts. *J. Chem. Phys.*, 113(10):4454–4464.

[158] Zhang, Z., Pfaendtner, J., Grafmüller, A., and Voth, G. A. (2009). Defining coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models. *Biophys. J.*, 97:2327–2337.

[159] Zheng, W. (2008). A unification of the elastic network model and the gaussian network model for optimal description of protein conformational motions and fluctuations. *Biophys. J.*, 94(10):3853–3857.

[160] Zheng, W. and Brooks, B. (2005). Identification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model. *J. Mol. Biol.*, 346(3):745–759.

[161] Zheng, W., Brooks, B. R., and Thirumalai, D. (2006). Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci. USA*, 103(20):7664–7669.

[162] Zhou, H. and Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, 11(11):2714–2726.

[163] Zhou, L. and Siegelbaum, S. A. (2008). Effects of surface water on protein dynamics studied by a novel coarse-grained normal mode approach. *Biophys. J.*, 94(9):3461–3474.