Understanding protein motions by computational modeling and statistical approaches

Lei Yang
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/rtd
Part of the Bioinformatics Commons, Biophysics Commons, and the Biostatistics Commons

Recommended Citation
Understanding protein motions by computational modeling and statistical approaches

by

Lei Yang

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Robert Jernigan, Co-major Professor
Zhijun Wu, Co-major Professor
Drena Dobbs
Karin Dorman
Vasant Honavar

Iowa State University
Ames, Iowa
2008

Copyright © Lei Yang, 2008. All rights reserved.
I would like to dedicate this thesis to my wife Lifeng You, without whose support I would not have been able to complete this work. I would also like to thank my family and friends for their continuous guidance and assistance during the writing of this work.
# TABLE OF CONTENTS

LIST OF TABLES ................................................................. vii
LIST OF FIGURES ................................................................. viii
ABBREVIATIONS ............................................................... x
ABSTRACT ............................................................... xi

CHAPTER 1. OVERVIEW ......................................................... 1
  1.1 General Introduction .................................................. 1
    1.1.1 X-Ray and NMR Structures ...................................... 1
    1.1.2 Molecular Dynamics (MD) ....................................... 2
    1.1.3 Normal Mode Analysis (NMA) .................................. 3
    1.1.4 Elastic Network Model (ENM) .................................. 3
    1.1.5 Principal Component Analysis (PCA) ......................... 3
  1.2 Organization of Dissertation ....................................... 4

CHAPTER 2. CLOSE CORRESPONDENCE BETWEEN THE PROTEIN MOTIONS FROM PRINCIPAL COMPONENT ANALYSIS OF MULTIPLE HIV-1 PROTEASE STRUCTURES AND ELASTIC NETWORK MODES . 6
  2.1 Abstract ........................................................... 6
  2.2 Introduction ........................................................ 7
  2.3 Methods ............................................................ 9
    2.3.1 Datasets ......................................................... 9
    2.3.2 Principal Component Analysis (PCA) ....................... 10
    2.3.3 Anisotropic Network Model (ANM) ......................... 11
CHAPTER 4. THE EXTENT OF IMPROVEMENT ON PROTEIN B-FACTOR PREDICTION BY COMBINING RIGID BODY MOTIONS AND INTERNAL FLEXIBILITIES

4.1 Abstract

4.2 Introduction

4.2.1 Protein Dataset

4.2.2 Protein Rigid Body Motions by the PFP Model

4.2.3 Protein Internal Flexibility by the CN Models

4.2.4 The Mechanical Model — Gaussian Network Model (GNM)

4.2.5 Comparing Computed B-factors with Experimental Values

4.2.6 Multiple Linear Regression Models

4.2.7 Protein Rigid Motions from the PFP Model

4.2.8 Protein Intrinsic Flexibility from Different CN Models

4.2.9 B-factor Predictions by the GNM

4.2.10 B-factor Predictions by the Regression Models

4.2.11 Two Protein Examples

4.3 Conclusions

CHAPTER 5. COMPARISONS OF EXPERIMENTAL AND COMPUTED PROTEIN ANISOTROPIC TEMPERATURE FACTORS

5.1 Abstract

5.2 Introduction

5.3 Methods

5.4 Results: Prediction of Anisotropic Mean-Square Fluctuations

5.4.1 Magnitude and Anisotropy Prediction Using ANM

5.4.2 Motion Directions Predicted by ANM

5.4.3 Correlation Coefficients between Experimental and Theoretical Results

5.5 Discussions

5.6 Conclusions
CHAPTER 6. HOW WELL CAN WE UNDERSTAND LARGE-SCALE PROTEIN MOTIONS USING NORMAL MODES OF ELASTIC NETWORK MODELS? ................................................................. 72

6.1 Abstract ........................................................................................................... 72
6.2 Introduction ..................................................................................................... 73
6.3 Methods ........................................................................................................... 73
  6.3.1 Protein Dataset ......................................................................................... 73
  6.3.2 Identifying Rigid Domains ....................................................................... 74
  6.3.3 Constructing the Elastic Network with Rigid Domains – Domain-ENM . 78
  6.3.4 The Improved Overlap Measure ............................................................... 78
6.4 Results and Discussion .................................................................................. 81
  6.4.1 Initial Analysis of the Protein Dataset .................................................... 81
  6.4.2 The 170 Transitions Analyzed ............................................................... 81
  6.4.3 How Large a Conformational Change Can Be Predicted by Mode Motions? 84
  6.4.4 Dimensionality Reduction: Proteins Move as Rigid Domains ............ 87
  6.4.5 Where ENM Fails: the Limitation of Using Mode Motions to Study Conformational Transitions ................................................................. 88
6.5 Conclusions .................................................................................................. 93

ACKNOWLEDGEMENTS ..................................................................................... 94

BIBLIOGRAPHY ................................................................................................. 95
LIST OF TABLES

Table 2.1  Comparison of PCs and ENM modes ........................................ 26

Table 3.1  The overlap between the first 5 PCs and 20 modes ....................... 35

Table 6.1  Classification of protein motions in the database ......................... 74

Table 6.2  Conformational transitions by the motion types ........................ 87
LIST OF FIGURES

Figure 2.1  Coarse-grained representations of HIV-1 protease structure ............... 8
Figure 2.2  RMSD with respect to the reference structure ............................... 15
Figure 2.3  The fraction of variance represented by the first 6 PCs .................. 16
Figure 2.4  Distribution of structures along pairs of the first 3 PCs .................. 18
Figure 2.5  Residue positional fluctuations of the first 3 PCs ......................... 20
Figure 2.6  Visualizations of the dominant PCs and the most corresponding ENM modes 21

Figure 3.1  Distribution of residue numbers and model numbers .................... 32
Figure 3.2  Distribution of the fraction of variance explained by the first 5 PCs 33
Figure 3.3  Correlation between NMR pseudo B-factors and ENM predicted B-factors 34
Figure 3.4  Mode index distribution of MO for the first 5 PCs ....................... 36
Figure 3.5  RMSIP between the first 5 PCs and 20 modes ............................. 37
Figure 3.6  Protein examples of predicting NMR revealed dynamics by ENMs ....... 38

Figure 4.1  Correlations between the experimental B-factors and computed B-factors 48
Figure 4.2  The distribution of the optimal power values .............................. 48
Figure 4.3  B-factor predictions using different regression models ................ 50
Figure 4.4  Improvement of B-factor predictions using regression models ......... 51
Figure 4.5  Examples of B-factor predictions by different models ................. 52

Figure 5.1  Comparison of anisotropic B-factor tensors from experiment and theory 60
Figure 5.2  The correlations between experimental and computed isotropic B-factors 63
Figure 5.3  The anisotropy differences between experimental values and predictions 64
Figure 5.4  The anisotropy predictions of two protein examples .................... 66
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5</td>
<td>The directional preference predictions</td>
<td>67</td>
</tr>
<tr>
<td>5.6</td>
<td>The directional preference predictions of two protein examples</td>
<td>68</td>
</tr>
<tr>
<td>5.7</td>
<td>The percentage of residues with ncc above 1 for all the proteins</td>
<td>69</td>
</tr>
<tr>
<td>5.8</td>
<td>The cc/ncc values at the residue level for two protein examples</td>
<td>69</td>
</tr>
<tr>
<td>6.1</td>
<td>Global and infinitesimal local motions</td>
<td>79</td>
</tr>
<tr>
<td>6.2</td>
<td>Distributions of protein sizes, resolutions and RMSDs</td>
<td>82</td>
</tr>
<tr>
<td>6.3</td>
<td>Histogram of the reduced DOF</td>
<td>83</td>
</tr>
<tr>
<td>6.4</td>
<td>Concertedness of conformational transitions</td>
<td>84</td>
</tr>
<tr>
<td>6.5</td>
<td>Maximum overlaps using ENM</td>
<td>86</td>
</tr>
<tr>
<td>6.6</td>
<td>Comparison of domain-ENM with uniform ENM</td>
<td>89</td>
</tr>
<tr>
<td>6.7</td>
<td>Cumulative square overlaps using different models</td>
<td>90</td>
</tr>
<tr>
<td>6.8</td>
<td>Relationship between the overlap and reduced DOF</td>
<td>92</td>
</tr>
</tbody>
</table>
ABBREVIATIONS

ADP: anisotropic displacement parameter
ANM: anisotropic network model
CN: contact number
CO: cumulative overlap
CSO: cumulative square overlap
DOF: degree of freedom
ENM: elastic network model
GNM: Gaussian network model
MD: molecular dynamics
MO: maximum overlap
NMA: normal model analysis
NOE: nuclear Overhauser enhancement
OCN: optimized contact number
PC: principal component
PCA: principal component analysis
PDB: protein data bank
PFP: protein fixed-point
RMSD: root mean-square distance
RMSIP: root mean-square inner product
WCN: weighted contact number
ABSTRACT

Because of its appealing simplicity, the elastic network model (ENM) has been widely accepted and applied to study many molecular motion problems, such as the molecular mechanisms of chaperonin GroEL-GroES function, allosteric changes in hemoglobin, ribosome motions, motor-protein motions, and conformational changes in general. In this dissertation, the ENM is employed to study various protein dynamics problems, and its validity is also examined by comparing with experimental data. First, we apply principal component analysis (PCA) to identify the essential protein motions from multiple structures (X-ray, NMR and MD) of the HIV-1 protease. We find significant similarities between the first few of these key motions and the first few low-frequency normal modes from the ENM, suggesting that the ENM provides a coarse-grained and structurally-based explanation for the experimentally observed conformational changes. Second, we extend these approaches from a single protein (HIV-1 protease) to thousands of proteins whose multiple NMR structures are available. We also find close correspondence between the experimentally observed dynamics and the ENM predicted ones, indicating the validity of using the ENM to computationally predict protein dynamics. Third, we develop a regression model for the isotropic B-factor predictions by combining the protein rigid body motions with the ENM. The new model shows significant improvements in B-factor predictions. Fourth, we further examine the validity of using the ENM to study protein motions. We use the anisotropic form of ENM to predict the anisotropic temperature factors of proteins. It presents a timely and important evaluation of the model, shows the extent of its accuracy in reproducing experimental anisotropic temperature factors, and suggests ways to improve the model. Finally, we apply the ENM to study a dataset of 170 protein pairs having “open” and “closed” structures, and try to address how well a conformational change can be predicted by the ENM and how to improve the model. The results indicate that the applicability of ENM for explaining conformational changes is not limited by either the size of the studied protein
or even the scale of the conformational change. Instead, it depends strongly on how collective the transition is.
CHAPTER 1. OVERVIEW

1.1 General Introduction

Protein motions play important roles in their functions since many protein functions are realized by the conformational changes of their structures. In general, protein motions are not easy to study by experiments, thus, using computational methods to simulate, model and analyze protein motions is of great importance. Based on the available experimental structures, many computational methods including molecular dynamics (MD) and normal mode analysis (NMA) have been widely used for studying many complicated biological problems, such as protein folding, protein transition pathways, signal transduction, and enzyme catalysis.

Protein dynamics is the link between structure and function. It plays important roles in many biological processes, which often require conformational changes of structures. The Protein Data Bank (PDB) [1] provides a valuable resource for the study of protein dynamics, and contains thousands of structures that have been solved by X-ray crystallography, NMR spectroscopy and other techniques. Although more and more structures are being determined and deposited into the PDB, the structures themselves are not fully informative. In order to comprehend the functions of proteins, there is a keen need for a better understanding of their dynamics.

1.1.1 X-Ray and NMR Structures

The Protein Data Bank [1] grows rapidly – as of September 2007, there are over 40,000 protein structures deposited. These structures are mainly determined by two experimental methods, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy [2].

In X-ray crystallography, X-ray diffraction is applied to the crystal grown from the purified protein to develop the electron density map. Then the atomic structure containing the 3-D coor-
coordinates of the protein is solved from the electron density map. The X-ray structure is static and lacks the dynamic properties, though the available multiple structures of a given protein may be considered to be snapshots of protein conformational changes sampling the conformational space. Normally the crystallographers report the thermal B-factors for each atom which is a measure of its disorder, but this is not fully informative about what the variety of structures included actually are.

In NMR spectroscopy, protein structure is determined based on the magnetic property of atomic nuclei. A number of restraints including distance, angle and orientation restraints, are determined experimentally. Computational algorithms are then used to solve the structure by satisfying as many of the restraints as possible. Unlike the static structure determined by X-ray crystallography, NMR results are presented as an ensemble of structures. NMR may also yield information on the dynamic properties of the protein by measuring relaxation times to determine order parameters. This technique can provide detailed information on the 3-D protein structure in solution instead of in the crystalline state.

Although neither X-ray nor NMR can provide a large set of information about protein motions, they nonetheless are very useful at least in providing the initial conformations for the application of computational approaches for the study of protein motions.

1.1.2 Molecular Dynamics (MD)

Molecular Dynamics (MD) [3–5] is a widely used computational method to study protein motions. Using a force field to approximate the potential energy of a given protein, MD can calculate the time-dependent behavior of the molecular system and provide detailed information on the atomic fluctuations and conformational changes of the protein. At present, MD is widely used for modeling various problems such as the thermodynamics of ligand binding and the folding of proteins. MD is also useful for experimental procedures including X-ray crystallography and NMR structure determination. Since a MD trajectory can provide a large set of conformations from a single protein structure, it enables us to study protein motions when only a limited number of structures (or even a single structure) are available.
1.1.3 Normal Mode Analysis (NMA)

Although MD simulations can provide much detailed information about protein motions, the computational cost is significantly high. So a more computationally efficient method called Normal mode analysis (NMA) [6–8], is also popular for protein motion study. In NMA, the protein’s concerted motions are expressed in terms of a set of collective variables – normal modes. Tirion [9] adopted a single-parameter elastic model with Hookean potentials for the pairwise interactions of all atoms in the crystal structures for NMA and was able to obtain large-amplitude elastic motions.

1.1.4 Elastic Network Model (ENM)

Based on Tirion’s simplified NMA, Bahar et al. [10] further simplified the model by using a single-parameter harmonic potential together with a coarse-grained protein model having one point mass per residue. This so-called Gaussian network model (GNM) [11] is more efficient to produce the mode motions at much lower cost of computation. The computed X-ray B-factors are correlated well to the experimental values for many proteins. Instead of just considering the magnitude of fluctuations in GNM, by taking the directionality of fluctuations into account, Atilgan et al. [12] extended the GNM to an anisotropic model, called anisotropic network model (ANM) [12], which permits the directions of motions to be computed. Both GNM and ANM are referred to as elastic network models (ENMs).

1.1.5 Principal Component Analysis (PCA)

Protein motions are complicated systems which generally are of very high dimensionality. Principal component analysis (PCA) is a powerful statistical method for dimensionality reduction based on covariance analysis that can transform the original space of correlated variables into a reduced space of independent variables (i.e., principal components or PCs). By this transformation, most of the system variance can be captured by a few PCs. PCA can be used to analyze the trajectory data from MD simulations. In recent years, this so-called essential dynamics (ED) has been widely used to study protein motions.
1.2 Organization of Dissertation

The objective of this dissertation is to combine all the above methods to reveal the details and properties of protein motions, compare the computational results with experimental ones, and moreover, evaluate the applicability of ENM to study proteins motions. The dissertation is organized as follows.

In Chapter 2, we apply principal component analysis (PCA) to analyze the motions apparent among multiple structures (X-ray, NMR and MD) of the HIV-1 protease. Key motions revealed in the principal components reflect the important dynamic behaviors of the protein. We find significant similarities between the first few of these key motions and the first few low-frequency normal modes from the elastic network model (ENM). The results suggest that ENM provides a coarse-grained, structurally-based explanation for the experimentally observed conformational changes and, a sufficiently large number of experimental structures can directly provide important information about protein dynamics.

In Chapter 3, we extend the similar approaches in Chapter 2 from a single protein (HIV-1 protease) to thousands of proteins whose multiple NMR structures are available. We use PCA to mine the principal dynamics from the NMR ensembles and apply the ENM to predict the protein dynamics. We found that the pseudo B-factors from NMR ensembles are well correlated with those predicted from the ENM and the principal dynamics from the NMR ensembles have high overlap with the normal modes from ENM for many proteins. The results suggest that the NMR ensembles contain valuable information on protein dynamics, and such experimental dynamics can be well captured by the elastic network modes, indicating the validity of using the ENM to computationally predict protein dynamics.

In Chapter 4, we develop a regression model for isotropic B-factor predictions by combining the protein rigid body motions with the ENM (GNM). Using the new model, we obtain significant improvements of B-factor predictions for 1,220 protein structures with high resolution in a large non-redundant dataset. Compared with results from the GNM, the B-factor correlations of over 80% of the proteins are improved by at least 5% with the use of the regression model, suggesting the incorporating of protein rigid body intermolecular motions significantly improves the accuracy
of B-factor predictions.

In Chapter 5, we further examine the validity of using the ENM to study protein motions. In this study, we use the anisotropic form of ENM (ANM) to predict the anisotropic temperature factors of proteins. The rich and directional anisotropic temperature factor data available for hundreds of proteins in the Protein Data Bank (PDB) are used as validation data to closely test the ANM model. The significance of this work is that it presents a timely, important evaluation of the model, shows the extent of its accuracy in reproducing experimental anisotropic temperature factors, and suggests ways to improve the model.

Finally, in Chapter 6, we apply the ENM to study a dataset of 170 pairs having “open” and “closed” structures from Gerstein’s protein motion database [13, 14](http://www.molmovdb.org/), to try to address the following questions: how well can a conformational change be predicted by the mode motions? Is there a way to improve the model to gain better results? Our results show that the conformational transitions fall into three categories: (i) transitions of proteins that can be explained well by ENM, (ii) the transitions that were not explained well by ENM but whose results are significantly improved after considering the rigidity of some residue clusters, and (iii) those transitions having only a low degree of collectivity that prevents their conformational changes from being represented well with the low frequency modes of any elastic network models. Our results thus indicate that the applicability of ENM for explaining conformational changes is not limited by either the size of the studied protein or even the size of the conformational change. Instead, it depends strongly on how collective the transition is.
CHAPTER 2. CLOSE CORRESPONDENCE BETWEEN THE PROTEIN MOTIONS FROM PRINCIPAL COMPONENT ANALYSIS OF MULTIPLE HIV-1 PROTEASE STRUCTURES AND ELASTIC NETWORK MODES

Reference:

2.1 Abstract

The large number of available HIV-1 protease structures provides a remarkable sampling of conformations of the different conformational states, which can be viewed as direct structural information about its dynamics. After structure matching, we apply principal component analysis (PCA) to obtain the important apparent motions, including bound and unbound structures. There are significant similarities between the first few key motions and the first few low-frequency normal modes calculated from a static representative structure with an elastic network model (ENM), strongly suggesting that the variations among the observed structures and the corresponding conformational changes are facilitated by the low-frequency, global motions intrinsic to the structure. Similarities are also found when the approach is applied to an NMR ensemble, as well as to molecular dynamics (MD) trajectories. Thus, a sufficiently large number of experimental structures can directly provide important information about protein dynamics, but ENM can also provide similar sampling of conformations.
2.2 Introduction

In this chapter we present an approach that can be applied to find the essential protein motions from multiple structures of the same protein, in contrast to using just the two “open” and “closed” conformations in Krebs et al.’s and Tama et al.’s studies. To demonstrate our approach, we use HIV-1 protease for the application, an enzyme that plays a critical role in the life cycle of HIV [15], since there are abundant experimentally determined structures, and the size of the protein is relatively small. The HIV-1 protease functions as a homodimer with a single active site and has three domains: the terminal domain (residues 1-4 and 95-99 of each chain), which is important for the dimerization and stabilization of an active HIV-1 protease; the core domain (residues 10-32 and 63-85 of each chain), which is useful for dimer stabilization and catalytic site stability; and the flaps domain, which includes two solvent accessible loops (residues 33-43 of each chain) followed by two flexible flaps (residues 44-62 of each chain), and is important for ligand binding interactions. The conserved ASP25-Thr26-Gly27 active site triad is located at the interface between parts of the core domains. The active site of HIV-1 protease is formed by the homodimer interface and capped by the two flexible flaps. A large conformational change occurs during the process of ligand binding consisting of the opening and closing of the flaps over its binding site. Such principal motions were identified by applying PCA to multiple HIV-1 protease structures, including a set of about 150 crystal structures and a set of conformations generated by MD simulation [16, 17].

Many computational studies of the motions of this protein have been carried out. Zoete et al. [18] performed MD and NMA studies on a dataset containing 73 X-ray structures of HIV-1 protease inhibitor complexes. They found that the backbone RMSD differences of these X-ray structures showed the same variation as those obtained from MD, NMA and reflected in the X-ray B-factors. They also found that inter-domain motions observed from the X-ray dataset agree with those from MD and NMA. These results suggested that the observed structural fluctuations may be used for measuring the intrinsic protein flexibility. Kurt et al. [19] studied the dynamics of HIV-1 protease by using GNM on observed X-ray structures and MD simulated snapshots. They found that the GNM mode motions from different conformations of the HIV-1 protease are conserved along the MD simulations. The conservation of overall dynamic behavior supports the applicability
Figure 2.1 Cartoon representation (a) and alpha carbon trace (b) of the HIV-1 protease structure. Blue – the flap domain; green – the core domain; cyan – the terminal domain; yellow – other residues. The red spheres represent the conserved Asp25-Thr26-Gly27 active site triad. The figure was created using PyMOL (DeLano Scientific).
of GNM for protein motion studies. Chen and Bahar [20] utilized the GNM (a scalar ENM) motions to identify the most conserved residues within three sub-families of proteases.

In the present study, essential motions are first identified by PCA from a large set of X-ray structures of HIV-1 protease, from an NMR ensemble, and from a conformational ensemble generated from an MD simulation. Next, we calculate the normal modes from elastic network model (ENM) using a representative structure closest to the center of each dataset. Significant similarities are found between these essential motions for all three datasets and the low-frequency normal modes calculated from ENM, strongly suggesting that the dynamics encoded in these datasets is facilitated by the low-frequency, global motions that are intrinsic to the structure. ENM thus provides a coarse-grained, structure-based explanation for the experimentally observed conformational changes upon inhibitor binding or the conformational changes found through MD simulations.

2.3 Methods

2.3.1 Datasets

2.3.1.1 X-ray Structure Dataset I (X-ray-I)

The X-ray structures of the HIV-1 protease were downloaded from the Protein Data Bank [1]. Those structures with missing residues are excluded, and the remaining 164 structures form our X-ray dataset. We adopt a coarse-grained simplification in which each C$\alpha$ atom is used to represent its corresponding residue. The representative structure is chosen after aligning all the structures to a reference structure. For the alignment, it matters little which structures are used as the beginning structures since these structures are all quite similar to one another. Since averaging would result in physically unrealistic structures, we use the structure that is the nearest to the average, in this case the PDB 1ebw structure is taken to be the reference structure for subsequent normal mode calculations and MD simulations. The use of slightly different structures for normal mode calculation has little effect upon the results (data not shown). That is due to the insensitivity of the ENM calculations to structural details.
2.3.1.2 X-ray Structure Dataset II (X-ray-II)

As will become clear from the initial analysis of X-ray structures, there are eight X-ray structures, namely 1b6l, 1b6m, 1b6p, 1mtr, 1rq9, 1rv7, 1rpi, and 1aid, which are significantly different from the remainder of the X-ray structures and represent outliers for the PCA. We therefore create a separate slightly smaller dataset named X-ray-II that is a subset of X-ray-I dataset, by excluding these eight outliers. This modified dataset thus contains 164 - 8 = 156 structures. The reference structure is chosen using the same procedure as for the X-ray-I dataset, and it actually leads to the same structure (PDB code: 1ebw) as for the X-ray-I dataset.

2.3.1.3 NMR Structures

One PDB file 1bve including 28 NMR structures of the HIV-1 protease is obtained from the Protein Data Bank [1]. Similarly as for the X-ray case, these NMR structures are aligned and averaged. The structure nearest the average (Number 19 in the ensemble) is used as the reference structure for the normal mode calculation.

2.3.1.4 MD Structures

The initial structure for the MD simulation is taken to be the same as the reference structure (1ebw) of the X-ray dataset. The simulation was performed with the NAMD2 program [21] using the CHARMM27 force field [22]. The simulation was carried out in a TIP3 water box using periodic boundary conditions. Electrostatic interactions were treated with a particle mesh Ewald integration [23, 24]. After 100 ps initial equilibration, the simulation was continued for 10 ns at 300 K and 10,000 structures are collected from the MD trajectory. The structure near the middle of the trajectory is found to be closest to the average of the 10,000 structures (Number 1,583 along the trajectory) is chosen as the reference structure for the normal mode calculation.

2.3.2 Principal Component Analysis (PCA)

PCA is performed on the X-ray, NMR and MD datasets respectively. The input is an $n$ by $p$ coordinate matrix $X$ where $n$ is the number of structures and $p$ is 3 times the number of
residues [16, 17]. Each row in $X$ represents the $C^\alpha$ coordinates of each structure. From $X$ the elements of the covariance matrix $C$ are calculated as

$$c_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle,$$

(2.1)

where averages over the $n$ structures are indicated by the brackets $\langle \rangle$. The covariance matrix $C$ can be decomposed as

$$C = P \Delta P^T,$$

(2.2)

where the eigenvectors $P$ represent the principal components (PCs) and the eigenvalues are the elements of the diagonal matrix $\Delta$. The eigenvalues are sorted in descending order. Each eigenvalue is directly proportional to the variance it captures in its corresponding PC.

### 2.3.3 Anisotropic Network Model (ANM)

ANM is used to calculate the normal modes on the reference structures for the X-ray, NMR and MD datasets. In ANM, the potential energy $V$ is a function of the displacement vector $D$

$$V = \frac{\gamma}{2} DHD^T,$$

(2.3)

where $\gamma$ is the force constant for all spring interactions of residues (here we used a cutoff distance of 13 Å to establish the spring connections between residues), and $H$ is the Hessian matrix containing the second derivatives of the energy function, which is assumed to be harmonic.

For a structure with $n$ residues, the Hessian matrix $H$ contains $n$ by $n$ super-elements of size 3 by 3. The $i^{th}$ super-elements of $H$ is given as

$$H_{ij} = \begin{bmatrix}
\frac{\partial^2 V}{\partial X_i \partial X_j} & \frac{\partial^2 V}{\partial X_i \partial Y_j} & \frac{\partial^2 V}{\partial X_i \partial Z_j} \\
\frac{\partial^2 V}{\partial Y_i \partial X_j} & \frac{\partial^2 V}{\partial Y_i \partial Y_j} & \frac{\partial^2 V}{\partial Y_i \partial Z_j} \\
\frac{\partial^2 V}{\partial Z_i \partial X_j} & \frac{\partial^2 V}{\partial Z_i \partial Y_j} & \frac{\partial^2 V}{\partial Z_i \partial Z_j}
\end{bmatrix},$$

(2.4)

where $X_i$, $Y_i$, and $Z_i$ are the positional components of residue $i$, and $V$ represents the harmonic potential between residues $i$ and $j$, given that residues $i$ and $j$ are in contact and that there is a
Hookean spring connecting them. Thus, \( V \) can be expressed as

\[
V = \frac{\gamma}{2} (s_{ij} - s_{ij}^0)^2 = \frac{\gamma}{2} \left( [(X_j - X_i)^2 + (Y_j - Y_i)^2 + (Z_j - Z_i)^2]^{1/2} - s_{ij}^0 \right)^2, \tag{2.5}
\]

where \( s_{ij}^0 \) is the equilibrium distance between residues \( i \) and \( j \), and \( \gamma \) is the spring constant \([12]\).

The Hessian matrix \( H \) can be decomposed as

\[
H = \Lambda M M^T, \tag{2.6}
\]

where \( \Lambda \) is a diagonal matrix of the eigenvalues and the eigenvectors form the columns of the matrix \( M \). This decomposition generates \( 3n-6 \) normal modes (the first 6 modes account for the rigid body translations and rotations of the system) reflecting the vibrational fluctuations. The eigenvalues are sorted in descending order. Each eigenvalue represents the importance as well as the frequency of the corresponding mode, while the corresponding eigenvector represents the directions and relative magnitudes of the motions of residues.

### 2.3.4 Overlaps between PCs and Normal Modes

The alignment between the directions of a given PC and a given normal mode is measured by their overlap, which was defined by Tama and Sanejouand \([25]\)

\[
O_{ij} = \frac{|P_i \cdot M_j|}{\|P_i\| \cdot \|M_j\|}, \tag{2.7}
\]

where \( P_i \) is the \( i^{th} \) PC and \( M_j \) is the \( j^{th} \) normal mode. A perfect match yields an overlap value as 1. We define the cumulative overlap (CO) between the first \( k \) normal modes and a given PC \( i \) as

\[
CO(k) = \left( \sum_{j=1}^{k} O_{ij}^2 \right)^{1/2}, \tag{2.8}
\]

which measures how well the first \( k \) modes together can capture the motion of a single PC.
2.3.5 Relating the PC and Mode Spaces

The overlap between the motion spaces of the first \( I \) PCs and the first \( J \) low-frequency modes is defined by the root mean-square inner product (RMSIP) \([26, 27]\) as

\[
RMSIP(I, J) = \left( \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{J} (P_i \cdot M_j)^2 \right)^{\frac{1}{2}},
\]

(2.9)

where \( P_i \) is the \( i^{th} \) PC and \( M_j \) is the \( j^{th} \) normal mode. This RMSIP indicates how well the motion space spanned by the first \( I \) PCs is represented by the first \( J \) modes.

2.4 Results and Discussion

2.4.1 The RMSD Distribution in the Three Datasets

The initial X-ray dataset (X-ray-I) contains 164 X-ray structures. The RMSD with respect to the reference structure is shown in Figure 2.2(a). There are 4 structures 1b6l, 1b6m, 1b6p and 1mtr that are especially close to each other (RMSD < 0.22 Å), but quite far from the reference structure (RMSD > 3.31 Å). These 4 structures are complexes bound to macrocyclic peptidomimetic inhibitors. Three structures 1rq9, 1rv7 and 1rpi are close to each other (RMSD < 0.58 Å), but far from the reference structure (RMSD > 1.98 Å). These 3 structures are multidrug-resistant HIV-1 protease. The structure 1aid is 1.40 Å from the reference structure, and is 1.38 Å and 3.81 Å from the average of the groups (1b6l, 1b6m, 1b6p and 1mtr) and (1rq9, 1rv7 and 1rpi), respectively. The aforementioned eight structures, which are the same ones we have excluded by defining the X-ray-II dataset, appear to be quite different from the rest of the structures in their RMS distances to the reference structure. However, the reason why they are considered to be outliers is more evident from the PCA scatter plot analysis in the next section. The structural differences between these outliers and the rest are likely due to the different ligands they bind, the mutational differences or the experimental conditions, etc. For instance, the first group (1b6l, 1b6m, 1b6p and 1mtr) all have a macrocyclic or cyclic inhibitor bound to the enzyme, while the 3 structures in the second group (1rq9, 1rv7 and 1rpi) are multidrug-resistant mutants, each having an expanded active-site cavity. The NMR dataset is an ensemble with 28 conformations. The RMSD with respect to the
reference structure is shown in Figure 2.2(b). MD is carried out using NAMD2 and 10,000 structures are obtained from the MD trajectory. The RMSD of each conformation with respect to the starting structure for the MD simulations is shown in Figure 2.2(c). The RMSD with respect to the reference structure is shown in Figure 2.2(d). So, immediately it can be seen that each of our datasets includes a range of conformations having rather similar extents of deviations from their characteristic conformation.

2.4.2 Dimensionality Reduction by PCA

PCA is performed on the X-ray dataset. The fraction of variance and the cumulative fraction of variance explained by the first 6 PCs are shown in Figure 2.3(a) and (b). It can be seen that the first 2 PCs explain 50% and 16% of the variance respectively and the first 6 PCs together explain over 85% of the variance for X-ray-I dataset. For X-ray-II dataset, the first 2 PCs explain 28% and 15% of the variance respectively and the first 6 PCs together explain over 67%. PCA is also performed on the 28 NMR structures. The fraction of variance and the cumulative fraction of variance explained by the first 6 PCs are shown in Figure 2.3(c). It can be seen that the first 2 PCs explain 38% and 23% of the variance respectively. The first 6 PCs together explain over 79% of the variance. Lastly, PCA is performed on the MD simulated structures. From the fraction of variance and the cumulative fraction of variance plots (Figure 2.3(d)), it can be seen that the first 2 PCs explain 22% and 10% of the variance respectively, and the first 6 PCs together account for about 55% of the variance. The above results indicate that most of the internal motions of the protein can be captured by only a few principal motions (the first several PCs). It is also noted that the first 6 PCs capture variance better for X-ray and NMR structures than for MD structures.

2.4.3 PCA Scatter Plots

The PCA scores can provide a simple overview of all the structures in the dataset. Scatter plots of two PCA scores show the distribution of the actual structure’s deviations from the characteristic structure plotted along the directions of these two PCs. An ideal representation by the PCs will have the structures quite uniformly distributed about the center of these plots. For the X-ray-I dataset, the scatter plot of PC 1 and PC 2 (Figure 2.4(a)) shows that most structures are close to
Figure 2.2 RMSD with respect to the reference structure for: (a) X-ray dataset, with the RMSD values sorted in ascending order. X-ray-II dataset is the same as X-ray-I, excluding the eight structures that have significantly larger RMSD values than the rest. (b) NMR dataset, sorted by the RMSD values in ascending order. (c) MD dataset, shown in the order of the time steps along the 10 ns simulation. (d) MD dataset, sorted by the RMSD values in ascending order.
Figure 2.3 The fraction of variance (‘o’) and cumulative fraction of variance (‘x’) represented by the first 6 PCs for: (a) X-ray-I dataset. (b) X-ray-II dataset. (c) NMR dataset. (d) MD dataset.
the reference structure and are clustered into one group. The classified small groups (1b6l, 1b6m, 1b6p and 1mtr), (1rq9, 1rv7 and 1rpi) and 1aid appear as outliers, which is consistent with their RMSD distributions seen earlier. The scatter plot of PC 1 and PC 3 (Figure 2.4(b)) further confirms the above classification. The scatter plots for the X-ray-II dataset, after excluding the outliers are shown in Figures 2.4(c) and (d). In the NMR case, the scatter plot of PC 1 and PC 2 (Figure 2.4(e)) and the scatter plot of PC 1 and PC 3 (Figure 2.4(f)) show the 28 structures distributed along the 2-PC projection. In the MD case, the scatter plot of PC 1 and PC 2 (Figure 2.4(g)) and the scatter plot of PC 1 and PC 3 (Figure 2.4(h)) show the 10,000 structures (represented by 100 data points) distributed along the 2-PC projection. It is seen that the results from the unpruned X-ray dataset (X-ray-I) are characteristically different from the others, which are more comparable to one another. The first two PCs of the unpruned X-ray dataset mainly reflect the characteristics of those eight outliers whose large RMS deviations enable them to dominate the rest of X-ray structures in influencing the directions of the first two PCs. Therefore, it has been necessary to exclude these and to form a separate dataset X-ray-II in order to identify the key motions of the remaining 156 X-ray structures. Unless otherwise specified, X-ray-II is the dataset we will use for the X-ray structures henceforth.

2.4.4 Identification and Visualization of the Principal Motions

Because most of the protein displacements, in terms of the variance of the structures, can be captured by only a first few principal components (PCs), these PCs can thus be used to characterize the dominant dynamical behaviors of the protein. The X-ray dataset is direct experimental evidence (snapshots) of protein dynamics. PCA enables us to analyze these experimental data and identify a few key directions of motions, i.e., those along the first few PCs. Note that most X-ray structures of HIV-1 protease have some drug molecules bound and thus their conformational displacements reflect the effects of such ligand binding. Therefore, the key directions of motions identified after applying PCA to the X-ray data may provide valuable insights for drug design, such as what the available conformational subspace is, the geometry variance of the binding site, the accessibility of the binding site and the potential pathways for a candidate ligand to reach it [28, 29].

Figure 2.5 shows the residue fluctuations of the first 3 principal motions (the first 3 PCs) of each
Figure 2.4 Distribution of individual structures along pairs of the first three principal component directions. Shown are the planes of PC 1 and PC 2 and of PC 1 and PC 3 for X-ray-I, X-ray-II, NMR and MD datasets respectively. (For the MD dataset, the 10,000 data points are represented by 100 data points by coarse-graining.)
dataset. As mentioned earlier, the first 2 PCs of the original X-ray dataset (X-ray-I) mainly reflect the deviations of the eight outliers (namely 1b6l, 1b6m, 1b6p, 1mtr, 1rq9, 1rv7, 1rpi, and 1aid) and their distinct features of motions. For the PC 1 motion, the second half of each protein chain has significantly larger amplitudes of fluctuations than the first half and is nearly symmetric for the two chains that form the dimer. Since structures 1b6l, 1b6m, 1b6p, and 1mtr have a dominant PC 1 component (see Figure 2.4(a) and (b)), the PC 1 motion mainly reflects their “motions” (or deviations) relative to the reference structure. For the PC 2 motion, there are large amplitudes of fluctuations at the two flaps and is again nearly symmetric for the two chains, which is a feature distinguishing structures 1rq9, 1rv7, 1rpi, and 1aid from the rest. The symmetry between the two chains of this homodimer, however, is much less obvious, sometimes even hardly visible, in the PC 1 and PC 2 fluctuation plots (and higher PCs as well) for the other datasets such as X-ray-II dataset (see Figure 2.5 (X-ray-II)), where the amplitudes of the conformational displacements are much smaller. The decreased data/noise ratio is the main reason for the apparent loss of symmetry. Visualization of the first dominant motion direction (PC 1) of X-ray-II is shown in Figure 2.6(a) together with that of the ENM mode that closely resembles it (see Figure 2.6(b)).

Similarly, PCA is also applied to the NMR ensemble and the MD dataset to identify the key motions. An NMR ensemble can be more advantageous than a single X-ray structure in that it provides more than the mean-square fluctuations of each atom, but also may provide some directional information on protein dynamics. In our case, the NMR ensemble for HIV-1 protease (PDB code: 1bve) includes 28 conformers. A few key directions of motion are revealed and visualized (see Figure 2.6(c) for PC 1), which may represent the dominant motion directions of the protein in solution. Interestingly, the direction of PC 1 aligns extremely well with one mode predicted by ENM, which is shown in Figure 2.6(d). PCA applied to the MD dataset (10,000 structures) reveals the dominant motions of the protein in simulation (see Figure 2.6(e) for the visualization of PC 1 of MD dataset). One advantage of MD is that it can easily be used to generate many structures by computer simulation, but on the other hand to its disadvantage it is difficult to know how well the conformational space is represented or how biased the data may be. However, we also see significant matches between the dominant directions identified by PCA and those calculated from ENM (see Figure 2.6(e) and (f) and more in the next section).
Figure 2.5  Residue positional fluctuations of the first 3 PCs in each dataset. Note that the PC 1 and PC 2 in the X-ray-I dataset have symmetrical fluctuations for the two protein chains (the first chain: residues 1-99; the second chain: residues 100-198). But no symmetrical fluctuations are observed for the X-ray-II, NMR and MD datasets.
Figure 2.6 Visualizations of the motions of the dominant PCs (left column) and the most similar corresponding modes predicted by ENM (right column). In the X-ray-II dataset, the overlap between (a) PC 1 and (b) Mode 2 is 0.52. In the NMR dataset, the overlap between (c) PC 1 and (d) Mode 2 is 0.91. In the MD dataset, the overlap between (e) PC 1 and (f) Mode 1 is 0.74. Blue – the flap domain; green – the core domain; cyan – the terminal domain. The motions of PCs and modes are shown as red sticks with the directions indicated. The stick lengths represent the relative amplitudes of fluctuations of corresponding residue.
It is noted that the fluctuation profiles of the first several PCs are quite different between the datasets (X-ray, NMR and MD), see Figure 2.5. Such differences in the fluctuation profiles reflect the difference in dynamics among the datasets. The principal component axes in one dataset may not perfectly align with those in another dataset. For instance, it is not expected that the PC 1 of an X-ray dataset would match perfectly with the PC 1 in the NMR dataset, but rather it may be expressed as a combination of a few PCs of the NMR dataset. Yet, as will be seen later, these distinct PC profiles can all be described by a set of low-frequency ENM modes. As shown in Table 2.1(c), the subspace of the first several PCs can be well captured by the first several low-frequency ENM modes for all the datasets. This is quite remarkable, and it suggests that the ENM normal modes have captured well the essential motions found in all datasets, although there are some differences in dynamics encoded in the different datasets.

2.4.5 Large Overlaps between PCs and Normal Modes – A Structure-Based Explanation of Observed Motions

The dominant directions of motions represented by the first few PCs have been obtained by direct principal component analysis (PCA) of experimental data (X-ray or NMR) and MD trajectories. In this section, we will investigate whether there are structure-based and physics-based explanations for these directions of motions. In other words, are there intrinsic reasons why these directions of motions are preferred?

For this purpose, we compare these directions of motions with the computationally predicted mode motions by ENM. We calculate the overlaps between the first few PCs and low-frequency modes according to Equation 2.7, for the 3 datasets. In all the cases, we observe some large overlap values between the first several PCs and a few low-frequency modes. The results imply that the observed structures and the corresponding conformational changes are likely facilitated by the low-frequency, global motions that are intrinsic to the structure. ENM thus provides a coarse-grained, structure-based explanation for the experimentally observed conformational changes taking place mostly upon inhibitor binding (for the X-ray structures), as well as for the dynamics revealed from both the NMR ensemble and the simulated MD dataset.

In addition to providing a structure-based explanation for the experimentally observed confor-
mational changes, the mode motions of the protein from ENM can also be used to predict the collective motions of the protein that have not been detected in crystal or NMR structures, and when combined with the experimentally observed conformational changes, can deepen our understanding of the dynamics of the protein, and provide specific information regarding the dynamics in the vicinity of the binding site, e.g., the motion of the flaps. Such an understanding (and visualization) of the dynamics may provide key insights for better ways to design new drugs for protein targets.

2.4.5.1 Matching a Single PC with a Single Mode

The overlaps between the first 3 PCs and the first 3 low-frequency modes (Mode 1-3) are shown in Table 2.1(a). In the X-ray-II dataset, the largest overlap is 0.52, between PC 1 and Mode 2. The overlap between PC 2 and Mode 3 is 0.51. In the NMR dataset, the largest overlap is 0.91, between PC 1 and Mode 2. The overlap between PC 2 and Mode 1 is 0.88. In the MD dataset, the largest overlap is 0.74, between PC 1 and Mode 1. The overlap between PC 3 and Mode 3 is 0.65. These results indicate that the principal motions (i.e., the first few PCs) can be explained well by a single low-frequency normal mode in each of the X-ray, NMR and MD cases.

The largest overlaps found for the first two PCs of the NMR ensemble are highly significant, at 0.91 and 0.88 respectively (see Table 2.1(a)). This significance has two implications. On one hand, as mentioned above, the dynamics revealed from applying PCA to the NMR ensemble yields a structure-based explanation. On the other hand, the NMR ensembles promise improved agreements over the X-ray structures, so that the dynamics revealed may provide an important validation tool of the accuracy of the ENM modes of motion. The large overlaps suggest that the ENM, even though coarse-grained, can capture well the essential dynamics of protein in solution (for the NMR case). In a recent study by Yang et al. [30], they applied GNM to both X-ray structures and NMR ensembles of the same proteins, and find GNM is able to reproduce the residue fluctuations in NMR structures better than that from X-ray structures. These results also support the applicability of ENM to capture the dynamics of NMR structures.

However, we also see that the larger overlap for the third PC of the NMR dataset is far smaller (0.30). This is mainly because there are only 28 structures in the NMR ensemble, which means
that higher PCs may quickly become unreliable. Therefore, a larger ensemble or more ensembles are desired. Unfortunately, there is no other NMR structure available for HIV-1 protease in the Protein Data Bank. A more thorough study using other NMR ensembles of structures is underway.

### 2.4.5.2 Principal Motion (PC) Represented by A Few Modes

Since ENM is a coarse-grained model, it is possible that each individual mode may not be so precise. The details of each normal mode will of course depend on the force field details. However, the subspace of the low-frequency modes is much less affected by such details [31,32], and it has been shown that the overall shape is dominant in determining the motions of the slower modes [33–35]. Therefore, it is worthwhile to determine how well a given principal motion (PC) can be represented by a few low-frequency normal modes collectively. To do so, we calculate the cumulative overlap (CO) for each PC with the subspace defined by the first few low-frequency normal modes.

The results in Table 2.1(b) show that even with 3 modes, overlap values are usually significantly improved. More improvements are gained across the board when the first 20 low-frequency modes are used. The cumulative overlap for PC 3 of the NMR set remains relatively low. As pointed out earlier, this is mainly due to the small size of the NMR ensemble, which renders its high PCs undependable. In summary, the principal motions determined from PCA can be well captured by a small number of low-frequency normal modes.

### 2.4.5.3 Overlaps between PC and Mode Subspaces

The first few PCs collectively capture the majority of the total variance. So the subspace spanned by these PCs reflects the dominant motion space of the protein. To measure how well this motion space can be captured by the first several low-frequency normal modes, we calculated the RMSIP (see Equation 2.9) between the two spaces. Intuitively, RMSIP measures the percentage of the PC subspace that is covered by the subspace spanned by the selected low-frequency modes.

Table 2.1(c) lists the RMSIP values between the subspaces spanned by the first 6 PCs with those spanned by the first 3, 6, and 20 modes. Large RMSIP values are seen even with 3 modes, and marginal improvements are achieved as more modes are included, until the RMSIP values reach about 0.7 (or 70%) when the first 20 low-frequency modes are considered. These results
suggest that the majority of the dynamics displayed in these datasets can be explained by a small set of the ENM modes. This, in addition to ENM’s success in interpreting the crystal B-factors of X-ray structures and the NMR ensembles [30], confirms the validity of using ENM to study protein dynamics. And, these include the dynamics from a broad range of cases, that in crystals, in solution, or from MD simulations.

Though ENMs are coarse-grained models, their usefulness in capturing the collective dynamics of macromolecules has been proved over the last decade. Here we can see again in Table 2.1(c) that the subspace spanned by the first 20 low-frequency modes of the ENM matches quite well with the subspace spanned by the PCs of the X-ray and the NMR structures, as well as that of the MD trajectory.

2.4.6 Significance Test of Overlap Values

To test whether the large overlaps we have obtained in Table 2.1(a) are statistically significant, we have conducted a permutation test. In the following, we carry out a test on the overlap between PC 1 and Mode 2 (0.52) of X-ray-II dataset to demonstrate our approach. In the test, at each iteration, the order of the columns in the coordinate matrix $X$ is permuted randomly. PCA is then performed on the permuted $X$ and the overlap is computed. The simulation is carried out 1,000 times and an empirical distribution of overlaps is generated. This empirical distribution plays the role of the null distribution for hypothesis testing and enables us to estimate the probability of observing an overlap at least as large as the one observed if in fact there were no association between the motion spaces estimated under the two approaches. Based on the simulation the observed value 0.52 is significantly larger than from the permutation test, corresponding to a p-value below 0.0001.

2.5 Conclusions

In this study we have identified the key directions of motion of the HIV-1 protease from crystal structures, in solution, and from MD simulations. This is accomplished by applying PCA to the more than 150 available X-ray structures of the protein, an NMR ensemble (28 models), and the simulated structures generated from a 10 ns MD simulation. These key motions reveal some
Table 2.1 Comparison of PCs and ENM modes. (a) The overlaps between the first 3 PCs and the first 3 low-frequency normal modes. The bold values are the largest values for each dataset. (b) The cumulative overlap (CO) between the first 3 PCs and a set of low-frequency normal modes. (c) The RMSIP between the PC and mode spaces.

<table>
<thead>
<tr>
<th></th>
<th>X-ray-II</th>
<th>NMR</th>
<th>MD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PC1</td>
<td>PC2</td>
<td>PC3</td>
</tr>
<tr>
<td>Mode1</td>
<td>0.46</td>
<td>0.27</td>
<td>0.24</td>
</tr>
<tr>
<td>Mode2</td>
<td><strong>0.52</strong></td>
<td>0.31</td>
<td>0.20</td>
</tr>
<tr>
<td>Mode3</td>
<td>0.17</td>
<td>0.51</td>
<td>0.31</td>
</tr>
</tbody>
</table>

(a) Overlap between a single PC and one mode

<table>
<thead>
<tr>
<th></th>
<th>X-ray-II</th>
<th>NMR</th>
<th>MD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PC1</td>
<td>PC2</td>
<td>PC3</td>
</tr>
<tr>
<td>3 modes</td>
<td>0.71</td>
<td>0.66</td>
<td>0.43</td>
</tr>
<tr>
<td>6 modes</td>
<td>0.74</td>
<td>0.68</td>
<td>0.48</td>
</tr>
<tr>
<td>20 modes</td>
<td>0.84</td>
<td>0.73</td>
<td>0.62</td>
</tr>
</tbody>
</table>

(b) Overlap between one PC and a set of modes

<table>
<thead>
<tr>
<th></th>
<th>X-ray-II</th>
<th>NMR</th>
<th>MD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3 PCs</td>
<td>6 PCs</td>
<td>3 PCs</td>
</tr>
<tr>
<td>3 modes</td>
<td>0.61</td>
<td>0.53</td>
<td>0.78</td>
</tr>
<tr>
<td>6 modes</td>
<td>0.65</td>
<td>0.57</td>
<td>0.80</td>
</tr>
<tr>
<td>20 modes</td>
<td>0.74</td>
<td>0.68</td>
<td>0.83</td>
</tr>
</tbody>
</table>

(c) Overlap between PC and mode subspaces
important dynamic behaviors of the protein and thus should be able to provide valuable new insights for drug design.

Moreover, large overlaps between the first few of these key motions (or PCs) and the first few low-frequency normal modes of ENM are seen, suggesting that the observed structures and the corresponding conformational changes are facilitated by the low-frequency, global motions that are intrinsic to the structure. ENM thus provides a coarse-grained, structure-based explanation for the experimentally observed conformational changes. This, in addition to ENM’s success in interpreting the crystal B-factors of X-ray and NMR structures, confirms its validity for studying protein dynamics. And the dynamics can be that in crystals, or in solutions, or from simulations. Even though the dynamics encoded in these different datasets are not necessarily fully identical, nonetheless the ENM normal modes have been shown to capture well the essential motions found in all of these datasets (see Table 2.1(c)).

Our approach may also help identify which modes contribute most to the functional motions. For example, from the normal mode calculations alone, it cannot be directly established which normal mode is actually the most important one functionally. By using our approach, one may first employ PCA to obtain the principal motions, and then identify the most important normal mode(s) by comparing them with the principal motions - the modes having the largest overlaps being the obvious candidates.
CHAPTER 3. EXPLORING THE DYNAMICS OF PROTEINS FROM NMR STRUCTURES AND VALIDATING ELASTIC NETWORK MODES

Reference:
L. Yang, G. Song and R. L. Jernigan. “Exploring the dynamics of proteins from NMR structures and validating normal modes”, to be submitted.

3.1 Abstract

The wealth of NMR ensemble data in the Protein Data Bank (PDB) provides a valuable resource for the study of protein dynamics. By using the dimensionality reduction technique of principal component analysis (PCA), we mine the dynamics from the NMR ensembles in a large protein dataset. Unlike the experimental isotropic B-factors preserved in X-ray structures, such dynamics from NMR ensembles provides not only information about the magnitudes, but also the directions of the fluctuations. We also apply the elastic network models (ENMs) to predict the protein dynamics. We found that the pseudo B-factors from NMR ensembles are well correlated with those predicted from the ENMs. We also found that the principal dynamics from the NMR ensembles has a high overlap with the normal modes from ENM for many proteins. The results suggest that the NMR ensembles contain valuable information on protein dynamics, and that such experimental dynamics can be well captured by the elastic network modes, indicating that NMR provides a useful validation of the computationally predicted dynamics by ENMs.

3.2 Introduction

An X-ray structure is the static picture of crystals. The only information of dynamics in an X-ray structure is provided by its B-factors. One limitation of such dynamic information is
that the B-factors are generally isotropic and do not provide information about the directions of fluctuations. Although for X-ray structures with ultra-high resolutions, it is possible to measure the anisotropic B-factors, currently, such structures are still rare in the PDB. Another limitation is that the real protein motions normally occur in a solution environment, but the static X-ray structure in the crystals may not reflect the real protein dynamics in solutions and thus may be physically unrealistic.

The NMR structures can be a good complement to the X-ray structures. In NMR spectroscopy, the experimental data include the restraints from nuclear Overhauser enhancement (NOE), dipolar couplings or chemical shifts. From such restraints, the structure calculation (i.e., constrained-MD or simulated annealing) is repeated many times to generate an ensemble of structures, each of which is in agreement with the experimental restraints within an acceptance range of variation. The NMR data are directly derived from the protein in solutions. Thus, the NMR structures provide direct dynamic information on the protein in a solution environment. Moreover, in an NMR ensemble, not only the magnitudes, but also the directions of fluctuations can be obtained directly, which are more informative than the isotropic B-factors from X-ray structures. In the PDB, some NMR data are averaged structures. We suggest that an ensemble of NMR structures provides more dynamics information, and so is better than an averaged NMR structure [36, 37]. Thus, the NMR ensembles can be a valuable resource to mine the important feasible protein dynamics under real solution conditions.

While a significant number of NMR structures have been published in the PDB, until now, a systematic and comprehensive study of their dynamics is still lacking. In this study, we will mine the principal motions from the multiple models of NMR ensembles by principal component analysis (PCA) [38–40]. In the previous chapter, using the similar approach we have already successfully identified the essential motions of HIV-1 protease from multiple structures (X-ray, NMR and MD). After mining the principal dynamics from the NMR ensembles, we will use the ENMs to predict the protein dynamics by normal modes and compare both of them. Moreover, a comparison of the ENM modes with the principal motions from NMR ensembles will provide a way to validate them.
3.3 Methods

3.3.1 NMR Dataset

By March 2007, there were over 5,000 NMR entries deposited in the PDB [1]. Using the limiting criteria that the number of models must be more than 5, the number of residues more than 50 and the sequence identity less than 90%, the 2,257 NMR ensembles meeting the criteria form our NMR dataset. For each ensemble, the multiple structures are aligned and averaged. The structure closest to the average is used as the reference structure for ENM calculation.

3.3.2 Pseudo B-factors of NMR Ensembles

The dynamics of NMR structures can be obtained from the reported models in the ensemble by defining pseudo B-factors that are derived from atomic distances between the individual NMR models and their average [41]. The pseudo B-factors are defined as

\[
B_i^{\text{pseudo}} = k \left\langle \| x_i - \langle x_i \rangle \| ^2 \right\rangle,
\]

where \( x_i \) contains the coordinates of the \( i^{th} \) atom in the \( j^{th} \) NMR model, and \( k \) is a scaling constant. The averages \( \langle \rangle \) are computed over all reported models. For the NMR ensemble, such pseudo B-factors can be viewed as experimental values, analogous to experimental B-factors for the X-ray structures.

3.3.3 Gaussian Network Model

The GNM was used to calculate motion modes of the reference structure. For the GNM, each residue of the protein is represented by its corresponding alpha carbons, and interacts only with other residues within a cutoff distance (here we use 8 Å). For each pair of interacting residues, their connections are each simplified as harmonic forces with equal spring constants. The Kirchhoff
matrix $\Gamma$ is given by:

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

(3.2)

for a pair of residues $i$ and $j$, and $r_c$ is the cutoff distance. The mean-square fluctuations of the individual residue $i$ are found as

$$
\langle u_i^2 \rangle = (3k_BT/\gamma)[\Gamma^{-1}]_{ii}.
$$

(3.3)

Since the inverse of the Kirchhoff matrix can be expressed by the eigenvalue $\lambda_k$ and eigenvector $u_k$ of $\Gamma$, the mean-square fluctuations of residue $i$ can also be written as

$$
\langle u_i^2 \rangle = (3k_BT/\gamma)\sum_k \lambda_k^{-1} u_k u_k^T.
$$

(3.4)

There are a total of $n-1$ modes generated by the GNM (the smallest eigenvalue, representing rigid body motion being zero) with the eigenvalues represent the magnitudes of fluctuations of the modes given in detail by the eigenvectors, without information regarding the directions of fluctuations.

### 3.3.4 Correlation of Predicted B-factors with the NMR Pseudo B-factors

We compare the B-factors predicted by GNM and the pseudo B-factors from the NMR ensemble. Their correlation is given by

$$
corr(B^{exp}, B^{theo}) = \frac{(B^{exp} - \langle B^{exp} \rangle) \cdot (B^{theo} - \langle B^{theo} \rangle)}{\|B^{exp} - \langle B^{exp} \rangle\| \cdot \|B^{theo} - \langle B^{theo} \rangle\|}.
$$

(3.5)

where averages $\langle \rangle$ are computed over the $n$ residues. The B-factors can also be computed by ANM from the decomposition of the Hessian matrix, and the correlation with experimental values is generally slightly worse than for the GNM.

The calculations of anisotropic network model (ANM), principal component analysis (PCA) and overlaps between PCs and modes (single overlap, CO and RMSIP) are the same as those defined
3.4 Results

3.4.1 The NMR Ensemble Dataset

As mentioned in the Methods section, there are 2,267 NMR ensembles in our dataset. Figure 3.1(a) shows the distribution of model numbers and Figure 3.1(b) shows the distribution of residue numbers of NMR ensembles. It can be seen that most of the proteins have residues less than 200 which is due to the incapacity of NMR spectroscopy for large macromolecules. It is also noted that 20 is a typical number of structure models reported, which indicates that most NMR experimentalists report an arbitrary number of models as solutions for the NMR structures.

3.4.2 Dynamics Observed from NMR Ensembles

The NMR ensembles contain direct information on protein dynamics. The dynamics can be interpreted by the pseudo B-factors that have been defined in the Methods section. These pseudo B-factors reflect the magnitudes of the residue fluctuations of the protein. We also applied PCA to the ensemble for identification of the principal motions of the protein. The extent of dimensionality reduction by PCA can be measured by the fraction of total variance explained by the first several
Figure 3.2  Distribution of the fraction of variance explained by the first 5 PCs of the NMR ensembles.

PCs. Our results show that for all the NMR ensembles we have studied, the average fraction of variance explained by the first 5 PCs is 82.8%, with a standard deviation of 9.4% (see Figure 3.2). This dramatic dimensionality reduction indicates that dynamics preserved in the NMR ensembles can be well represented by only a few distinct motions with PCA.

3.4.3 Dynamics Predicted by Elastic Network Modes

3.4.3.1 Prediction of Pseudo B-factors

The GNM is used to predict the B-factors and to compare them with the pseudo B-factors from the NMR ensemble. Figure 3.3(a) shows correlations between the NMR pseudo B-factors and the GNM predicted ones. For most proteins, the GNM is able to reproduce the pseudo B-factors very well. For the whole dataset, the mean correlation is 0.76 with a standard deviation of 0.18. We also use the ANM to predict B-factors and compare with the NMR pseudo B-factors. Figure 3.3(b) shows that the ANM is also good for predicting the NMR pseudo B-factors, giving a mean correlation of 0.70 with a standard deviation of 0.22. It is in agreement with the previous
findings for X-ray structures that the ANM is not as good as the GNM for B-factors predictions [42]. That may be due to the fact that for the ANM, more is required – not only the magnitudes, but also the directions of fluctuations are modeled.

3.4.3.2 Prediction of Principal Motions

The principal motions of the proteins can be identified by the PCA from the NMR ensembles. We use the ANM to calculate the normal modes and compare these modes with the principal motions. In our previous study [43] of HIV-1 protease, we have found that the principal motions can be captured well by the ANM modes. Here we apply the same approaches to our large NMR ensemble dataset.

First, for each of the first 5 PCs, we calculate the overlap between each PC and each ANM mode, and determine which mode gives the maximum overlap (MO, see Eq. (2.7)). From Table 3.1 we can see that for a given PC from the NMR ensemble, there generally is a single ANM mode that has a large overlap with that PC. The means of the MO values for the first 5 PCs are in descending order with the PC index increasing. Figure 3.4 shows the distributions of the indices of the mode giving the MO for each PC. It can be seen that for most of proteins, it is generally the low-index mode giving the MO. There are a large number of proteins for which the first mode gives the MO.
Table 3.1  The MO and CO(20) between the first 5 PCs and the first 20 low-frequency modes.

<table>
<thead>
<tr>
<th></th>
<th>PC 1</th>
<th>PC 2</th>
<th>PC 3</th>
<th>PC 4</th>
<th>PC 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MO</td>
<td>0.59</td>
<td>0.54</td>
<td>0.47</td>
<td>0.43</td>
<td>0.40</td>
</tr>
<tr>
<td>CO</td>
<td>0.87</td>
<td>0.84</td>
<td>0.78</td>
<td>0.75</td>
<td>0.71</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>(0.16)</td>
<td>(0.15)</td>
<td>(0.14)</td>
<td>(0.12)</td>
<td>(0.11)</td>
</tr>
</tbody>
</table>

Second, for each of the first 5 PCs, we calculate the cumulative overlap (CO, see Eq. (2.8)) using first 20 low-frequency modes, to see how well each principal motion (PC) can be represented by a few modes. From Table 3.1 it is seen that each PC can be captured well by the first 20 low-frequency modes. The means of the CO(20) values for the first 5 PCs are in descending order as the PC index increases.

Finally, we calculate the overlap (root mean-square inner product, RMSIP, see Eq. (2.9)) between PC and mode subspaces. Figure 3.5 shows the distribution of the RMSIP between the first 5 PCs and the first 20 low-frequency ANM modes. The mean and standard deviation of the RMSIP values are 0.80 and 0.09 respectively. It is impressive to see that for most proteins, the principal motions (i.e., the first several PCs) directly from the NMR ensembles can be captured well by using only a few low-frequency modes, suggesting the capability for the ANM to predict protein dynamics.

3.4.4 Several Protein Examples

Here we show two protein examples (PDB codes: 1pu3 and 1y9j) to illustrate the dynamics identified from the NMR ensembles and how well such dynamics can be predicted by the elastic network modes. Figure 3.6(a) shows the pseudo B-factors and GNM/ANM reproduced ones for 1pu3. The correlations are 0.82 and 0.72 for GNM and ANM prediction respectively. Figure 3.6(b) shows the pseudo B-factors and GNM/ANM reproduced ones for 1y9j. The correlations are both 0.81 for GNM and ANM prediction. In these cases, the pseudo B-factors from NMR ensembles can be predicted quite well by either the GNM or the ANM.

PCA was applied to these two proteins to identify their principal motions. For 1pu3, the first
Figure 3.4 Mode index distribution of MO for the first 5 PCs. (The last bar represents the number of all the modes with MO index equal to or greater than 20.)
5 PCs can account for 69.3% of the total variance. The maximum overlap for the first PC (PC 1) is 0.62, given by the first low-frequency ANM mode (Mode 1). The CO for PC 1 using 20 low-frequency modes is 0.93. The RMSIP between the first 5 PCs and the first 20 low-frequency modes is 0.84. The eigenvectors of PC 1 and Mode 1 for 1pu3 match with each other well, as shown in Figure 3.6(c). For 1y9j, the first 5 PCs can account for 75.7% of total variance. The maximum overlap for PC 1 is 0.68, given by Mode 1. The CO for PC 1 using 20 low-frequency modes is 0.93. The RMSIP between the first 5 PCs and the first 20 low-frequency modes is 0.78. The eigenvectors of PC 1 and Mode 1 for 1y9j match with each other well, as shown in Figure 3.6(d). These results show that both the pseudo B-factors and the principal motions from NMR ensembles can be predicted well by ENMs (GNM/ANM) for these two proteins.
Figure 3.6 Comparison of dynamics identified from the NMR ensembles and predicted from the ENMs for two proteins. (a) The NMR pseudo B-factors and the GNM predicted ones for 1pu3. (b) The NMR pseudo B-factors and the GNM predicted ones for 1y9j. (c) PC 1 and ANM Mode 1 for 1pu3. (d) PC 1 and ANM Mode 1 for 1y9j.
3.5 Discussion

3.5.0.1 NMR Ensemble vs. Averaged Structure

In the PDB, sometimes NMR structures are deposited as an averaged structure instead of an ensemble of multiple models. One advantage of using such averaged structure to represent the whole data is that the structural features can be visualized much more easily from a single representative structure than from an ensemble. However, there is the risk that the averaged structure is not fully representative or is at least less informative than all the models, or even misleading when highly disordered regions are present. Moreover, during the averaging process, the dynamics information present in the ensemble will be lost. Thus, in our opinion, an ensemble is generally more informative than an average structure for representing the dynamics in the original NMR data.

From the results, we also note that most NMR experimentalists seem to choose an arbitrary number of solved models to deposit in the PDB. Therefore, the completeness of conformational space sampled by the deposited NMR models may greatly affect the quality of the dynamics contained in these models. A thorough and comprehensive examination of the completeness would be helpful for the evaluation of the dynamics. A possible way to avoid the incompleteness is to identify the dynamics directly from the original NMR data (for example, the restraints of NOE data), rather than from the solved NMR models.

3.5.0.2 The Validity of the Elastic Network Modes

During the past decade, the ENMs (both the GNM and the ANM) have been widely used for modeling and analyzing protein dynamics. Although there are many successful cases of using the GNM to reproduce the experimental B-factors, however, a thorough test of the validity of using the elastic network modes to represent real protein dynamics has not yet been carried out. Here, our work demonstrate the significance of the ENMs by comparing the elastic network modes with the real dynamics identified from NMR ensembles. Strong correlations between the NMR pseudo B-factors and the ENM predicted ones and significant overlaps between the principal motions from NMR ensembles and the first few low-frequency elastic network modes indicate that ENMs provide a coarse-grained, structure-based explanation for the experimentally observed protein dynamics,
thus, confirming the validity of using ENMs to study protein dynamics.

3.6 Conclusions

In this work, we have identified the protein dynamics from NMR ensembles. We found that the NMR pseudo B-factors are predicted well by the ENMs, and the principal motions are also captured well by the low-frequency elastic network modes for many proteins. The results suggest that the NMR ensembles can serve as a valuable and informative resource for studying protein dynamics. Moreover, the success of the elastic network modes in capturing the experimental dynamics observed from the NMR ensembles confirms the validity of using the ENMs for protein dynamics studies.
CHAPTER 4. THE EXTENT OF IMPROVEMENT ON PROTEIN B-FACTOR PREDICTION BY COMBINING RIGID BODY MOTIONS AND INTERNAL FLEXIBILITIES

Reference:

4.1 Abstract

The B-factors of protein crystal structures reflect the fluctuations of atoms about their average positions and provide important information about protein dynamics. Computational approaches that are able to predict such thermal motions can be useful for analyzing the dynamical properties of proteins. Previous studies have demonstrated that the success of the coarse-grained Gaussian network model (GNM) on B-factor predictions can be attributed to its reasonable modeling of the intramolecular motions. Other contact number-based models have also been shown useful for B-factor predictions. In this work, taking the rigid body motions of the crystalline proteins into account and combining them with the GNM or other contact-based models, we have developed regression models for predicting protein B-factors. We test our method on a large non-redundant dataset of protein structures of high resolution and obtain significantly better results on B-factor predictions. Compared with results using GNM alone, the B-factor correlations of over 70% of the proteins are improved by at least 5% using the regression model, suggesting protein rigid body motions play a significant role in the observed thermal motions of crystalline proteins. We also find a simple contact number-based model (optimized contact number or OCN) performs better in B-factor prediction than the computational expensive GNM model.
4.2 Introduction

Protein dynamics is the link between protein structures and their functions. Most functions of proteins are generally realized through conformational changes of the structures. In X-ray structures, the information of thermal dynamics is provided by the Debye-Waller temperature factors or B-factors, which are proportional to the mean square fluctuations of atoms in a crystal.

Thus, an accurate prediction of crystalline B-factors may be helpful for understanding the functional dynamic properties of proteins. A number of computational and statistical approaches have been proposed to predict protein B-factors from protein sequence [44–50], atomic coordinates [10, 32, 51–54], and electron density maps [35].

Sequence-based B-factor prediction methods include using a sliding window averaging technique to average the B-factors of neighboring atoms within a window [44,45]. Machine learning techniques such as support vector regression (SVR), support vector machine (SVM) and neural network (NN) methods [47–50] are also sequence-based predictions. The atomic coordinate-based methods [10, 32, 51–54] such as molecular dynamics (MD) [3–5] and normal mode analysis (NMA) [6–8] are also used for B-factor predictions [55–57].

One popular structure-based method for B-factor predictions is the Gaussian network model (GNM) [10, 11, 58], which is a simplified NMA by using a single parameter harmonic potential. In the GNM, the protein motions are expressed as a set of collective variables, known as the normal modes. In the past decade, the GNM has been widely used for B-factor predictions. In many cases, the GNM predicted B-factors are in quite good agreement with experimental B-factors determined by crystallographers [10,51–53].

Kundu et al. [52] studied 113 X-ray protein structures and found that the GNM is able to predict the experimental B-factors well in most cases. Their average correlation between prediction and experiment was about 0.59. Sen et al.’s results have shown that the correlations between experimental B-factors and the GNM predict ones are quite similar at either coarse-grained or atomic levels [54].

Other efforts have been made recently to improve predictions of protein B-factors. Erman did a fitting of B-factors to the experimental values by iteratively changing the spring constants of the
Kirchoff matrix of GNM [59]. One concern is that such a fitting may not have a physical sourced basis, even though the correlation achieved between experiment and prediction may be high. Thus, a proper way of incorporating the intrinsic structural properties into the prediction methods is more appealing and represents a stronger connection to structural reality. In fact, in Kundu et al.’s [52] paper, they have mentioned that by including the crystal contacts and neighbors in the GNM, the predictions are slightly improved. Hinsen [60] developed an elastic network model (ENM) for the whole protein crystal to consider the influence of crystal packing effects on the atomic thermal fluctuations. He found that the crystal packing could affect the atomic fluctuations considerably. Hinsen and Kneller studied the solvent effects on protein dynamics by comparing molecular dynamics simulations of solvated and unsolvated lysozyme [61]. Their results show that solvent effects have significant influence on the low-frequency motions. The above findings suggest that considering intrinsic structural properties such as packing and contact density may be helpful for improving B-factor predictions.

The structural intrinsic packing density can be estimated by the contact numbers (CN) of the structure. The CN models [62, 63] have been used to compute the atomic contact numbers within a given distance threshold. It is found that the inverse of CN is approximately linearly correlated to the protein B-factors [62]. Later Hwang’s group proposed a weighted contact number (WCN) model [64] based on the real values of the contact numbers for each residue. Using the WCN model, they find a good correlation between the calculated WCN and the experimental B-factors. The main difference between the WCN and CN models is that the WCN model does not use any cutoff, but considers the contact between any pair of residues weighted by the inverse of their squared distance. One advantage of the WCN model is that it avoids choosing an arbitrary cutoff distance which may affect the quality of B-factor predictions. Although the GNM is also based on the atomic contacts of the structure, these CN models are different since they do not involve any mechanical modeling as the GNM does. Thus, the computational cost of the CN models is much less than the GNM.

In recent papers by Hwang’s group [65, 66], they have modeled the residue fluctuations as the square distance of a given residue from the centroid (center of mass) of the structure. They found that the correlation between such distance profiles and the experimental B-factors is comparable
to other existing prediction methods. Actually, their so-called protein fixed-point (PFP) model is the same as some previous works [67, 68] for B-factor predictions. The physical meaning of the distance profiles in the PFP model is that they reflect the intramolecular rigid body motions of the proteins. It can be shown that this model is related to the radius of gyration in polymer theory of statistical mechanics [69], which reflects the overall packing density of the protein.

Since both the contact-based models (the CN models and the mechanical GNM model) and the rigid body motion-based PFP model can reasonably predict the experimental protein B-factors, it is of interest to see how much improvement we can gain by combining these two approaches. In this article, we combine the contact-based models (the OCN and the mechanical GNM model) with the rigid body motions of a protein (the PFP model) and develop regression models for B-factor predictions. We test our models using a large non-redundant dataset containing 1,220 X-ray structures. Our results show a significant improvement for B-factor predictions with the regression models – over 70% of the proteins have a gain of at least 5% in their B-factor correlations in comparison with the GNM.

4.2.1 Protein Dataset

We use PDB-PEPRDB [70] to select protein structures from the Protein Data Bank (PDB) [1]. We choose protein structures determined by X-ray crystallography at resolutions better than 2.0 Å and with R-factors better than 0.2. We exclude protein fragments or membrane proteins. All protein sizes are at least 50 residues with sequence identities no higher than 25% and structure similarities differ by more than 10 Å. We exclude structures that only have backbone atoms or alpha carbons. We also remove a structure which does not provide experimental B-factors. Finally, we obtain 1,220 protein structures for our dataset.

4.2.2 Protein Rigid Body Motions by the PFP Model

The protein intramolecular rigid body motions are calculated by the PFP model of the structures. In the PFP model, the residue fluctuations are calculated as the square distance of all structure points from the centroid (center of mass) of the structure.
4.2.3 Protein Internal Flexibility by the CN Models

The protein internal flexibility is obtained directly from the contact information of the protein structures. It can be calculated by different CN models (CN and WCN models). In the CN model, the residue fluctuations are computed as the number of interacting neighboring atoms within a given cutoff distance. In the WCN model, no cutoff distance is used – the atomic fluctuation is inversely proportional to the sum of the weighted contact numbers, which is taken as the inverse of the square distance between any other residues and the given residue.

In the WCN model, the power $p$ of contact distance is set to 2 (see Eq. (4.1)), which may not be the optimal value. We do a fitting to choose the best power to maximize the B-factor correlation (see Eq. 4.5) with the experimental values. This optimized contact number (OCN) is used as the protein intrinsic flexibility to predict the B-factors.

$$B_{i}^{OCN} = \frac{1}{\sum_{j=1,j\neq i}^{n} (r_{ij})^{-p}}$$  \hspace{1cm} (4.1)

where $n$ is the number of residues and $r_{ij}$ is the distance between residues $i$ and $j$.

4.2.4 The Mechanical Model — Gaussian Network Model (GNM)

The GNM [10] is a mechanical model based on the contact information of a protein structure. Given a structure, the GNM simplifies the system by modeling it with its alpha carbons only and attaching springs with uniform constants to all contacting alpha carbon pairs. Alpha carbon pairs are considered to be in contact when their separation distance is smaller than a preset cutoff distance (here we use 7.3 Å). All springs are taken to be at equilibrium for the input structure. One advantage of this approach is that the fluctuations of each atom around its equilibrium position and their cross-correlations can be expressed in analytical forms. To determine the atomic fluctuations, we first calculate the Kirchhoff matrix based on the contact information:

$$\Gamma = \begin{cases} 
-1 & \text{if } i \neq j \text{ and } r_{ij} \leq r_{c} \\
0 & \text{if } i \neq j \text{ and } r_{ij} > r_{c} \\
-\sum_{i,j \neq i} \Gamma_{ij} & \text{if } i = j
\end{cases}$$  \hspace{1cm} (4.2)
where $r_{ij}$ is the distance between atoms $i$ and $j$, and $r_c$ is the cutoff distance. The mean square fluctuations of each atom are given by:

$$< \Delta R_i^2 > = (3k_B T / \gamma)[\Gamma^{-1}]_{ii}$$  \hspace{1cm} (4.3)

Then the theoretical B-factors can be conveniently expressed as:

$$B_i^{GNM} = \frac{8\pi^2}{3} < \Delta R_i^2 >$$  \hspace{1cm} (4.4)

where $\gamma$ is the spring constant. It is straightforward to extend the method to the all atom model, even though there might not be much gained for the increase in complexity [54].

4.2.5 Comparing Computed B-factors with Experimental Values

The correlation between experimental and computed B-factors is given by:

$$corr(B^{exp}, B^{cmp}) = \frac{(B^{exp} - < B^{exp} >) \cdot (B^{cmp} - < B^{cmp} >)}{\|B^{exp} - < B^{exp} >\| \|B^{cmp} - < B^{cmp} >\|}$$  \hspace{1cm} (4.5)

A perfect correlation between the two vectors gives a value of 1 while a perfect anti-correlation gives -1. Here, the computed B-factors can be from the PFP, any CN model, the GNM or the following regression models.

4.2.6 Multiple Linear Regression Models

Although a simple linear least squares fitting can be used to combine contact-based OCN model (or GNM) with the PFP model in B-factor predictions, the disadvantage is that it would not include any interaction effects between the two terms, which might be significant. Thus, the multiple regression model is a better choice. The first regression model (REG1) is built using predicted B-factors from the OCN (eq.(4.1)) and the PFP model for rigid body motions as predictor variables:

$$B_i^{REG1} = \alpha_0 + \alpha_1 B_i^{OCN} + \alpha_2 B_i^{PFP} + \alpha_3 B_i^{OCN} \cdot B_i^{PFP} + \epsilon_i$$  \hspace{1cm} (4.6)
where $\epsilon_i$ is the observational error, and $\beta_j$'s ($j=0,1,2,3$) are the parameters of the model. Since these two predictor variables may not be independent, we also include the interaction effect between the OCN and the PFP in the model. The second regression model (REG2) is similar, by using predicted B-factors from the GNM (eq.(4.4)) and the PFP model for rigid body motions as predictor variables:

$$B_{i}^{\text{REG2}} = \beta_0 + \beta_1 B_i^{\text{GNM}} + \beta_2 B_i^{\text{PFP}} + \beta_3 B_i^{\text{GNM}} \cdot B_i^{\text{PFP}} + \zeta_i$$  \hspace{1cm} (4.7)

where $\zeta_i$ is the observational error, and $\beta_j$'s ($j=0,1,2,3$) are the parameters of the model.

4.2.7 Protein Rigid Motions from the PFP Model

The protein rigid motions are represented by distance profiles from the PFP model. In the PFP model, the percentage of proteins having correlation values higher than 0.5 is about 59%.

4.2.8 Protein Intrinsic Flexibility from Different CN Models

The protein intrinsic flexibility is calculated from the atomic coordinates of the structures by different CN models (CN and WCN). The histogram of B-factor predictions from these models is shown in Figure 4.1. It shows that the WCN model is better for B-factor prediction than the other two models. In the WCN model, about 76% proteins have correlation values over 0.5, while in the CN model, the value is 43%.

As mentioned in the Methods section, the power $p$ in WCN model is 2 which may not be the optimal value. We do a fitting to find the best power for the optimized contact number (OCN) model. From Figure 4.2 it can be seen that the mean of the best power is about 2.3. The results from the OCN model and other CN models (CN and WCN) are compared in Figure 4.1. In the CN and WCN model, the percentage of proteins having correlation values higher than 0.5 is about 43% and 76%, while for the OCN model, this percentage is increased to 84%.

From the above results, it is clearly seen that the OCN model is the best for capture the atomic fluctuations of proteins. Thus, we will use only this model to represent the structural contact-based model.
Figure 4.1 Correlations between the experimental B-factors and the B-factors predicted by the contact-based models (CN, WCN, OCN and GNM).

Figure 4.2 The distribution of the optimal power values ($p$) (the last bar represents all the cases greater than 10).
4.2.9 B-factor Predictions by the GNM

The GNM is also based on the contact information, but it requires mechanical modeling which is computationally more expensive. The correlation between the computed B-factors and the experimental values are computed for the 1,220 protein structures in our dataset. The percentage of proteins having correlation values over 0.5 is about 67%, which is better than the CN model, but not as good as the WCN and OCN models (Figure 4.1).

In the work of Kundu et al. [52], the mean of the correlation between experimental B-factors and the GNM predicted values is 0.59. From our dataset, we obtain a mean correlation of 0.55, which is similar to their results. However, we need to point out that we must be cautious to use such “mean values” to interpret the results. The reason is that each protein structure is not a replicate of another, and they cannot be looked upon as observations of the same population. Thus, their average may not be meaningful. The histogram of the correlation distributions is a more proper way to interpret the results.

4.2.10 B-factor Predictions by the Regression Models

Using the OCN as the contact-based model and the PFP to represent the protein rigid body motions, we build a multiple linear regression model (REG1) based on these two models. The advantage of a regression model over a simple linear fitting is that the OCN and PFP are not necessarily independent of each other, and the effect of their interaction can be significant (see Eq. 4.6). We also build another regression model (REG2) by combining the contact-based mechanical model, the GNM and the PFP model for the rigid body motions (see Eq. 4.7).

The results from the OCN, PFP and REG1 are shown in Figure 4.3(a). It is obvious that the REG1 model gives the best results for B-factor predictions. Over 89% of protein structures have correlations above 0.5 with the REG1 model, while for the OCN and PFP, the values are 84% and 59% respectively. The results from the GNM, PFP and REG2 are shown in Figure 4.3(b). It is seen that the REG2 model gives the best results for B-factor predictions. Over 84% of protein structures have correlations above 0.5 with the REG model, while for the GNM and PFP, the values are 67% and 59% respectively.
Figure 4.3  (a) Distribution of the correlations between the experimental B-factors and those predicted by the OCN, PFP and REG1 models. (b) Distribution of the correlations between the experimental B-factors and those predicted by the GNM, PFP and REG2 models.

To further compare the REG1 and GNM models, we calculate the percentage of B-factor correlations improved by using the REG1 over the OCN model (Figure 4.4(a)). It is shown that for about 29% of protein structures, the REG1 correlation is at least 5% better than the OCN correlation. We also calculate the percentage of B-factor correlations improved by using the REG2 over the GNM (Figure 4.4(a)). It can be seen that for about 70% of protein structures, the REG2 correlation is at least 5% better than the GNM correlation.

It is of interest to see that the PFP is more complementary to the GNM than the OCN. One possible explanation is that the OCN model predicts the B-factors better than the GNM (Figure 4.1), i.e., the OCN captures a larger portion of the atomic fluctuations.

4.2.11 Two Protein Examples

From the results shown in the previous section, we can see that for most of proteins in the dataset, the REG predictions are better than the GNM results. We select the following two proteins as examples to illustrate the extent of improvements on B-factor predictions using the REG model over the GNM. In both cases, the GNM predictions are poor, while the REG model gives significantly better results.
4.2.11.1 Superoxide Reductase

Figure 4.5(a) shows the B-factors predicted by the OCN, GNM, REG1 and REG2 models, along with the experimental values for the superoxide reductase (1vzi:A). The correlation between the experimental B-factors and the OCN predicted ones is 0.66, while the correlation between the experimental B-factors and the REG1 predicted ones is 0.69. The correlation between the experimental B-factors and the GNM predicted ones is 0.26, while the correlation between the experimental B-factors and the REG2 predicted ones is 0.75. We can see significant improvement of the REG2 prediction over the GNM.

4.2.11.2 HU Protein

Figure 4.5(b) shows the B-factors predicted by the OCN, GNM, REG1 and REG2 models, as well as the experimental values for the HU protein from the thermophilic bacterium Bacillus stearothermophilus, BstHU (1hnu:A). The correlation between the experimental B-factors and the OCN predicted values is 0.78, while the correlation between the experimental B-factors and the REG1 predicted values is 0.80. The correlation between the experimental B-factors and the GNM predicted values is 0.44, while the correlation between the experimental B-factors and the REG2
predicted values is 0.69. We can see that the results from the REG2 model corresponds to the experimental results more closely.

4.3 Conclusions

In this work, we combine the protein rigid body motions with the internal flexibilities from the OCN and GNM, and develop regression models for B-factor predictions. For a large non-redundant dataset of protein structures of high resolution, we have obtained better results using the regression models for B-factor predictions. Compared with the GNM results, over 70% of proteins in the dataset gain at least 5% improvement in B-factor correlations, indicating the combining of protein rigid body motions and internal flexibility significantly improves the B-factor predictions. We also observe that the OCN model performs better in B-factor prediction than the mechanical GNM model, though both of them are based on the contact information of the structure. This may account for the fact that only about 29% of protein structures have the REG1 correlation at least 5% better than the OCN correlation. Thus, for B-factor prediction purpose, the regression model combining the PFP and OCN seems to be the best.
CHAPTER 5. COMPARISONS OF EXPERIMENTAL AND COMPUTED PROTEIN ANISOTROPIC TEMPERATURE FACTORS

References:


5.1 Abstract

Because of its appealing simplicity, the anisotropic network model (ANM) has been widely accepted and applied to study many molecular motion problems: such as ribosome motions, the molecular mechanisms of GroEL-GroES function, allosteric changes in hemoglobin, motor-protein motions, and conformational changes in general. However, the validity of the ANM has not been closely examined. In this work, we use ANM to predict the anisotropic temperature factors of proteins obtained from X-ray and NMR data. The rich, directional anisotropic temperature factor data available for hundreds of proteins in the Protein Data Bank (PDB) are used as validation data to closely test the ANM model. The significance of this work is that it presents a timely, important evaluation of the model, shows the extent of its accuracy in reproducing experimental anisotropic temperature factors, and suggests ways to improve the model. An improved model will help us better understand the internal dynamics of proteins, which in turn can greatly expand the usefulness of the models, which has already been demonstrated in many applications.
5.2 Introduction

Functional proteins are not static structures and most of their functions are generally realized through protein motions. It is of great interest to know how these bio-machines work. Understanding the underlying detailed mechanisms can have a broad practical impact.

One of the most intuitive approaches for the study of molecule motions is molecular dynamics (MD) [71, 72]. By using a force field to approximate the atomic interactions of a given protein, MD can compute the time-dependent behavior of the molecular system and provide much detail about the atomic fluctuations and conformational changes of the molecular system being studied. It is an important tool and has been used extensively in protein structure determination and refinement, simulating (un)folding pathways, dynamics and fluctuations of folded proteins, etc. The major challenge in applying MD to study the motions of large macromolecules is the limits of computational power. In general, there is a huge gap between the feasible simulation time duration and the time required for a real biological process to take place, e.g., the folding of a moderately large protein. Moreover, MD is governed by the interactions among the individual atoms and does not explicitly consider the overall concertedness in motion which is commonly seen in the dynamics of folded proteins.

Atomic normal mode analysis (NMA) is an ideal alternative method for the study of the collective motions of proteins. Basically, NMA reflects simple harmonic oscillations about a local energy minimum. To apply NMA, an energy minimization has to be first applied to the input structure. The new, energy-locally-minimized structure may make significant changes from the original structure. After the minimization, the second derivative of the potential energy, the Hessian matrix, has to been calculated and then diagonalized. But there are problems with NMA too, especially with large systems. The necessary initial energy minimization process not only requires time and memory but also can distort the input structure significantly, which casts doubt on the validity of the analysis or the structure. In addition, the diagonalization of the Hessian matrix can become prohibitive as the size of the system increases.

Therefore, a more efficient method was needed in order to study the collective motions of larger systems. Tirion [9] showed that a single-parameter Hookean potential for all the pairwise
interactions between atoms, without the energy minimization step, is able to produce similar low
frequency modes to those from the original NMA. This was a big step forward since it allowed
the direct analysis of crystal coordinates. Bahar et al. [10,11] and Hinsen [58] took the simplifica-
tion one important step further. They demonstrated that a single parameter harmonic potential
together with a simplified protein model that represents each residue by a point mass was able
to produce the correct low frequency normal modes and predict reasonably well the equilibrium
isotropic fluctuations of several proteins. Such models are referred to as elastic network models
(ENM). Specifically, the ENM for isotropic fluctuations is usually called the Gaussian network
model (GNM) [11], where only the magnitudes of the fluctuations are computed. Its anisotropi-
counterpart, where the directions of the collective motions are examined, is called the anisotropic
network model (ANM) [12].

Because of its appealing simplicity and efficiency, ANM has been widely accepted and applied
to study many motion problems: such as ribosome motions [73], the molecular mechanisms of the
GroEL-GroES function [74], allosteric changes in hemoglobin [75], motor-protein motions [76], and
conformational changes in general [77–79].

However, the validity of ANM has not been sufficiently examined. In reproducing the isotropic
B-factors, it had been noticed that ANM actually performs slightly worse than GNM (see [52] for
example), which raised a warning signal. ANM was also used to interpret conformational changes
for some proteins [77], but the data about the conformational changes alone was insufficient to fully
verify the model.

In the present work, we use ANM to predict the anisotropic temperature factors of proteins.
The dataset containing hundreds of proteins with directional anisotropic temperature factors can
be used as validation data to closely test the ANM model. The significance of this work is that it
presents a timely, important evaluation of the model and shows how accurately the experimental
anisotropic temperature factors can be reproduced. It also draws attention to the need for an
improved model to help us better understand the internal dynamics of proteins and expand the
usefulness of the model, which has already been seen in many applications.

Anisotropic B-factors, or anisotropic displacement parameters (ADPs), have become available
recently thanks to improvements in crystallographic data collection techniques that make the de-
termination of atomic or near atomic resolution structures (resolution better than 1.2 Å) available. In the PDB file, these are denoted with ANISOU, followed by six numerical values that are the elements of a symmetric tensor, see PDB data format for details [1] (http://www.pdb.org). As of December 1997, there were only 10 protein structures in the PDB with such entries [80]. By now, however, there are hundreds of protein structures with ANISOU entries. Some recent works have shown the usefulness of normal mode-based methods for predicting [81, 82] and refinement of anisotropic thermal motions in X-ray structures [83, 84].

Besides the high-resolution X-ray structures, NMR ensembles provide another good resource of structural and dynamic information. Recent studies have shown that the dynamics from NMR ensembles is less tainted by the surroundings and agrees better with computational results than that from X-ray data [30, 43], the latter of which may be strongly affected by crystal packing. Here we thus include in our study a dataset containing hundreds of NMR ensembles as well.

5.3 Methods

X-ray Dataset. We choose to include in our dataset all protein crystals with atomic or near-atomic resolution (resolution equal or better than 1.2 Å) currently available in the Protein Data Bank (PDB) that have anisotropic temperature factors, or ANISOU, entries. There were 341 such structures in our dataset.

NMR Dataset. Based on our previous study of NMR ensembles (Yang, Song and Jernigan, unpublished), we select NMR ensembles whose conformers are representative and sufficient in covering the conformational space. Technically, for each ensemble, we check the correlations between the first three principal components (PCs) calculated from all the conformers and those from a reduced number of conformers (half as many, randomly chosen), and only keep the ensembles that have high enough correlations for all the three PCs (PC1: > 0.9, PC2: > 0.8, PC3: > 0.7). We also set the criteria that the number of conformers in each ensemble is no less than 20 and the protein size is no less than 50. Finally, we obtain 455 ensembles to form our NMR dataset.

Isotropic and Anisotropic B-factors. X-ray diffraction data of a protein crystal usually provide information about protein dynamics in the form of isotropic temperature factors $B_i$, which relate
to the mean-square fluctuation $< \Delta R_i^2 >$ of atom $i$ from its average coordinate by:

$$B_i = \frac{(8\pi^2/3)}{3} < \Delta R_i^2 >$$ (5.1)

The $B_i$'s, one for each non-hydrogen heavy atom, is determined by fitting the X-ray diffraction data during the structural determination and refinement process. The fluctuation of atoms, as we know, is generally not isotropic. A more accurate description of the fluctuations is to use the anisotropic B-factors, or anisotropic displacement parameters (ADP). Anisotropic B-factors $B_{aniso}$ are represented as a $3 \times 3$ symmetric tensor $U$ to represent both the magnitude and the directionality of the fluctuations, i.e.,

$$B_{aniso} = \begin{pmatrix}
U_{11} & U_{12} & U_{13} \\
U_{12} & U_{22} & U_{23} \\
U_{13} & U_{23} & U_{33}
\end{pmatrix}$$ (5.2)

In essence, these describe the probability distribution of the electron density using a 3-dimensional Gaussian function. For a fixed probability value, the distribution is ellipsoidal and has a directional preference. The more deformed the shape is from a sphere, the more anisotropic is the fluctuation. We will measure this using a term called anisotropy, to be defined later.

Similarly to the isotropic B-factors, $B_{aniso}$ relates to the fluctuation $\Delta R_i$ of atom $i$ as:

$$B_{aniso}^i = \frac{(8\pi^2)}{3} < \Delta R_i \Delta R_i > = \frac{(8\pi^2)}{3} \begin{pmatrix}
< \Delta x_i^2 > & < \Delta x_i \Delta y_i > & < \Delta x_i \Delta z_i > \\
< \Delta x_i \Delta y_i > & < \Delta y_i^2 > & < \Delta y_i \Delta z_i > \\
< \Delta x_i \Delta z_i > & < \Delta y_i \Delta z_i > & < \Delta z_i^2 >
\end{pmatrix}$$ (5.3)

From the anisotropic B-factors, we can obtain the corresponding isotropic B-factors, since they are related by:

$$B_i = \text{trace}(B_{aniso}^i)/3$$ (5.4)

For NMR ensembles, the pseudo anisotropic B-factors are calculated by averaging the residue fluctuations between all conformer pairs in the ensemble, using Eq.(5.3).

**GNM and ANM.** Given a protein structure, GNM [10] simplifies the system by modeling it with its alpha carbons only and attaching springs with uniform constants to all contacting alpha
carbon pairs. Alpha carbon pairs are considered to be in contact when their separation distance is smaller than a preset cutoff distance, usually 7 to 8 Å. All springs are set at equilibrium for the input structure. One advantage of this approach is that the fluctuations of each carbon around its equilibrium position and their cross-correlations can be expressed in analytical forms. To determine the atomic fluctuations, we first write down the Kirchhoff matrix based on the contact information,

$$\Gamma = \begin{cases} 
-1 & \text{if } i \neq j \text{ and } r_{ij} \leq r_c \\
0 & \text{if } i \neq j \text{ and } r_{ij} > r_c \\
-\sum_{i,j \neq i} \Gamma_{ij} & \text{if } i = j
\end{cases} \quad (5.5)$$

where $r_{ij}$ is the distance between atoms $i$ and $j$, and $r_c$ is the cutoff distance. The mean square fluctuations of each atom and the theoretical B-factors can be conveniently expressed as:

$$< \Delta R_i^2 > = (3k_BT/\gamma)[\Gamma^{-1}]_{ii} \quad (5.6)$$

$$B_i = 8\pi^2 < \Delta R_i^2 > /3 \quad (5.7)$$

where $\gamma$ is the spring constant.

In ANM [12], the counterpart of the $N \times N$ Kirchhoff matrix $\Gamma$ is a $3N \times 3N$ Hessian matrix $H$ (see [12] for details). As a result, the inverse of $H$ contains $N \times N$ super-elements, whereas the $ii^{th}$ super-element of $H^{-1}$, a 3 by 3 matrix, describes the self correlations between the components of $\Delta R_i$, i.e.,

$$< \Delta R_i \cdot \Delta R_i > = (3k_BT/\gamma)H^{-1} \quad (5.8)$$

The coarse-grained alpha-carbon model is normally used for both ANM and GNM. In this work, we set the cutoff distance to be 13 Å for ANM [12] and 7.3 Å for GNM [52].

Now it is straightforward to extend the method to the all atom model, even though there might not be much gain with the increased complexity [54]. It is also easy to treat the backbone contacts, which are covalent bonds, differently by assigning them a larger spring constant. Though it has been shown that this has little effect in reproducing isotropic B-factors [10], it is possible that it might give a more pronounced effect when using ANM to produce anisotropic B-factors. For the
NMR ensembles, the GNM/ANM is applied to the reference structure which is chosen as the one closest to the average among all the conformers in the ensemble [43].

**Calculating Anisotropic B-factors from ANM.** From Eq.(5.8), it is straightforward to obtain theoretical anisotropic B-factors $B_{i}^{\text{theo}}$ by:

$$B_{i}^{\text{theo}} = (8\pi^2) \langle \Delta R_i \Delta R_i \rangle = 8\pi^2 \frac{3k_BT}{\gamma} H_{ii}^{-1}$$  \hspace{1cm} (5.9)

The single parameter $\gamma$ will serve as a scaling factor.

**Comparing Theoretical Anisotropic B-factors with Experimental Data.** Isotropic B-factors are scalars. The most commonly used method for comparing experimental and calculated isotropic B-factors is the correlation between these two arrays. However, anisotropic B-factors are tensors. The comparison of tensors is more complex. A naive comparison of two tensors by converting them to arrays and then calculating their correlation is not appropriate, since the elements of the tensor are not independent [81]. Instead, the following approach is used. Each tensor represents a 3-dimensional distribution, which can be visualized as an ellipsoid. Therefore, comparing two tensors can be done by comparing the two corresponding ellipsoids. We want to compare their size (or magnitude), their shape, and their orientation. To do this, we first diagonalize the tensors. The magnitude and shape are represented by the eigenvalues, while the directional preferences of the fluctuations are captured by the eigenvectors. The three eigenvectors of a tensor represent an orthonormal frame, and the orientations of two ellipsoids can be compared by measuring how the two corresponding orthonormal frames align with one another.

We use five metrics we use in comparing the anisotropic B-factors:

- The magnitude of the fluctuation. For this, we use the trace of the tensors, which is related to the isotropic B-factors.

- The shape of the ellipsoids, or how anisotropic they are. To this end we define two terms: (1) *first anisotropy*, $\kappa$, which is the ratio of the smallest eigenvalue to the largest eigenvalue. The ratio ranges from 0 to 1, with 1 being spherical and 0 being extremely non-spherical; and (2) *second anisotropy*, $\chi$, which is the ratio of the middle eigenvalue to the largest eigenvalue.
Figure 5.1 Comparison of anisotropic B-factor tensors from experiment and theory. The first anisotropy $\kappa$ is defined by the ratio $\lambda_3/\lambda_1$; the second anisotropy $\chi$ is defined as $\lambda_2/\lambda_1$; $\theta$ is the angle between the first principal axes of the two tensors; $\phi$ is the angle between the second principal axes after the first principal axes are aligned.

- The orientation of the ellipsoids, or the directional preference of the fluctuations. For this, we use polar angles: (1) the angle $\theta$, which is the angle between the first principal axes of the two tensors being compared (see Figure 5.1). (2) the angle $\phi$: the angle between the second principal axes after the first are aligned (see Figure 5.1).

The comparison process between the theoretical and experimental anisotropic B-factors of a given protein can be summarized as follows:

1. retrieve experimental anisotropic B-factors from the PDB (ANISOU entries) and calculate the theoretical anisotropic B-factors using ANM (Eq. (5.9));

2. for each residue $i$ ($1 \leq i \leq N$), based on its experimental and theoretical anisotropic tensors, determine $B_i$, $\kappa_i$, $\chi_i$ for both experiment and theory, and $\theta_i$ and $\phi_i$. 
3. for isotropic B-factors, or $B_i$’s, calculate the correlation between experiment and theory by:

$$corr(B_{\text{exp}}, B_{\text{theo}}) = \frac{(B_{\text{exp}} - < B_{\text{exp}} >)}{\|B_{\text{exp}} - < B_{\text{exp}} >\|} \cdot \frac{(B_{\text{theo}} - < B_{\text{theo}} >)}{\|B_{\text{theo}} - < B_{\text{theo}} >\|}$$

(5.10)

A perfect correlation between two vectors gives a value of 1 while a perfect anti-correlation gives -1. Others fall in between.

4. for the first anisotropy $\kappa_i$’s and second anisotropy $\chi_i$’s, calculate the difference between experiment and theory, i.e., set $\Delta \kappa_i = \kappa_{\text{exp}} - \kappa_{\text{theo}}$, and likewise $\Delta \chi_i$ ($i$ is the residue index).

To measure how well overall the first anisotropy (and second anisotropy) is predicted by theory for a given protein, we use $< \Delta \kappa > = \text{mean}(\Delta \kappa_i)$ and its standard deviation $\sigma(\Delta \kappa) = \text{std}(\Delta \kappa_i)$, and express the difference as $< \Delta \kappa > \pm \sigma(\Delta \kappa)$. Similarly $< \Delta \chi > \pm \sigma(\Delta \chi)$ is used for the second anisotropy.

5. Similarly, we use $< \Delta \theta >$ and $< \Delta \phi >$ to measure how well overall the directions of the fluctuations are predicted for a protein.

Another Measure – The Correlation Coefficients for Comparison of Anisotropic B-factors. Besides the above comparison of experimental and predicted anisotropic B-factors, there is another method to compare two tensors [85]. Let $U$ and $V$ be two tensors (anisotropic B-factors), the correlation coefficient between them is derived from their electron-density maps as follows:

$$cc(U, V) = \frac{(\text{det}U^{-1}\text{det}V^{-1})^{1/4}}{[(1/8)\text{det}(U^{-1} + V^{-1})]^{1/2}}$$

(5.11)

The normalized correlation coefficient is given by:

$$ncc(U, V) = \frac{cc[U, (U_{eq}/V_{eq})V]}{cc(U, U_{iso})cc(V, V_{iso})}$$

(5.12)

where $U_{iso}$ and $V_{iso}$ describe a pair of isotropic atoms, with $U_{iso}^{11} = U_{iso}^{22} = U_{iso}^{33} = U_{eq} = \text{trace}(U)/3$ and similarly for $V_{iso}$. This normalized correlation coefficient $ncc$ will be greater than 1 if two atoms described by $U$ and $V$ are more similar to each other than to an isotropic atom, and will be no more than 1 otherwise. Thus, the $ncc$ provides an excellent measure to compare the size, orientation and
direction of two tensors. In practice, a simple ratio of how many atoms in a structure have their normalized correlation coefficient values larger than 1 and the total number of atoms would give a good measure of the quality of an anisotropic B-factor prediction.

5.4 Results: Prediction of Anisotropic Mean-Square Fluctuations

As we discussed in the Methods section, the anisotropic B-factors, or anisotropic displacement parameters (ADPs), are symmetric tensors for each atom. We diagonalize the tensors to find the eigenvalues and principal axes (eigenvectors). The eigenvalues indicate the magnitude of the fluctuations and the shape of the atom displacements, which in general are anisotropic and therefore ellipsoidal instead of spherical. On the other hand, the eigenvectors of a given ADP tensor tell us the directionality of the fluctuation. The fluctuation is usually not isotropic and is biased toward the direction of the principal axis corresponding to the largest eigenvalue (in other words, along the longest axis of the ellipsoid).

For the magnitude/shape of the fluctuation, we look at three terms: the magnitude, which is equivalent to the isotropic B-factors $B_i$’s; the first anisotropy $\kappa$ and the second anisotropy $\chi$, which measure the shape of the atomic displacements.

We perform these comparisons for all the proteins in our dataset and give the results below.

5.4.1 Magnitude and Anisotropy Prediction Using ANM

Isotropic B-factors. The correlation between experimental and calculated isotropic B-factors gives us a good measure of how well a model can reproduce/predict these values. As shown in Figure 5.2, the quality of prediction using ANM is comparable to that from GNM. For the X-ray dataset, the mean correlation obtained by using ANM is about 0.51, which is slightly lower than what is obtained with GNM, about 0.58. For the NMR dataset, it is 0.70 and 0.77 for ANM and GNM respectively. Using either model (ANM or GNM), significantly better correlations are found with NMR dataset, as observed in previous studies [30,43].

Anisotropy Prediction. Figure 5.3(a) shows the mean first anisotropy difference $<\Delta \kappa>$ (see Methods section) between experiment and calculation, for the X-ray dataset. From the figure we
can see that ANM on average is able to predict fairly well the overall level of the first anisotropy. For most proteins, $<\Delta \kappa>$ is within the range of [-0.2,0.2]. However, we see the standard deviation $\sigma(\Delta \kappa)$ is fairly large, about 0.2, and is strikingly similar for all the proteins. This means that for an individual residue, the first anisotropy predicted by ANM on average deviates by about 0.2 from experimental values, for all these proteins. The results for the second anisotropy $\chi$ are similar (see Figure 5.3(b)) – the second anisotropy predicted by ANM also deviates by about 0.2. For most proteins, since the anisotropy distribution among all residues/atoms is roughly normal with a mean value around 0.5 [80] and the mean value for the second anisotropy is about 0.7 based on our calculations, the discrepancy of 0.2 means that the anisotropy predictions of ANM differs from experimental values by about 0.2/0.5 = 40% for the first anisotropy and 0.2/0.7 = 30% for the second anisotropy. For the NMR dataset, the results of first and second anisotropy prediction are shown in Figure 5.3(c) and (d). The results are slightly better than those of the X-ray dataset.

Figure 5.4 shows, at the residue level, the difference between the experimental anisotropies and the values predicted by ANM. For the X-ray structure of the rubredoxin (PDB id: 1IR0), it is seen from Figure 5.4(a) and (b) that the shape of the fluctuation of each residue is reproduced.

Figure 5.2  The correlations between experimental isotropic B-factors and those predicted by GNM (shown in dashed line) and ANM (shown in solid line), for all the proteins in the (a) X-ray dataset and (b) NMR dataset (the results are sorted by the GNM correlation values). The quality of prediction using ANM is quite similar to that of GNM.
Figure 5.3  The anisotropy differences between experimental values and predictions: (a) the first anisotropy difference $\Delta\kappa$ for X-ray dataset, (b) the second anisotropy difference $\Delta\chi$ for X-ray dataset, (c) the first anisotropy difference $\Delta\kappa$ for NMR dataset, and (d) the second anisotropy difference $\Delta\chi$ for NMR dataset.
reasonably well (in terms of the first anisotropy and second anisotropy values). The results for
the NMR ensemble of the poxvirus complement control protein (PDB id: 1E5G) are shown in
Figure 5.4(c) and (d). Again, the residue fluctuations are well reproduced.

5.4.2 Motion Directions Predicted by ANM

Anisotropic B-factors (or ADPs) provide not only the magnitude, but also, of even greater
interest, directional information about atomic fluctuations. Direct comparison between a model
and experimental data can help uncover some further details about atomic fluctuations and identify
collective modes of motion that could be important for function. The experimental anisotropic B-
factors thus can provide more extensive experimental validation of a model, such as ANM. If good
agreement is found, such validations can provide justification for applying a model to study other
aspects of protein dynamics, in order to understand how large scale protein conformation transitions
take place.

As we defined earlier (see Methods section), the $\theta$ value measures the angle between the experi-
mental and calculated directions of fluctuations, while the $\phi$ value measures the rotation needed to
align the two sets of principal axes after their largest axes are aligned (see Figure 5.1). The $<\theta>$
value thus gives an overall estimation of the performance of the model (here ANM) in predicting
the directions of fluctuations for a given protein.

Figure 5.5(a) shows the $<\theta>$ values for the proteins in the X-ray dataset. It is seen from the
figure that $<\theta>$ and $<\phi>$ are consistently quite large, around 50°. Slightly better results are
obtained for the NMR dataset, as shown in Figure 5.5(b), where the average of $<\theta>$ is about 40°.

Using one protein (again the rubredoxin, PDB id: 1IR0) as an example, Figure 5.6(a) shows the
$\theta$ and $\phi$ values of individual residues, specifically the alpha carbons. Since $\theta$ and $\phi$ measure how
well the directions of the fluctuations are predicted and the lower the $\theta$ and $\phi$ values, the better the
prediction, Figure 5.6(a) indicates that the quality of the prediction for the directions of the atomic
fluctuations varies significantly from residue to residue. While for some residues the directions of
fluctuations calculated from ANM match well with those deduced from experimental anisotropic
B-factor data, for many other residues the predictions are quite poor, some even differing by nearly
90 degree. A possible explanation for the latter is that the first principal direction of the fluctua-
Figure 5.4 The values of the the first anisotropy $\kappa$ (a) and the second anisotropy $\chi$ (b) for the alpha carbons of rubredoxin (X-ray structure, PDB id: 1IRO), and the first anisotropy $\kappa$ (c) and the second anisotropy $\chi$ (d) for the alpha carbons of poxvirus complement control protein (NMR ensemble, PDB id: 1E5G). The experimental values are shown in solid lines, while the values predicted by ANM are shown in dashed lines. In ANM, each residue is represented by its alpha carbon.
Figure 5.5 The $<\theta>$ and $<\phi>$ values for all the proteins in the (a) X-ray dataset and (b) NMR dataset. The $<\theta>$ and $<\phi>$ measure how well overall the directions of the fluctuations are predicted for a protein. A perfect prediction renders both $<\theta>$ and $<\phi>$ as 0.

5.4.3 Correlation Coefficients between Experimental and Theoretical Results

The unnormalized and normalized correlation coefficients ($cc$ and $ncc$, see Eqs. (5.11) and (5.12)) are used to compare the experimental anisotropic temperature factors with those predicted by ANM. From Figure 5.7 we can see that for most X-ray structures, the percentage of residues with $ncc$ above 1 (which means the prediction is good) is quite high, with an average value of about 68%. For the NMR dataset, the results are significantly better, with an average value of about 89%. These results demonstrate that there exists high similarity between the experimental (or derived, for NMR case) anisotropic B-factors and ANM predicted ones. And in general, the prediction is more successful for NMR data than for X-ray. The difference likely comes from the crystal packing effects that are not accounted for in the ANM model but exist for X-ray structures. Using X-ray
The \( \theta \) and \( \phi \) values at the residue level: for the alpha carbons of (a) X-ray structure of rubredoxin (PDB id: 1IRO) and (b) NMR ensemble of poxvirus complement control protein (PDB id: 1E5G). The \( \theta \) and \( \phi \) measure how well the directions of the fluctuations are predicted for a given atom/residue. A perfect prediction renders both \( \theta \) and \( \phi \) as 0.

structure (1IRO) and NMR structure (1E5G) again as examples, Figure 5.8 shows the cc and ncc distributions at the residue level. It is seen that most residues have ncc values above 1 for the chosen X-ray structure, while for the selected NMR structure, all of its residues have ncc values larger than 1.

5.5 Discussions

From the comparisons between results from ANM and experimental data shown above, we see that ANM is able to predict moderately well the relative fluctuation magnitudes of individual residues and even their anisotropies. Its prediction of the directional aspect of the fluctuations using the \( \theta \) and \( \phi \) measures, on the other hand, appears to deviate quite significantly from experimental values. However, it is quite likely that many of these deviations result from the artifact of the misalignment of the principal axes, as discussed in Section 5.4.2. Indeed, the results using normalized correlation coefficient as the measure show that most of the predictions are correct (see Section 5.4.3). In particular, the results from NMR dataset are consistently better than those from X-ray data.
Figure 5.7  The percentage of residues with ncc above 1 for all the proteins in the (a) X-ray dataset and (b) NMR dataset.

Figure 5.8  The cc/ncc values at the residue level: for the alpha carbons of (a) X-ray structure of rubredoxin (PDB id: 1IRO) and (b) NMR ensemble of poxvirus complement control protein (PDB id: 1E5G).
A natural question to ask is why this occurs, and then what can be done to improve the model. First, experimental anisotropic B-factors, or anisotropic displacement parameters (ADPs), are found by fitting X-ray diffraction data of protein crystals. These parameters thus may describe static disorder (atomic coordinate differences between unit cells), dynamic disorder (since the diffraction data represent a time average of protein motion), rigid-body motion of the protein, internal motion of the protein, and lastly, refinement errors and uncertainties. And it is not clear how much the internal motion contributes to the total observed fluctuations. It has been proposed that the external rigid-body motions of proteins may contribute up to 60% of the total fluctuations [86,87]. If this is true, ANM, as a coarse-grained model that only considers the internal motion of a protein, may have missed this important component in the comparison with experimental data.

On the other hand, the coarse-grained nature of the ANM itself may account for some of the differences between experimental and theoretical results shown earlier. ANM normally simplifies each residue by representing it with its alpha carbons. It ignores the other atoms on the backbone and even the side chains, which likely strongly influences how atoms fluctuate locally. It also normally does not take any bound ligands into account. ANM uses a uniform spring constant and cutoff distance for every residue/atom. While in reality, the interaction strength may be residue specific and distance/orientation dependent. Therefore, these details may contribute significantly to the anisotropy of the atomic fluctuations. For example, ANM uses the same spring constant for the backbone contacts as for the rest of the contacts. While this has been shown not to affect the isotropic B-factors much, i.e., the fluctuation magnitude [10], one may wonder whether it might have a more pronounced effect on the directional aspects of the fluctuations. Even though a protein molecule in a crystal can be in close contact with other molecules in neighboring cells, ANM usually treats a protein as an isolated molecule and ignores any effects of the crystal environment. It has been shown [32, 52] that including some neighboring effects helps improve such models to some extent.

It is notable that the prediction is better for the NMR data than for the X-ray data. One possible explanation is that the NMR structures are determined in a solution environment that is free from the crystal packing effects that X-ray structures have. Similar phenomena are observed in our previous comparison of NMR and X-ray structures of HIV-1 protease [43] and also reported
in Yang et al.’s work [30].

5.6 Conclusions

In this work, we have used ANM to compute the anisotropic temperature factors of a large set of high resolution protein structures. The rich experimental anisotropic temperature factor data in turn are used as validation data to closely test the ANM model. We employed five terms to compare the experimental and theoretical anisotropic tensors: (1) isotropic B-factors, (2) first anisotropy κ, (3) second anisotropy χ, and (4 and 5) directional preferences θ and φ. As a separate measure of similarity, we also calculated the (normalized) correlation coefficients between the experimental and calculated anisotropic tensors. Our results show that for the X-ray data: (i) the correlation for isotropic B-factors predicted by ANM is about 0.51, (ii) the anisotropy predictions differ from experimental values by about 30% to 40%, (iii) the directions of fluctuations are different by about 50 degrees on average, (iv) using normalized correlation coefficient as the measure, over 68% of the residue anisotropic tensors are predicted well by the ANM. For the NMR data, the prediction results are even better. These results further confirm the validity of ANM for predicting the anisotropic temperature factors. On the other hand, there still exist some differences between the experimental and predicted results, indicating improvements to the model are needed to resolve these differences and to obtain a more accurate understanding of protein motions and dynamics.
CHAPTER 6. HOW WELL CAN WE UNDERSTAND LARGE-SCALE PROTEIN MOTIONS USING NORMAL MODES OF ELASTIC NETWORK MODELS?

Reference:


6.1 Abstract

In this work, we apply a coarse-grained elastic network model (ENM) to study conformational transitions. We address the following questions: how well can a conformational change be predicted by the mode motions? Is there a way to improve the model to gain better results? To answer these questions, we use a dataset of 170 pairs having “open” and “closed” structures from Gerstein’s protein motion database. Our results show that the conformational transitions fall into three categories: (i) the transitions of these proteins can be explained well by ENM, (ii) the transitions were not explained well by ENM but the results are significantly improved after considering the rigidity of some residue clusters and modeling it accordingly, and (iii) the intrinsic nature of these transitions, specifically the low degree of collectivity, prevents their conformational changes from being represented well with the low frequency modes of any elastic network models. Our results thus indicate that the applicability of ENM for explaining conformational changes is not limited by either the size of the studied protein or even the scale of the conformational change. Instead, it depends strongly on how collective the transition is.
6.2 Introduction

The widely used elastic network models (ENMs) are based on a harmonic potential so that the mode motions they produce yield only the small local fluctuations of atoms. Therefore, they are good for reproducing the temperature B-factors of proteins, usually representing small-scale fluctuations, as first demonstrated by Bahar et al., and followed by others [10, 51, 52]. But, are they suitable for understanding the larger-scale molecular motions?

In this chapter, we aim to address several questions. We want to know how large are the conformational changes that can be predicted well with the mode motions? And for the proteins exhibiting poor overlaps between conformational changes and mode motions, is there anything we can do to improve the ENM to gain better results?

To answer these questions, we use a dataset of 170 pairs of “open” and “closed” structures that were obtained from Gerstein’s protein motion database [13, 14](http://www.molmovdb.org/). These protein sizes range widely from tens of residues to near a thousand residues, and their conformational displacements can be as high as 28 Å. Our results show that the conformational transitions of these 170 proteins fall into three categories: (i) the transitions can be explained well by ENM, (ii) the transitions were not explained well by ENM but the results are significantly improved after considering the rigidity of some residue clusters and modeling them accordingly, and (iii) the intrinsic nature of these transitions, i.e., typically having a low degree of collectivity, prevents their being interpreted with the low frequency modes of elastic network models. Our results thus indicate that the applicability of ENM for explaining conformational changes is not limited by either the size of the studied protein or even by the scale of the conformational change. Instead, it depends strongly on how collective the transition is.

6.3 Methods

6.3.1 Protein Dataset

In this study, we use a protein dataset that is obtained from Gerstein’s Macromolecular Movements Database [13, 14] (http://www.molmovdb.org/). There are about 200 pairs of structures in Gerstein’s database, classified by the motion scales and types of pairwise structures. A few
Table 6.1 Classification of protein motions in Gerstein’s Database of Macromolecular Movements. The categories in motion scale and motion type are the same as those used in the Gerstein’s database.

<table>
<thead>
<tr>
<th>Motion Scale</th>
<th>Motion Type</th>
<th># of Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Motions of fragments</td>
<td>A. Motion is predominantly shear</td>
<td>11</td>
</tr>
<tr>
<td>smaller than domains</td>
<td>B. Motion is predominantly hinge</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>C. Motion can not be fully classified at present</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>D. Motion is not hinge or shear</td>
<td>6</td>
</tr>
<tr>
<td>II. Domain motions</td>
<td>A. Motion is predominantly shear</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>B. Motion is predominantly hinge</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>C. Motion can not be fully classified at present</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>D. Motion is not hinge or shear</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>E. Motion involves partial refolding of tertiary structure</td>
<td>6</td>
</tr>
<tr>
<td>III. Larger movements</td>
<td>A. Motion involves an allosteric transition</td>
<td>9</td>
</tr>
<tr>
<td>than domain movements</td>
<td>B. Motion does not involves an allosteric transition</td>
<td>4</td>
</tr>
<tr>
<td>involving the motion of subunits</td>
<td>C. Complex Protein Motions</td>
<td>5</td>
</tr>
</tbody>
</table>

Structures are excluded here since their PDB entries are not specified. The remained 170 pairs of structures are used in our analyses (Table 6.1 lists the number of proteins in each motion category). The number of pairs in each motion category ranges from 2 to 59. The 340 PDB files are downloaded from Protein Data Bank (http://www.pdb.org/). For each pair of structures, the residues not having corresponding partners in both structures are removed and the alpha carbon coordinates are then extracted for further analysis.

6.3.2 Identifying Rigid Domains

Given two experimentally stable structures of a protein, our goal is to identify the relatively most rigid portions between the two structures. A number of computational methods have been developed for this purpose. In Ref. [88], a difference-distance matrix based method was proposed to determine sets of residues such that the distance between any pair of residues within the set has the same distance in the two structures. One drawback of difference-distance based approaches is their low tolerance to the imprecision in the atomic coordinates. To overcome this, Wriggers and
Schulten [89] developed a method that extracts the rigid domains by iteratively superposition of the protein structures. The preserved geometry (which is used to identify domains) defined by such a superposition process is generally insensitive to the local fluctuations of individual atoms. In Ref. [31], Hinsen et al. proposed an approach using the so-called “deformation energy”. The idea is that residues in the rigid regions are hardly deformed. In addition, deformation energy provides a scale of how rigid a certain region of the protein is locally. Once all the rigid residues are identified, they are then clustered to form domains. Here we present a simple method which utilizes RMSD calculations. In this sense it relates most closely to the work by Wriggers and Schulten. The idea is to separate the local fluctuations (intrinsic “noise” in the X-ray or NMR structures) from the global transitions. Since the local fluctuations is on a scale below 1-2 Å, we define a set of residues to be rigid between the two structures if the RMSD between the two corresponding sets of coordinates is below 2 Å. However, there are a significant number of transitions among the 170 pairs of proteins in our dataset whose scale (i.e., the RMSD between the “open” and “closed” forms of the protein) is around 2 Å and or even smaller. For these protein pairs (specifically $scale < 4$ Å), since using a threshold of 2 Å would cause more or less the whole structure to be considered as rigid, we use a smaller threshold that is dependent on the translation scale, which is 1 Å if $2 \leq scale \leq 4$ Å, 0.5 Å if $1 \leq scale \leq 2$ Å, and so on.

For convenience, we make the following definitions.

**Definition 1.** Given two structures of the same protein, a subset of its residues is considered to form a **rigid domain** if the RMSD of that group between the two structures is smaller than a predefined threshold. A **rigid segment** is defined as a rigid group made up of consecutive residues. A smaller threshold is used in searching for rigid segments and is set to be $\frac{3}{4}$ (a parameter) of the threshold set for rigid domain.

The method has two major steps. In the first step, we calculate a set of rigid segments by comparing the two structures. In the second step, we combine the rigid segments as much as possible to form larger rigid groups. We merge two rigid groups together if and only if the combined group is still rigid by the above definition. The iteration continues until no more new rigid groups can be formed. The resulting rigid groups are then identified as the rigid domains. Note that there
are usually residues that don’t belong to any of these rigid groups. They normally fall into the “hinge” regions and are the ones connecting between the rigid groups. They are much more flexible in nature compared to the residues in the rigid groups. For the remainder of the paper, we refer to these as hinge residues.

**Algorithm A**

**Input:** two structures of a protein  
**Output:** a set of non-overlapping rigid domains  
**Steps:**

1. For any $i$ (1 ≤ $i$ ≤ $N$, where $N$ is the number of residues), find the longest rigid segment starting with residue $i$, i.e., find the largest $j$ that $RMSD(X_{\text{open}}(i : j), X_{\text{closed}}(i : j)) < threshold$. Save all these segments in a set by $Q$.

2. Create an empty set $S$.

3. Among all the segments in $Q$, find the longest one, remove it from $Q$ and move it into set $S$. Update the remaining segments in $Q$ so that they don’t overlap with any segment in the set $S$. This means that some segments in $Q$ must be shortened or discarded.

4. Repeat step (3) until the set $Q$ is empty. Return the set $S$.

5. Starting with the segments in the set $S$ as separate rigid groups, iteratively merge them with one another to form larger rigid groups until no new groups can be formed. (At each iteration, a greedy algorithm is applied to select a pair of rigid groups to merge. The selected pair is the one that, once merged, has the smallest RMSD change between the “open” and “closed” structures than any other choice of pairs would yield. The iteration stops when the smallest RMSD found is larger than the preset threshold.)

6. Lastly, “absorb” as many free residues (those not in any rigid group) as possible into the rigid groups. A similar greedy algorithm to that in the previous step is used to select the best rigid group-free residue pair to merge. Again the iteration stops when the selected rigid group would result in a RMSD larger than the preset threshold if absorbing the selected free
residue. The resulting rigid groups are returned as rigid domains and the free residues as hinge residues.

The rigid groups defined by this algorithm are then considered as the rigid domains of the proteins. With such modeling, the degree of freedom (DOF), or simply $\delta$, of a protein is reduced approximately from $\delta_{\text{original}} = 3N$ to $\delta_{\text{reduced}} = 6 \times n_{\text{domain}} + 3 \times n_{\text{hinge}}$, where $N$ is the protein size - the number of residues, $n_{\text{domain}}$ is the number of rigid domains, and $n_{\text{hinge}}$ is the number of hinge residues. Compared with $\delta_{\text{original}}$, $\delta_{\text{reduced}}$ serves as a metric indicating how collective the transition between the “open” and “closed” form is, i.e., the smaller $\delta_{\text{reduced}}$, the more collective the transition is. Indeed, $\delta_{\text{reduced}}/6$ gives an estimate of how many rigid domains there are. In the extreme case when there is just one single rigid domain, the motion of the protein will be fully collective.

We thus define collectivity as:

**Definition 2.** The collectivity, $\chi$, of a protein transition is defined as the inverse of $\delta_{\text{reduced}}/6$, the estimated number of its rigid domains. In short, $\chi = 6/\delta_{\text{reduced}}$.

The collectivity thus defined is unitless and falls into the range of $(0,1]$, where $\chi = 1$ means complete collectivity, while a smaller $\chi$ means the transition is less collective.

We also would like to define a variable to measure on average how many residues move together, or how large the average domain size is. We thus define concertedness as the collectivity scaled by the proteins size.

**Definition 3.** The concertedness of a motion, $\kappa$, is defined as the collectivity $\chi$ times the size of the protein, i.e., $\kappa = N \times \chi$.

Realize that $\kappa = N \times \chi = N \times 6/\delta_{\text{reduced}} = 2 \times \delta_{\text{original}}/\delta_{\text{reduced}}$. Therefore, the concertedness $\kappa$ also measures the extent of reduction in the degrees of freedom.

In the following section, we describe how to build a special kind of ENM, namely Domain-ENM, once we know where the rigid domains and hinge residues are located.
6.3.3 Constructing the Elastic Network with Rigid Domains – Domain-ENM

In Ref. [90], we presented a new way for constructing elastic network for domain-swapped proteins which is called Domain-ENM. In Domain-ENM, we assign a larger spring constant for intra-domain contacts. This conveniently and effectively encodes domain rigidity with a single parameter. It also enables rigid body domain motions to be separated from the low amplitude fluctuations of each rigid domain, thereby making the dominant rigid body domain motions more easily captured than with uniform ENMs.

Another way to incorporate the rigidity is to use the block normal mode analysis (BNMA) or the rotation-translation block (RTB) method [77, 91]. These methods normally work by modeling a small number of consecutive residues (e.g., 6 residues) as a rigid block. To adapt such methods to our case where the residues within a rigid cluster are not necessarily consecutive in sequence, one may artificially reorder the residues to treat them as if they were consecutive. After the vibration modes or the fluctuation patterns of each residue are obtain, one can reconstruct the modes so that they reflect the original residue sequence order.

6.3.4 The Improved Overlap Measure

The commonly used definition of “overlap” [25, 92] is a measure of the similarity between the direction of global conformational displacement and the direction given by one normal mode, that is,

\[ O_{j}^{\text{original}} = \frac{|M_j \cdot \Delta X|}{\|M_j\| \cdot \|\Delta X\|}, \]

(6.1)

where \(M_j\) is the \(j^{th}\) eigenvector and \(\Delta X\) is the displacement between the “open” and “closed” forms after the two structures are superimposed.

However, the global conformational displacement is a finite motion, whereas the mode motions are infinitesimal motions. The two are not directly comparable, especially when large-scale rotations are involved. In such a case, the initial motion direction, which is comparable with the mode motions, may little resemble what is depicted in the global conformational displacement (illustrated in Figure 6.1) [90].

In light of this, in Ref. [90] we proposed a new measure for calculating overlaps for domain-
Figure 6.1 A simple illustration of the pathway difference for a global conformational change $\Delta X$ and the initial moving direction $\Delta X_0$ when translation is utilized to represent a rotation, as a rigid stick rotates counterclockwise $90^\circ$ from position A to B.
swapped proteins. This improved overlap definition was originally designed for domain-swapped proteins with two distinct domains, but it can easily be extended to systems consisting of multiple rigid domains. For such a system, the global conformational change for each domain can always be expressed as

$$\Delta X^{(i)} = R(k_i, \theta_i) \cdot X_i + T_i - X_i \quad 1 \leq i \leq N_r,$$

(6.2)

where $T_i, R(k_i, \theta_i)$ are the translation and rotation for the $i^{th}$ domain, $k_i$ and $\theta_i$ are the rotational axis and rotational angle, $X_i$ contains the coordinates of the residues in the $i^{th}$ domain relative to its center of mass, and $N_r$ is the number of rigid domains. In order to make a fair comparison with the infinitesimal motions of the modes, we use an infinitesimal motion extracted from the global conformational changes in Eq. 6.2. In other words, we use

$$\Delta X^{(i)}_0 = R(k_i/\lambda, \theta_i/\lambda) \cdot X_i + T_i/\lambda - X_i \quad 1 \leq i \leq N_r,$$

(6.3)

as the infinitesimal version of the global conformational displacement, where $\lambda$ is a large positive number corresponding to the step size (e.g., $\lambda = 100$). For any residue $m$ that is not in any domain, we use plain linear interpolation. Now we form a new directional vector $V$ to obtain an approximate overlap measure. For each residue, the motion direction is:

$$V(m) = \begin{cases} \Delta X^{(i)}_0 & \text{if residue } m \text{ is in domain } i \\ (X_{closed}(m) - X_{open}(m))/\lambda & \text{otherwise}, \end{cases}$$

(6.4)

and hence the overlap is:

$$O_{\text{improved}}^j = \frac{|V \cdot M_j|}{\|V\| \cdot \|M_j\|}.$$  

(6.5)

$O_{\text{improved}}^j$ measures how well the two directions, the initial moving direction $\Delta X_0$ and the direction of the $j^{th}$ mode $M_j$, line up, by calculating the cosine of the angle between them. A perfect agreement in directions corresponds to an overlap value of 1.

Based on the above overlap definition, we define the maximum overlap between a conformational displacement with any mode as:

$$O_{\text{max}} = \max_j (O_j).$$  

(6.6)
We also define the cumulative square overlap (CSO) of the first \( k \) vibrational modes as:

\[
CSO(k) = \sum_{j=1}^{k} O_j^2.
\]  

(6.7)

While maximum overlap indicates how the best mode overlaps with the conformational displacement, it is often helpful to use CSO of the first \( k \) modes to measure how well the first \( k \) modes together can capture the whole transition.

6.4 Results and Discussion

6.4.1 Initial Analysis of the Protein Dataset

The histogram of protein sizes is shown in Figure 6.2(a). From the figure we can see that the sizes of the 170 pairs of proteins fall over a wide range, from tens of residues to near a thousand residues. Out of the total of 340 protein structures in our dataset, 34 are NMR structures. The resolutions for the remaining 306 X-ray structures are shown in Figure 6.2(b) giving a mean of 2.28 Å and a standard deviation of 0.48 Å. The histogram of pairwise RMSDs is shown in Figure 6.2(c). It can be seen that more than 50% of pairs of structures have a RMSD value within 4 Å.

6.4.2 The 170 Transitions Analyzed

Before we apply any mode analysis method to interpret the transitions, it is important for us to analyze these transitions first in order to gain a better understanding of the characteristics of these transitions, especially the collectivity (Definition 2). This is because for all mode analysis methods, from fine-grained all-atom models to coarse-grained models that, for example, represent each residue with its alpha carbon only (as is usually with ENM), they all aim to describe the motions using a small number of collective variables, i.e., the low frequency modes. In order for a motion to be well described with a small number of collective variables, it is necessary that the motion is intrinsically highly collective.

While neither the displacement between the “open” and “closed” forms nor the motion direction as defined in Eq. 6.4 directly tells us how collective a transition is, the collectivity we have defined
Figure 6.2 Characterization of the protein dataset: distributions of protein sizes, resolutions and pairwise RMSDs. (a) Histogram of protein sizes. (b) Histogram of protein resolutions for X-ray structures. (c) Histogram of pairwise RMSDs.
in the Methods section (see Definition 2) does. It gives us a simple measure of how likely residues are to move together, or separately. This intrinsic property of the transition thus poses an inherent limit on how well any NMA-like method, even before it is applied, can interpret the transition. For transitions with low collectivity, mode analysis methods have little chance to perform well. While for those transitions that do display large collectivity, there is clearly the possibility that a properly chosen mode analysis method could provide an excellent representation of how the transition may take place. How to choose a proper model in such a case will be addressed later.

For many proteins, the intrinsic nature of their transitions are not collective. This is demonstrated in Figure 6.3 which shows the reduced degree of freedom $\delta_{\text{reduced}}$ of the proteins. As we can see, some significant number of proteins still possess high degrees of freedom, indicating that the level of collectivity for their transitions should be quite low.

Besides the collectivity of a transition, we are also interested in knowing the average number of residues that move together collectively, i.e., the concertedness as in Definition 3. Figure 6.4
Figure 6.4  Concertedness of conformational transitions for 170 pairs of proteins. For category II.B (see Table 6.1) domain hinge motions (with proteins indexed from 76 to 134, see the black bar at the top of the figure), there typically exists a higher concertedness than for the other motion classes.

shows the dimensionality reduction, or concertedness of all 170 transitions after rigid domains are identified and modeled accordingly. We can see from the figure that there is a large dimensionality reduction (concertedness) especially for domain hinge motions.

With the inherent limit to mode representations in mind, we are now ready to explore how we may best explain the transitions.

6.4.3 How Large a Conformational Change Can Be Predicted by Mode Motions?

In Ref. [25], Tama and Sanejouand looked at the “open” and “closed” structures of 20 proteins and studied the overlap of the mode most involved in the conformational changes. Krebs et al. [93] performed NMA on the Macromolecular Movements Database [13,14], and found that most of 3,814 known protein motions can be described well by a small number of low-frequency normal modes. These works resonate with the previous works by Harrison [94], Brooks and Karplus [6], Gibrat and
Go [95], and Marques and Sanejouand [92] with the findings that a low frequency mode motion, not necessarily the very lowest one, compares well with the conformational changes that these proteins make upon ligand binding.

One question naturally arises is, how large a conformational change can the mode motion predict well? Is there a limit? Since the modes are based on the local equilibrium vibrations of a structure, it is reasonable to expect that the motions predicted by modes are only locally meaningful. And one may reasonably doubt any attempt to use mode motions to analyze large-scale conformational transitions, say over 10 Å, or even 5 Å.

Using the dataset of 170 pairs of “open” and “closed” structures that we created based on Gerstein’s Database [13, 14], with the scale of conformational changes ranging from less than 1 Å to 28 Å (see Figure 6.2(c)), we are ready to look into this question. Based on a previous study by Tama and Sanejouand [25], the normal modes calculated from the “open” form generally have better overlap with the conformational change than those obtained from the “closed” form. In this paper, we only show results for the normal modes obtained from the “open” form. We also did the same analysis using the normal modes calculated from the “closed” form and the results are quite similar to those obtained from the “open” form (see supplemental materials). Figure 6.5(a) shows the distribution of the best overlaps versus the scale of conformational changes (i.e., RMSD between the “open” and “close” structures). From the figure we can see that the overlap is quite significant even for a number of proteins having large conformational displacements. Figure 6.5(b) displays the histogram of the best overlaps found for each protein. One can see that there are a significant number of proteins with overlaps larger than 0.7, though more proteins have overlaps less than 0.5.

Though one may expect that as the scale of conformational displacement increases, the quality of the match (in terms of overlap values) would decrease, this is not completely evident from Figure 6.5(a). Even though the overlap value for the last few proteins (with largest conformational change) is relative small, there are too few of them to draw such a strong conclusion. Instead, the data suggest that, up to about 15 Å, the mode motions can perform fairly well in interpreting the conformational transitions.

However, for many other proteins, we do see that the overlap between conformational changes
Figure 6.5  Maximum overlaps using ENM. (a) Maximum overlap as a function of the transition scale, the RMSD between the “open” and “closed” structures. (b) Histogram of maximum overlaps.
Table 6.2 Analyses of the conformational transitions by the motion types. The numbers shown are the mean values over all the structure pairs in each motion type.

<table>
<thead>
<tr>
<th>Motion Type</th>
<th>I. Fragments</th>
<th>II.A Shear</th>
<th>II.B Hinge</th>
<th>II. Other</th>
<th>III.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Pairs (170 total)</td>
<td>48</td>
<td>27</td>
<td>59</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Concertedness ($\kappa$)</td>
<td>23.9</td>
<td>37.4</td>
<td><strong>99.7</strong></td>
<td>51.8</td>
<td>46.0</td>
</tr>
<tr>
<td>Reduced DOF $\delta_{\text{reduced}} (6/\chi)$</td>
<td>81</td>
<td>107</td>
<td>68</td>
<td>79</td>
<td>113</td>
</tr>
<tr>
<td>Original Maximum Overlap</td>
<td>0.37</td>
<td>0.50</td>
<td><strong>0.59</strong></td>
<td>0.38</td>
<td>0.43</td>
</tr>
<tr>
<td>Improved Maximum Overlap</td>
<td>0.50</td>
<td>0.58</td>
<td><strong>0.67</strong></td>
<td>0.46</td>
<td>0.50</td>
</tr>
<tr>
<td>Original CSO(20)</td>
<td>0.35</td>
<td>0.53</td>
<td><strong>0.67</strong></td>
<td>0.42</td>
<td>0.46</td>
</tr>
<tr>
<td>Improved CSO(20)</td>
<td>0.56</td>
<td>0.70</td>
<td><strong>0.79</strong></td>
<td>0.61</td>
<td>0.60</td>
</tr>
</tbody>
</table>

and mode motions is rather small (say, less than 0.5). We are prompted to ask whether such poor overlaps are due to any inappropriateness in how the proteins are modeled or something more intrinsic, such as the inherent collectivity of the transition as discussed earlier. The answer to this question will help us determine the applicability and limits of ENMs in understanding conformational transitions. In the following sections, we will show how an enhanced ENM can significantly improve the overlap values for some proteins, while for some others, the intrinsic nature of their conformational transitions prevent their displacements from being explained by low frequency, collective mode motions.

6.4.4 Dimensionality Reduction: Proteins Move as Rigid Domains

In our previous study of domain-swapped proteins [90], one key conclusion we arrived at is that in order to better understand the large scale domain swapping motions, it is helpful to take domain rigidity into account and to apply the more appropriate overlap calculation that was first proposed in [90] and extended here to systems having multiple rigid domains. With this in mind, we use Algorithm A (see Methods section) to identify rigid domains and then apply domain-ENM (see Methods section) to study all the transitions. Table 6.2 lists the average dimensionality reduction (or concertedness) for the different motion types. One notable point is that for hinge domain motions (category II.B), the concertedness is apparently higher than for other groups.

Consequently, we see significant improvements in the overlap values for a large percentage of
protein pairs, and this is true even for those structure pairs having very large conformational displacements. Table 6.2 shows that there is a significant increase in maximum overlap and CSO for all motion types, all with a similar extent of improvement. The apparent reason why results for domain hinge motions (category II.B) do not have a more significant improvement than the other types of motions, despite their larger dimensionality deduction, is that some of the concertedness of these transitions have already been utilized by the uniform ENM. This is confirmed by their apparent larger overlap values even before domain rigidity is taken into account.

Figures 6.6(a) and (b) compare the scatter plots of the maximum overlaps and CSOs from uniform ENM (without domain rigidity) and domain-ENM (with domain rigidity) calculations. From the figures we can see that for most protein pairs, domain-ENM is able to improve overlap (maximum overlap and CSO) by a significant amount. Figure 6.7 gives a few examples of proteins with their CSO distributions. It is seen clearly that both rigid domain modeling and the improved definition need to be applied to achieve a truly significant improvement.

Why certain residues form a rigid group and how rigid the group is are not easy to discern. Our analysis of domain-swapped proteins [90] may imply that the rigidity comes from strong hydrophobic interactions and hydrogen bonding, which is the basis of the \textit{FIRST} rigidity analysis method [96]. As explained in the Methods section, here we determine the rigid groups within a protein by directly comparing its “open” and “closed” structures. For simplicity and consistency with the coarse-grained ENM, we assign a uniform, but larger, spring constant for the contacts within all rigid domains without considering their specific, detailed interactions [90].

6.4.5 Where ENM Fails: the Limitation of Using Mode Motions to Study Conformational Transitions

Despite the improvement in overlap values that comes from domain-ENM, there remain a significant number of proteins whose overlap values remain small. This is reflected in the points at the lower left corner of Figures 6.6(a) and (b). For these protein pairs and their transitions, neither uniform ENM nor domain-ENM is able to produce modes that have large overlaps with their conformational displacements. Is there an intrinsic reason for this? From our earlier analysis, we can more or less guess the answer – that the low frequency modes from ENMs is good at describing only
Figure 6.6 Comparison of the new model (domain-ENM) with the old (uniform ENM). (a) Scatter plot of the maximum overlaps. (b) Scatter plot of the CSO(20)s. The lines, along the direction of the arrow, indicate where the increasing scales of improvement are.
Figure 6.7 Cumulative square overlaps (CSOs) for some proteins using different models. These models include uniform ENM, ENM with rigid domains, ENM with the improved overlap definition (see Equation 6.5), and ENM with both rigid domains and the improved overlap definition (i.e., domain-ENM). The first 6 modes account for the rigid body translation and rotation of the system.
the collective motion of a system, but not localized, uncorrelated motions. Therefore, those points with small overlap values probably correspond to proteins exhibiting non-collective transitions.

This intuition is confirmed in Figure 6.8, which shows the correlations between the overlaps (maximum overlap and CSO) and the inverse of collectivity \(\delta_{\text{reduced}}\) for both uniform ENM and domain-ENM (which uses the improved overlap definition), as well as the correlations between the overlaps and the protein size. In contrast to the fact that there is little correlation (about 0.1) between the overlap and the protein size (Figure 6.8(c)), there is a strong correlation between the overlap and the inverse of collectivity for both uniform ENM and domain-ENM (Figure 6.8(a) and (b)).

For ENM, the correlation values are about 0.5 (0.49 between the maximum overlap and \(\delta_{\text{reduced}}\) and 0.55 between CSO(20) and \(\delta_{\text{reduced}}\)). It is remarkable that ENM, with a uniform spring constant, is able to capture the potential collective behavior of a protein rather accurately from a single structure (see Figure 6.8(a)). This suggests it might be possible to use ENM to identify protein domains [97].

Domain-ENM is a better model than ENM when the rigidity of domains can be determined and explicitly taken into account in the model (as is the case here) and is more suited for studying the collective motions of a protein. Indeed, we see much better correlations between the overlaps and the inverse of collectivity (0.65 between the maximum overlap and \(\delta_{\text{reduced}}\) and 0.70 between CSO(20) and \(\delta_{\text{reduced}}\)) in Figure 6.8(b). This strong correlation between the overlap and the inverse of the collectivity demonstrates that it is the inherent collectivity of a transition that limits the effectiveness of using normal modes to interpret protein conformational transitions – it is not the size of the protein, nor the scale of the conformational transition, for both of which we see low correlations (see Figures 6.8(c) and 6.5(a)). Note that a similar conclusion could be drawn from the results of ENM (especially Figure 6.8(a)). However, for ENM it would be less conclusive because the correlation between the overlap and the collectivity is obscured to some extent due to the inaccurateness of the modeling, especially since the stronger interactions within a domain are not explicitly treated.

It is useful to predict the collectivity of a protein from a single structure (here it is done by comparing two structures). Then for the proteins with high collectivity, we might be able to use
Figure 6.8 Relationship between the overlap and $\delta_{\text{reduced}}$. The overlap (maximum overlap or CSO(20)) and $\delta_{\text{reduced}}$ (the inverse of collectivity) are compared by using the original overlap definition and ENM (a), the improved overlap definition and domain-ENM (b), and their dependence on protein size $N$ (c). There is a strong correlation between overlap and collectivity (0.49 and 0.55 in (a) and 0.65 and 0.70 in (b), from left to right), while there is almost no correlation between the overlap and the protein size (0.11 and 0.16 in (c), from left to right).
ENM (or domain-ENM) to predict reliably their conformational transitions.

6.5 Conclusions

In this paper we carry out a study on a large protein dataset (170 pairs of “open” and “closed” protein structures) to investigate how well conformational changes can be explained with normal mode motions. Our results show that the 170 pairs of structures and their conformational transitions fall into three categories: (i) the transitions of these proteins can be explained well by the uniform ENM, (ii) the transitions are not explained well by the uniform ENM but the results are significantly improved after considering the rigidity of domains and modeling it accordingly, and (iii) the intrinsic nature of these transitions, i.e., low degree of collectivity, prevents them from being explained with the low frequency modes of either ENM.

Our results indicate that the applicability of ENM for explaining conformational changes is not limited by either the size of the protein studied or even by the scale of the conformational change. Therefore, the answer to the question posed in the title of this article - how well can we understand large-scale molecular motions using normal modes, really depends strongly on how collective the motion is. As shown in this article, the collectivity of a transition can be estimated by comparing the “open” and “closed” forms of the studied protein. The collective nature of ENM low frequency modes makes it unsuitable for explaining non-collective transitions. Perhaps an investigation of packing densities and atomic interactions can be used to predict the collectivity of a structure [98, 99].

For this reason, ENMs show extremely promising results in understanding large-scale, collective motion, such as that of the ribosome [73]. Yet on the other hand, it is not an appropriate method in simulating protein folding, since that process is not always collective [100, 101].
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with my research. First and foremost, I would like to thank my advisor, Dr. Robert L. Jernigan for his guidance, discussion and support throughout this work. Without his encouragement, help and continuous support, it would have been impossible to finish this work. Special thanks are also expressed to my co-worker, Dr. Guang Song, for his important guidance and suggestions throughout the course of this research. I would also like to thank my co-major professor Dr. Zhijun Wu and other committee members Dr. Drena Dobbs, Dr. Karin Dorman and Dr. Vasant Honavar for their valuable advice and comments to this work. I would additionally like to thank Dr. Xuefeng Zhao for computer technical support, Dr. Andrzej Kloczkowski for valuable discussion and suggestions, and Ms. Mary Jane McCunn for administrative support. Finally, I would like to thank all the other people in the Jernigan group.


