2016

Computational Study on the Protein Conformational Transitions and Their Pathways

Jie Liu
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Bioinformatics Commons

Recommended Citation
Liu, Jie, "Computational Study on the Protein Conformational Transitions and Their Pathways" (2016). Graduate Theses and Dissertations. 15955.
https://lib.dr.iastate.edu/etd/15955

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Computational study on the protein conformational transitions and their pathways

by

Jie Liu

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 L. Jernigan, Co-Major Professor
Guang Song, Co-Major Professor
Heike Hofmann
Zhijun Wu
Drena Dobbs

Iowa State University

Ames, Iowa

2016

Copyright © Jie Liu, 2016. All rights reserved.
DEDICATION

To my husband Hongchang Wang and our parents for their love and support.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF FIGURES</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>xi</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>xii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xiv</td>
</tr>
<tr>
<td>CHAPTER 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 General Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 A Guide to This Dissertation</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 2 APPLICATION OF DIRECTED SINGLE FORCE FROM ATP HYDROLYSIS DRIVES THE GROEL CONFORMATIONAL CHANGE</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Abstract</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Introduction</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Methods</td>
<td>16</td>
</tr>
<tr>
<td>2.3.1 Dataset</td>
<td>16</td>
</tr>
<tr>
<td>2.3.2 Coarse-graining</td>
<td>19</td>
</tr>
<tr>
<td>2.3.3 Elastic network models</td>
<td>20</td>
</tr>
<tr>
<td>2.3.4 Linear response theory</td>
<td>23</td>
</tr>
<tr>
<td>2.3.5 Elastic network based force application</td>
<td>24</td>
</tr>
<tr>
<td>2.3.6 Computing the internal distance changes</td>
<td>29</td>
</tr>
<tr>
<td>2.4 Results</td>
<td>30</td>
</tr>
<tr>
<td>2.5 Conclusions</td>
<td>46</td>
</tr>
<tr>
<td>CHAPTER 3 MONTE CARLO FORCE APPLICATION SIMULATION ON GROEL WITH TRANSITION PATHWAYS ON THE FREE ENERGY LANDSCAPE</td>
<td>47</td>
</tr>
<tr>
<td>3.1 Abstract</td>
<td>47</td>
</tr>
<tr>
<td>3.2 Introduction</td>
<td>48</td>
</tr>
<tr>
<td>3.3 Methods</td>
<td>55</td>
</tr>
<tr>
<td>3.3.1 Datasets</td>
<td>55</td>
</tr>
<tr>
<td>3.3.2 Principal component analysis</td>
<td>56</td>
</tr>
<tr>
<td>3.3.3 Metropolis Monte Carlo simulation</td>
<td>57</td>
</tr>
<tr>
<td>3.3.4 Knowledge-based potentials</td>
<td>60</td>
</tr>
<tr>
<td>3.3.5 Computation of entropy from the elastic network model</td>
<td>61</td>
</tr>
</tbody>
</table>
3.3.6 Free energy and the construction of free energy landscape ............... 62
3.3.7 Sampling along other principal components .................................. 64
3.4 Results ............................................................................................. 65
3.5 Conclusions ..................................................................................... 92

CHAPTER 4 EXPERIMENTAL STRUCTURE-BASED CONFORMATIONAL SAMPLING AND THE CONSTRUCTION OF FREE ENERGY LANDSCAPES. 94
4.1 Abstract .......................................................................................... 94
4.2 Introduction ....................................................................................... 94
4.3 Method ............................................................................................. 102
  4.3.1 Datasets ....................................................................................... 102
  4.3.2 Weighted contribution method (WCM) ........................................ 106
  4.3.3 Knowledge-based entropies ............................................................ 106
  4.3.4 The construction of the free energy landscape .............................. 107
4.4 Results ............................................................................................. 107
  4.4.1 A case study for GroEL ............................................................... 107
  4.4.2 A case study for SERCA .............................................................. 112
  4.4.3 A case study for F1-ATPase .......................................................... 117
  4.4.4 Some other cases ...................................................................... 122
4.5 Conclusions ..................................................................................... 123

CHAPTER 5 A NEW BASIS FOR COMBINING ENM NORMAL MODES FROM DIFFERENT CONFORMATIONAL STATES ......................... 125
5.1 Abstract .......................................................................................... 125
5.2 Introduction ....................................................................................... 125
5.3 Method ............................................................................................. 127
  5.3.1 Dataset ....................................................................................... 127
  5.3.2 Construction of new basis .............................................................. 128
  5.3.3 Comparing normal modes and principal components .................. 129
  5.3.4 Free energy landscape on the mode space ................................. 130
5.4 Results ............................................................................................. 130
  5.4.1 A case study on GroEL ............................................................... 130
  5.4.2 A case study - adenylate kinase .................................................. 138
  5.4.3 Results from 50 cases ................................................................. 148
5.5 Conclusions ..................................................................................... 150
CHAPTER 6  SUMMARY AND CONCLUSIONS .............................................. 151
  6.1 General Discussion ............................................................................... 151
  6.2 Recommendations for Future Research ............................................ 155
APPENDICES .......................................................................................... 156
  APPENDIX A  WEIGHTED CONTRIBUTION METHOD ............................. 156
  APPENDIX B  CONSTRUCTION OF NEW BASIS .................................. 164
REFERENCES ......................................................................................... 173
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The structures of ATP molecule and ADP molecule.</td>
<td>11</td>
</tr>
<tr>
<td>2.2</td>
<td>The ATP hydrolysis mechanism</td>
<td>12</td>
</tr>
<tr>
<td>2.3</td>
<td>Coarse-grained representation of a polypeptide with a one bead model.</td>
<td>20</td>
</tr>
<tr>
<td>2.4</td>
<td>The GroEL/GroES allosteric cycle.</td>
<td>32</td>
</tr>
<tr>
<td>2.5</td>
<td>The structure of Chaperonin GroEL (PDB id: 1AON).</td>
<td>33</td>
</tr>
<tr>
<td>2.6</td>
<td>Domains and hinge residues of GroEL obtained by computing the internal distance changes with ENM.</td>
<td>34</td>
</tr>
<tr>
<td>2.7</td>
<td>Elastic network model for a GroEL subunit.</td>
<td>36</td>
</tr>
<tr>
<td>2.8</td>
<td>Structural difference vector between the GroEL in closed and open forms.</td>
<td>37</td>
</tr>
<tr>
<td>2.9</td>
<td>Probing of random forces.</td>
<td>38</td>
</tr>
<tr>
<td>2.10</td>
<td>Probing of GroEL with random forces from different directions.</td>
<td>40</td>
</tr>
<tr>
<td>2.11</td>
<td>A schematic description of the force perturbation protocol.</td>
<td>43</td>
</tr>
<tr>
<td>2.12</td>
<td>Conformational transitions of the GroEL subunit.</td>
<td>44</td>
</tr>
<tr>
<td>2.13</td>
<td>Potential energy of transition intermediates after iteratively applying forces on a closed GroEL subunit.</td>
<td>45</td>
</tr>
<tr>
<td>3.1</td>
<td>Energy landscape.</td>
<td>52</td>
</tr>
<tr>
<td>3.2</td>
<td>An illustration of principal component analysis.</td>
<td>57</td>
</tr>
<tr>
<td>3.3</td>
<td>Metropolis Monte Carlo method.</td>
<td>59</td>
</tr>
<tr>
<td>3.4</td>
<td>The procedure of Metropolis Monte Carlo force application.</td>
<td>60</td>
</tr>
</tbody>
</table>
Figure 3.5: RMSD and free energy in a MMC closed to open transition simulation of GroEL subunit. ................................................................. 67

Figure 3.6: Displacement profiles of pairwise residues in the initial-target structure pair and the initial-final structure pair of GroEL. .............. 70

Figure 3.7: Residue-by-residue difference of two root mean square displacements, one computed from the initial-target structural pair, another from the initial-final response pair ................................................................. 71

Figure 3.8: Principal component analysis of 38 GroEL subunits. ......................... 73

Figure 3.9: Visualization of the first two principle components. .......................... 74

Figure 3.10: The projection of GroEL experimental structures onto the free energy landscape constructed by sampling of conformations in the PC1-PC2 space for the two directions of transition pathways generated by MMC force applications. ................................................................. 75

Figure 3.11: RMSDs and energies for 40 MMC simulations. .............................. 79

Figure 3.12: The means and standard deviations of the pairwise RMSDs among all 40 simulations for each frame ................................................................. 80

Figure 3.13: The projection of GroEL experimental structures onto the free energy landscape constructed by the DIW-LV method in the PC1-PC2 space for the two directions of transition pathways generated by MMC force applications. ............... 82

Figure 3.14: The overlaps of 10 combined normal modes with structural change vectors from each force application step during the open to closed transitions (left 3 panels) and the closed to open transitions (right 3 panels) for GroEL.......... 84
Figure 3.15: Overlaps of the individual combined normal modes with the structural change vectors in each MMC simulation step during the GroEL open to closed transitions.......................................................... 86

Figure 3.16: Overlaps of the individual combined normal modes with the structural change vectors in each MMC simulation step during the GroEL closed to open transitions.......................................................... 88

Figure 3.17: Cumulative overlaps for the first N (N=3, 6, 10) combined normal modes and the mean overlaps of transitions for each mode. ............................................. 90

Figure 3.18: Cooperativity in a 7-subunit GroEL ring during a MMC simulation. ... 92

Figure 4.1: The interpolation of scattered data over a uniform grid. ..................... 97

Figure 4.2: A transmembrane SERCA in different conformational states. ............. 99

Figure 4.3: An overview of an F-ATPase structure. ............................................. 101

Figure 4.4: Visualization of a 3N dimensional surface in 3D space for the GroEL specified by three PCs. .......................................................... 109

Figure 4.5. Components of the knowledge-based potential terms in the free energies for GroEL. .......................................................... 110

Figure 4.6: Two profiles of entropies. .......................................................... 111

Figure 4.7: Free energy landscape of GroEL using (A) GNM entropy and (B) AACC entropy. .......................................................... 111

Figure 4.8: Projection of SERCA structure ensemble on the PC1-PC2 space labelled by bound molecules................................. 113
Figure 4.9: The knowledge-based potentials and the AACC entropy terms in the free energy landscape of SERCA................................. 114

Figure 4.10: Free energy landscape of SERCA and the conformational transition from the open to the closed form by a MMC force application simulation........ 116

Figure 4.11: MMC of SERCA ........................................................................................................ 117

Figure 4.12: Distribution of pairwise RMSDs among 78 αβ-subunits of F1-ATPase. ........................................................................................................ 118

Figure 4.13: The projection of the experimental structures F1-ATPase onto the PC1-PC2 space and the PC1-PC3 space.................................................. 120

Figure 4.14: Animation of PC1 and PC2................................................................. 121

Figure 4.15: Transition pathways on the free energy landscape for F1-ATPase...... 122

Figure 4.16: Free energy landscapes from the 50 cases............................. 123

Figure 5.1: Overlaps of ENM normal modes from the closed form (left) and the open form (right) and cumulative overlaps of the first K modes (K=3,6,10,20) with the first three PCs........................................... 131

Figure 5.2: Overlaps of the individual normal modes from the closed (open) form with the normal mode space from the open (closed) form......................... 133

Figure 5.3. Overlaps of the first 10 normal modes from the GroEL closed and open forms. .................................................................................................................. 134

Figure 5.4: Overlaps of each combined mode from the new basis with original normal modes................................................................. 135
Figure 5.5: Comparison of combined modes from the new basis with the first three PCs. .............................. 136
Figure 5.6: Essential Dynamics extracted from the ADK structure ensemble. ....... 139
Figure 5.7: Overlap and CO values of the closed modes, the open modes, and the combined modes with the first 3 PCs. .................................................. 141
Figure 5.8: Free energy landscape in the mode1-mode2 space. ...................... 144
Figure 5.9: The conformational states for ADK. .................................................. 144
Figure 5.10: Components of the knowledge-based potentials for ADK. .......... 146
Figure 5.11: The GNM entropy and the knowledge-based (AACC) entropy used for computing the energy landscape. .................................................. 147
Figure 5.12: Free energy landscape of ADK on the combined mode space for a conformational transition from the open form to the closed form by a MMC force application simulation. .................................................. 148
Figure 5.13: Overlaps of PCs with ENM normal modes and combined modes. ...... 149
LIST OF TABLES

Table 2.1: A collection of Chaperonin GroEL structures with their resolution, experimental method, ligands bound to the structure, the number of mutated residues, and whether GroES is bound to the GroEL. ............................................................ 18

Table 2.2: The performance of applying iteratively small forces multiple times compared with the performance of applying a large magnitude force only once to the structure........................................................................................................... 27

Table 4.1. List of proteins, the number of residues for each structure, the number of structures for each protein, and the source organism for the 50 protein structure ensemble. .................................................................................. 104

Table 5.1: Match of mode indices from the closed form with individual open mode that yields the highest overlaps. The mode indices of open modes in bold are unchanged. ........................................................................................................... 134

Table 5.2: Overlaps of the combined modes with the first three PCs............... 137
ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to those who helped and supported me in my research and dissertation. First, I want to thank my major advisor, Dr. Robert Jernigan for his support, guidance, patience and tolerance throughout my Ph.D. career. His insight and erudition have shaped me into a professional and independent researcher. It has been my great pleasure to work with him and learn from him. Second, I would like to thank my co-advisor Dr. Guang Song and other committee members Dr. Zhijun Wu, Dr. Drena Dobbs, and Dr. Heike Hofmann, for their suggestions and helps in my POS meeting, my preliminary exam, group meetings and individual talks. They have also influenced me a lot from their lectures, department activities, collaborative meetings and seminars. Many other faculty members at Iowa State University have also played important roles in my graduate life, including Dr. Dianne Cook, Dr. Amy Andreotti, Dr. Xuefeng Zhao, Dr. Laura Ziegla, and Dr. Gordan Miller. In addition, our beloved Trish Stauble in BCB program has really helped me a lot. If it were not for her help and advice, my Ph.D. life would not have been nearly as smooth as it was.

I would also like to thank my former and current group members in Dr. Jernigan's lab for making my time at Iowa State University a wonderful experience. Dr. Michael T. Zimmermann spent a lot of his time walking me through research areas and topics when I joined the lab. He also offered me much help when I encountered difficulties at the beginning of my Ph.D. career. Dr. Sumudu Pamoda Leelananda, a
very smart, persevering and diligent lady, as was like my big sister in the lab. We always had conversations on our research progress as well as all aspects of life. Special thanks go to my current colleagues, Kannan Sankar, Yuan Wang, Kejue Jia and Sambit Mishra. I have been working with them for more than four years and the discussions and conversations among us were invaluable for my professional development.

Besides, my parents have been supportive and caring all the time. They always teach me to be persistent and independent. They encourage me each time when I am down or feel depressed. They build up my strength to achieve my goal.

Last but not least, I want to thank my husband Hongchang Wang, my lifetime friend and lover. He has a Doctorate Degree in Business and knows everything about happiness and hardship during a Ph.D. His encouragement means everything to me. He is my teacher, my consultant, and my supporter. Without him, this dissertation would not have been possible.
ABSTRACT

A large number of protein structures have had their structures determined. However, there is still little information about their dynamics and functions. It is widely accepted that protein functionality is usually accompanied by conformational changes, and that there exists an ensemble of many structures of a given protein with different conformational states, yet the underlying mechanisms for the transitions between these states are still unclear.

Here we take a novel approach and investigate such transitions by applying forces to the structures, originating from the exothermic reactions such as ATP hydrolysis. We use the directed force application approach as well as the Metropolis Monte Carlo force application simulations within the framework of elastic network models (ENMs). The directed force application reveals the existence of strongly preferred directions of forces that can drive a protein structure towards its known end state.

When forces are applied more randomly with the Metropolis Monte Carlo method, the initial form is usually able to pass over energy barriers toward the target form and can finally achieve RMSDs of around 4 Å, matching the resolution level of the coarse-grained models themselves. Our free energy landscape agrees with the concept that native structures mostly fall in low energy regions. These landscapes are generated by computing the free energies by sampling conformations interpolated from known experimental structures, and are able in this way to suggest possible conformational transition pathways. The transitions by the application of forces are then projected onto the landscapes, and it is observed that they follow relatively low energy pathways that
are overall energetically favorable. The comparison of these conformational transitions with the ENM normal modes for GroEL demonstrates that these transitions correspond well to following the low frequency modes.
CHAPTER 1 INTRODUCTION

1.1 General Introduction

Many important biological processes are regulated by changes in protein conformation. The basic functions of proteins, such as catalysis, regulation, transportation, and binding depend largely on their structures and conformational changes. Conformational transitions between two forms are often important for understanding the relationship between structure and function. Therefore, studying conformational transitions is useful for understanding biological mechanisms. Nowadays many protein structures have been determined to have multiple conformations (1), but what causes the transition between their different states? The underlying mechanisms for these conformational transitions are still unclear. Since experimental methods are not able to capture the conformational transition pathways of proteins, computational research on this problem is of significant importance because it can supply missing details. For years, scientists have devoted significant effort to study the mechanisms of these transitions (2–11). There are several widely used computational approaches to study conformational transition pathways. These approaches include molecular dynamics (MD) simulation (12, 13), normal mode analysis (NMA) (14, 15) and its simplified coarse-grained version, i.e., elastic network model (ENM) (16, 17), and Monte Carlo (MC) simulations (18), as well as interpolation of coordinates (19). MD simulations are proven to be able to handle large and complex systems, and its many types of force fields are coming to a consensus
However, large MD simulations are computationally costly. Elastic network models (ENMs), instead, are capable of obtaining the dominant functional motions of biomolecules at significantly lower computational cost (21–27). It is especially useful for studies of the largest-scale global motions of large proteins structures. The accuracy of using ENMs has been validated by numerous comparisons against crystallographic temperature factors (17), principal components (PCs) from experimental structure sets (25) and essential dynamics from MD simulations (28, 29). Coarse-graining is essential to understand the dynamics of protein conformational transitions. Both PCs from experimental structures and the coarse-grained ENMs popularized by the our group (30–32) and many others (33–38) offer fast and convenient ways for representing the details of flexibility of large protein systems needed to compute their responses to specific forces. In this dissertation we utilize the coarse-grained ENMs for the study of conformational transitions.

Bridging between the molecular and the cellular requires significant information presently missing regarding the forces that are exerted by various biological events. This is particularly clear when we begin to learn the details of the high level of organization of the cell, e.g. the recent parking ramp structure of the endoplasmic reticulum (39). Many researchers are now studying cell mechanics (40–42), and yet we are missing some foundational information that is essential to connect these to the molecular level. Force is a push or pull that originates in an object’s interactions with another object, and force is a vector with directionality. There are different types of forces: 1) those that originate from contacts and the derivatives of the
energy landscape, 2) those that are allosteric, resulting from action-at-a-distance, i.e., from remote forces, and 3) highly directed forces that are ballistic in nature, which can originate from a specific chemical event, such as ATP hydrolysis. The first type of force is usually considered as the negative slope of the potential energy, i.e., a force that pushes towards the minimum energy, under the constraints present in the neighborhood. Allosteric forces are those forces that are initiated remotely in structures by an effector or other distant changed site within the same structure and transmitted through the densely packed structure by a communication network. The third type of force could originate from an event either near or far away, but primarily from close by events, given the dense packing within cells. Forces between cells are known to be large enough to change the cadherin intermolecular interactions (43–48). Exothermic chemical reactions, such as ATP hydrolysis, could be the origin of forces. We hypothesize that some chemical reactions releasing exothermic energy can exert specific forces, which may cause large conformational transitions in a protein structure.

It is well established that there are a relatively limited number of functionally characteristic motions for a given protein, and that its dynamics must follow certain prescribed pathways. We and others have often computed the most likely pathways for such transitions, but not previously in response to a specific force. Our recent breakthroughs in computing cooperative coarse-grained energies and entropies (49–51) yield an unprecedented view of the coarse-grained free energy landscapes. These gains in understanding protein conformational transitions come from understanding the limitations imposed by the energy landscape, reflecting the structural details. We start
our investigation of the transition pathways by applying forces to coarse-grained models using ENMs and driving the protein across the energy landscape. In many cases we have a structure prior to the exothermic reaction, and one after the exothermic event. This provides with a starting point and an end point. The details of the microscopic forces are not yet known, but will be uncovered in this dissertation. We will solve for the directions of these forces by investigating what directions for forces will move the protein from its starting structure to its ending structure. In some cases we have intermediate structures along a conformational pathway, which will provide important checkpoints for validating predicted conformational pathways.

In this dissertation, the transition pathways are further visualized and analyzed from the perspective of the free energy landscape. We hypothesize that the location of the experimental structures on the energy landscape tells something about the likely pathway (52–68). When viewed on the free energy landscapes these structures normally lie in the most favorable regions of free energy. If two different structures represent the endpoints for a transition pathway, then the challenge is to determine what forces directly cause this transition. In this dissertation, we study a wide-ranging variety of structures and the findings enable us to learn about forces on the molecular level. Outcomes include new, more useful representations of protein structure dynamics, mechanisms and allostery. One important specific outcome is simple dynamics models to produce intermediate protein structures along the pathway, which are obtained in response to specific forces arising from exothermic chemical reactions. These descriptive models of the resulting conformational transitions can be used to inform
researchers about highly directed functional dynamics in simple ways, and will provide guidance for a wide variety of types of simulations utilizing forces. This innovative study may also provide an important computational framework for treating the forces that drive conformational transitions.

1.2 A Guide to This Dissertation

This dissertation focuses on an unexplored area in biomolecular behaviors in which individual small molecular projectiles can exert strong forces. The overall aim of this study is to delineate the transition pathways for conformational transitions effected by exothermic reactions, which will ensure more coherent pathways for the transitions. The content of this dissertation covers various topics of protein conformational changes affected by molecular forces. Chapter 2 and 3 investigate the application of forces in the molecular chaperonin GroEL to study the transition pathways of its large conformational changes. These transition pathways are illustrated on the free energy landscape. The free energy landscape is improved to better delineate the transition trajectories with additional cases studied such as SERCA, F1-ATPase, and ADK in Chapter 4. The transition pathways are further compared with the normal modes from the elastic network models (ENMs) in Chapter 5.

One of our two main areas of interest is to explore the conformational changes in a protein. In Chapter 2, we investigate the forces originating from ATP hydrolysis, a common chemical reaction and one that occurs in GroEL, by applying forces originating from the reaction site. The Elastic Network Model (ENM) is widely used
for modeling the functional dynamics of macromolecules. We utilize ENMs to apply directed forces to study the conformational changes. This novel approach is developed based on linear response theory (LRT). It provides a framework for treating the forces and may lead to improvements in predicting protein conformational transitions and pathways. The exploration of forces applied to the structure in all directions reveals that there can be strongly preferred directions for the forces to drive a protein structure from its initial form towards a known target form. By adding forces originating from the ATP ligand, we can always generate defined intermediate conformations in response to the forces. In Chapter 3, the Metropolis Monte Carlo (MMC) simulation is incorporated into the force application approach to introduce randomness and to include the effects of the free energy on the transitions. Conformations with lower free energy are more likely to exist than those with higher free energy. Therefore, we apply forces randomly at the active site where ATP is bound and evaluate the free energy of the response conformation. Transitions with new conformations that are downhill in free energy are always accepted while transitions with uphill free energy are accepted with the Metropolis criterion, i.e., a Boltzmann factor of the free energy. The molecular chaperonin GroEL is well studied in this Chapter, followed by the results from case studies on other structures in Chapter 4. Our MMC force application approach is also applied to the GroEL ring structure to reveal positive cooperativity among subunits, which may possibly guide future studies on allostery to understand the mechanisms in molecular machines. The use of the MMC method and the calculation of free energy in the elastic network based force application are conducive to the study of protein
conformational transitions and the prediction of new conformations of a protein structure.

Another goal that we aim to achieve in this dissertation is to delineate the protein conformational transition pathways on the free energy landscape (FEL). In Chapter 4, we propose a simple and efficient method of structural sampling for landscape construction that requires only a small computation. The sampling of conformation is based on the Cartesian coordinates of all X-ray experimental structures of a protein that are used in principal component analysis (PCA) to capture the characteristic motions of a structure. Free energy is evaluated for each and every sampled conformation to form a complete landscape. It is commonly accepted that the lowest energy forms of a protein structure are usually presumed to be more native-like. The free energy landscapes of GroEL, SERCA, ADK, and many others from this study agree well with this concept. Our results show that most of the experimental structures fall in low free energy regions on the free energy landscape. The transition pathways of protein conformational changes can therefore be delineated from this perspective of free energy landscapes. This is achieved by projecting the transition intermediates onto the coordinate space of the landscape. The visualization of pathways on the free energy landscapes indicates that the conformational transitions can indeed occur by following relatively low energy paths, while the force enables hill climbing to overcome barriers. To further analyze and validate the transition pathways, in Chapter 5 we compare the transition steps with the normal modes from ENMs for GroEL. A combined normal mode space is computed to characterize the dynamics of multiple conformational states...
of a protein and are used to compare with the structural changes in the transition pathways. Significant overlap in direction is seen between the low frequency modes and the observed transitions, indicating that the transition pathways show strong agreement with the most important normal modes from ENMs. Chapter 6 summarizes the findings in Chapters 2, 3, 4 and 5 and suggests directions for future studies.
CHAPTER 2  APPLICATION OF DIRECTED SINGLE FORCE FROM ATP HYDROLYSIS DRIVES THE GROEL CONFORMATIONAL CHANGE

2.1 Abstract
Protein functional mechanisms usually require conformational changes, and there are often known structures for the different conformational states. However, usually neither the origin of the driving force for the change nor the underlying pathways for these conformational transitions are known. Exothermic chemical reactions are one important source of forces that can drive conformational changes. Here we investigate this type of force originating from ATP hydrolysis in the chaperonin GroEL, by applying forces originating from the chemical reaction. For this purpose, we utilize coarse-grained elastic network models. Specifically, we apply directed forces to drive the GroEL conformational changes and learn that there is a highly directed force that can drive the closed form to the open form. The knowledge-based potentials of this conformational change fueled by forces follow an increasing trend during the transition, and the RMSD of final response structure after force application is 7.7 Å, reaching approximately half way through the transition.

2.2 Introduction
Conformational transitions between multiple forms of a protein often play an important role in the relationship between structure and function. Therefore, comprehending conformational transitions is critical for understanding biological mechanisms, such as those of protein machines. Although a large number of protein structures have been determined by X-ray crystallography to have multiple
conformations (69, 70), there still remains little information about what causes these transitions (71). In many previous investigations we and others have observed that the directions from open to closed forms can occur spontaneously and have reported that both forms are included in the conformational ensembles sampled, by use of elastic network models (72–77) or molecular dynamics (78, 79), but transitions in the opposite direction, i.e., from closed to open forms are more difficult to achieve. These transitions usually entail the breaking of energetically favorable interactions, and exactly how this occurs has been difficult to determine. The underlying causes for these conformational transitions are so far unclear. We hypothesize that exothermic reactions, in particular ATP hydrolysis, can cause conformational transitions in general, and we further hypothesize that some of the forces generated during such reactions may have a strong directional character and that the dependence for this directionality depends on the details of the structure. Here we specifically investigate how highly directed forces emanating from the site of hydrolysis can directly cause the closed to open conformational transition in the GroEL chaperonin protein. We propose that the active projectile for transmitting these forces should be the leaving phosphate group.

ATP was discovered in 1929 by Lohmann (80), Fiske and Subbarow (81), and has long been identified as the source of biological energy. Energy is released in the form of heat when the terminal phosphoanhydride bond of ATP is ruptured by hydrolysis to produce ADP and a separate inorganic phosphate Pi (82). Historically the pyrophosphate bond energy is thought to diffuse through the cell where it is converted into mechanical force (83) and drives a variety of essential metabolic reactions, motor
molecule functions, ion transport processes, and biosynthetic reactions (84). Exothermic chemical reactions, especially ATP hydrolysis, are the likely driving forces for many protein conformational transitions (85–89), although this remains relatively unstudied. Figure 2.1 displays the general structure of Adenosine-5’-triphosphate (ATP) and Adenosine-5’-diphosphate (ADP). ATP is comprised of an adenine ring, a ribose sugar, and three phosphate groups. ADP is comprised of an adenine ring, a ribose sugar, and two phosphate groups. In a neutral solution, ATP has negatively charged groups that usually bind to metals such as $\text{Mg}^{2+}$ for stabilization. ATP is an unstable molecule which hydrolyzes to ADP and inorganic phosphate when it is in equilibrium with water. ATP has been viewed as an energy source for a long time, and it is often used for energy transfer in the cell. The high energy of ATP comes from the two high-energy phosphate bonds. A free energy ($\Delta G$) of -30.5 kJ/mol is released from the ATP hydrolysis chemical reaction. Figure 2.2 displays the general chemical mechanism of the ATP hydrolysis reaction.

![Figure 2.1: The structures of ATP molecule and ADP molecule.](image)

(A) The ATP molecule. (B) The ADP molecule. The adenine ring is at right, which is connected to a ribose sugar (in the middle place) which is further connected to the phosphate groups. The ATP and ADP molecules have three and two phosphate groups, respectively.
Figure 2.2: The ATP hydrolysis mechanism (90).

In single molecule studies, forces are applied and their effects on molecular structures are observed (91). The preponderance of molecular information is on static structures, either at or near equilibrium. Most of the forces that have been investigated to date are those arising either from Brownian motions (92) or from derivatives of potential functions. More specific forces that occur can be mechanical such as the arm of a protein moving with acceleration to impact other molecules with a large force, i.e., “hammer-like”. Strongly exothermic chemical reactions can give rise to driving forces with specific directionality (93), such as the expulsion of products from an enzyme active site. The relative rigidity of enzyme active sites means that products are usually expelled in directions away from the active site in a strongly directional way. Protein conformational transitions are essential for mechanistic investigations of chemical reactions (94). We start to investigate “molecular ballistics” which may account for some protein conformational transitions, based on the putative “bullets” originating in
the chemical products of exothermic reactions. One place to begin this study is to investigate forces that originate from chemistry in the cell. One of the most important such reactions is the widespread hydrolysis of ATP, which in textbooks is described as the most important source of biological energy. But what does this actually mean at the molecular level? What happens to this energy? Is it manifested as forces and why are these important on the cellular level?

To tackle this problem, it is straightforward to apply forces to a protein structure. In recent years, there have been significant studies of forces perturbing protein structures. Ikeguchi et al. (95) published linear response theory (LRT) and revealed how proteins change their conformations upon ligand binding. In their study the Hessian for computing linear force response is built by both performing molecular dynamics (MD) and using elastic network models (ENMs). C. Atilgan and A.R. Atilgan developed the perturbation response scanning (PRS) method to study the ligand binding mechanism of ferric binding protein (96). Zheng and Tekpinar also used a similar method to identify key residues in protein dynamics (97), followed by Gerek and Ozkan who utilized the PRS method to study allostery in PDZ domains (98). They developed and used a dynamic flexibility index based on the PRS method to quantify the dynamic responses of individual residues in the protein. One way to perform the PRS is to construct the Hessian from MD simulations (99, 100), a powerful approach for the study of mechanical responses. Normal mode analysis (NMA) from MD using all-atom empirical potentials is often used to follow the dynamics of proteins, and it has demonstrated significant successes (101, 102). However, the use of atomic
approaches becomes computationally inefficient for increasing sizes of a system (103). An alternative way to apply PRS is to compute the Hessian for coarse-grained ENMs. ENMs use a representation of a biological structure as an elastic bead-and-spring network to study and understand structure-based dynamics, which reflect the packing densities and yield the important motions. The earliest ENM was proposed by Tirion (17) in 1996, assuming that the interactions of both bonded and non-bonded contacts in proteins can be represented by a single universal spring. In 1997, Bahar and others developed the Gaussian network model (GNM) (16, 104) to describe vibrational fluctuations of coarse-grained models. Later Atilgan et al. (72) and Doruker et al. (105) developed the anisotropic network model (ANM) which is an extension of GNM that accounts for directionality, permitting direct visualization of the direction of motions on the structures. Due to their simple nature, coarse-grained elastic network approaches have proven to be computationally efficient compared to atomistic MD simulations (106), and they are able to capture most of a structure’s important motions.

The mixed coarse-grained elastic network model is a type of ENM developed to investigate the collective dynamics of proteins described as a compromise: part of the system remains in atomic detail with the remainder coarse-grained (107). With mixed resolution, we are able to analyze molecular effects on motions such as chemical modifications, mutations, drug binding, proline isomerization, or post-translational modifications, while retaining nearly the computational efficiency of coarse-grained ENM. In the mixed coarse-grained elastic network models, each node represents either a heavy atom in high-resolution or a single residue in the low-resolution region, and
the neighboring node pairs within a cutoff distance are linked with harmonic springs (108). It is necessary and important to adjust the cutoff distance to successfully perform uniform coarse-graining of the network at the hierarchical level (106). Similarly, different cutoffs and interaction parameters need to be assigned to the high and low resolution regions of the mixed-resolution network (103). This represents a process of matching the interaction energies to maintain the same extent of cohesiveness for models at the various levels of coarse-graining. The force constant for the interface region between the high-resolution and the low-resolution parts should be carefully adjusted to properly account for the cohesiveness. For the coarse grained models, structural responses are not very sensitive to the details because each point usually interacts with a large number of other sites. The use of mixed coarse-grained ANM can provide insights into the performance of each atom in the ATP molecule. Therefore, mixed coarse-grained elastic network based force application is an appropriate approach for studying large proteins such as the molecular chaperonin GroEL.

Numerous observations have been made on proteins with multiple conformations that can undergo significant changes. There are many cases of multiple structures of the same protein, and these structures with different ligands ATP/ADP bound in catalytic sites represent two end points for the transition. Here we ask the question: can application of some specific forces drive these changes? We apply directed forces from the ATP ligand to investigate how forces might drive the initial closed form of a protein structure towards its target open form.
2.3 Methods

2.3.1 Dataset

There are many homologous structures of the molecular chaperonin GroEL. We collected 34 intact crystal structures from the PDB and perform a thorough analysis on each structure. Table 2.1 displays a detailed description of our dataset with the structure’s resolution, the experimental method, the ligands bound to the structure, the total number of mutated residues, and whether GroES is bound to the GroEL. Most of these 34 GroEL structures are from the organism *Escherichia coli* with the exceptions 1IOK from *Paracoccus denitrificans* and 4V4O from *Thermus thermophilus*. The structures determined by X-ray have good resolutions, approximately around 3 Å. The structures solved by the Electron Microscopy (EM) method have poorer resolutions, mostly around 8 Å. The structures bound with different ligands distinguish themselves as having different conformational states. We choose 1KP8 and 1SX3 as possible ATP-binding forms and 1AON, 1PCQ, 1PF9, 4V4O, 1SVT and 1SX4 as possible ADP-binding forms by taking resolution and structural completeness into consideration. Pairwise structural alignment is performed with Pymol since all the structures have high sequence similarity (Pymol Align algorithm does a good job on proteins with reasonable high sequence similarity > 30%). Pairwise root mean square deviations (RMSD) are computed for each ATP-bound and ADP-bound structure pair. The structure pair with the best resolution and the largest RMSD value is the ideal. Chain A from 1KP8 is used for the reference ATP bound state (in the closed form) and chain A from 1AON is used for the reference ADP bound state (in the open form). The two
conformational states are trimmed with both having 523 corresponding residues, and these are aligned for their 22 ATP binding residues, which are within 5 Å of the ligand. These residues are 30-33, 51-53, 87-91, 414-416, 478-481, 454, 493, and 495. After we superimpose the open form 1AON on the closed form 1KP8, the RMSD of structural difference between the two states is 14.7 Å.
Table 2.1: A collection of Chaperonin GroEL structures with their resolution, experimental method, ligands bound to the structure, the number of mutated residues, and whether GroES is bound to the GroEL.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Resolution (Å)</th>
<th>Method</th>
<th>Ligand</th>
<th>Mutation(s)</th>
<th>GroES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GRL</td>
<td>2.8</td>
<td>X-ray</td>
<td>none</td>
<td>21</td>
<td>N</td>
</tr>
<tr>
<td>1OEL</td>
<td>2.8</td>
<td>X-ray</td>
<td>None</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>1AON</td>
<td>3.0</td>
<td>X-ray</td>
<td>ADP,Mg</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>1IOK</td>
<td>3.2</td>
<td>X-ray</td>
<td>none</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>1GR5</td>
<td>7.9</td>
<td>EM</td>
<td>none</td>
<td>28</td>
<td>N</td>
</tr>
<tr>
<td>1GRU</td>
<td>12.5</td>
<td>EM</td>
<td>none</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>1KP8</td>
<td>2.0</td>
<td>X-ray</td>
<td>ATP,Mg,K,SO4</td>
<td>42</td>
<td>N</td>
</tr>
<tr>
<td>1MNF</td>
<td>3.0</td>
<td>X-ray</td>
<td>peptide substrate</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>1PCQ</td>
<td>2.81</td>
<td>X-ray</td>
<td>ADP,AF3,K,Mg</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>1PF9</td>
<td>2.99</td>
<td>X-ray</td>
<td>ADP,Mg</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>4V43</td>
<td>3.52</td>
<td>X-ray</td>
<td>none</td>
<td>56</td>
<td>N</td>
</tr>
<tr>
<td>4V4O</td>
<td>2.8</td>
<td>X-ray</td>
<td>ADP,DMS,Mg</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>1SS8</td>
<td>2.7</td>
<td>X-ray</td>
<td>none</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>1SVT</td>
<td>2.81</td>
<td>X-ray</td>
<td>ADP,AF3,K,Mg</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>1SX3</td>
<td>2.0</td>
<td>X-ray</td>
<td>ATP,K,Mg</td>
<td>28</td>
<td>N</td>
</tr>
<tr>
<td>1SX4</td>
<td>3.0</td>
<td>X-ray</td>
<td>ADP,Mg</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>1XCK</td>
<td>2.92</td>
<td>X-ray</td>
<td>K,MPD,PEG,SO4</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>2C7C</td>
<td>7.7</td>
<td>EM</td>
<td>none</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>2C7D</td>
<td>8.7</td>
<td>EM</td>
<td>none</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>2C7E</td>
<td>9.7</td>
<td>EM</td>
<td>ATP,K,Mg</td>
<td>28</td>
<td>N</td>
</tr>
<tr>
<td>2CGT</td>
<td>8.2</td>
<td>EM</td>
<td>none</td>
<td>0</td>
<td>Y</td>
</tr>
<tr>
<td>2EU1</td>
<td>3.29</td>
<td>X-ray</td>
<td>none</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>2NWC</td>
<td>3.02</td>
<td>X-ray</td>
<td>none</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>3C9V</td>
<td>4.7</td>
<td>EM</td>
<td>none</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>3CAU</td>
<td>4.2</td>
<td>EM</td>
<td>none</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>3E76</td>
<td>3.94</td>
<td>X-ray</td>
<td>none</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>2YEY</td>
<td>4.5</td>
<td>X-ray</td>
<td>none</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>2YNJ</td>
<td>8.4</td>
<td>EM</td>
<td>AGS,MG,TL</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>4AAQ</td>
<td>8.0</td>
<td>EM</td>
<td>ATP,Mg,PO4</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>4AAR</td>
<td>8.0</td>
<td>EM</td>
<td>ATP,Mg,PO4</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>4AAS</td>
<td>8.5</td>
<td>EM</td>
<td>ATP,Mg,PO4</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>4AAU</td>
<td>8.5</td>
<td>EM</td>
<td>ATP,Mg,PO4</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>4AB2</td>
<td>8.5</td>
<td>EM</td>
<td>ATP,Mg,PO4</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>4AB3</td>
<td>8.5</td>
<td>EM</td>
<td>ATP,Mg,PO4</td>
<td>14</td>
<td>N</td>
</tr>
</tbody>
</table>


2.3.2 Coarse-graining

Coarse-graining is a way to represent a structure by a reducing the number of degrees of freedom. Due to the reduction in the degrees of freedom and the elimination of fine details, a coarse-grained (CG) system requires less computational resources than an atomic system. In the study of protein dynamics, many popular and widely-used techniques, e.g. Molecular Dynamics simulation for protein folding and structure function relationships, Monte Carlo simulation for structural sampling etc., require considerable computational power. Modelling of protein structures by using coarse-graining significantly reduces the complexity of the system as well as enhances the computational efficiency. There are many different coarse-grained models such as the one-bead model, two-beads model, Go model, and harmonic network (109). The one-bead model is simple and often used to modelling proteins when simulations are beyond the scope in terms of size or time scale for atomic MD. In the one-bead model, the backbone $C^\alpha$ is represented as beads on a string while the remainder of atoms are ignored. A schematic representation of coarse-graining is shown in Figure 2.3. In all of the topics in this thesis, the coarse-graining uses only the $C^\alpha$ atom to represent each residue.
**Figure 2.3: Coarse-grained representation of a polypeptide with a one bead model.**

(A) The structure with all heavy atoms for a sequence of amino acid chains taken from the Protein Data Bank PDB structure 1usg residues 1-10. Atom C, N, O are colored green, blue, and red respectively. The backbone Cα is colored in magenta. The gray spheres on the Cα atoms are the beads used to represent a coarse-grained residue. (B) A bead and string model. The coarse-graining model for a protein structure includes all Cα atoms as beads.

### 2.3.3 Elastic network models

To model protein structures with ANM, we coarse-grain the structures by representing each residue by only its Cα atom. An elastic network is built by placing springs between any two Cα atoms if they are within a certain cutoff distance. The springs are treated as being harmonic for all interactions. In this study, a cutoff distance of 12 Ångstroms (Å) is used, which has been validated by previous studies showing that the cutoff distance does not much affect the global motions extracted from NMA significantly. The cutoff is typically set between 12 Å and 15 Å. Support for the adoption of the $r_c$ values of 12-15 Å in the ANM is provided by agreement between the histograms of vibrational fluctuations and X-ray crystallographic temperature factors (72). We use a typical force constant of 1.0 kcal / (mol $\cdot$ Å²) for all springs.

This section on the ENM method, especially emphasizing the ANM, provides full details on the common ANM methods utilized in the following Chapters. To model
protein structures using ANM, we need to coarse-grain the structure where each residue is represented by its $C^\alpha$ atom. The coarse-grained structure is connected between consecutive $C^\alpha$ atoms. An elastic network is built by placing springs between any two $C^\alpha$ atoms if they are within a certain cutoff distance measured in Ångstrom ($10^{-10}$ m). The springs are treated as harmonic for all interactions. The cutoff distance and spring constant are the two main parameters in the ANM model. Thus, the global protein structure is represented as an elastic network of interacting nodes connected with identical springs. In ANM, the potential energy $V$ is given as a summation over all the interacting pairs of atoms $i$ and $j$ in Equation 2.1

$$V = \left(\frac{\gamma^2}{2}\right) \left[ \sum_{i,j}^N (R_{ij} - R_{ij}^0)^2 h(R_c - R_{ij}^0) \right]$$

(2.1)

where $R_{ij}$ and $R_{ij}^0$ are the instantaneous and equilibrium distance between two nodes $i$ and $j$, $\gamma$ is the spring constant, and $h$ is the Heaviside function that has a value of 1 when the pair of nodes are closer together than $R_c$ and 0 otherwise. The overall potential energy is the sum of harmonic potentials between all interacting nodes and the potential of a structure with $N$ pairwise interactions is given in matrix notation in Equation 2.2

$$V = \left(\frac{\gamma^2}{2}\right)DHDT$$

(2.2)

where energy $V$ is a function of the square of the displacement vectors $D$ of each node in the structure, $\gamma$ is the identical spring constant for all closely interacting points in the structure determined by a cutoff distance, and $H$ is the Hessian matrix containing the second derivatives of the energy with respect to each of the coordinates $x$, $y$, $z$. For
a structure with \( n \) geometric points, the Hessian matrix contains \( n \times n \) super-elements each with size \( 3 \times 3 \). The \( (i,j)^{th} \) element of the Hessian matrix is given in Equation 2.3

\[
H_{ij} = \begin{bmatrix}
\frac{\partial^2 V}{\partial x_i \partial x_j} & \frac{\partial^2 V}{\partial x_i \partial y_j} & \frac{\partial^2 V}{\partial x_i \partial z_j} \\
\frac{\partial^2 V}{\partial y_i \partial x_j} & \frac{\partial^2 V}{\partial y_i \partial y_j} & \frac{\partial^2 V}{\partial y_i \partial z_j} \\
\frac{\partial^2 V}{\partial z_i \partial x_j} & \frac{\partial^2 V}{\partial z_i \partial y_j} & \frac{\partial^2 V}{\partial z_i \partial z_j}
\end{bmatrix}
\]  

(2.3)

where \( V \) is the harmonic potential between node \( i \) and \( j \), and \( X_i, Y_i, Z_i \) are the positional components of node \( i \). The Hessian matrix therefore, is the second derivatives of the potential energy with respect to the mass-weighted atomic Cartesian coordinates and can be further decomposed as

\[
H = \Lambda \Lambda^T
\]

(2.4)

where \( \Lambda \) is a diagonal matrix comprised of the eigenvalues with the individual eigenvectors forming the columns of the matrix \( \Lambda \). For a structure with \( n \) points, this decomposition generates \( 3n - 6 \) normal modes (the first 6 modes are the rigid body translations and rotations of the system), which inform about the directions of the fluctuations of each bead in the structure with the corresponding frequencies specified by the eigenvalues. The normal modes are ordered by the eigenvalues which ranges from low frequency to high frequency modes.

In other Chapters that utilize the ENM, and the cutoff distance to construct the contact matrix for the ENM from a protein structure is normally set in the range of 10-13\( \text{Å} \).
2.3.4 Linear response theory

In ENM based linear response theory, a protein is constructed as a residue network of N nodes which are represented by their $C^\alpha$ atoms. Any pair of nodes within a certain cutoff distance is connected with a harmonic potential. When there are no external forces applied on the system, the equilibrium condition for each residue $i$ is that the summation of internal interaction forces for each residue $i$ is zero, as defined in Equation 2.5

$$b\Delta f_i = 0$$

(2.5)

where $b$ is a $3 \times m$ coefficient matrix in which the direction cosines of each force representing the residue-residue interaction, $\Delta f_i$ is a $m \times 1$ column vector of forces corresponding to the bond between two interacting residues. Generalizing Eq. 2.5 to the whole system of N nodes and M interactions gives

$$B\Delta f_i = 0$$

(2.6)

where $B$ is the $3N \times M$ direction cosine matrix which can be generated from the topology of the native structure with a specified cutoff value and $\Delta f$ is a $M \times 1$ column vector of residue-residue interaction forces. Under a force perturbation, the equilibrium condition for the system is that the summation of the residue-residue interaction forces must equal to the external force, which gives Equation 2.7

$$B_{3N\times M}\Delta f_{M\times 1} = \Delta F_{3N\times 1}$$

(2.7)

where $\Delta F$ is an external force. Under the external force, each residue is moved by a displacement $\Delta R$, which is a $3N \times 1$ positional displacement vector. The bond distance
changes between any two residues in the amount of $\Delta r$ are in correspondence with the positional displacement $\Delta R$ (110), as

$$B_{M \times 3N}^T \Delta R_{3N \times 1} = \Delta r_{M \times 1}$$ (2.8)

Within the ENM in which each residue is connected to their neighbors with linear elastic springs, the residue interaction forces $\Delta f$ are related to the changes in the contact distances $\Delta r$ through Hooke’s law

$$K_{M \times M} \Delta r_{M \times 1} = \Delta f_{M \times 1}$$ (2.9)

where the coefficient $K$ is an $M \times M$ diagonal matrix. Rearranging Equation 2.7-2.9 gives Equation 2.10

$$B K B^T \Delta R = \Delta F$$ (2.10)

where $\Delta F$ is a column vector which signifies external forces applied on the residues, the $B K B^T$ matrix is equivalent to the Hessian (111) and its inverse has six zero eigenvalues, which correspond to the global translational and rotational degrees of freedom of the system. The elements of the inverse of the Hessian, $G = H^{-1}$, is an $M \times M$ matrix whose $ij^{th}$ element is the $3 \times 3$ matrix of correlations between the $x$-, $y$-, and $z$-components of the fluctuations $\Delta R_i$ and $\Delta R_j$ of residues $i$ and $j$

$$G_{ij} = \begin{bmatrix} \langle \Delta X_i \Delta X_j \rangle & \langle \Delta X_i \Delta Y_j \rangle & \langle \Delta X_i \Delta Z_j \rangle \\ \langle \Delta Y_i \Delta X_j \rangle & \langle \Delta Y_i \Delta Y_j \rangle & \langle \Delta Y_i \Delta Z_j \rangle \\ \langle \Delta Z_i \Delta X_j \rangle & \langle \Delta Z_i \Delta Y_j \rangle & \langle \Delta Z_i \Delta Z_j \rangle \end{bmatrix}$$ (2.11)

### 2.3.5 Elastic network based force application

Our force application approach builds on the Hessian Matrix computed from coarse-grained ANM and generates a displacement vector in response to an external force perturbation vector based on LRT. The displacement vector is computed as
$G_i \cdot F_i = \Delta R_i$ \hspace{1cm} (2.12)

where matrix $G$ has dimensions $3N \times 3N$ and is equivalent to the inverse Hessian.

Eq. 2.12 is the abbreviation of

\[
\begin{bmatrix}
g_{x_1x_1} & g_{x_1y_1} & g_{x_1z_1} & \cdots & \cdots & g_{x_1x_N} & g_{x_1y_N} & g_{x_1z_N} \\
g_{y_1x_1} & g_{y_1y_1} & g_{y_1z_1} & \cdots & \cdots & g_{y_1x_N} & g_{y_1y_N} & g_{y_1z_N} \\
g_{z_1x_1} & g_{z_1y_1} & g_{z_1z_1} & \cdots & \cdots & g_{z_1x_N} & g_{z_1y_N} & g_{z_1z_N} \\
\vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots & \vdots \\
g_{x_Nx_1} & g_{x_Ny_1} & g_{x_Nz_1} & \cdots & \cdots & g_{x_Nx_N} & g_{x_Ny_N} & g_{x_Nz_N} \\
g_{y_Nx_1} & g_{y_Ny_1} & g_{y_Nz_1} & \cdots & \cdots & g_{y_Nx_N} & g_{y_Ny_N} & g_{y_Nz_N} \\
g_{z_Nx_1} & g_{z_Ny_1} & g_{z_Nz_1} & \cdots & \cdots & g_{z_Nx_N} & g_{z_Ny_N} & g_{z_Nz_N}
\end{bmatrix}_{3N \times 3N}
\begin{bmatrix}
\Delta R_{1x} \\
\Delta R_{1y} \\
\Delta R_{1z} \\
\vdots \\
\Delta R_{Nx} \\
\Delta R_{Ny} \\
\Delta R_{Nz}
\end{bmatrix}_{3N \times 1}
\]

In the matrix $G$, $g_{lm}$ is an element in $G$ where $l, m = x_j, y_j, z_j$ ($j \in \{1, \ldots, N\}$ and $j$ is the residue index) which is the second order partial differential of total energy with respect to directions. The $3N \times 1$ vector $F_i$ is the external force vector which is applied to residue $i$ and is specified by the direction $(\Delta F_{xi}^i, \Delta F_{yi}^i, \Delta F_{zi}^i)$. $\Delta R_i$ is a $3N \times 1$ displacement vector denoting the response obtained by applying force to residue $i$.

We develop a pipeline to perform force application on the nodes that we are interested in. For the aim of exploring the thermodynamic effect of the ATP hydrolysis reaction, in this study we mainly focus on the role of the leaving phosphate group which may
drive the conformational transition of initial structure towards the target structure. Since a large magnitude of force could even cause bond ruptures and this may fail to rationally represent the true relationship of the connection between two nodes during a conformational change, we adopt the procedure of applying iteratively small forces on the same node where we are pushing. In this way we can not only maintain connections between pair of nodes still being connected by springs in the elastic network when they are actually undergoing large distance changes and should in principle undergo bond breakage after the force perturbation, but also allow new contacts between two nodes to form if they are moving closer together during the transition. A preliminary study on the effect of iterative small forces in comparison with a large single force supports the feasibility and reliability of the iterative approach. In Table 2.2, five GTP-binding structure pairs are aligned and forces are added on the initial starting structure. The overlap value after force application is computed with Eq.2.13, and the RMSD from the target structure is computed with Eq.2.14 in the following section. The five cases in Table 2.2 are GTPase ImapFamily2, Rab6A, Hras, lyase, Hras with an energy minimization treatment by KoBaMinimization (112–114). The results in Table 2.2 show that the overlap value of response vector with base-target structure difference vector is quite similar for these two types of force-application schemes. There is not much difference whether we add one single force or smaller forces iteratively applied, in terms of the overlap value of force response vector and structural difference vector as well as the final RMSD of the response structure with the target structure. However, the approach of adding small forces iteratively on the structure avoids extreme changes
during protein structure conformational changes, which seems physically more reasonable. In addition, the minimum RMSD that the response form can reach for iterative force application scheme is slightly smaller than a large one-time push.

Table 2.2: The performance of applying iteratively small forces multiple times compared with the performance of applying a large magnitude force only once to the structure.

<table>
<thead>
<tr>
<th>GTP Case</th>
<th>Overlap (single force)</th>
<th>Overlap (iterative small forces)</th>
<th>RMSD (single force) (Å)</th>
<th>RMSD (iterative small forces) (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.44</td>
<td>0.44-0.48</td>
<td>2.69</td>
<td>2.66</td>
</tr>
<tr>
<td>2</td>
<td>0.34</td>
<td>0.33-0.36</td>
<td>1.64</td>
<td>1.62</td>
</tr>
<tr>
<td>3</td>
<td>0.69</td>
<td>0.68-0.69</td>
<td>1.69</td>
<td>1.68</td>
</tr>
<tr>
<td>4</td>
<td>0.55</td>
<td>0.46-0.53</td>
<td>5.28</td>
<td>5.29</td>
</tr>
<tr>
<td>5</td>
<td>0.69</td>
<td>0.69-0.70</td>
<td>1.68</td>
<td>1.66</td>
</tr>
</tbody>
</table>

We therefore apply small forces to the structures, realign the resulting structure to the current state for the invariant domain to eliminate rigid body motions, as well as to obtain the internal structural response, and repeat this process multiple times. The response vector is re-computed by subtracting the previous structure from the new response structure. The force iteration proceeds until the RMSD from the target structure has converged and no longer decreases. After applying forces to the structure, we have to evaluate the response intermediates to decide whether or not the conformation is acceptable. We develop several criteria, i.e., overlap, overall RMSD, correlation of residue-wise RMSD and mean square error, to evaluate the response structure after force perturbation iteration so as to select the most favorable force direction for the current structure to move towards the target structure. The criteria to evaluate the response intermediate to force application are:
(1) Overlap (Correlation Cosine) of response vector with the aligned current form and target form structural difference vector

\[ O_r = \frac{\Delta R_j \cdot \Delta D^T}{|\Delta R_r| |\Delta D|} = \frac{x_1y_1 + x_2y_2 + \ldots + x_ny_n}{\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \sqrt{y_1^2 + y_2^2 + \ldots + y_n^2}} \]  

(2.13)

where \( r \) indicates the \( r^{th} \) force direction, \( \Delta R = \{x_1, x_2, \ldots, x_n\} \) is the response vector to application of force, and \( \Delta D = \{y_1, y_2, \ldots, y_n\} \) is the structural difference vector of the aligned current form and target form structure pair. Usually a high value of \( O_r \) is expected since we aim to push the initial structure toward the target structure.

(2) Root mean square deviation of the response structure with the target structure shown in Equation 2.14

\[ \text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta_i^2} \]  

(2.14)

where the one-dimensional vector of the response structure is computed by adding the one-dimensional vector of the initial form and the response vector, denoted as: \( \text{Vec}_{\text{response}} = \text{Vec}_{\text{initial}} + \Delta R \). In Eq.2.14, \( N \) is the total number of nodes in the model, \( i \) is a specific node among all \( N \) nodes, and \( \delta \) is the distance between a pair of corresponding equivalent \( i^{th} \) nodes (\( C^\alpha \) atoms for our coarse-grained ENM). A decreasing RMSD is expected since the initial structure keeps undergoing conformational changes toward the target structure.

(3) Correlation of root mean square displacement vector for each and every equivalent node between structure pairs of initial-target and initial-response
\[ \delta_{\text{Initial-Target}} = [\delta_1^0, \delta_2^0, \delta_3^0, ..., \delta_N^0] \]

\[ \delta_{\text{Initial-Response}} = [\delta_1, \delta_2, \delta_3, ..., \delta_N] \]

\[
\text{Correlation} = \text{Corr}(\delta_{\text{Initial-Target}}, \delta_{\text{Initial-Response}})
\]

where \( \delta_{\text{Initial-Target}} \) (\( \delta_{\text{Initial-Response}} \)) is the distance between a pair of corresponding equivalent atoms starting from atom 1, atom 2, atom 3 through atom N. \( \delta_{\text{Initial-Target}} \) and \( \delta_{\text{Initial-Response}} \) generate two one-dimensional vectors for which the correlation measures how effectively the application of forces has driven the initial structure to the target structure.

(4) Mean square error (MSE) for each and every equivalent node between structure pairs of initial-target and initial-response

\[
\text{MSE} = \sum_{i=1}^{N} (\delta_i^0 - \delta_i)^2
\]

where \( N \) is the total number of nodes in the ENM, \( \delta_i^0 \) is the displacement between a pair of corresponding equivalent nodes \( i \) in the initial-target structure pair, and \( \delta_i \) is the displacement between a pair of corresponding equivalent nodes \( i \) in the initial-response structure pair. Measurement of MSE also reveals how near to the target structure the response structure has finally been driven.

2.3.6 Computing the internal distance changes

The mean square internal distance (ID) change is informative about the internal conformational change in the structure. We use the displacements of the positions of each node in ANM to compute the ID change. The mean square change in the internal distance is
\begin{equation}
\langle (\Delta R_i - \Delta R_j)^2 \rangle = \langle \Delta R_i^2 \rangle + \langle \Delta R_j^2 \rangle - 2 \langle \Delta R_i \Delta R_j \rangle \tag{2.17}
\end{equation}
which can be generated from the inverse of the Hessian Matrix $\Gamma$

\begin{equation}
\langle (\Delta R_i - \Delta R_j)^2 \rangle = (3k_B T / \gamma) \cdot [I_{ii}^{-1} + I_{jj}^{-1} - 2I_{ij}^{-1}] \tag{2.18}
\end{equation}
where $k_B$ is the Boltzmann constant, $T$ is the temperature, $\gamma$ is the spring constant for ANM.

2.4 Results

ATP hydrolysis is a chemical reaction in which chemical energy stored and transported in a high-energy bond in ATP is released. Since this reaction is exothermic and releases around 7.3 kcal mol$^{-1}$ (115), the energy released by this reaction may drive conformational changes and propagate forces throughout a structure. We hypothesize that a force is initially applied at the reaction site and that the released phosphate group can itself be the agent for applying the force. From our study, we find a specific direction of forces that can drive the GroEL closed form towards its targeted open form.

Chaperonins provide assistance to the process of folding newly translated or newly translocated polypeptides (116), which play an important role in protein folding. They facilitate the production of the native state of protein structures under conditions in cases where the native form would otherwise not be achieved, to prevent misfolding of polypeptides. The timing and synchronization in this highly allosteric system are currently recognized, and this allosteric event coordinates the binding and hydrolysis of ATP, the binding of GroES, and the binding and release of polypeptide. However, the interaction and interplay among GroEL, GroES, ATP binding and hydrolysis in an
integrated fashion are still unclear. During the function of GroEL assisted by its co-chaperonin GroES, it undergoes a series of structural transitions among multiple conformational states triggered by ATP binding and hydrolysis (117). Study of the GroEL transition is a straightforward approach to help understand the allostery of the system. Without binding of the nucleotides, both rings in GroEL would remain in the closed T state. The binding of ATP to the subunits in one of the rings (cis ring) drives the conformational change of these subunits to the open R state. In this state, the unfolded peptide (substrate) is encapsulated in the cylindrical chamber and the cap GroES is attached thereafter. ATP hydrolysis provides the energy needed to process the substrate and leads to the R’’ state. See Figure 2.4 for details.
Figure 2.4: The GroEL/GroES allosteric cycle.
The GroEL ring structure consists of two rings, i.e., cis and trans (identified in the upper left panel), and three states. The T state is ATP-free state shown as 1GR5, the R state is ATP-binding state shown as 1KP8, and the R” state is ADP-binding state shown as 1AON.

The crystal structure of E. coli GroEL is composed of two 7-fold rotationally symmetrical rings stacked back-to-back with dyad symmetry, each ring having 7 identical subunits (see Figure 2.5). Each subunit has three identifiable domains: a large equatorial domain that forms the foundation of the assembly, a small intermediate domain at its waist, and a loosely structured intermediate-size apical domain (118).
Figure 2.5: The structure of Chaperonin GroEL (PDB id: 1AON).
Two rings, i.e., the cis ring at the top and the trans ring at the bottom, stack back-to-back on each other (GroES is not shown here). The equatorial domain is in blue (residues 1-133 and 409-523), the intermediate domain is in green (residues 134-190, and 377-408), and the apical domain is in red (residues 191-376).

The internal distance changes describe the changes within a structure for a certain pair of nodes in a protein structure using ANM. A small internal distance change for a given node pair indicates that the two nodes move together rigidly in the protein dynamics. Hinge residues are identified in Figure 2.6 by computing the internal distance changed from ANM by following Eq.2.18 to validate the invariant domains in the GroEL subunit for the purpose of structural alignment. The hinge residues identified are: Ser135, Val136, Val190, Glu191, Ala377 in the ATP bound form and Ser135, Ile150, Ser151, Asn153, Ala377, Leu494, Pro496 in the ADP binding form, compared to the hinge residues Gly192, Gly375, Pro137 and Gly410 obtained from the database MolMovDB (119) from Mark Gerstein’s laboratory (www.molmovdb.org).
Figure 2.6: Domains and hinge residues of GroEL obtained by computing the internal distance changes with ENM.

(A) Three domains are identified in the ATP-bound form in the contact matrix (blue marks shows residues closes to one another. These domains are in colored circles. The apical domain is marked with the red circle, the intermediate domain is marked with the green circle, and the equatorial domain is marked with the blue circle. These are the same as the domains in Figure 2.5. (B) Domains identified in the ADP binding form. (C) Hinge residues identified in the ATP-binding form. These residues are color-coded with their internal distance change value and are labeled by their residue name. (D) Hinge residues identified in the ADP-binding form.

In our study, we use mixed coarse-grained Elastic Network Models (120) to model the dynamics of the GroEL system in order to investigate the collective
dynamics of proteins described as a combination of atomistic (high-resolution) and coarse-grained (low-resolution) regions (see Figure 2.7). The detailed mixed coarse-graining modelling treatment to our system is as follows: (1) The interesting part around the ATP is described at atomistic level in which each atom represents a node. $C^\alpha$ atoms are used to represent low-resolution parts of the structure, (2) For coarse-grained nodes, all node pairs within a cutoff distance are connected with identical harmonic springs. Similarly we treat fine-grained nodes with a smaller cutoff distance to build a contact matrix. The interactions between the fine-grained and coarse-grained nodes are connected with a cutoff distance that is their geometric mean. The system we construct in this chapter adopts mixed coarse-grained ENM for the GroEL subunit, with a cutoff distance 8 Å ($r_c1$) in the atomic part for the ATP ligand and 12 Å ($r_c2$) in the coarse part. The cutoff distance between the atomic and the coarse part is computed by taking the geometric mean of $r_c1$ and $r_c2$ ($\sqrt{r_c1 \cdot r_c2}$), which is 9.8 Å.
**Figure 2.7: Elastic network model for a GroEL subunit.**

ATP is described by atoms with each atom representing a node, with a cutoff distance 8Å to connect a pair of nodes. Amino acid residues are represented as low-resolution nodes by using only their $C^\alpha$ atoms, with cutoff distance 12Å. The cutoff between the atomic and the coarse-grained nodes is 9.8 Å.

To start our force investigation, we align the open structure at the invariant equatorial domain (residues 1-133, 409-523) on the closed one validated in the previous Figure 2.6. The structural difference between the closed form and the open form of GroEL subunit shows an RMSD of 14.7 Å. The structural difference vector is computed by taking the coordinate difference between the closed and the open states shown as grey arrows in Figure 2.8.
Figure 2.8: Structural difference vector between the GroEL in closed and open forms.

The two conformational states of the GroEL subunit are aligned at the invariant equatorial domain in the square box (residues 1-133, 409-523). The closed form is displayed in blue while the open form is in pink. The grey arrows are structural difference vectors for each of the corresponding residue pairs in the two conformational forms. The atomic ligand ATP is bound in the equatorial domain (rectangular box) and is shown as spheres with its atoms C, N, O colored green, blue and red.

The overlap of the structural difference vector with force response vector is measured by calculating as the dot product of two vectors - the difference vectors and the cumulative force response vectors defined in Equation 2.13. Previous studies of Ikeguchi (95), Atilgan (96), and Gerek (98) indicate that proteins respond to the forces applied to individual residues in the nucleotide-binding regions with conformational changes. According to Duttmann, the responses are sensitive to the choice of the residues to which the forces are applied (121). Therefore, one particularly critical factor in force application is the direction in which the forces are applied. Initial probing is carried out to test the effect of applying a single force in different directions in order to
see how effective a certain direction is to effect the conformational transition. We start by scanning a set of random force directions originating from the hydrolysis site in the closed form of the GroEL subunit. Figure 2.9 presents a schematic illustration of these random forces in all directions applied to a certain node in the structure model. The red point in the center of sphere is the node to which we add force. The blue points are end points of random force vectors and the connection of blue-red point pointing from red to blue is a force direction. Since computational time is highly dependent on iteration times, we have a trade-off between computational efficiency and reproducibility.

![Figure 2.9: Probing of random forces.](image)

The force vector is from the central red node, pointing toward the blue end points on the surface of the sphere. The blue points are the end points of random force vectors and the connection of blue-red point pointing from red to blue indicates the force direction. In practice, we generate 1000 random force directions to select the favorable force directions.

Tests on the number of random force vectors generated ranged from 200, 500, 1000, 5000, 10000, and up to 100000 to help determine the random sample of force vectors. In all our application of forces afterwards, we use 1000 random force directions to select the best force direction, which gives a sufficient sample and yet is
computationally effective in pushing the initial form to its target form. By scanning all the directions of forces applied to the ATP in GroEL subunit, we find that only a small fraction of the force directions yields a high overlap value between the response displacement and the experimental displacement. Figure 2.10 shows the overlap values of conformational response vectors in response to forces from all directions with the structural difference vector. Figure 2.10 (A) shows that the highest overlap between these random response vectors and the structural difference vector is above 0.8 (in red) and that the lowest overlap is around -0.8 (in blue). Since we do not want the response structure to undergo any translation and we would like to accentuate the response of moving part to observe biological relevance of changes, we realign the response structure back to the current state in the ATP binding regions which are conserved and invariant. Then we re-compute the response vector by subtracting the structural vector of the current state from the new structural vector of the aligned response structure. Each force direction is colored according to the overlap value between the response vector and the structural difference vector from the current state and the target form. We observe that the high overlap regions (overlap > 0.8) are clustered within a small range of angles defining a cone, and are opposite in direction to those regions that show the lowest overlaps. Figure 2.10 (B) shows that the favorable force vectors in magenta can push the closed form toward its targeted open form. The force direction vectors in magenta yield high overlaps (overlap value > 0.8) between the response vectors and the forces in all directions that yield low overlaps (overlap value ≤ 0.8) are colored in gray. Among 10,000 random force directions, only a small fraction (~2%) of forces
yield high overlaps between the response displacement and the experimental
displacement, demonstrating that a highly directed force can indeed drive the closed
form toward the open form.

Figure 2.10: Probing of GroEL with random forces from different directions.
(A) Overlap values between the conformational response vectors in response to forces
from all directions and the structural difference vector. Force directions are random.
The overlap value of the response vector and the structural difference vector ranges
from a highest value above 0.8 to the lowest value near -0.8, colored in red and blue
correspondingly. (B) A favorable cone of the force vectors (in magenta) that
successfully pushes the closed form of GroEL subunit toward its targeted open form.
All random force direction vectors are shown in gray, with only a small fraction
yielding high overlap values (> 0.8), shown colored in magenta.

The iterative force application approach is as follows:

1. We extract the coarse-grained representations of the starting and the target structure
   (ATP binding form and ADP binding form) $S_{base}$, $S_{target}$ and compute the structural
difference vector set $\Delta S$ after the target form has been aligned to the base starting form
in the ligand binding regions, $\Delta S = S_{target} - S_{base}$.
2. In the $k^{th}$ force perturbation iteration, we start from the base form to add a force on node $i$ which represents the phosphate group leaving with a certain force direction $F_r^i$ selected among a pool of force directions ($\{F_r^i | r = 1, 2, \ldots, 1000\}$) and compute the displacement vector $\Delta R$. The pool of force directions is illustrated in Figure 2.9.

3. We update the intermediate response by adding the response displacement vector to the previous form $\Delta S_k = \Delta S_{k-1} + \Delta R$. Since we realign response structure $\Delta S_k$ to its previous conformation $\Delta S_{k-1}$ in the ATP binding regions, we get the new response coordinate $\Delta S_k'$ and re-calculate the displacement vector $\Delta R' = \Delta S_k' - \Delta S_{k-1}$.

4. We compute the overlap $O_r$ (correlation cosine of two vectors defined in Eq.2.13) of the displacement vector $\Delta R'$ ($\Delta R'$ reflects the response of the structure to the random force $\Delta F_r$ applied on node $i$) with the structural difference vector of the current state and the target form $\Delta S_{diff} = S_{target} - \Delta S_{k}'$ in the $k^{th}$ force perturbation iteration.

5. We repeat steps 2-4 to find the maximum overlap value $O_{max}$ and its corresponding force direction $\Delta F_{max}$ which yields a response intermediate with the maximum overlap between the structural difference vector of the current state and the target form $\Delta S_{diff} = S_{target} - \Delta S_{k}'$ after all $r$ iterations ($r = 1000$ in practice) of random force application.

6. We add force once again using the same approach described above with the force vector $\Delta F_{max}$ on node $i$ to get the intermediate response $\Delta S_{max}$, the displacement vector $\Delta R_{max}'$, and the aligned response structure $\Delta S_{max}'$ for the $k^{th}$ iteration.
7. We compute RMSD of the intermediate response $\Delta S_{\text{max}}'$ with the target form in the $k^{th}$ iteration denoted as $rms_k$. If $rms_k > rms_{k-1}$, the force perturbation process stops. This means that the new current state deviates more from the target form compared to the previous iteration, and we do not expect to observe the transition of the initial base form to the target form going too far. The algorithm describing the whole force application procedure is shown in Figure 2.11.
Figure 2.11: A schematic description of the force perturbation protocol.
The diagram includes a sequence of force perturbation cycles (indexed as $k$), and each cycle consists of $r$ repetitions of force applied on node $i$. 
We apply forces iteratively on the γ-phosphate in the leaving phosphate group of ATP of the initial starting form with a force direction that yields maximum overlap in each iteration step. It is observed that the RMSD between the closed-open states decreases from 14.7 Å to 7.7 Å after adding 10 iterative small forces. The mean overlap of the response displacement vectors with the initial-target structural difference through all the iterations is 0.82. Pearson correlation between the two root mean square displacement vectors of the initial-target state and the initial-final state is as high as 0.90. The application of iterative forces to the leaving phosphate group of ATP in the closed structure of the GroEL subunit includes multiple conformations as the transition progresses from the initial closed blue form to the target open red form (See Figure 2.12).

![Figure 2.12: Conformational transitions of the GroEL subunit.](image)

The transition goes from the closed form to the open form by iteratively applying forces on the leaving phosphate group of ATP. It moves from the initial closed form in dark blue to the final form in red with light blue and light red intermediates. ATP is shown as sticks located in the lower equatorial domain.
By simply applying forces at the ATP phosphate group, we reach approximately half way through the transition. As we will see subsequently when we view the free energy landscape in Chapter 3, this approach will be enough to reach the peak of the free energy barrier, and the remainder of the transition will be downhill in free energy. The four-body potentials for each intermediate state are computed with our own knowledge-based potential server (49), which will be introduced in detail in Chapter 3. Results are shown in Figure 2.13 that most of these moves during the partial opening transition are uphill. If the leaving phosphate group is the active player, it might be anticipated that points of impact during the latter part of the transition could be further away from the original binding site.

**Figure 2.13: Potential energy of transition intermediates after iteratively applying forces on a closed GroEL subunit.**
Force application starts from frame 1 and ends at frame 10 for the final response state. Optimized knowledge-based four-body potential is computed for each frame and the potential mostly increases during the transition from the closed form (frame 1, in blue) to the open form (frame 11, in red).
2.5. Conclusions

In this study, we apply single directed forces on the GroEL to investigate its conformational responses. Our work provides evidence that force propagation is likely to originate at the reaction site in exothermic chemical reactions. Elastic network based force application reveals strongly preferred directions for the forces that can drive a protein structure towards its known targeted end point. By adding forces on the leaving phosphate group directly, we can always generate defined intermediate conformations as responses to forces and obtain a transition pathway for conformational changes. The novel methods to compute conformational transitions originating from exothermic chemical reactions, especially ATP hydrolysis, lead to significant improvement in predicting protein conformational transitions. Besides, this study can add a new area to the already stunning progress in protein structure prediction achieved over recent years. The work also makes computational contributions for advancing computational methods in studying protein conformational transitions and pathways. Some improvements can be made to this approach, such as considering the energies of transition intermediates and the randomness of forces. These issues will be discussed in Chapter 3.
CHAPTER 3  MONTE CARLO FORCE APPLICATION SIMULATION ON GROEL WITH TRANSITION PATHWAYS ON THE FREE ENERGY LANDSCAPE

3.1 Abstract

Single iterative forces based on linear response theory and an elastic network with strong directional bias applied has achieved success in driving a protein structure towards its known targeted end point. However, we are also interested in the mechanism of protein conformational transitions from the closed to the open states as well as from the open to the closed states, which may not be possible with a single directed force. We improve our force application method by combining it in a Monte Carlo simulation to introduce some randomness in the application of forces. We further visualize the transition trajectories on the free energy landscape constructed by performing principal component analysis and structural interpolation. Molecular chaperonin GroEL is a simple and well-studied structure that is a good case to investigate. Principal component analysis performed on 34 GroEL experimental structures yields their most important motions, and these are used in structural interpolation for the construction of the free energy landscape. We validate that these experimentally structure-derived motions captured by the PCs can be achieved by applying forces in specific directions; they can be used for constructing the free energy landscape. We adopt the landscape idea and compute energy landscapes for the study of protein conformational transitions. We predict and confirm that the native structures fall in low energy regions and the energy landscape can be used to identify a transition
pathway between two structural states of a protein. We demonstrate pathways for the closed-open conformational transition in both directions by computing trajectories and display these on the energy landscape. The initial RMSD between the open and closed forms of the subunit is 14.7 Å and the final forms from our simulations have an average RMSD of 3.6 Å from the targeted forms, closely matching a level of resolution to be expected for this coarse-grained model.

3.2 Introduction

Applying forces to protein structures has achieved significant success in explaining protein conformational changes, ligand binding mechanisms (95)(96), and allosteric mechanism (98) in the past few years. In Chapter 2, we have developed our elastic network based force application approach and applied it to push the GroEL subunit from its initial closed form to the targeted open form. We have proposed possible conformational transition pathways using the force application response intermediates and observed an increase in energy during the transition. However, forces applied in Chapter 2 are strongly directional and there is a strong bias in the choice of force directions. In the study in this chapter, we introduce randomness in forces and carry out a thorough analysis on the potentials and energies for each intermediates. The Metropolis Monte Carlo method is used to randomly select the response intermediate biased by the energy landscape. We hypothesize that the conformational transitions will follow a relatively low energy pathway and that they will still be able to go across energy barriers. The free energy landscape is also constructed to display transitions from the viewpoint of the energies of intermediates.
One particularly critical factor in force application is the direction in which the forces are applied. We use a biased Monte Carlo method to introduce some randomness in the direction of force application. Monte Carlo (MC) is one of the generic types of sampling methods which are used to estimate values of mathematical functions such as integrals by using random sampling (122, 123). The Monte Carlo approach was first developed in the 1940s and 50s by researchers at Los Alamos working on nuclear weapons (124). It is now widely applied throughout science, as well as to model molecular systems such as proteins (125) and membranes (126) and has become well recognized as a useful computational tool for molecular simulations and in many other fields. Several advantages of MC are (i) MC can sample conformations independently of time; (ii) MC only requires energies and not their derivatives; (iii) MC can explore conformational space broadly. While a MC method offers great flexibility in choosing random steps by which the system evolves, the Metropolis Monte Carlo (MMC) (127) in particular is a more efficient method of sampling that follows a series of steps that are primarily downhill on the energy landscape, but sometimes uphill so as to provide an unbiased sample. Instead of accepting all steps during a simulation, MMC accepts some and rejects others using a decision criterion. In this study, we utilize the MMC method in our force application approach to generate transition pathways for protein conformational changes.

Proteins in multiple conformations that undergo significant changes have been reported for many years. There are many cases of proteins with two or more different structures in different conformational states, and the Cartesian coordinates of these
structures are nowadays available in the Protein Data Bank (PDB) (www.rcsb.org), as well as in the database MolMovDB from Mark Gerstein’s laboratory (www.molmovdb.org). One straightforward and common approach has been the extraction of protein dynamics from a set of structures by using principal component analysis (PCA), a dimensionality reduction statistical method based on covariance analysis (128–130) that the Jernigan lab and others have popularized for studying protein dynamics. PCA is a technique to capture variations in the dataset for easy exploration and visualization. PCA transforms the original space of correlated positions within a set of structures into a greatly reduced space of the most important variations, i.e., principal components (PCs). By performing PCA, most of a system’s variance is captured by a relatively small subset of PCs (131), which is the primary advantage of PCA. Currently PCA is a widely accepted and common approach to characterize collective protein motions from a set of structures. It has been applied widely and most often it has been used to analyze trajectory data from MD simulations to extract the essential dynamics (132–134). Howe (135) used PCA to classify calculated structures in NMR ensembles automatically according to correlated structural variation across the population. Teodoro et al. (136) applied PCA to a dataset composed of many conformations of the same protein HIV-1 protease which transformed the original high-dimensional representation of protein motions into a lower dimensional representation that captures the dominant modes of motion of proteins. Others have shown that PCs from experimental structure sets capture the most important motions of proteins and that these also correspond closely to the normal modes from ENM (137, 138) as well
as to the essential dynamics from MD trajectories (102). PCA can be realized easily by using various supporting libraries, software and packages, e.g., MAVEN (139), Bio3d (140), ProDy (141), Scikit-learn (142), Weka (143), etc. In this study, we perform PCA from sets of experimental structures, which are available in the Protein Data Bank. We validate that these experimental structure-derived motions can be driven by adding forces. The principal components from PCA are also utilized in calculating the free energy landscape.

The term “free energy landscape” (FEL) has achieved common usage in protein folding (144–146). A FEL is a mapping of all possible conformations of a protein structure in a system together with their corresponding free energies (147). Although theoretically proteins can exist in an infinite number of conformations on its energy landscape, in reality proteins fold into 3D structures that only have the lowest possible free energy. In protein thermodynamics, the free energy landscape model was employed initially in the late 1980s to understand protein folding through the work of Ken Dill in 1985 (148), followed by Byrngelson and Wolynes in 1987 (149), among others. The key concept in the energy landscape approach to protein folding is that the energy landscape of a foldable protein looks like a rugged funnel (150, 151) (See Figure 3.1). Recent papers on the energy landscape of biomolecular function open up a new perspective for understanding the functions of biomolecules, which is “the next revolution in physicochemical biology” (152). The free energy maps reveal that the clusters of experimental structures often occur around the regions of local free energy minima (153).
Protein dynamics is often characterized by the energy landscape (154), which is a common measurement employed to describe the complex relationships among structure, enthalpy and entropy of proteins. Entropy ($\Delta S$) is a measurement of how many microstates are available for a system to sample, while enthalpy ($\Delta E$) is the thermodynamic potential for a system (155). The free energy ($\Delta G$) of a conformation is calculated as $\Delta G = \Delta E - T\Delta S$ by combining enthalpy ($\Delta E$) and entropy ($\Delta S$) scaled by temperature $T$. Free energy landscapes can clearly be used to study protein conformational transitions from the perspective of free energetics, and this has even been proposed as a second revolution in molecular biology for understanding protein functions (152). There are many commonly used approaches for constructing a free energy landscape. Some researchers collect dynamics information based upon a given 3D experimental structure by MD or MC simulations and then compute the energy for each sampled conformation. However, despite proven successes (156–162), such
methods are computationally time consuming and require very intensive computation. We value and learn from these ideas and compute similar energy landscape to ultimately study protein conformational transitions from the perspective of free energy. Our hypothesis is that the native protein crystal structures lie in low energy regions and we can use the energy landscape to identify transition pathways between two conformational states of the protein. In this study, we propose a new method of combining the PCs with our previously successful free energy estimates (163) to construct the free energy landscape to indicate possible conformational transitions derived from MMC force application.

The empirical statistical potentials, usually termed knowledge-based potentials, are calculated based on the relative preferences for contacts between different types of amino acids based on a database of known structures, under the assumption that the global free energy minimum is the native structure of the protein. Two-body statistical contact potentials were pioneered by Tanaka and Scheraga (164) and subsequently developed extensively by Miyazawa and Jernigan (165, 166) and Sippl (167). Three-body (168, 169) and four-body potentials (169, 170) were developed later to account for the three dimensional interactions among interacting amino acid clusters in densely packed protein structures. The Jernigan group developed four-body potentials (49, 50) that consider protein interactions in a more cooperative model and offer a better characterization of the interactions occurring in the densely-packed protein environments. The potential of a structure is computed as an optimized linear combination of the sequential four-body potential, the non-sequential four-body
potential, and the short-range potential (171), with each individual term computed by using our knowledge-based potential server (49).

Previously, vibrational entropy was computed based on the frequencies of the normal modes (172), or based on the mean square fluctuations from GNM as a direct measure of entropy (163). Recent methods for calculating entropy include adaptive integration and the hypothetical scanning molecular dynamics method (173). Our group also developed a new method to compute knowledge-based entropies using amino acid contact changes (unpublished work). By using a dataset of pairs of conformations of 167 diverse proteins, we extract information about the frequencies of amino acid contact changes occurring in proteins during conformational transitions. This is a procedure analogous to the way that frequencies of amino-acid contacts in a large dataset of known protein structures are used to parameterize knowledge-based potential functions: the patterns of contact changes between amino acids provide information about the entropies in protein structures by indicating which types of residue pairs are more likely to fluctuate. We compute entropies based on these amino acid contact change patterns. The free energy of a conformation is therefore obtained by combining the optimized four-body potentials described above with the new knowledge-based entropies. In this way we are able to compute free energies of experimental structures, however, other points on the PC space are required for a complete free energy landscape. This is readily carried out by interpolating the points on the PC space. Free energies are computed for all interpolated conformations on the PC space to form a complete energy landscape (174). The landscape includes sufficient
information to indicate possible transition pathways between two conformational states and any other possible protein conformational transitions.

Allostery is crucial for understanding the mechanisms of molecular machines. It is likely that the actions of most machines are driven by conformational changes due to binding and hydrolysis and that their cooperativity is responsible for their coordinated allosteric movements (175). It is well known that the intermolecular contacts in a crystal structure affect dynamics (176), and the dynamics of GroEL are affected by its overall ring type structure. GroEL has been shown to display positive cooperativity within the ring (177), however, the detailed mechanism for such cooperativity remains unknown. In the final section of our study, we show preliminary results investigating the cooperativity of GroEL subunits by MMC force application on a complete ring structure.

3.3 Methods

3.3.1 Datasets

We use GroEL subunits as structural ensemble for the construction of the energy landscape. First, we search for intact molecular chaperonin macromolecule in the Protein Data Bank to get structures with high sequence similarity (>95%). These structures are: 1aon, 4v43, 1kp8, 1kpo, 1mhf, 1oel, 1pcq, 1pf9, 1ss8, 1svt, 1sx3, 1sx4, 1we3, 1wf4, 1xck, 2eu1, 2nwc, 2yey, 3e76, and 3wvl. Owing to the fact that both cis and trans rings in GroEL are comprised of 7 identical subunits, we only take chain A in the cis ring to represent the conformations of all 7 subunits (chain A, B, C, D, E, F, and G), and chain H in the trans ring to represent chain H, I, J, K, L, M, and N. For
each structure, we collect two subunits if both cis and trans rings exist. There are two that are exceptions; protein structures 1oel and 1ss8 have only the cis ring available, so we only extract one subunit from each of these. The structure ensemble we finally obtain consists of 38 subunits. The structures are trimmed at the beginning and end residues to ensure identical protein sequence length as well as identical residues in each position. These subunits form the dataset that is also used in the construction of the free energy landscape.

3.3.2 Principal component analysis

We perform PCA on the structure ensemble and extract the most important principal components. The input for PCA is an $n \times p$ coordinate matrix $X$, where $n$ is the number of structures and $p$ is three times (for the Cartesian coordinates) the number of residues ($p = 3N$) in each structure from a set of aligned protein structures (178). Each row in the coordinate matrix $X$ represents the $\text{C}^a$ coordinates in a sequential array of length $3N$ ($x_{k1}, y_{k1}, z_{k1}, x_{k2}, y_{k2}, z_{k2}, \ldots, x_{kN}, y_{kN}, z_{kN}$) for all residues of the $k^{th}$ structure ($k = 1, 2, 3, \ldots, n$). From $X$, the elements of the covariance matrix $C$ are calculated as $c_{ij}$

$$c_{ij} = \langle (x_{ki} - \langle x_i \rangle) \cdot (x_{kj} - \langle x_j \rangle) \rangle, (i, j = 1, 2, 3, \ldots, 3N)$$

where averages over the $n$ structures are indicated by the $<>$ brackets. In the equation, $x_{ki}$ refers to the value of the $i^{th}$ variable for the $k^{th}$ structure in the dataset $X$ and $\langle x_i \rangle$ refers to the mean of the $i^{th}$ variable. The covariance matrix, $C$, can be decomposed as $C = P \Lambda P^T$, where the matrix $P$, which is comprised of eigenvectors represents the
principal components (PCs) and the eigenvalues are the elements of the diagonal matrix $\Lambda$. The eigenvalues are sorted in descending order. Each eigenvalue is directly proportional to the variance it captures in its corresponding PC. Figure 3.2 is an illustration of the PCA method.

**Figure 3.2: An illustration of principal component analysis.**

PCA transforms the dataset in the original coordinate system to a new coordinate system which captures the most variance in the dataset based on covariance analysis, thus facilitating dimension reduction. (A) A dataset in its original coordinate system. (B) A dataset in the new coordinate system specified by three arrows as three new coordinates colored in blue, green, and red. (C) The variation of the dataset in the original coordinate system. (D) The PC scores are calculated as projections of the mean centered data onto the PCs, which displays the variation of the dataset in the dimensional reduced PC coordinate system centered at zero.

### 3.3.3 Metropolis Monte Carlo simulation

We integrate the MMC method into our force application approach. The Metropolis decision criterion uses the free energy of the newly generated state $n$ in comparison with the free energy of the previous state $n-1$ to compute the probability of accepting a conformational sampling, as is defined in Equation 3.2
\[ p = \begin{cases} 
1, & G_n \leq G_{n-1} \\
\exp\left(-\frac{G_n - G_{n-1}}{kT}\right), & G_n > G_{n-1} 
\end{cases} \] (3.2)

where \( p \) is the probability of accepting the newly generated conformation in a MMC simulation, \( G_n \) is the free energy of the newly deformed conformation in the current state \( n \), \( G_{n-1} \) is the free energy of the conformation in the previous state \( n-1 \), \( k \) is the Boltzmann constant, and \( T \) is the temperature. The details to compute free energies will be introduced in the following sections. The probability computed from the Metropolis criterion is then compared with a random number in the range from 0 to 1. If the random number is smaller than the acceptance probability \( p \), we accept the new state \( n \). Otherwise the new state \( n \) is rejected since the energy increase from state \( n-1 \) to state \( n \) is unfavorable. In MMC, the downhill steps in terms of free energy are always accepted while uphill steps are only accepted with a certain probability based on the Metropolis criterion. Steps are updated regardless of acceptance or rejection of a current step. A schematic illustration of how Metropolis Monte Carlo method works is shown in Figure 3.3. The two square boxes on the left denote two states: The preceding state \( m \), and the subsequent state \( n \). A positional change in a particle in the box differentiates the two states by free energy \( E_m \) and \( E_n \). A probability is computed based on the two free energies according to equation 3.2. If the current free energy state is lower than the previous one, i.e. \( E_n < E_m \), the current state \( n \) is always accepted. If the free energy of the current state \( n \) is higher, i.e.\( E_n > E_m \), the probability is computed by the equation

\[ p = e^{-\frac{\delta E_{nm}}{kT}} \], where \( \delta E_{nm} \) is the free energy difference for the two states, i.e. \( E_n - E_m \).
A metropolis criteria $p = e^{-\frac{\delta E_{nm}}{kT}}$ is used to dictate whether to accept the current step.

In our procedure of MMC force application, a random node is first selected to add force to. A response intermediate is obtained by force application and its free energy is evaluated. A Metropolis decision criterion specifies the probability of accepting the current step according to the free energy of the current conformation, i.e., a response conformation with decreased free energy is always accepted while those with increased free energy are accepted with a probability. We start from the beginning of the procedure regardless of acceptance or rejection of the current step. Figure 3.4 below presents the details of our MMC procedure.
Figure 3.4: The procedure of Metropolis Monte Carlo force application. The force direction applied on the structure is the one that yields the best overlap to push the current conformation to its target form. The iteration stops when the RMSD between the current conformation and the target form stops decreasing, or stops when the total number of simulation steps are completed.

3.3.4 Knowledge-based potentials

The potential for the structures are estimated in Equation 3.3 as an optimized linear combination of three knowledge-based statistical potential functions: the four-body sequential potential (50), the four-body non-sequential potential (180), and the short-range potential (171). An introduction to four-body construction and the potential generation by Feng et al. is explained in details in reference (163).

\[ V_{opt} = V_{4-body\ seq} + 0.28 \times V_{4-body\ non-seq} + 0.22 \times V_{short\ range} \]  \hspace{1cm} (3.3)

The weights for the four-body sequential and the four-body non-sequential potential terms were obtained previously by minimizing the RMSD of the best decoys from homology modeling targets of CASP8 to their corresponding native structures using
particle swarm optimization (PSO) (181–183). Refer to the previous work in the Jernigan lab (181) for more details about how the weights for each of the potential terms was optimized.

3.3.5 Computation of entropy from the elastic network model

A detailed introduction of Elastic Network Model, including Gaussian Network Model (GNM) and Anisotropic Elastic Model (ANM) is presented in Chapter 2.3.3. As an extension, this study mainly discusses how entropies are computed from ENM.

Since ENMs have proven to be very successful in capturing the global most important motions of protein structures (184), it is reasonable to expect that they can also capture the entropies of diverse structures relatively well. The protein structure in ENM is represented as beads on strings, in which the beads are the $C^\alpha$ atoms of each amino acid residue and the strings specify the interaction between two spatially close beads (The cutoff distance for the determination of two beads as being spatially adjacent is usually 7 Å). Gaussian network model is the simplest type of ENM, which was developed by Bahar and Haliloglu (104) to describe vibrational fluctuations of a structure, as discussed earlier in Chapter 2.3.3. The GNM stiffness matrix (Kirchhoff Matrix $\Gamma$) describes how resistant a node is to the deformation in the context of the whole structural system while the node is restrained by the interactions with other nodes. Therefore, a deformation exerted on the system with a certain amount of energy ($k_BT$, a scaling factor for energy values in molecular systems) can determine how far each point will be displaced. Our lab’s previous work used vibrational entropies based on the frequencies of the normal modes (185), but we found significant gains by
utilizing the mean square fluctuations computed from the ENM as a direct measure of entropy (163). To compute GNM entropies, the mean square fluctuations from the ENM are used as a direct measure of entropy

\[ \Delta S \propto \Gamma^{-1} = \sum_{i=2}^{N} \frac{1}{\lambda} (Q_i Q_i^T) \]  

where \( Q \) is a normal mode vector, \( \lambda \) is the corresponding square frequency, \( \Gamma \) is the system’s Hessian, and \( \Gamma^{-1} \) is its pseudo-inverse.

### 3.3.6 Free energy and the construction of free energy landscape

As indicated earlier, we calculate the free energy as \( \Delta G = \Delta E - T \Delta S \). We simply state that \( \Delta G = \Delta E - \Delta S \) by combining the four-body potential with the ENM entropy. We sum up the sequential four-body potential, the non-sequential four-body potential, and the short range potential with weights 1, 0.28, and 0.22 respectively (163).

The experimental structures can be used directly for computing the energies; however, other points on the PC space are required for evaluation. This is more easily carried out by distorting an experimental structure in the direction of the two predominant PC vectors. PCs are combined linearly with a reference crystal structure, the one with minimum RMSD from all the others, to interpolate intermediate conformations in the two-dimensional (2D) PC space. The reference structure we use is chain H subunit of 1MNF. The combination of four-body potentials for these coarse-grained landscape conformations can be computed using our Knowledge-Based Potential Server as energy, including the sequential four-body potential, the non-
sequential four-body potential, and the short range potential, and the entropy is computed directly from the GNM.

First, for each grid point on the PC1-PC2 space, we need to transform the PC1 and PC2 score coordinates into a $3N$-dimensional structure in the original coordinate system that contains all experimental data. Coordinates of an experimental structure are necessary as a reference structure. We select the reference structure by computing the pairwise RMSD among all experimental structures and chose the one that has the smallest RMSD sum over all other structures. The coordinate of the reference structure $(R_0)$ in the $3N$-dimensional space is defined as $R_{01\times3N}$, while in the $2D$ PC1-PC2 space its coordinate is denoted as $(R_{0x}, R_{0y})$. The $3N$-dimensional coordinates $R_{1\times3N}$ (where $N$ is the total number of residues in the structure) of a representative conformation $R$ on the $2D$ PC space grid at position $(R_x, R_y)$ are obtained from

$$R_{1\times3N} = R_{01\times3N} + (R_x - R_{0x}) \times e_1 + (R_y - R_{0y}) \times e_2 \quad (3.5)$$

where $e_1$ and $e_2$ are the eigenvectors corresponding to PC1 and PC2.

After interpolating the $3N$-dimensional coordinates of all the conformations in the $2D$ PC space, the free energy of each interpolated conformation is evaluated for all conformations in the $2D$ PC space. As indicated in section 3.3.6, we calculate the free energy for construction of the free energy landscape as $\Delta G = \Delta E - T\Delta S$ by combining the knowledge-based potential energies with the GNM-based entropies.
3.3.7 Sampling along other principal components

The projection of the mean centered data onto the most important principal components PC1 and PC2 specifies the coordinates in the $2PC$ coordinate system for each experimental structure in the PCA dataset. However, this dimensional-reduction projection does not contain any information about the other PCs, which can also contribute to the variations in the structure ensemble. We have further refined this method to take the less important PCs into consideration to include the effect of these PCs on the PC1-PC2 free energy landscape. We sample each conformation taken from a grid point on the $2D$ landscape that is transformed into a $3N$-dimensional coordinate along the PC3 vector in both positive and negative directions with a scaling ranges from -50 to 50 and a stride of 5 (The scaling includes a stride of zero which indicates that the current conformation on the PC1-PC2 space is also included in the sampling). The sampling conformation is formed as

$$V_{sampling} = V_{ij} + 5k * PC3$$  \hspace{1cm} (3.6)

where $V_{sampling}$ is the current sampled conformation vector in $3N \times 1$ dimension, $V_{ij}$ is the conformation vector in $3N \times 1$ dimension on the $i^{th}$ row and $j^{th}$ column of PC1-PC2 space, $k$ is the current iteration which ranges from -10 to 10 in our sampling procedure, and $PC3$ is a $3N \times 1$ column vector. The larger the range, the better the sampling should be. However this takes extra computational time and a compromise range may be considered. As for the stride, 5 is quite rough and the smaller the stride the better the sampling results. However, it is still possible to find the conformation
with minimal free energy. Therefore, 21 conformations are sampled along PC3 for each 3N-dimensional conformation by adding up PC3 vectors multiplied by the scaling amount. The potentials and entropies are computed with the optimized weights assigned in Equation 3.3, and we select one conformation from the sampled structures along PC3 that has the minimum free energy. Sampling along PC4, or other less dominant PCs, follows the same procedure.

3.4 Results

By applying forces with greater randomness using MMC on the GroEL subunit, we obtain intermediate conformations and are able to successfully delineate the conformational transition pathways. The trajectories projected on an energy landscape also shed light on how the protein conformations change occurs.

We introduce random choices of points for application, but still close by, and apply these in random directions by utilizing the MMC method. By doing this we find that the resulting structures are significantly closer to the target form. For each force application step, we randomly choose a node from a pool of residues within 5 Å of the ATP ligand to add forces on. The direction of force is finally selected as the one that moves the structure most effectively towards the target, i.e., yielding maximum overlap between the structure difference vector and the response vector. This is a greedy approach. A Boltzmann factor of the free energy (a combination of knowledge-based potentials and entropies with optimized weights assigned) is used as the Metropolis criterion to decide whether to accept or reject the new conformation during the simulation. In order to control the simulation step, we also set RMSD between the
response and target forms as another Metropolis criterion which determines the end of process. After each iteration of force application, the response conformation is re-aligned onto the current state to remove the effects of any rigid body motions. The iteration repeats from the beginning of the process and ends when the Metropolis decision criterion for RMSD rejects the new conformation. After this MMC simulation, the RMSD between the response intermediates and the target form is evaluated in Figure 3.5.
Figure 3.5: RMSD and free energy in a MMC closed to open transition simulation of GroEL subunit.

(A) The decrease of RMSD between the conformational intermediates in response to MC force application and the target form. Force application starts at frame 1 with the RMSD equal to the difference between the starting closed form and the target open form (14.7 Å), and ends at frame 176 with an RMSD value of 3.9 Å between the final response state and the target form. (B) Free energy for each frame for intermediates during MC simulation. The values of the free energies are scaled in order to set the minimum value of energy equals zero with a unit of $kJ/(K \cdot mol)$. (C) The changes in RMSD between the conformational intermediates and the target form for each of the three domains of GroEL. RMSD decreases from as large as 23.9Å to 4.4Å in the loosely structured apical domain shown in the red line, while for the intermediate domain the RMSD decreases by only 1.2Å shown in the green line, and in the equatorial domain the RMSD increases slightly by 0.8Å shown in the blue line at the bottom of the panel.
The overall RMSD decreases faster at the beginning stages of simulation during frame 1 to frame 80, and then slows after frame 100. The RMSD of the response structure at the final step of the simulation from the target form is 3.9 Å shown in Figure 3.5 (A). It is worth noting that the number of simulation steps is allowed to be up to 2000 to ensure that the conformation can potentially undergo sufficient steps to reach its target due to the fact that we do not have any prior knowledge of how long it will take the initial form to reach the target form. Besides, the iterative forces are assigned a magnitude as small as 1 in the simulation. Under most circumstances the RMSD threshold does not function as a stop signal for the simulation, instead the simulation will keep going until the last step. After inspection of the trajectory and the conformational transition animation from the initial to the target form, it is seen that most steps after 200 steps do not contribute significantly to the transition. Instead, the structure undergoes local motions with subtle changes for some residues and domains which are usually not as important as the earlier steps. In the meanwhile, the free energy increases and decreases without any clear pattern. We therefore could truncate the trajectory by discarding those insignificant steps and focus on the first few hundred steps. An important issue to mention is how low the RMSD can go in these coarse-grained simulations. In terms of the RMSD that we can achieve, there is a limit because of the coarse-grained nature of the structure being used. It is known that the resolution in terms of electron density is a measurement of the resolvability in the electron density map of a molecule. The virtual bond length between \( C_i^\alpha \) and \( C_{i+1}^\alpha \) is fixed at 3.8 Å. Since in our study we use a coarse-grained protein structure comprised of only \( C^\alpha \)
atoms for the force application method, the overall RMSD achieved by the Monte Carlo force application simulation in Figure 3.5 (A) should be expected to reach approximately this value, and indeed this is what has been achieved. Further RMSD decreases does not represent any significant improvements in terms of resolution. In Figure 3.5 (B), the starting structure first moves uphill with increased energy, presumably to overcome a barrier and then begins sampling a set of conformations with somewhat lower energies. Figure 3.5 (C) shows RMSD decreasing in each one of three GroEL domains during the transition. RMSD decreases from 23.9 Å to 4.4 Å in the loosely structured apical domain. In the intermediate domain RMSD decreases by 1.2 Å and in the equatorial domain RMSD increases slightly by 0.8 Å. The most variant part of the GroEL subunit is its apical domain. The apical domain contains a large number of hydrophobic binding sites for unfolded protein substrates that undergo large conformational changes. The RMSD value in the equatorial domain increases subtly which can be viewed as an invariant domain due to the attribute of coarse grained system (only the Cα atom of each residue is used to build the ANM model). As for the intermediate domain which mainly functions as a linker between the apical domain and the equatorial domain which is invariant, the small decrease in RMSD is reasonable. The RMSD converges and stops decreasing after a certain step, at which point the simulation is stopped. Figure 3.6 shows the displacements of each residue for the superimposed initial-target pair (black) and the initial-final pair (red). It can be seen that the two displacement profiles overlap closely throughout the structure, and especially well for residues 200-350, which are the residues in the most variant apical
domain of the GroEL subunit. The two lines do not overlap well in the beginning or at the end of residue regions which form invariant equatorial domain. However, since we re-align the invariant domain after each force application, we can presume that this domain remains rigid during transition and does not undergo large conformational changes. When the RMSD reaches 3.9 Å, the Pearson correlation coefficient of the two root mean square displacement vectors is 0.99.

Figure 3.6: Displacement profiles of pairwise residues in the initial-target structure pair and the initial-final structure pair of GroEL.

The displacements for each two corresponding residues in the initial-target pair are computed and colored black. The displacements for pairwise residues in the initial-final pair, i.e., final response conformation after all the iterations of Metropolis Monte Carlo simulation, are colored red. The two displacement vectors are shown together for all residues and the y-axis is the displacement value in units of Ångstrom. The Pearson correlation of the two displacement vectors is 0.99.

In addition to calculating the correlation of the two displacement vectors, we also compute absolute residue-by-residue displacement differences between the two
displacement vectors, and the values for each residue are mapped onto the structure shown in Figure 3.7 (A). The details of the RMS displacements for each pair of residues are shown in Figure 3.7 (B). The mean of the absolute residue-by-residue displacement differences is 1.5 Å.

Figure 3.7: Residue-by-residue difference of two root mean square displacements, one computed from the initial-target structural pair, another from the initial-final response pair.

(A) Differences of residue-wise displacements mapped on the structure. The residue-by-residue absolute displacement difference is measured in Ångstrom for each residues in the subunit structure and the values are mapped onto the structure for each corresponding residue which are color-coded blue-white-red (blue = low, red = high).

(B) Differences of residue-wise displacements for each residue pair. The x-axis identifies the residues, and the y-axis shows the RMS displacement difference values measured in Ångstrom. The values are mapped onto the structure in part (A).

Before and after the Metropolis Monte Carlo force application, most of the displacement differences lie in the intermediate domain from the helices that link the terminal residues of equatorial domain as well as in the lower part of apical domain.
(shown as red and pale pink at the bottom of topmost part of the structure) and the topmost part in the apical domain (at the upper part of the structure). The mean square error of the displacements is 6.1 Å².

Principal component analysis is performed for the purpose of constructing the free energy landscape. We first collect the structural ensemble of GroEL and extract subunits from each of the two back-to-back symmetrical rings. We only extract one subunit from each ring structure since in each ring the seven subunits are identical. We align 38 GroEL subunit structures using MUSTANG (186). PC1 well separates the open state and the closed state in Figure 3.8 (A). PC1 is distinctively dominant; it captures 94.5% of the variance and almost completely accounts for the open to closed transition while PC2 captures only 3.9% of the variance. PC3 and PC4 include only 0.52% and 0.41% of the variance in Figure 3.8 (B). The crystal structures for PCA are projected onto the 2D PC1-PC2 space, with corresponding PC scores specifying the coordinates.
Figure 3.8: Principal component analysis of 38 GroEL subunits.

(A) The projection of crystal structures onto the PC1-PC2 space. The open forms are on the left side of the PC space with small PC1 values, and the closed forms are on the right side with large PC1 values. The two groups are indicated with arrows. (B) Percentage of variance captured by the first 10 PCs from the GroEL structures. The points on the red dashed line denote the marginal variance captured by each PC, and the points on the black dashed line denote the cumulative variance explained by a collective of PCs from PC1 to the current PC. PC1 and PC2 are dominant which explain almost 99% of the variance in the structure ensemble.

By deforming a representative structure along a PC vector, we can visualize the motions captured by the most important PCs (See Figure 3.9 (A) and (B) for motions of PC1 and PC2).
Figure 3.9: Visualization of the first two principle components.
(A) Visualization of PC1 as an opening-closing motion of the apical domain (red) and the equatorial domain (blue) moving towards each other. (B) Visualization of PC2 as the twisting motion.

In order to build a landscape on the PC space, we need to sample a broad ensemble of conformations by using these principal components from PCA. Since the first two principal components capture almost all variability in the dataset, it is reasonable and reliable to perform structural interpolation by using only PC1 and PC2. We perform deformations of the reference structure along PC1 and PC2 in PC space, i.e., each sampling conformation on the PC1-PC2 space grid is interpolated by linearly adding PCs to the reference experimental structure, as described in the method section 3.3.6. For each conformation on the PC space grid, we also perform sampling along PC3 and PC4 to find the conformation with the minimum free energy. We compute the free energy of the conformation with the combined potentials and ENM based entropies.
which correspond to the free energy used in the Metropolis Monte Carlo force application simulation.

**Figure 3.10:** The projection of GroEL experimental structures onto the free energy landscape constructed by sampling of conformations in the PC1-PC2 space for the two directions of transition pathways generated by MMC force applications.

(A) The open to closed transition with white arrows showing the transition direction.
(B) The closed to open transition with red arrows marking the transition direction. The sampling procedure starts with the combination of PC1 and PC2 and then includes PC3 and PC4 to obtain the energy minima. The free energy landscape is created by computing free energies for every interpolated landscape conformation constructed for the linear combination of PC1, PC2 with weights 1, 0.28, 0.22 for potential terms and -1 for entropy. These interpolated conformations are further sampled along PC3 and PC4 and the conformation with lowest free energy is selected. The free energies are color-coded (blue=low, red=high) and are in units of $kJ/(K \cdot mol)$. The white open circles are the experimental structures projected onto the PC space. The structures at the upper right side are in the closed form and the ones at the upper left side are in the open form. The magenta open circles mark the two states of the GroEL subunit that are used as the starting and target structures, i.e., the closed form on the right side and the open form on the left side. Three individual transition trajectories are shown for both directions as red, magenta and black dots projected onto the landscapes.
The weights for each potential term are (1.0, 0.28, 0.22) which follows our lab’s previous published study on the free energies. Entropies are computed from the Gaussian network model and the free energy landscape is computed according to the equation of Gibbs free energy. The intermediate response conformations from MMC simulations are projected onto the energy landscape.

On the PC1-PC2 space, we observe that PC1 provides a distinct separation between the closed form and the open form, and the two clusters fall in low energy regions shown in dark blue. Only a few structures deviate from these two clusters along PC2, however they still lie in relatively low energy regions. This result corresponds well to the notion that native structures generally adopt low energy states. Here three MMC simulation trajectories are shown for both the closed to open transition and the open to closed transition in Figure 3.10. Forces are iteratively applied to each intermediate response structure in MMC simulations, and these simulations were repeated three times for both directions. Each of the intermediate conformation from the MMC simulations is projected onto the energy landscape, showing a transition trajectory as a series of points. The red, magenta and black dots form the individual transitional pathways. We observe that the intermediate conformations overcome an energy barrier with a gradual increase in the free energy until near the mid-point, and beyond this point the free energy then decreases as the conformation approaches the target form. For the open to closed transition, the transition starts from the open state in the left cluster, moves mostly along PC1, and in the last stages reaches the closed state at the upper right side of the landscape by moving also along PC2. A similar
pattern is observed for the closed to open transition where the closed state on the right side of the landscape first moves along PC1 before a final traversal that includes some motion along PC2 also. By comparing the groups of three individual trajectories, the transition band formed by the three trajectories of closed to open transition shows a wider variability along PC2. This can be understood by the fact that it is usually easier for a structure to go from the open to the closed state, whereas the open state is less uniquely defined and is usually higher in energy. Although on the 2D PC space there are many outliers around the open state and randomness may yield some differences in intermediates, the overall trend of conformational change remains similar regardless of the starting random seeds for the Monte Carlo simulation. The final RMSDs reached are always below 4 Å from the target, which is an excellent final RMSD for a protein with more than 500 residues.

To test the robustness of this procedure, we further perform 40 MC runs for both the open to closed and the closed to open transition directions. The final RMSDs achieved after 200 iterations of MC steps in 40 runs range from 3.84 Å to 4.06 Å. Figure 3.11 and Figure 3.12 display the resulting statistics these 40 MC runs. Although there is randomness in the forces applied at each step, the overall trends in the transition are similar and consistent. Figure 3.11 (A) shows the RMSD decrease between the force application response intermediates and the target form during the 200 steps in the closed to open transitions. The RMSD decreases in large steps for the first 100 steps, and the same trend is found for the mean RMSD among 40 runs. The inset plot on the left is the RMSD decrease for the apical (red), the intermediate (green) and the
The equatorial domain (blue). The inset plot on the upper right side shows the changes in energies. Changes in the free energies, the entropies and the potentials are in red, green, and blue respectively. The energies generally increase during the simulation and then decrease slightly until the end of the simulation. The average final RMSD achieved for the closed to open transition is 3.95 Å. Figure 3.11 (B) shows similar results for the open to closed transition. The average final RMSD achieved for the closed to open transition is 3.8 Å. The decrease in RMSD is more convergent for all three domains as well as for the whole structure, which corresponds well to the transition trajectory on the energy landscape.
Figure 3.11: RMSDs and energies for 40 MMC simulations.
(A) The open to closed transitions. (B) The closed to open transitions. The inset plot on the left shows the decrease of RMSDs in each domain. The inset plot on the right is the energy profile for 40 simulations.
The means and standard deviations of the pairwise RMSDs among all 40 simulations are computed for each simulation step in both the open to closed and the closed to open transitions. In Figure 3.12 (A), the means of RMSD are mostly below 1 Å, which indicates that the transition band in the open to closed transitions is narrow, and there is not much variability compared to the closed to open transitions, which can be seen from the larger mean RMSD values in Figure 3.12 (B). The standard deviation values of RMSDs for the closed to open transition are also larger than those from the open to closed transitions.

**Figure 3.12:** The means and standard deviations of the pairwise RMSDs among all 40 simulations for each frame. (A) The open to closed transitions. (B) The closed to open transitions. The x-axis shows simulation steps, and the y-axis shows RMSD values measured in Ångstrom.

Apart from linearly combining PCs for structural interpolation, we also develop an alternative structural sampling approach, the DIW-LV method (distance-inverse-weight and linear-varying method), to interpolate the grid points on the PC space. This
method takes the original coordinates of all the experimental structures into consideration and constructs a hyperplane in $3N$ dimensions to interpolate the coordinates of the grid points on the PC space. Each experimental structure’s coordinate contributes to a certain grid point, and the contribution depends on the inverse distance of the experimental structure to the grid point on the PC space, which makes the function “distance-inverse-weight”. The linear-varying function is similar to the interpolation approach in which PCs are combined linearly. The details of the DIW-LV method will be covered in Chapter 4. Here in Figure 3.13 we show this alternative free energy landscape by using the DIW-LV method. While the multiple trajectories exhibit some variabilities, the trends of conformational changes are similar. The final average RMSDs achieved after 300 MC steps in the 3 runs of the open to closed and the closed to open transition are 3.3 Å and 3.9 Å.
Figure 3.13: The projection of GroEL experimental structures onto the free energy landscape constructed by the DIW-LV method in the PC1-PC2 space for the two directions of transition pathways generated by MMC force applications. The white open circles on the landscape are the projections of the 38 subunit structures from crystal structures onto the PC space. The structures at the upper right side are in the closed form and the ones at the upper left side are the open forms. Free energies (in the units of $kJ/(K \cdot mol)$) are computed for all interpolated conformations, which are color-coded (red = high, blue = low) and visualized as a contour plot. The magenta open circles are the two end states. The transition trajectories from the open form to the closed state move progressively from the left to the right, and they are shown as white dots. The transition trajectories from the closed to the open form progress from the right to the left, and they are shown as red dots. The whole process of these transitions can also be seen in the movies.
Compared with the single directed force application, the Metropolis Monte Carlo force application simulation introduces randomness in application of forces and takes the energies of intermediate forms into consideration. Free energy is used as the Metropolis criterion for acceptance of each step and the simulation can undergo free energy increases during the transition to overcome possible energy barriers. The trajectories generated from the MMC force application performed on GroEL has also been projected onto the PC1-PC2 free energy landscape to reveal the possible transition pathways.

We further compare the transition intermediate differences with the ENM normal modes. The transition intermediates start from the initial form, undergo a series of structural changes and finally reach the target form. Throughout the 300 iteration steps performed in Figure 3.13, we compute the structural change between each intermediate and its precedent and compare the structural change vectors with the ENM normal modes. The overlaps are calculated between the structural change vectors and the first 10 normal modes for both the closed to open transitions and the open to closed transitions, shown in Figure 3.14. The normal mode space we use is a combined mode space that takes normal modes from both the closed form and the open form into consideration (See Chapter 5 for the details of this method). Almost always the top overlaps are with the first 3 modes and the lower contributions are from the higher blue modes.
Figure 3.14: The overlaps of 10 combined normal modes with structural change vectors from each force application step during the open to closed transitions (left 3 panels) and the closed to open transitions (right 3 panels) for GroEL.
In Figure 3.14, subplots A1, A2, A3 (subplots B1, B2, B3) correspond to the three trajectories for the open to closed (the closed to open) transitions shown in Figure 3.13, shown in magenta, red, and black respectively. The left three panels display the overlaps of 10 most significant combined normal modes with the structural change vectors from the three open to closed transition trajectories (A1), (A2), (A3). The same is shown on the right three panels for the three closed to open transition trajectories (B1), (B2), (B3). In each subplot, the x-axis shows 300 transition steps, the y-axis shows the overlap values of individual combined normal mode with the structural change vector at each transition step. Each individual mode index that is used to compute overlap is color-coded as shown on the color bar.

Figure 3.15 and Figure 3.16 further divide the 10 normal modes into 1st - 5th modes and 5th - 10th modes for a clearer visualization. Figure 3.15 shows the overlaps of the normal modes and the structural differences for the open to closed transition while Figure 3.16 shows the overlaps in the opposite transition direction. As for the open to closed transitions in Figure 3.15, mode 1 from the combined normal modes is dominant for the initial steps of the MMC force application. It yields the highest overlaps with the structural change vectors of each iteration step for the initial 50 steps, which is an opening-closing hinge bending motion. The rest modes follow the trend in the way that the more important (lower frequency) the normal mode is, the higher the overlap it yields with the structural change vectors. For the closed to open transitions in Figure 3.16, mode 1 is dominant however mode 2 and mode 3 are also important for the initial 50 steps, but in later steps mode 2 and mode 3 are more dominant.
Figure 3.15: Overlaps of the individual combined normal modes with the structural change vectors in each MMC simulation step during the GroEL open to closed transitions.
In Figure 3.15, subplot (A) shows the overlaps for the first 5 combined normal modes, mode 1 to mode 5. The three trajectories of the MMC simulations are shown in subplots (A1), (A2), (A3) respectively. Subplot (B) shows the overlaps for modes 6 to mode 10. The three trajectories for the separate MMC simulations are shown in subplots (B1), (B2), (B3) respectively. In each subplot, the x-axis shows 300 transition steps, y-axis shows the overlap values of individual combined normal modes with structural change vectors for each transition step. Each individual mode index overlap is color-coded as shown on the color bar.
Closed to open transition

(A) Mode 1 – Mode 5

(B) Mode 6 – Mode 10

Figure 3.16: Overlaps of the individual combined normal modes with the structural change vectors in each MMC simulation step during the GroEL closed to open transitions.

(A) Mode 1 to mode 5. (B) Mode 6 to mode 10.
In general, the first three combined normal modes have the highest overlap values with the structural change vector for almost all MMC simulation steps, which dominate the conformational transitions.

To understand the relationship between the structural changes and the collective normal modes, we further compute the cumulative overlaps of the structural change vectors with the first N (N=3, 6, 10) combined modes (See Figure 3.17 (A)). We can see from the figure that the first 10 modes yield the largest overlap with the structural change vectors in almost all trajectory frames (red lines with circles). This reveals that although mode 1 is dominant in conformational transitions, the low frequency modes, such as mode 2 and mode 3, are also important in capturing structural changes. Averaging over the overlap values during a certain duration of transition frames for each individual combined normal mode (See Figure 3.17 (B)) shows that mode 1 has the largest mean overlaps among all the rest modes during 1-100 simulation frames, as well as during all transition frames.
Figure 3.17: Cumulative overlaps for the first N (N=3, 6, 10) combined normal modes and the mean overlaps of transitions for each mode.

(A) Cumulative Overlaps for the first N (N=3, 6, 10) combined normal modes in the open to closed transitions (left) and the closed to open transitions (right). The cumulative overlaps for the first 3 modes are colored in black, for the first 6 modes are colored in blue, for the first 10 modes are colored in red. (B) Averaging over the overlap values for ranges of transition frames for each individual normal mode. The red dotted line gives the mean overlap values over the first 100 transition steps in the MMC simulation. The green dotted line takes the middle 101-200 steps, and the blue dotted line takes the last 100 steps, from 201 to 300. The black dotted line shows the mean overlaps over all 300 transition steps for each mode.
To investigate allostery in response to the impact of forces in the GroEL ring structure, we carry out a MMC simulation for one ring of GroEL, which consists of 7 identical subunits in their closed forms (See Figure 3.18, where subunits are named as 1AON in the PDB file). Each closed subunit has an RMSD of 14.7 Å from its corresponding open form. Forces are applied to the single subunit A at the ATP binding pocket, mimicking the effect of ATP hydrolysis on subunit A to steer it towards its target open form. The method used here closely follows the MMC force application procedure for a single subunit of GroEL. We observe that the RMSD of subunit A drops down to 7.7 Å, which is similar as the result of single directed force application for a subunit. In addition, the forces influence the neighboring subunits. The RMSDs of B and G (two neighbors of subunit A) are affected with a decrease to 12.6 Å and 11.7 Å, respectively. Our result indicates that after pushing subunit A to its open conformation, its two neighboring subunits will start to open as well, which suggests some positive cooperativity.
Figure 3.18: Cooperativity in a 7-subunit GroEL ring during a MMC simulation. Forces are applied to subunit A, and changes are observed throughout the ring. The seven subunits are colored based on the final RMSD of the subunits from their corresponding open conformation (units of Ångstrom). The opening of subunit A in response to the applied force causes both neighboring subunits G and B to partially open. The effects are asymmetric, with a smaller effect on subunit B and a larger effect on subunit G.

3.5 Conclusions

This study shows that the experimental structure-derived motions of the GroEL subunit captured by the principal components can be driven by adding forces. Elastic network based force application reveals strongly preferred directions for the forces that can drive structure towards its known end point. By adding forces to the leaving phosphate group (Chapter 2), we can always generate defined intermediate conformations as well as a final conformation in response to forces and obtain a transition pathway for conformational changes. In this Chapter, by adding forces at the active site residues around the phosphate group randomly with the Metropolis Monte
Carlo force application method, we are able to reach a final state that is nearer to the target structure for both the open to closed and the closed to open transitions. The final states fit well into their corresponding target forms. A comparison between the transition intermediate differences and the ENM normal modes reveals that low frequency normal modes have the highest overlaps with the structural change vectors from consecutive steps in the simulation for both transition directions. In our MMC simulations, the transitional pathway follows a majority of downhill steps with a small fraction of uphill steps to pass over energy barriers and moves from one low-energy region to another. The free energy landscape on which almost all experimental structures are in low energy regions validates the notion that most crystal structures tend to adopt low energy conformation. Besides, the interpolation method along with the principal components involved in the construction of the free energy landscape developed in our work is also informative to suggest possible conformational transition pathways between the open-closed conformational states of a structure. Our study on the ring structure of GroEL raises the issue of molecular interactions between subunits and the cooperativity between subunits. The approach that we develop in this study may guide future studies on protein conformational transitions to learn more about the cooperativity in multi-subunit proteins. The strengths of the springs used in ENM are likely to affect these results, future studies can investigate different elastic network models having different types of springs, e.g., whether the subunit-subunit interactions are better represented by different strength springs than the intra-subunit interactions.
CHAPTER 4 EXPERIMENTAL STRUCTURE-BASED CONFORMATIONAL SAMPLING AND THE CONSTRUCTION OF FREE ENERGY LANDSCAPES

4.1 Abstract
In this study, we develop an efficient sampling method that performs structural interpolations for the construction of a free energy landscape based on the information intrinsic to the structural ensemble. This approach basically takes the Cartesian coordinates of all experimental structures into account and interpolates the new conformations according to their proximity to existing experimental structures on a two-dimensional free energy landscape. We investigate specific cases for the chaperonin GroEL, F1-ATPase, sarcoplasmic reticulum Ca\textsuperscript{2+} ATPase, HIV1 protease, and several other selected examples. This method successfully overcomes the restraint of dimensions and computational time in the sampling approach that samples along multiple PCs, which makes it far more efficient. The results show that the native structures are always lying in the low energy regions on the free energy landscape. Metropolis Monte Carlo force application simulations are performed on SERCA and F1-ATPase to study the conformational transitions pathways in the context of a free energy landscape. The possible pathways for these transitions are also relatively low in free energies. The RMSD achieved for SERCA with 994 coarse-grained residues is around 4 Å, and for F1-ATPase an average of 1.2 Å.

4.2 Introduction
The energy landscape has been proposed as a second revolution in molecular biology to understand the protein structure-function paradigm based on the fact that
biomolecules are not just static snapshots that exist in the expanding PDB, but instead are dynamics objects that are always interconverting among a variety of structures with different energies (152). People often characterize protein dynamics by the energy landscape of proteins (187). The concept of free energy landscape is important for portraying protein functions and dynamics because it can depict the changes in the experimental structure ensemble under varying conditions. It ultimately provides information important for studying the mechanisms and functions of biomolecules. Nowadays, people construct energy landscapes with a variety of approaches to study a wide range of biomolecular problems including but not limited to protein folding (188–194), protein dynamics and functions (195, 196), misfolding (197), binding events and mechanisms (198–200), receptors (201), regulation (202), allostery and signaling (203).

The main issue involved in the construction of an energy landscape is thoroughly sampling the conformational space. Currently, there are many computational techniques for the efficient conformational sampling of proteins. One of the most commonly used methods is to collect dynamics information based upon a given experimental structure by performing molecular dynamics simulations or Monte Carlo simulations. Zhou and Berne used an improved Monte Carlo algorithm called Smart-Walking for Boltzmann sampling of protein conformations (204). Replica Sampling (RS) Monte Carlo method was developed and is widely used. Later, Zhang et al. developed a parallel hyperbolic sampling (PHS) Monte Carlo algorithm on a side-chain-only protein model (205). This method logarithmically flattens local high-energy
barriers and allows the simulation to tunnel more efficiently through energetically inaccessible regions to access low-energy valleys. Kwak and Hansmann performed sampling of protein structures by using model hopping to study the protein folding problem (206), and Thomas Hamelryck et al. used local structural bias to sample realistic protein conformations (207). With these sampled conformations, energies can be easily computed and they eventually form a landscape. Despite the success of the above studies, the sampling method by MD or MC is usually computationally time consuming. However, the use of coarse-grained models enables a more efficient conformational sampling, enhanced by orders of magnitude (208).

In our previous studies, we built energy landscape by sampling conformations on the two-dimensional ($2D$) PC space with linear combination of PCs extracted from sets of experimental structures which are added onto a representative coarse-grained structure. Although this approach is efficient, it has several limitations. If the two PCs, usually PC1 and PC2, are not the only PCs that are important to account for the total variance in the structural ensemble, the combination of PC1 and PC2 can merely represent a subset of the motions of the dataset. Instead, some additional PCs may also contribute to the dynamics of the ensemble. Therefore, an ideal solution is to linearly combine PCs in an $N$ dimensional space where $N$ denotes the first $N$ PCs that explain nearly 100% of the cumulative variance in the dataset and then project the $N$ dimensional space with energies onto a $2D$ space (usually PC1-PC2 space), in which each grid on the $2D$ space adopts the conformation with the minimum free energy for combinations of all of the rest of PCs. This is, however a daunting task, constrained by
the computational power. A compromise approach was developed in Chapter 3, where conformational sampling is performed along other less significant PCs one at a time based on the order of their corresponding eigenvalues. Instead of exploring the whole $N$-dimensional PC space, we can only sample along a few PCs, e.g., sampling along PC1-PC4 to find energy minimum conformations for a certain grid point on the PC1-PC2 space. This method is simple and straightforward; however, it is still extremely time consuming and requires significant computational power since we need to deform each conformation on the landscape along PC3 and PC4 to compute the energies for each sampled conformation.

There are many algorithms available for interpolating unknown data given a set of data points. The griddata function in Matlab provides a variety of methods for interpolation, see Figure 4.1 as an example. However, none of these appears to be suitable for high dimensional interpolation for 3N-dimensional Cartesian coordinates of protein structures.

![Figure 4.1](image)

**Figure 4.1: The interpolation of scattered data over a uniform grid.**
The sample points are taken from the surface $z(x, y) = xe^{-x^2-y^2}$. 
In this study we propose a novel method to construct a surface to capture all experimental structures with $3N$ dimension ($N$ is the number of atoms for atomic structure or the number of $C^{\alpha}$ atoms for a coarse-grained model). The interpolation is based on the surface function and the $2D$ space on which we want to project the interpolated conformations for visualization. We conduct several case studies to show free energy landscapes by using our interpolation method that takes the Cartesian coordinates of all experimental structures into consideration. Results are shown for the molecular chaperonin GroEL, Sarcoplasmic Reticulum Ca$^{2+}$ ATPase, F1-ATPase, and other few examples selected from the 50 cases in our lab’s previous study.

SERCA (sarcoplasmic endoplasmic reticulum Ca$^{2+}$ ATPase) is a membrane protein which pumps Ca$^{2+}$ ions from the cytoplasm into the sarcoplasmic reticulum lumen of muscle cells by ATP hydrolysis. SERCA consists of a single polypeptide chain with 994 amino acids and is formed by four domains. Figure 4.2 shows the three cytoplasmic domains in SERCA. They are the nucleotide binding domain N with ATP binding site (in green), the phosphorylation domain P (in blue), and the actuator domain A (in red). The transmembrane domain is formed by ten helices that go across the membrane and the Ca$^{2+}$ binding sites are located in this domain (in light grey). Two representative forms are shown here. For the closed form on the left (PDBid: 1T5S), ATP binds in the nucleotide binding domain which is close to the actuator domain. For the open form on the right (PDBid: 1SU4), the nucleotide domain and the actuator domain undergo large conformational changes and move toward each other upon ligand binding.
Figure 4.2: A transmembrane SERCA in different conformational states.
The two grey thick bars define the membrane, with the lumen of sarcoplasmic reticulum below the bottom bar, membrane lipids in between the bars, and cytoplasm above the top bar.

Another structure that is studied in this Chapter is F1-ATPase. ATP, the main object of biological energy currency, is synthesized from ADP and inorganic phosphate by ATP synthase in an energy-requiring reaction (209). F1-ATPase, a water-soluble portion of F0F1-ATP synthase, is a rotary molecular motor made of a single protein molecule that reversibly can convert the adenosine triphosphate (ATP) hydrolysis free energy into mechanical work of the rotation of the central stalk (210). Since the first structure was solved, there have been around 30 F1-ATPase structures (without F0) deposited in the Protein Date Bank. Some have different sequences, some have different bound ligands, and others were crystallized under different conditions. Although all structures in general resemble each other, these structures actually differ to some extent. Given this situation, the set of crystal structures can be regarded as an ensemble of conformations. A systematic analysis of the ensemble may provide new
insights into the structure-dynamics-function relationship for F1-ATPase. The first X-ray structure of the functional F1-ATPase complex 1BMF (comprised of α3β3γ-subunits) resolved in 1994 provides us with rich information. In F1-ATPase, the α and β subunits alternate to make a hexametric ring and the rod-like γ-subunit is located at the center of the ring (211). Figure 4.3 shows an overview of the F1-ATPase structure. The three β-subunits take different conformations depending on the bound nucleotides. The β-subunit without nucleotide takes on a hinge-open conformation, which is denoted as βE. The β-subunit with AMP-PNP (ATP analog) takes on a closed conformation, denoted as βTP. The β-subunit with adenosine diphosphate (ADP) takes on almost the same conformation as βTP, but its catalytic αβ-interface is more tightly packed than that in αβTP (denoted as βDP). F-ATPase uses these conformational changes to produce torque to rotate the central γ-subunit.
Figure 4.3: An overview of an F-ATPase structure.

(A) A schematic illustration of the F-ATPase (Both F0 and F1 sectors). F-ATPase is composed of a catalytic F1 and a proton channel F0. The rotation of γ-stalk couples catalysis and proton transport (212). (B) The structure of the F1 complex (PDB id: 1E79) with the three αβ3-subunits and a central stalk. The α and β subunits are colored red and blue respectively. The central axle stalk consists of subunits γ (orange), ε (green), and δ (gray) and originates from the center of αβ3 complex.

Among all these cases, conformational sampling is performed based on the PCs of a dataset collected from a protein with multiple conformational states. Free energies are computed for these sampled conformations that form an energy landscape on the 2D PC space. We show GroEL, SERCA, F1-ATPase and some other cases. We further apply forces on SERCA and F1-ATPase by following the procedure in Chapter 3 using the Metropolis Monte Carlo force application simulation method. The trajectories are projected onto the free energy landscape, revealing a possible transition pathway between the two distinct conformational states.
4.3 Method

4.3.1 Datasets

4.3.1.1 GroEL

We use a dataset of 38 subunits comprised of chain A and chain H of all GroEL ring structures. See Chapter 3.3.1 for the detailed information. The reason for using only one subunit instead of seven from each ring is that the interpolation method used in this Chapter which creates a high dimensional surface for interpolation is smooth everywhere except at the sample points. The surface will not be smooth when there are many dense sample points, which will create discrete patches on the 2D PC landscape. In addition, the 7 subunits in the GroEL ring structure are identical. Therefore we can use one subunit as a representative from each ring.

4.3.1.2 Sarcoplasmic reticulum Ca\(^{2+}\) ATPase

We collect the structures of SERCA: 51 X-ray structures from 2 organisms (3TLM from *Bos taurus* and the rest 50 from *Oryctolagus cuniculus*), all from the Protein Data Bank. Some structures include multiple conformers, among which 7 structures are eliminated due to a number of missing residues. 3TLM from *B. taurus* is also removed to ensure that all the structures come from the same organism. We finally collect 56 structures for PCA. They include both chain A and chain B of 1iwo, 1vfp, 2agv, 2c9m, 2zbe, 3b9r, 3fgo, 3n5k, 3w5a, and 4bew; chain A, chain B and chain C of 1wpg; and chain A of the following structures: 1su4, 1t5s, 1t5t, 1xp5, 2by4, 2c88, 2c8k, 2c8l, 2dqs, 2ear, 2eas, 2eat, 2eau, 2o9j, 2oa0, 2yfy, 2zbd, 2zbf, 2zbg, 3ar2, 3ar3, 3ar4,
3ar5, 3ar6, 3ar7, 3ar8, 3ar9, 3b9b, 3ba6, 3fpb, 3fps, 3n8g, 3tlm, 3w5b, 3w5c, 3w5d, 4h1w, 4j2t, 4kyt, and 4nab.

4.3.1.3 F1-ATPase

In the PDB, we only search for eukaryotic F-ATPase structures. Among them, 27 are from bovine and 11 from yeast. We use only the bovine structures. The PDB ids for these experimentally solved crystal structures are: 1bmf, 1cow, 1e1q, 1e1r, 1e79, 1efr, 1h8e, 1h8h, 1nbm, 1ohh, 1w0j, 1w0k, 2ck3, 2jdi, 2jiz, 2jj1, 2jj2, 2v7q, 2w6e, 2w6f, 2w6g, 2w6h, 2w6i, 2wss, 2xnd, and 4asu. Apart from 2w6j (resolution 3.8 Å), 2w6e, 2w6f, 2w6g, 2w6h and 2w6i that have resolutions worse than 4 Å, and these 6 structures align well (pairwise RMS is around 1.3 Å for 2979 residues). We take 2w6j as a representative and remove the rest. For each F1-ATPase which is composed of three pairs of αβ-subunits that sandwich ATP catalytic sites, we split it into 3 states. The three states are assigned according to the chain ID specified during protein crystallography. Subunits with chain A and chain E are specified as αβ_E state, chain B and chain F as αβ_TP state, chain C and chain D as αβ_DP state according to Ikeguchi’s study (95). The experimental structures 2jiz, 2jj1, 2jj2 and 2wss are composed of two F1-ATPase complexes, while each complex can be split into 3 states, i.e., the αβ_E state, αβ_TP state, and the αβ_DP state. We finally obtain 78 αβ-subunit structures and unambiguously assign each αβ-subunit as a specific state among αβ_E, αβ_TP, αβ_DP. Our dataset differs from Takada's study (213) in that we take all our structures from the same organism. We also include two more recently published structures: 2xnd and 4asu, making the analysis of F1-complex ensemble more rigorous and comprehensive.
In their study they performed a systematic comparison of 29 X-ray crystal structures of F1-complexes and identified two motions that dominate the variations. They also identified the partially closed conformation $\alpha\beta_{HC}$ and found that the $\gamma$ rotation highly correlates with loosening of $\alpha\beta_E$ interface and $\alpha\beta_{DP}$ hinge motions.

### 4.3.1.4 50 other cases

We compute the free energy landscapes for the 50 protein structure ensemble studied in our previous work (174). Table 4.1 presents the details of the dataset with protein names, number of residues, number of structures, and the organism. Experimental X-ray structures for each case are collected from the PDB. Multiple structural alignment (MSA) is performed by MUSTANG followed by PCA.

Table 4.1. List of proteins, the number of residues for each structure, the number of structures for each protein, and the source organism for the 50 protein structure ensemble.

<table>
<thead>
<tr>
<th>Case</th>
<th>Protein Name</th>
<th>#Residue</th>
<th>#Structure</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sarcoplasmic/endoplasmic reticulum calcium ATPase 1 isoform (SERCA1a)</td>
<td>995</td>
<td>63</td>
<td><em>Oryctolagus cuniculus</em></td>
</tr>
<tr>
<td>2</td>
<td>Peptidyl-prolyl cis-trans isomerase A</td>
<td>159</td>
<td>136</td>
<td><em>Homo sapiens</em></td>
</tr>
<tr>
<td>3</td>
<td>Human Lysozyme C</td>
<td>131</td>
<td>218</td>
<td><em>Homo sapiens</em></td>
</tr>
<tr>
<td>4</td>
<td>B. anthracis Dihydrofolate reductase (DHFR)</td>
<td>162</td>
<td>76</td>
<td><em>Bacillus anthracis</em></td>
</tr>
<tr>
<td>5</td>
<td>Cytochrome c peroxidase, mitochondrial</td>
<td>292</td>
<td>165</td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td>6</td>
<td>HLA class II histocompatibility antigen, D-R alpha chain</td>
<td>172</td>
<td>108</td>
<td><em>Homo sapiens</em></td>
</tr>
<tr>
<td>7</td>
<td>Thaumatin I</td>
<td>202</td>
<td>80</td>
<td><em>Thaumatococcus</em></td>
</tr>
<tr>
<td>8</td>
<td>FK506-binding protein</td>
<td>108</td>
<td>59</td>
<td><em>Homo sapiens</em></td>
</tr>
<tr>
<td>9</td>
<td>Human serum albumin (HSA)</td>
<td>555</td>
<td>99</td>
<td><em>Homo sapiens</em></td>
</tr>
<tr>
<td>No.</td>
<td>Protein Name</td>
<td>Length</td>
<td>E Values</td>
<td>Organism</td>
</tr>
<tr>
<td>-----</td>
<td>--------------------------------------------------</td>
<td>--------</td>
<td>----------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>10</td>
<td>Phi6 RNA-directed RNA polymerase</td>
<td>665</td>
<td>55</td>
<td>Pseudomonas phage Phi6</td>
</tr>
<tr>
<td>11</td>
<td>Squalene synthase</td>
<td>332</td>
<td>61</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>12</td>
<td>Camphor 5-monoxygenase</td>
<td>402</td>
<td>134</td>
<td>Pseudomonas</td>
</tr>
<tr>
<td>13</td>
<td>Azurin</td>
<td>129</td>
<td>202</td>
<td>Pseudomonas</td>
</tr>
<tr>
<td>14</td>
<td>Proteinase K</td>
<td>280</td>
<td>61</td>
<td>Engyodontium</td>
</tr>
<tr>
<td>15</td>
<td>Beta-lactamase</td>
<td>359</td>
<td>143</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>16</td>
<td>Hepatitis C RNA-directed RNA polymerase</td>
<td>548</td>
<td>162</td>
<td>Hepatitis C virus</td>
</tr>
<tr>
<td>17</td>
<td>Tankyrase-2</td>
<td>186</td>
<td>64</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>18</td>
<td>Heparin-binding growth factor</td>
<td>122</td>
<td>61</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>19</td>
<td>Casein kinase II subunit alpha</td>
<td>326</td>
<td>78</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>20</td>
<td>Thioredoxin 1</td>
<td>104</td>
<td>80</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>21</td>
<td>H-2 class I histocompatibility antigen, alpha chain</td>
<td>272</td>
<td>89</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>22</td>
<td>T4 lysozyme</td>
<td>163</td>
<td>183</td>
<td>Enterobacteria</td>
</tr>
<tr>
<td>23</td>
<td>GTPase HRas</td>
<td>165</td>
<td>100</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>24</td>
<td>Heparin-binding growth factor</td>
<td>121</td>
<td>130</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>25</td>
<td>Aldose reductase</td>
<td>309</td>
<td>120</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>26</td>
<td>Phosphopentomutase</td>
<td>390</td>
<td>60</td>
<td>Bacillus cereus</td>
</tr>
<tr>
<td>27</td>
<td>MHC class I antigen</td>
<td>274</td>
<td>64</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>28</td>
<td>Carboxypeptidase B</td>
<td>304</td>
<td>58</td>
<td>Sus scrofa</td>
</tr>
<tr>
<td>29</td>
<td>HLA class I histocompatibility antigen, A-2 alpha chain</td>
<td>276</td>
<td>256</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>30</td>
<td>Chemotaxis protein CheY</td>
<td>115</td>
<td>109</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>31</td>
<td>DNA polymerase beta</td>
<td>326</td>
<td>154</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>32</td>
<td>Human Dihydrofolate reductase</td>
<td>183</td>
<td>74</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>33</td>
<td>Glucosylceramidase</td>
<td>488</td>
<td>64</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>34</td>
<td>D-alanyl-D-alanine Carboxypeptidase</td>
<td>461</td>
<td>72</td>
<td>Actinomadura sp.</td>
</tr>
<tr>
<td>35</td>
<td>WD repeat-containing protein 5</td>
<td>294</td>
<td>80</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>36</td>
<td>LeuT Transporter</td>
<td>503</td>
<td>45</td>
<td>Aquifex aeolicus</td>
</tr>
<tr>
<td>37</td>
<td>Cathepsin S</td>
<td>217</td>
<td>58</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>38</td>
<td>Thermolysin</td>
<td>317</td>
<td>122</td>
<td>Bacillus</td>
</tr>
<tr>
<td>39</td>
<td>Polymerase</td>
<td>458</td>
<td>55</td>
<td>Human poliovirus 1</td>
</tr>
<tr>
<td>40</td>
<td>Hen –egg white lysozyme</td>
<td>130</td>
<td>586</td>
<td>Gallus gallus</td>
</tr>
<tr>
<td>41</td>
<td>Beta-2-microglobulin</td>
<td>100</td>
<td>242</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>42</td>
<td>Phospholipase A2</td>
<td>122</td>
<td>80</td>
<td>Daboia russellii</td>
</tr>
<tr>
<td>43</td>
<td>Beta-lactamase TEM</td>
<td>260</td>
<td>59</td>
<td>Escherichia coli</td>
</tr>
</tbody>
</table>
Table 4.1 Continued

<table>
<thead>
<tr>
<th></th>
<th>Enzyme Name</th>
<th>Number</th>
<th>Enzyme Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>Guanyl-specific ribonuclease T1</td>
<td>105</td>
<td>Aspergillus oryzae</td>
</tr>
<tr>
<td>45</td>
<td>E-coli Dihydrofolate reductase</td>
<td>160</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>46</td>
<td>Insulin-degrading enzyme</td>
<td>942</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>47</td>
<td>Cationic trypsin</td>
<td>224</td>
<td>Bos taurus</td>
</tr>
<tr>
<td>48</td>
<td>Elastase 1</td>
<td>241</td>
<td>Sus scrofa</td>
</tr>
<tr>
<td>49</td>
<td>Endothiapepsin</td>
<td>331</td>
<td>Endothia parasiatica</td>
</tr>
<tr>
<td>50</td>
<td>Macrophage metalloelastase</td>
<td>153</td>
<td>Homo sapiens</td>
</tr>
</tbody>
</table>

4.3.2 Weighted contribution method (WCM)

Please refer to Appendix A for details.

4.3.3 Knowledge-based entropies

Knowledge-based entropies are used in this Chapter for the construction of the free energy landscape. This is a novel method recently developed in the Jernigan group (unpublished work). By using a dataset of pairs of conformations of 167 diverse proteins from the molmovdb database from Gerstein’s group (69), information about the frequencies of amino acid contact changes occurring in proteins is extracted. This is a procedure analogous to the way people use frequencies of amino-acid contacts in a large dataset of known protein structures to parameterize knowledge-based potential functions. The patterns of contact changes between amino acids provide information about the entropies in protein structures by indicating which types of residue pairs are more likely to fluctuate. We compute entropies based on these amino acid contact change patterns, and it is termed amino acid contact change (AACC) entropy.
4.3.4 The construction of the free energy landscape

For the details on how to calculate energies, please refer to Chapter 3.3.6. In this study, the conformational sampling is performed as follows. For each sample conformation with grid point coordinate \((u, v)\) 2D landscape space, we construct a weight function \(f(u, v)\). The \(3N\) dimensional coordinate in the original space for this grid point \((u, v)\) is contributed by the given \(3N\) -dimensional coordinates of all experimental structure samples, each with a certain weight \(w\), where \(0 \leq w \leq 1\). The function to compute the coordinate for a grid point is shown in Appendix A Eq. A.1. The weight function for the grid point that is contributed by a certain experimental structure (sample \(i\)) is

\[
W_i = \frac{\text{inverse distance of grid point}(u,v) \text{ to sample } i}{\text{sum of inverse distance of gridpoint}(u,v) \text{ to all samples}}
\]  

(4.1)

where the maximum weight for sample \(i\) is 1 when the grid-point \((u, v)\) is at the same position as sample \(i\), and the minimum weight for sample \(i\) is 0 when the grid-point \((u, v)\) is at the position of another sample \(j \ (j \neq i)\).

4.4 Results

4.4.1 A case study for GroEL

For a collection of experimental structures which is a given set of points, we use the weighted contribution method (see Appendix A) to construct a surface that goes across all the experimental structures. The boundary of the surface is determined by the boundary of the \(2D\) space. We construct the landscape by using PC1 and PC2, which
are the same as were used in Chapter 3. $3N$-dimensional coordinates are constructed for each point on the surface which corresponds to the projection on the 2D PC space. We also compute the knowledge-based potentials as well as entropies for these coarse-grained conformations. This same procedure is used for all cases.

In Figure 4.4, we perform structural interpolation in PC1-PC2 space by computing the coordinates of each grid-point on the PC space using the DIW-LV (Distance-Inverse-Weight and Linear-Varying) in the WCM method to build a surface function for structural sampling. The high-dimensional surface can be visualized when projected onto a 3D space with a combination of two PCs, as is shown in Figure 4.4. Figure 4.4 (A), (B) and (C) form 3-dimensional PC1-PC2-PC3, PC1-PC2-PC4 and PC1-PC2-PC5 spaces correspondingly. We observe that the surface is smooth everywhere except in the regions of experimental structures. The color bar shows the projection score of a third PCs in each panel, i.e., PC3 in (A), PC4 in (B) and PC5 in (C) along the Z-axis from a high dimensional surface to a 3-dimensional space. Instead of using a linear combination of PC1 and PC2 to interpolate the conformations on a 2D landscape based on only one representative structure in Chapter 3, this method includes all PCs as well as uses the coordinates of all experimental structures for structural sampling.
Figure 4.4: Visualization of a 3N dimensional surface in 3D space for the GroEL specified by three PCs. (A) Projection of surface on PC1, PC2, and PC3. PC1 and PC2 determine the boundaries of the surface while coordinates along PC3 for each experimental structures are different. The color bar specifies values scores on PC3. (B) Projection of surface on PC1, PC2, and PC4. (C) Projection of surface on PC1, PC2, and PC5.

We then compute the free energies of each sampling conformation on the surface by combining the sequential four-body potential, the non-sequential four-body potential, the short-range potentials and the entropies. The details of each energy term are shown in Figure 4.5. The four-body potentials and the non-sequential four-body potentials show low energy regions along small PC2 values, while at high PC2 values...
these energies are relatively high. The short-range potential further brings down the energies at small PC2 values that span along PC1.

![Figure 4.5. Components of the knowledge-based potential terms in the free energies for GroEL.](image)

(A) Four-body potential, (B) non-sequential four-body potential, and (C) short-range potential for each interpolated conformation on the surface which is projected onto the 2D PC1-PC2 space. The x and y axes are for PC1 and PC2. The color code specifies low energies (in blue) and high energies (in red).

Both GNM and AACC entropies are computed for the sampling conformations (see Figure 4.6). The AACC entropies show higher entropy values at the middle of PC1 which span over along PC2 in Figure 4.6 (B). The free energy landscapes computed with two different entropy terms are shown in Figure 4.7. Since PC1 distinguishes two
distinct conformational states between the open form and the closed form, the addition of AACC entropy term in the free energy landscape calculation can bring down the barrier lying between the two states, as we can observe in Figure 4.7 (B).

Figure 4.6: Two profiles of entropies.
(A) GNM entropy. (B) AACC entropy. The x and y axis specifies PC1 and PC2. The AACC entropy uses a cutoff of 8 Å for the coarse-grained models.

Figure 4.7: Free energy landscape of GroEL using (A) GNM entropy and (B) AACC entropy.
The weight is 1 for the four-body potential, 0.28 for non-sequential four-body potential, and 0.22 for short-range potential. The weight for entropy is -1. The two clusters of different conformational states are labelled in white. The representative structures that distinguish different states are shown as magenta open circles. Using the AACC entropies permits the structures to pass between the open and closed forms more readily.
4.4.2. A case study for SERCA

The SERCA complex takes two major forms: the open form and the closed form. PCA is performed on the 56 structures by MUSTANG for those with no missing residues. Figure 4.8 shows the projection of SERCA structure ensemble on the PC1-PC2 space labelled by bound molecules. On PC1-PC2 space, the open forms all have $Ca^{2+}$ bound in the membrane helices which lie on the upper right of the PC space with high values of PC1 and PC2. The closed forms which lie on the lower right of PC space with high PC1 and low PC2 can bind both ADP, ATP and ACP (ATP analog). Another cluster with low values of PC1 and PC2 corresponds to structures where the A and N domains have rotated, but not opened. In the structure ensemble, there are many structures lying in the low PC1 region, centered around zero on the PC2 coordinate. The inset legend displays the variance explained by each principal component. The first three principal components capture dominant motions of structural dynamics. PC1, PC2, and PC3 explain 57.17%, 27.40% and 10.24% variance respectively. The cumulative variance explained by the first three PCs is 94.81%. Movies of the three dominant principal components show that PC1 captures the twist motion of the actuator domain A and PC2 explains the opening-closing motion of the nucleotide binding domain N.
Figure 4.8: Projection of SERCA structure ensemble on the PC1-PC2 space labelled by bound molecules. The binding residues are color-coded shown in the upper left legend. Variances captured by PCs are shown in the text box up in the middle of the figure.

The potentials and AACC entropy terms used in computing the free energies of each sampled conformation are shown in Figure 4.9. There is a large oval-shaped high energy barrier in the middle of PC1 in (A), however, the effect is counteracted by including the AACC entropy in (D). (B) and (C) show three low energy valleys, which agree well with the location of the clusters of experimental structures in Figure 4.8.
Figure 4.9: The knowledge-based potentials and the AACC entropy terms in the free energy landscape of SERCA. 
(A) Four-body potential. (B) Non-sequential four-body potential. (C) Short-range potential. (D) AACC entropy.

The free energy landscape of SERCA is show in Figure 4.10. We observe three major clusters that are separated by PC1 and PC2. A MMC force application simulation for the open to closed transition is performed with forces applied in the nucleotide binding domain N. We select E1.2Ca^{2+} (PDBid: 1SU4) as the starting structure and E1.ATP (1T5S) as the target structure. The simulation follows the procedure in Chapter 3 and runs for 3000 steps. Figure 4.11 (A) shows the results of only the first 100
simulation steps. The initial RMSD between the two states is 19.1 Å and the final RMSD for the initial-final pair is 4.1 Å (See the top panel in Figure 4.11 (A) ). During the first 50 steps, the directions of force can always yield large overlaps between the structural difference vector and the response vector, which in turn makes the starting structure undergo large conformational changes (RMSD decreases from 19.14 Å to 5.8 Å) to reach the target rapidly (shown as a large magenta circle in Figure 4.10). Later on, the application of force mainly contributes to smaller refinements. These simulation steps do not significantly change the transitional motions, and they only bring the RMSD down by an additional 1.7 Å (shown in smaller magenta circles in Figure 4.10). The transition follows a relatively low free energy path shown in the middle panel in Figure 4.11 (A). The RMSD between the current form with the target form for each domain in SERCA changes from 10.8 Å to 4.9 Å for the actuator domain, from 4.5 Å to 2.1 Å for the phosphorylation domain, from 35.8 Å to 2.5 Å for the nucleotide binding domain, and from 5.3 Å to 5.0 Å for the transmembrane domain (See the bottom panel in Figure 4.11 (A)). The Pearson correlation between the displacement vectors in the initial-target pair and the initial-final pair is 0.98, the mean of residue-by-residue displacement differences is 2.0 Å, and the mean square error of displacements is 12.4 Å² (See Figure 4.11 (B)). The transition trajectory goes through some midpoints, i.e., chain A and chain B of 3W5A and chain A of 3W5B that lie between the open and closed states. The RMSD between the intermediate conformation and these midpoints is around 5 Å, which suggests that these midpoints may play an
important role and serve as intermediate structures during the conformational transition.

**Figure 4.10:** Free energy landscape of SERCA and the conformational transition from the open to the closed form by a MMC force application simulation. The experimental structures are shown in white open circles. The two representative states are shown in magenta circles with black labels. Free energy values are color-coded from blue (lowest) to red (highest). The conformational pathway for the open to closed transition is the magenta points.
Figure 4.1: MMC of SERCA.
(A) RMSD and free energy in a MMC open to closed transition. The overall RMSD decreases from 19.1 Å to 4.1 Å during the transition. RMSD decrease for each domain is color-coded in the figure legend of the bottom panel. (B) Displacement profiles of pairwise residues in the initial-target structure pair and the initial-final structure pair. The displacements for each two corresponding residues in the initial-target pair are computed and colored in black. The displacements for pairwise residues in the initial-final pair are colored in red. The differences of residue-wise displacements are mapped onto the structure shown in the inset of (B). The largest differences are in the actuator domain and its connected membrane helices.

4.4.3 A case study for F1-ATPase

All 78 αβ-subunits in the dataset are atomic structures without solvent or ligand/ion binding. The MSA is carried out with MUSTANG. Generally the protein structure ensemble obtained from the PDB is not always of the same length or they always have the same corresponding residues. There are usually structures with residues with no coordinate information, resulting in gaps in the alignment. It is not desirable to include those structures that have too many gaps. However, since there are many structures with only a limited gaps in the α subunit and there are some structures such as 2v7q and 1ohh bound with an inhibitory protein IF1, we do not remove these
structures from the dataset. Instead, we trim the sequences from the larger structures to have the same dimension after MSA by MUSTANG. These properly aligned and carefully processed position coordinates for each residue in the ensemble serve as the input data for PCA. Figure 4.12 shows the distribution of pairwise RMSDs of MSA within the dataset. There are 3003 pairwise RMSD values for each pair of MSA aligned structures. From the distribution, we can observe three peaks centered approximately at RMSD 0.5Å, 2Å and 4.5Å, which separate the F1 complex into 3 distinct states. Although the peak reaches a maximum at RMSD of 4.5Å, the total counts of RMS values in three groups are roughly equal since among the 78 subunits, the amount of $\alpha\beta_E$, $\alpha\beta_{TP}$ and $\alpha\beta_{DP}$ states are equal. Slight difference in the three clusters could be explained by states with mutants or inhibitors.

![Histogram of pairwise RMSDs](image)

**Figure 4.12: Distribution of pairwise RMSDs among 78 $\alpha\beta$-subunits of F1-ATPase.**
The histogram quantifies pairwise RMSD values after MSA. RMSD values range from a minimum of 0.1Å to a maximum of 5.1Å, with an average of 2.7Å.

Figure 4.13 shows the projection of the experimental structures onto the PC1-PC2 space and the PC1-PC3 space. We examine the subunits in PDB files by the
SEQADV record which identifies conflicts between sequence information given in the SEQRES and the sequence database entry given in the DBREF. Records with engineered, modified residue or conflict are all defined as mutants. We find that out of the 78 subunits, there are only 6 native-sequence subunits, with the corresponding native structures 2v7q and 2w6j. The mutations mostly lie in the $\alpha$-subunit. There are some structures located at the middle of the PC1-PC2 space in Figure 4.13 (A). Two of them are in the $\alpha\beta_{DP}$ and $\alpha\beta_{TP}$ states from the native structure 2v7q and one in the $\alpha\beta_{DP}$ state from 1ohh. These three subunits are bound with IF1 inhibitory protein. In addition, in the lower left corner, there is a subunit in the $\alpha\beta_E$ state from 1h8e. This structure cannot be categorized into any of the three clusters because it binds ADP, while all the other $\alpha\beta_E$ states are ligand free. The first 3 most important PCs clearly identify three major clusters, i.e., $\alpha\beta_E$, $\alpha\beta_{DP}$ and $\alpha\beta_{TP}$. The cluster $\alpha\beta_E$ is located on the left side of PC space centered at zero along PC2 with a low PC1 score. The cluster $\alpha\beta_{DP}$ is on the lower right of PC space with low PC2 score and high PC1 score. The cluster $\alpha\beta_{TP}$ is on the upper right of PC space with high PC1 and PC2 scores. In Figure 4.13 (B), PC1 well separates $\alpha\beta_E$ from $\alpha\beta_{TP}$ and $\alpha\beta_{DP}$, while PC3 does not separate $\alpha\beta_{TP}$ and $\alpha\beta_{DP}$. PC3 plays a relatively subtle role in distinguishing the conformational states. Among the PCs, PC1 is dominant, accounting for 84.7% of the structural variance in the ensemble, which is an unusually high fraction of the total motions within one PC. PC2 and PC3 capture 12.0% and 0.72% of the variance. The cumulative variance captured by the first two PCs is 96.74%, while by 10 PCs it is 98.89%.
Figure 4.13: The projection of the experimental structures F1-ATPase onto the PC1-PC2 space and the PC1-PC3 space.

Native subunits are shown in blue filled circles and the mutants are in red open circles. The text labels $\alpha\beta_E$, $\alpha\beta_{DP}$ and $\alpha\beta_{TP}$ specify the three distinct states. (A) The projection of the dataset onto the PC1-PC2 plane. Inhibitor/ligand bound states 2v7q_D, 1ohh_D, 2v7q_T and 1h8e_E are labelled. (B) The projection of the dataset onto the PC1-PC3 plane.

Animation of PC1 and PC2 on the static snapshot of the structure provides us with visualization of the motions. PC1 in Figure 4.14 (A) captures the opening-closing hinge bending motion and it separates $\alpha\beta_E$ from the clusters of $\alpha\beta_{DP}$ and $\alpha\beta_{TP}$. The hinge is centered near the catalytic site where the ligand is bound. The biggest conformational changes in the opening-closing motion mainly occurs at the bottom part of the $\beta$ subunit. The $\alpha$ subunit remains rigid and only shows rigid body motions. PC2 in Figure 4.14 (B) further captures $\alpha\beta$ interface motions and it separates the cluster $\alpha\beta_{DP}$ from $\alpha\beta_{TP}$. The lower bottom part of the $\alpha\beta$ subunits under the ligand binding site shows significant interplay opposing each other, i.e. the two parts either move toward or against each other synchronously.
Figure 4.14: Animation of PC1 and PC2.
The motions are shown in black arrows. The $\alpha$ subunit is in blue and the $\beta$ subunit is in red. Ligands (ATP or ADP) are bound at the catalytic site of the $\beta$ subunit shown in colored spheres. (A) Hinge-bending motion of the $\beta$ subunit captured by PC1. (B) Interface motion of the $\alpha\beta$ subunit captured by PC2.

The free energy landscape of F1-ATPase is shown in Figure 4.15. The three states are distinguished by PC1 and PC2 and they fall in low energy regions on the landscape. The energy landscape provides us with a general picture of the energetics of crystal structures. It also points out possible low-energy pathways for conformational transitions. MMC force applications are performed from the $\alpha\beta_E$ state to the $\alpha\beta_{TP}$ state and from the $\alpha\beta_{TP}$ state to the $\alpha\beta_{DP}$ state. All three conformational states are extracted from 1BMF. The RMSD decreases from 6.6 Å (the initial-target pair) to 1.5 Å (the final-target pair) during the transition (See Figure 4.15 (A)) from $\alpha\beta_E$ to $\alpha\beta_{TP}$, and from 3.2 Å to 0.95 Å during the transition from $\alpha\beta_{TP}$ to $\alpha\beta_{DP}$ (See Figure 4.15 (B)).
Figure 4.15: Transition pathways on the free energy landscape for F1-ATPase.
The experimental structures are shown in the white open circles, and the three clusters are labelled. Representative structures for each state which are used in the MMC force application simulations are shown in the magenta open circles. The transition directions are shown with red arrows. (A) Transition from the ligand free form $\alpha\beta_E$ to the ATP-binding form $\alpha\beta_{TP}$. (B) Transition from the ATP-binding form $\alpha\beta_{TP}$ to the ADP-binding form $\alpha\beta_{DP}$.

4.4.4 Some other cases

Figure 4.16 shows some of the free energy landscapes selected from the 50 cases including Squalene synthase, GTPase HRas, Glucosylceramidase, and Peptidyl-prolyl cis-trans isomerase A. Refer to our previous work for details of these protein structures (174). We can observe that most of the experimental structures adopt energetically favorable conformations lying in low energy regions on the free energy landscapes.
Figure 4.16: Free energy landscapes from the 50 cases.
(A) Squalene synthase, (B) GTPase HRas, (C) Glucosylceramidase, and (D) Peptidyl-prolyl cis-trans isomerase A. The two distinct conformational states are shown in the magenta and the black open circles. The structures that have the minimum average RMSDs among all the structures are colored in magenta, and the structures that have the maximum average RMSDs among all the structures are colored in black.

4.5 Conclusions

Protein conformational sampling plays an important role in the construction of free energy landscapes. The interpolation method discussed in this study is successful to incorporate the Cartesian coordinates of various conformational states found in the
experimental structures of a protein. It interpolates to generate a full high-dimensional surface that passes through the available experimental structures. These newly generated conformations are further used to build a free energy landscape in a 2D perspective formed by the slowest two PCs. The method keeps the maximum information from the original X-ray dataset and is computationally efficient compared with the linear interpolation method using only PCs. The case studies suggest possible conformational transition pathways that go between one low energy regions to another through energy barriers. The MMC force application simulation on SERCA drives the conformational transition from the open state (E1) to the closed state (E2) successfully by bringing RMSD down from 19.1 Å to 4.1 Å. In F1-ATPase, the three experimental structure clusters fall into low energy regions on the landscape. The forces applied suggest that ATP-binding and hydrolysis may play an important role in the transitions among the three states. The study on the 50 cases shows that most of the crystal structures tend to fall into regions of relatively low free energies, and the new landscapes clearly help to predict conformational transition pathways.
CHAPTER 5  A NEW BASIS FOR COMBINING ENM NORMAL MODES FROM DIFFERENT CONFORMATIONAL STATES

5.1 Abstract

A straightforward way to further analyze and validate the transition pathways is to compare the transition steps from MMC simulation with the normal modes from ENM. This study brings up a novel method to account for protein dynamics by combining the dynamics from multiple conformations of a protein. To achieve this, a combined normal mode space is computed and used to compare with the structural changes observed on the MMC transition pathways. From the study on GroEL, fairly high overlap values between the low frequency modes and the transitions are observed, indicating that the transition pathways agree with the most important normal modes from ENM. The combined normal mode space, of which the first two modes are used as the landscape coordinates, can be utilized in the construction of the free energy landscape. The study on adenylate kinase, as well as previous results, show that most experimental structures fall in the low energy regions. These combined modes are also compared with PCs, suggesting that they can well capture the dynamics of a protein structure.

5.2 Introduction

Normal modes from ENM are widely used to delineate the dynamic motions of a protein structure. Proteins with specific conformational states, especially those with open and closed forms, also sometimes referred to as relaxed and tense forms, can give two sets of normal modes with ENM from each of the two structures that describe their
dynamic motions. These two set of normal modes usually are usually quite similar, i.e., high overlaps. However, indices of normal modes can be shuffled, and they are sometimes mismatched in their indices.

Both the modes from the closed form and the open form of a protein structure describe the dynamics. However, since these experimental structures are all snapshots from physical functional protein structures crystallized by X-ray, we hypothesize that they should all participate in protein dynamics. Therefore, the change of one conformation to another and the reverse process should take on similar motions, especially if the pathways for the two transitions show strong coincidence. Although some studies show that the forward and reverse transitions of two end states are not the same, the combined motions provided by the new method yields a set of mode vectors that always have higher overlap with modes of motions from both states, which is a straightforward solution when we need one set of motions to characterize the dynamics of a protein structure.

In this chapter, the application of this method is shown for GroEL. We use the combined normal modes to analyze and validate the trajectories from the MMC force application simulation for GroEL. We also perform a detailed study on adenylate kinase (ADK) and use the combined modes to construct a free energy landscape. ADK is a frequently and well-studied enzyme that displays large-scale motions of its LID and NMP-binding domains in the apo open form upon binding of its substrate (ATP/AMP) or an inhibitor (AP5A) and closes to form the holo form (214). It is a monomeric
phosphotransferase enzyme that catalyzes the interconversion of adenine nucleotides. 

*E. Coli* ADK consists of 214 amino acids and has 3 domains: the ATP binding domain-LID (residues 122-159), the NMP binding domain-NMP (residues 30-59), and the core domain (the remainder of the residues) (215). The LID and NMP domains open and close during its catalytic cycles, resulting in a 7.2 Å difference in RMSD between the open conformation (PDB: 4AKE) and the closed conformation (PDB: 1AKE). Finally we use this method to study more cases by comparing the overlaps between PCs and the modes.

5.3 Method

5.3.1 Dataset

5.3.1.1 Dataset of GroEL

In this study we use two conformational states from the molecular chaperonin GroEL to generate combined modes. The closed state is taken as chain A of 1KP8 and the open state as chain A of 1AON. The two conformational states are trimmed to have the same number of residues. The PCs are obtained from PCA performed in Chapter 3.3.1.

5.3.1.2 Dataset of adenylate kinase

We obtain 63 protein structures ADK from the PDB and separate chain A and chain B as two conformations. This yields 105 structures in total. However, since our study computes combined normal modes, missing residues may perturb the computations. Exception is made for those sequences with only missing residues at the
beginning and the end of the sequence. Therefore, we remove the structures with gaps in the sequence and finally we have a set of 63 structures. The dataset is comprised of both two chains (chain A and chain B) of 1ake, 1ank, 1e4v, 1e4y, 2eck, 2eu8, 2oo7, 2ori, 2osb, 2qaj, 3dl0, 3hpq, 3hpr, 3x2s, 4ake, 4jzk, 4mkf, 4pzl, 4qbh, 4qbi, 4tyq, 4x8l, and 4x8o, chain A of 1s3g, 1zin, 1zio, 1zip, 2p3s, 2xb4, 3dkv, 3fb4, 3l0p, 3l0s, 4k46, 4mkg, 4mkh, 4qbf, 4x8h and 4x8m, and chain B of 4qbg. Each structure contains 216 residues and MSA is performed by using MUSTANG.

5.3.1.3 50 other cases

The 50 cases used in this study come from our recent work (174). Please refer to Table 4.1 in Chapter 4.3.1.4 for details.

5.3.2 Construction of new basis

Given two orthonormal basis \( C = (c_1, c_2, c_3, \ldots, c_n) \) and \( O = (o_1, o_2, o_3, \ldots, o_n) \) in \( \mathbb{R}^n \). Denote \( K = \{k_{ij}\} \) where \((c_i, o_j) = k_{ij}\) and has properties

\[
O = CK, \quad C = OK^T
\]  

(5.1)

\( P = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \) is comprised of eigenvectors of \( KS \), where \( S \) is the sign matrix of matrix \( K \), i.e, the diagonal matrix \( S \) corresponds to the sign of the diagonal entries in \( K \). We will have

\[
L = PHDP
\]  

(5.2)

where \( L \) is the square root of matrix \( KS \), \( PH \) is the transpose and complex conjugate of \( P \), and \( D \) is the square root of eigenvalue matrix.
The new Basis $X$ that combines C and O is

$$X = CL = OSL^T \quad (5.3)$$

Please refer to Appendix B for more details.

### 5.3.3 Comparing normal modes and principal components

A normalized overlap that compares the $i^{th}$ mode (Mode with index $i$ in the mode space, i.e., $Mode_i$) with the $j^{th}$ PC (PC with index $j$ in the PC vectors matrix, i.e., $PC_j$) was suggested by Marques and Sanejouand (216). In this study we use this metric to compare the normal modes from ENM on a specific structure and the principal components from the experimental structure ensemble. The normal modes for a representative structure are represented as a set of vectors, which share the same dimensions as the principal components. The overlap value is given in Equation 5.4

$$O_{ij} = \frac{|Mode_i \cdot PC_j|}{|Mode_i||PC_j|} \quad (5.4)$$

The Cumulative Overlap (CO) is defined as the square root of the cumulative overlap values between the first $k$ normal modes and a specific individual principal component. CO indicates how well the first $k$ normal modes collectively capture the variance of structure ensemble in a certain principal component. It is given in Equation 5.5:

$$CO(k) = \sqrt{\sum_{i=1}^{k} O_{ij}^2} \quad (5.5)$$
But, this procedure is useful for comparing the directions of motion between any two representations of the dynamics.

5.3.4 Free energy landscape on the mode space

In addition to the approach of computing free energies for a $2D$ PC space in Chapter 4, a similar procedure can be carried out for a landscape projected onto a $2D$ mode space. Since low frequency normal modes capture global motions and are the most important ones, in this study we use the first and the second modes to construct a $2D$ landscape. The conformational sampling approach to interpolate structures on the $2D$ landscape is performed by using the method in Chapter 4. A surface is constructed for the experimental structure ensemble and each grid point on the $2D$ space corresponds to the projection of a high-dimensional conformation from the surface.

5.4 Results

5.4.1 A case study on GroEL

In Chapter 3, the overlaps between the structural difference vectors of two consecutive conformations during a MMC force application simulation and modes are computed for each simulation step (shown in Figure 3.14). The modes we use to perform the calculation is a combined mode space that takes both the normal modes from the closed form and those from the open form into consideration. The combined modes are generated by following section 5.3.2 to obtain a new mode space with orthonormal vectors.
Figure 5.1: Overlaps of ENM normal modes from the closed form (left) and the open form (right) and cumulative overlaps of the first $k$ modes ($k = 3, 6, 10, 20$) with the first three PCs.

(A) Overlaps of the first 10 ENM normal modes from the closed form with the first 3 PCs. The x-axis shows ENM mode indices, and the y-axis shows indices of PCs. (B) Overlaps of the first 10 ENM normal modes from the open form with the first 3 PCs. (C) Cumulative overlaps of the first $N$ ($N = 3, 6, 10, 20$) ENM modes from the closed form with the first PCs. (D) Cumulative overlap of the first $N$ ($N = 3, 6, 10, 20$) ENM modes from the open form with the first three PCs. The values above 0.2 are in white labels.

We first perform ENM on the GroEL single subunit in the closed and the open forms respectively. PCA is performed on the 38 subunit ensembles collected from the
experimentally crystallized structures. Refer to Chapter 2.3.3 for the details of the ENM method and Chapter 3.3.2 for PCA. Overlap is computed using the normal mode vectors and the first three principal components (see Eq. 5.4). Cumulative overlap is computed over the first $k$ normal modes for $k = 3, 6, 10, 20$ (see Eq. 5.5).

For the GroEL closed form, PC1 has its highest overlap with modes 3 and 4, and PC2 has an overlap value of 0.28 with mode 1 (see Figure 5.1 (A)). In Figure 5.1 (B), the overlap value of PC1 has the highest overlap with mode 1 of the open form, which describes the characteristic of the opening-closing motion. In Figure 5.1 (C) and Figure 5.1 (D), the cumulative overlaps increase slightly when more modes are considered. Comparing each item in the matrices of (C) and (D), we find that generally the collective modes from the open form capture relatively more variance of the structural ensemble represented by PC1 and PC3, while only the collective modes from the closed form capture more variance of in the structural ensemble represented by PC2. This is fully consistent with the results show in Figs 5.1 (A) and (B).

Before we shift the indices of normal modes from one conformational state to match the other, we want to inspect how much one mode space can represent an individual mode from another mode space. To understand the importance of normal modes in one space on each individual mode of another space, we compute the squared sum of overlaps (same as squared cumulative overlap, also referred to as CO) for the first 10 modes from the open form with each individual closed mode as well as CO for the first 10 modes from the closed form with each individual open mode.
Figure 5.2: Overlaps of the individual normal modes from the closed (open) form with the normal mode space from the open (closed) form.

The normal mode space for both the closed form and the open form is limited here to a consideration of the first 10 normal modes. The red line shows the sum of 10 squared overlaps of normal modes from the closed form with each individual open mode, which gives the projection of each individual normal mode from the open form onto the normal mode space of the closed form. The black line shows the sum of 10 squared overlaps of normal modes from the open form with each individual closed mode, which gives the projection of each individual normal mode from the closed form onto the normal mode space of the open form.

From Figure 5.2, we find that the projection of each individual mode from one conformational state onto a normal mode space from another conformational state is above 0.5 for nearly all modes for both the open form and the closed form. The exceptions are individual modes 8 and 9 of closed modes where the CO value is 0.34 and 0.40, respectively, as well as mode 9 of the open form, with the CO value of 0.49 for the mode space from the closed form. The high overlap values of individual mode in another mode space encourages us to shift indices in the two mode spaces to provide maximum overlap of each corresponding mode pairs, which can make the diagonal values in the matrix KS take their maximum values. Table 5.1 shows the indices of
matched mode pairs from the two mode space. We keep the indices of the modes from the open form and only shift the indices of the modes for the closed form.

Table 5.1: Match of mode indices from the closed form with individual open mode that yields the highest overlaps. The mode indices of open modes in bold are unchanged.

<table>
<thead>
<tr>
<th>Open Mode</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed Mode</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 5.3. Overlaps of the first 10 normal modes from the GroEL closed and open forms.
(A) Original overlaps of normal modes from the closed form with those from the open form. (B) Overlaps, after shifting of indices indicated in Table 5.3 of the normal modes from the closed form with those from the open form. Overlap values larger than 0.2 are shown with numbers. Notably the matrix in B is more nearly diagonal than the matrix in A.

Figure 5.3 and Table 5.1 show how shifting the indices of the modes affects the diagonal values in overlap matrix as well as the diagonal values of matrix $K_S$. Mode 1 and mode 2 from the two states of GroEL after shifting indices of the closed modes have overlap values of 0.60 and 0.62, respectively. Among the 10 modes, 5 mode pairs
have overlap values over 0.60: mode 1, mode 2, mode 4, mode 7, and mode 9. The 3rd and 5th mode pairs also have relatively high overlaps, which are 0.50 and 0.57. We compute a new basis that combines the two mode spaces derived from two conformational states of a protein. The new basis is also comprised of a set of column vectors that characterize the motion of each coarse-grained residue in the protein structure. We then take the first 10 combined modes from the new basis and compute their overlaps with the first 10 normal modes from the closed form and the open form. Results are shown in Figure 5.4. The overlap values on the diagonal of two matrices in (A) and (B) are exactly the same, indicating that the combined modes from the new basis have the best overlaps with both closed normal modes and open normal modes. Each corresponding overlap value is equal between the two mode spaces.

![Figure 5.4: Overlaps of each combined mode from the new basis with original normal modes.](image)

(A) Overlaps of the combined modes with the closed modes. (B) Overlaps of the combined modes with the open modes.
The set of vectors in the new basis set are further compared with PCs obtained in our previous study (See Chapter 3 for details). Overlap and CO values are computed likewise. Figure 5.5 (A) shows the overlap values between each PC and the individual combined mode. The overlap between PC1 and the 1st combined mode has the highest value 0.55. Mode 5 also has a relatively high overlap value 0.33 with PC1. The cumulative overlap value is computed over the first $N$ ($N = 3, 6, 10, 20$) collective modes in Figure 5.5 (B). PC1 has the highest COs with the first $N$ modes in comparison to PC2 and PC3, nearly double the values for PC2 and PC3. The CO value for PC2 is the lowest among the three PCs, with only 0.42 when all 20 combined modes are included.

**Figure 5.5: Comparison of combined modes from the new basis with the first three PCs.**

(A) Overlaps of each individual combined mode with the first three PCs. The x-axis shows the indices of the first 10 combined modes, and the y-axis shows the indices of PCs. (B) COs of the first $N$ ($N = 3, 6, 10, 20$) ENM modes from the closed form with the first three PCs. The values of the overlap or CO over 0.2 are shown on the grid in white labels.
Table 5.2 show details of overlap values for combined modes and PCs from Figure 5.5 (A). If the overlap value is increased for a specific PC-mode pair and is larger than either of the original overlap values from the two states, this number is labelled in bold, which means that combining modes from the two mode space increases their overlaps with PCs. If the overlap value is larger than the original overlap of one state and is smaller than the original overlap of another, the number is labelled italic, which means that this combined mode compromises the two mode spaces. The rest numbers in the table show no enhancement.

Table 5.2: Overlaps of the combined modes with the first three PCs

<table>
<thead>
<tr>
<th>PC \ Modes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.55</td>
<td>0.02</td>
<td><strong>0.12</strong></td>
<td>0.08</td>
<td>0.33</td>
<td>0.20</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td><strong>0.08</strong></td>
</tr>
<tr>
<td>2</td>
<td>0.03</td>
<td>0.22</td>
<td>0.003</td>
<td>0.19</td>
<td>0.02</td>
<td><strong>0.07</strong></td>
<td>0.08</td>
<td>0.08</td>
<td>0.02</td>
<td><strong>0.17</strong></td>
</tr>
<tr>
<td>3</td>
<td>0.24</td>
<td><strong>0.18</strong></td>
<td>0.04</td>
<td>0.18</td>
<td>0.03</td>
<td>0.02</td>
<td><strong>0.05</strong></td>
<td>0.06</td>
<td>0.05</td>
<td>0.03</td>
</tr>
</tbody>
</table>

In Table 5.2, the improvement in the individual combined modes is not significant. The motion of the open and closed modes has an impact on generating this combined mode which basically yields an overlap value between the two overlap values of original modes from two states with PCs. This is understandable since our combined mode computed has upper and lower boundaries which are provided by the two normal mode spaces from the two conformational states. Striking a balance between the two mode spaces is reasonable. Remarkably, the CO values not only increase compared to one set of the normal modes, but also increase compared to both of normal mode spaces with the first 20 modes, as labelled in bold. This suggests that combining normal modes from the two conformational spaces of a protein structure
may enhance its capability to capture the dynamics of the structure extracted from PCA compared to using a set of normal modes from only one of its conformational states.

5.4.2 A case study - adenylate kinase

PCA is performed on 63 ADK structures. Since the global dynamics of motions in ADK are mainly an opening-closing motions of the NMP and the LID domains, PC1 clearly distinguishes the open form from the closed form, as can be seen from Figure 5.6 (A). It presents the projection of ADK crystal structures onto the 2D PC1-PC2 space. The representative closed form is marked in blue and the open form is in red. The structures that are in the open states are: both chain A and chain B from 4ake and 4pzl, chain A from 4x8h, and chain A from 4x8m. Another cluster lying on the top right of the PC plot contain 3 structures, they are chain A of 2xb4, 3l0p, and 3l0s. The other 54 structures form the largest cluster, which are all in the closed form. From Figure 5.6 (B), we can see that PC1 is dominant and captures 81.5% of the variance. PC2 captures 9.2% of the variance in the ensemble, which captures the twisting motion in the LID domain. PC3 includes 5.3% of the variance while PC4 only accounts for 1.2%. The variance in the structural ensemble is mostly captured by the first two PCs.
Figure 5.6: Essential Dynamics extracted from the ADK structure ensemble.
(A) The projection of experimental structures (shown in black open circles) of ADK onto the PC1-PC2 space. The blue open circle islake that is the chosen representative closed state of ADK, and the red open circle islake that marks the chosen representative open state of ADK. (B) Percentage of variance captured by the first 10 PCs from the ADK structures. The red points on the red dashed line denote the marginal variance captured by each individual PC, and the blue points on the blue dashed line denote the cumulative variance explained by the collective PCs.

Similar to what was done in Section 5.4.1 (a case study on the GroEL), we compare the overlaps between the combined modes and PCs and the overlaps between the normal modes from the open/closed form and PCs. Figure 5.7 shows the overlap and CO values between the modes and PCs. Figure 5.7 (A1), (A2) and (A3) show overlap values between the closed form and PCs, the open form and PCs and the combined modes and PCs. In Figure 5.7 (A1) the overlap values of mode 1 and mode2 with PC1 and PC2 are 0.51 and 0.61, while in Figure 5.7 (A2) the values are 0.83 and
0.78, higher than those in the closed form. Remarkably, by combining the normal modes from the closed form and the open form, we observe that the overlap value with PC1 decreases to 0.75 while the overlap value with PC2 increases to 0.81. In Figure 5.7 (B1), (B2), and (B3), the CO values between modes and PCs are shown. Most of the overlap values from the combined modes fall between those of the closed and the open form, however they are very close to the open form, the conformational state that yields higher overlaps with PCs. This result suggests that the combined modes under most scenarios strike a balance between the closed and the open forms of a protein structure. However to some extent, the capability to capture structural dynamics extracted from PCA may also be improved by using a combination of normal modes from the two conformational states instead of using normal modes from one single conformation from the structure ensemble. This suggests, that if more than two characteristic structures are known, that combining modes from all of the structures could enhance the overlaps.
Figure 5.7: Overlap and CO values of the closed modes, the open modes, and the combined modes with the first 3 PCs.
In Figure 5.7, figure (A1), (A2), and (A3) are the overlap values of individual modes with PCs. (A1) shows the overlaps with the closed modes after index shifting, (A2) shows the overlaps with the closed modes in original index, and (A3) shows the overlap with the combined modes. Figure (B1), (B2), and (B3) show CO values with the closed modes, the open modes and the combined modes, corresponding to A1, A2 and A3, respectively.

In the previous studies in Chapter 3 and Chapter 4, the free energy landscape is created by using two of the most important PCs, referred to as the PC landscape. The same idea is adopted for the mode landscape. The free energy landscape on the 2D PC plane constructed in Chapter 4 is based on a surface function which interpolates the grid point on the 2D space from its projection, i.e., the projection of Cartesian coordinates onto the surface in a 3N-dimensional space (N is the total number of residues in a coarse-grained model) to a 2D space. Since a hyperplane can always be projected onto a lower-dimension plane, we can use the first two combined normal modes instead of PCs to construct the landscape. The advantage of using the combined normal modes is that these modes have the characteristics of the dynamics for both the closed and the open states, given that a protein structure with multiple conformations mainly clustered in two states. We use the combined modes for the construction of the free energy landscape, with a similar approach to that introduced in Chapter 4. The first two column vectors (other than the first 6 rigid body motions) in the matrix of the new basis represent the most favorable characteristic motions of the protein structure. Protein structures are projected onto the 2D mode space. The free energies of the
interpolated structures are computed based on the knowledge-based potentials and entropies (Refer to Chapter 3.3.4, Chapter 3.3.5 and Chapter 4.3.3 for details). We compute both GNM entropies and knowledge-based entropies in this study. The weights for the three knowledge-based potentials (the four-body potentials, the non-sequential four-body potentials and the short-range potentials) are 1, 0.28, and 0.22, while the weight for entropy is -1 following our lab’s previous work (163). The free energies are computed as in Eq. 3.3 from Chapter 3.

Figure 5.8 (A) and (B) show the free energy landscapes computed with GNM entropy and AACC entropy on a mode space. The free energy landscape with AACC entropy in Figure 5.8 (B) has an even lower energy valley for the closed form, with most of the patterns being similar in the two landscapes. Based on the projection of experimental structures onto the 2D mode space, we find that there are mainly three clusters on the mode space, and all of them lie in the low energy regions. The 1st combined mode is dominant which distinguishes the holo closed state from the apo open state, and it captures the opening-closing motion in the LID and NMP-binding domains of ADK. The 2nd combined mode further distinguishes a cluster of three structures which are also captured by PC2 (shown in Figure 5.6). These three structures (referred to as partial-open state here) lie in between the closed and the open state, where the most variant part is in their LID domains (See Figure 5.9).
Figure 5.8: **Free energy landscape in the mode1-mode2 space.**
(A) The free energy landscape computed with GNM entropy. (B) The free energy landscape computed with the knowledge-based AACC entropy. The experimental structures are projected onto the mode space shown in the white open circles. The characteristic closed and the open states are marked with magenta open circles.

Figure 5.9: **The conformational states for ADK.**
The LID domain is shown in red for the apo open state (PDBid: 4ake), blue for the holo closed state (PDBid: 1ake), and green for the half-open intermediate state (PDBid: 2xb4). The NMP domain is shown in pink for the open state, light blue for the closed state and light green for the half-open state. (A) Side view of the three structures. (B) Top view of the three structures.
After superimposing the closed state and the half-open state onto the open state in the CORE domain, we observe that the most significant changes of the half-open state and the open-state are in their LID domains. Therefore, the 2\textsuperscript{nd} mode from the combined normal mode space captures the twisting motions in the LID domain, which is verified by the animations of PC2 and mode 2.

The details of potentials and entropies are shown in Figure 5.10 and Figure 5.11. The four body potential contour in Figure 5.10 (A) shows low energy regions for the structures in the open form, and Figure 5.10 (B) shows lower energy regions for small mode 2 scores. The short-range potentials are very similar to the free energy landscape, which has the largest contribution to the free energies. The pathways from the half-open intermediate form to the open form are lowered further in free energies by the four-body potentials and non-sequential four-body potentials, which suggest a possible conformational change between these two conformations.
Two entropy contours used in the construction of the free energy landscapes are shown in Figure 5.11. The AACC entropy performs better in characterizing the entropies of conformations than the GNM entropy because the AACC values have higher entropies for native structures, especially in the closed form. Since we use the weight -1 for these entropies, the addition of AACC entropy terms can lower the free energies of the native structures. This is also observed in the comparison between the
GNM entropy and the AACC entropy in the GroEL free energy landscape (shown in Chapter 3) and the SERCA free energy landscape (shown in Chapter 4).

**Figure 5.1**: The GNM entropy and the knowledge-based (AACC) entropy used for computing the energy landscape. The color code specifies low entropies (in blue) and high entropies (in red). Both the GNM and AACC entropies use a cutoff of 8 Å to define the coarse-grained models.

MMC force application simulation are performed on the ADK open form. It is observed that forces can drive the open form towards the target closed form starting from 7.2 Å to reach a final RMSD around 3 Å (see Figure 5.12). The Pearson correlation of the RMS displacements between the initial-target and the initial-final pairs is 0.95. The mean of the absolute residue-by-residue displacement differences is 1.7 Å, and the mean square error of the displacements is 5.0 Å².
Figure 5.1: Free energy landscape of ADK on the combined mode space for a conformational transition from the open form to the closed form by a MMC force application simulation. The experimental structures are shown in white open circles. The two representative states are shown in magenta circles and labelled in white. Free energy values are color-coded from blue (lowest) to red (highest). The conformations in the transition are colored in magenta.

5.4.3 Results from 50 cases

We compute the combined modes for 50 cases and compare the overlaps between the modes and PCs. The two states, i.e., the base and the target, are selected by computing pairwise RMSDs. The base form has the smallest average RMSD with all the other structures, while the target form has the largest. Figure 5.13 shows the overlap values between PCs and modes for some representative cases. In Figure 5.13 (A), the overlap value between the 2nd combined mode and PC1 increases to 0.7, compared with 0.65 and 0.64 for the normal modes from the base form and the target form. Likewise, we observe an increase in the overlap values for the 1st combined mode with PC3 in (B), for the 3rd combined mode with PC3 in (C), and for the 1st combined
mode with PC3 in (D). Despite the improvements in the overlaps between PCs and the combined normal modes for some cases, under most circumstances, the overlap values strike a balance between those from the base and the target modes, implying a combination of motions from both of the two conformational states.

Figure 5.13: Overlaps of PCs with ENM normal modes and combined modes. (A) DNA polymerase beta. (B) Thaumatin I. (C) B. anthracis Dihydrofolate reductase (DHFR). (D) Human Lysozyme C.
5.5 Conclusions

The novel method for combining normal modes from the ENMs for different conformational states provides an approach to capture protein dynamics by incorporating the dynamics of multiple conformations. By comparing overlaps or COs between PCs and the combined modes as well as between PCs and the ENM normal modes, we find that the combined modes in some cases yield higher overlap values with PCs, implying that these modes may capture more of the system’s variance and better explain the observed experimental structures with their more comprehensive protein dynamics than the normal modes from ENM from only a single structure of a protein. These combined modes can further serve as a new coordinate system and can be used in the construction of a free energy landscape. The experimental structures of ADK are separated into several clusters by the combined modes and are mostly situated in the lowest energy basins on the mode landscape, which is usually observed in the native structures. MMC force application simulation from the open form to the closed form reaches a RMSD of 3 Å, pointing out a possible low-energy pathway for the transition.
CHAPTER 6  SUMMARY AND CONCLUSIONS

6.1 General Discussion

This dissertation explores the effect of mechanistic forces from ATP binding and hydrolysis events in proteins that undergo large conformational changes. The possible conformational transition pathways obtained by using the directed force application as well as by using the Metropolis Monte Carlo force application simulation method are further studied and observed upon the free energy landscapes. Principal component analysis is performed to extract the most important dynamic motions of a structure ensemble for landscape construction, and normal modes from ENMs are used to compare with the MMC transitions.

In Chapter 2, we applied directed forces on the phosphate group of ATP in the closed form of the GroEL subunit based on ANMs. The directed force application approach reveals strongly preferred directions for the forces, which can successfully drive a protein structure towards its known end point, i.e., the open form of GroEL in our study. The results provide strong evidence that highly directed forces, which may be ballistic in nature, can originate from a specific chemical event such as ATP hydrolysis, and can drive large conformational changes. The RMSD toward the target open form decreases by approximately half in the GroEL subunit. Despite the success, this greedy approach does not consider forces that would yield smaller gains for local refinements, which might also be important in the conformational transition. Besides, the energies for each intermediate in the transition are not taken into consideration. Therefore in Chapter 3, we have incorporated free energies and randomness into this
approach. Instead of applying forces directly on the leaving phosphate group, we apply forces on one of the residues nearby the ATP-binding sites to introduce randomness in each iteration. The Metropolis Monte Carlo method adopts the change in the Boltzmann factor of the free energy as a Metropolis criterion to decide whether to accept a new conformation or reject during each simulation step. A conformation generated at each MMC step with reduced free energy is always accepted, while a conformation with increased free energy is only be accepted with a certain probability, defined by the Metropolis criterion. The MMC generates relatively low energy pathways and is capable of driving the initial form near to the end form. In our MMC study on the GroEL subunit, we are able to reach a final state that is very close to its target state for both directions of the transitions, from the open to the closed form and from the closed to the open form. The final states almost fit into their corresponding target states with an average RMSD of 3.6 Å, which is an excellent result, given the resolution of the coarse-grained structures used in the simulations. The study of the GroEL ring structure sheds light on the intermolecular interactions and interplay between subunits. By pushing only one subunit in the GroEL to its open state, the two neighboring subunits start to open as well. Our results agree with the current findings that GroEL displays a positive cooperativity among its subunits in the ring structure, and may guide future studies on allostery to fully understand the mechanisms of molecular machines.

Similar results from the studies on other cases using the MMC force application method are shown in Chapter 4. The application of forces on the open form of SERCA drives the protein to undergo large conformational changes in the nucleotide binding
domain, which corresponds well to the essential dynamics captured by the PCs from the principal component analysis. Our findings agree with the current knowledge that the transition of SERCA from the open form to the closed form proceeds through the phosphorylated states and involves large conformational rearrangements. The transitions between different states in the F1-ATPase, i.e., the ligand-free state to the ATP-binding state and the ATP-bound state to the ADP-bound state, by using the MMC force applications also show remarkable success. The results provide strong evidence that ATP-binding and hydrolysis may play an important role in the transitions between the different conformational states in the F1-ATPase subunits. To better understand the conformational transitions by using MMC, in Chapter 4, we develop new free energy landscapes to delineate these transitions. The free energies are required for each intermediate conformation in response to force during a MMC simulation as well as for all the sampling conformations that contribute to the construction of the free energy landscape. We utilize our group’s well developed approaches for calculating the coarse-grained free energies in a combination of knowledge-based potentials and entropies. A simple sampling method which is computationally efficient is developed to interpolate conformations based on all the Cartesian coordinates of a given experimental structure ensemble. The free energies of all these sampled conformations become the states on the free energy landscape. Our results for GroEL, SERCA, F1-ATPase, ADK, and four representative cases from a 50 cases dataset show that the native structures mostly fall in low energy regions on the free energy landscape, which validates the notion that generally the crystal structures tend to adopt low free energy
conformations. The free energy landscape itself contains rich information for the known experimental structures as well as for new unknown possible structures, and it may suggest possible transition pathways. The projection of MMC force application intermediates on the free energy landscape reveals that these transition trajectories follow low energy pathways that are energetically favorable. The MMC delineates the transition pathways by driving the structure to pass over energy barriers on the free energy landscape and to find its targeted end state. To further analyze these transition trajectories, we compared the transition steps with the ENM normal modes in Chapter 5. The structural differences of all pairs of two consecutive steps in the transition trajectory from the MMC simulation are compared with the ENM normal modes by computing their overlaps. These normal modes combine two or more normal mode spaces derived from different structures of a protein to characterize the dynamics of multiple conformational states of the protein. Our findings indicate that the MMC transition pathways of GroEL corresponds well to its most important normal modes from ENM, while some high frequency modes may also contribute in the conformational transitions. The combined normal mode space, for which the first two modes are used as the landscape coordinates, can be further utilized in the construction of the free energy landscape. As was seen from the ADK free energy landscape in the mode space, experimental structures fall into the low free energy regions, implying that the combined modes can better capture the dynamics of a protein structure. In the 50 cases studied, it was found that these combined normal modes can also yield improved overlap values with the PCs extracted from the structure ensemble.
6.2 Recommendations for Future Research

The MMC force application approach in this dissertation has been applied to GroEL, SERCA, ADK, and F1-ATPase, and these can provide guidance for a wide variety of types of simulations utilizing the application of forces for other systems. This innovative approach can also provide an important computational framework for treating the forces that drive conformational transitions. In the Protein Data Bank, there exist more than 30 structure pairs with ATP or ADP binding in the catalytic sites and these structures that undergo conformational changes in the ATP binding and hydrolysis event, e.g., pyridoxal kinase, pyrophosphorylase, and shikimate kinase. The GTP binding structures may also be interesting subjects for study.

For the construction of the free energy landscape, the conformational sampling method can be further explored with other possible weighted functions and varying functions. Future investigations can be conducted on the DIW-LV function where the linear-varying function can be substituted by a variety of functions to generate smoother more continuous landscapes to offset the strong effect of the energetics of the individual experimental structures on the landscape. We can also include energetics representing some important properties of proteins such as packing density and geometry in the structural interpolation approach. New optimization of the free energy functions might generate more reliable free energy landscapes.
APPENDICES

APPENDIX A WEIGHTED CONTRIBUTION METHOD

A.1 The intuitive aspects

Let \( S = \{x_1, x_2, x_3, \ldots, x_m\} \) be a set of sample points in \( \mathbb{R}^n \). A general problem is: how to construct a surface \( \Omega \) thorough these points. We aim to obtain a hyperplane that goes through all experimental protein structures in a multi-dimensional space and then to interpolate to obtain other coordinates on the surface. The free energy landscape visualized on a 2D space can thus be constructed with these interpolated structures on the surface. There are actually infinite ways to achieve this. In this study we develop the weighted contribution method to solve this problem.

The idea of WCM is very intuitive. Imagine there are \( m \) electric charges on a plane, and we want to measure the electric field across the plane. The net field at a sample point is contributed by all these charges. How much a charge contributes depends on the distance and the direction to the charge. To get the total strength of the field, we can simply add up all the contributions. If a point on the field is very close to some charge, then the contributions from the other charges are not significant and are overwhelmed by that charge. If the total charge is zero and the charge distribution is somehow symmetric, then the field at a point close to the center of the distribution is near zero.

The weighted contribution method is derived from this concept described above. The coordinate of a point on the surface being constructed is contributed by the
coordinates of all sample points where each point on the surface represents a structural conformation, and how much they contributions will be determined by the weights. One of the most important requirements is that the surface must go through all the sample points, i.e., the experimental structural ensemble. In this way, the energies on the $2D$ free energy landscape for the projection location of the experimental structures are exactly the free energy values of these structures in the ensemble. To ensure the surface going through a sample point, e.g. $x_i$, we must define that the weight at the location of $x_i$ overwhelms the contributions from other points.

### A.2 The general surface equation for WCM

Let $\mathbb{R}^2 \rightarrow \mathbb{R}^n$ be the interpolation surface. Denote the pre-images of the given sample points in $S$ as $\Gamma^{-1}(x_i) = (u_i, v_i)$, and $\forall i \neq j, u_i \neq u_j, v_i \neq v_j$. $(u_i, v_i)$ is the coordinate in $\mathbb{R}^2$ for a sample point $x_i$ and $(u, v)$ corresponds to the coordinate in $\mathbb{R}^2$.

The general equation of $\Gamma$ is defined as

$$
\Gamma(u, v) = \sum_{i=1}^{m} w_i(u - u_1, u - u_2, \ldots, u - u_m, v - v_1, v - v_2, \ldots, v - v_m)(x_i + g_i(u - u_i, v - v_i))
$$

(A.1)

In Eq. A.1 $m$ is the number of given sample points. The two set of mappings $w_i : \mathbb{R}^2 \rightarrow [0,1]$ and $g_i : \mathbb{R}^2 \rightarrow \mathbb{R}^n$ have additional properties

$$
w_i(s_1, s_2, \ldots, s_m, t_1, t_2, \ldots, t_m) =
\begin{cases}
1, & s_i = t_i = 0, s_j \neq 0, t_j \neq 0, i \neq j \\
0, & s_j = t_j = 0, s_i \neq 0, t_i \neq 0, i \neq j
\end{cases}
$$

(A.2a)
\[ \sum_{i=1}^{m} w_i = 1 \quad \text{(A.2b)} \]
\[ g_i(0,0) = 0 \quad \text{(A.2c)} \]

where \( s_1 \) is short for \( u - u_1 \), \( s_m \) is short for \( u - u_m \), \( t_1 \) is short for \( v - v_1 \), and where \( t_m \) is short for \( v - v_m \). These properties make \( \Gamma(u, v) \) satisfy \( \Gamma(u_i, v_i) = x_i \). There are two terms, \( w_i \) and \( g_i \) in the function \( \Gamma(u, v) \) that requires definition. We call \( w_i \) the weight function and \( g_i \) the varying function.

A.3 Distance-inverse-weight functions and linear varying functions (DIW-LV)

There are many possible ways to choose the weight functions \( w_i \) and the varying functions \( g_i \). We hereby discuss distance-inverse-weight and linear-varying method.

We use the inverse of distance as the weight. According to the Coulomb’s law of electric charge, the magnitude of electric field increases when moving nearer to an electric charge and diminishes when moving away. This inspires us to use the inverse of distance from a random point on the surface to the sample point (the source) to define the weight

\[ w_i(s_1, s_2, ..., s_m, t_1, t_2, ..., t_m) = \frac{d_i^{-1}}{\sum_{j=1}^{m} d_j^{-1}} \quad \text{(A.3)} \]

Where \( m \) is the number of given sample points, \( d(s, t) = \sqrt{s^2 + t^2} \) is the common definition of the distance in \( \mathbb{R}^2 \) and \( d_i \) is short for \( d(s_i, t_i) \). Eq. A.3 have properties of Eq. A.2.
Generally, no other form is simpler than a linear function. When \( g_i \) is a linear function, we formulate it as a linear-varying functions

\[
g_i(s, t) = sa_i + tb_i
\]  

(A.4)

where \( a_i, b_i \in \mathbb{R}^n \).

A.4 General properties of DIW-LV functions

A.4.1. Smoothness

By the chain rule, we have

\[
\frac{\partial w_i}{\partial s_j} = \frac{\partial w_i}{\partial d_j} \cdot \frac{d_j}{s_j}, \quad \frac{\partial w_i}{\partial t_j} = \frac{\partial w_i}{\partial d_j} \cdot \frac{t_j}{d_j}
\]

Taking partial derivatives of the individual terms in Eq.A.3 with respect to \( d_j \) yields

Equation A.5

\[
\frac{\partial w_i}{\partial d_j} = \begin{cases} 
\frac{w_iw_j}{d_j}, & i \neq j \\
-\frac{w_i(1-w_i)}{d_i}, & i = j 
\end{cases}
\]  

(A.5)

We can see that \(-\infty < \frac{\partial w_i}{\partial d_i} |_{d_i=0} < 0\). Since \( \frac{t_i}{d_i} \) and \( \frac{s_i}{d_i} \) are discontinuous at \( t_i = s_i = 0 \), the weight functions are not smooth at the sample points. Therefore, DIW-LV surface is smooth almost everywhere except at the sample points.

From the chain rule, we also have

\[
\frac{\partial^2 w_i}{\partial s_j \partial s_k} = \frac{\partial^2 w_i}{\partial d_j \partial d_k} \cdot \frac{s_j s_k}{d_j d_k} + \frac{\partial w_i}{\partial d_j} \cdot \frac{1}{d_j} \left( 1 - \frac{s_j^2}{d_j^2} \right) \delta_{jk}
\]  

(A.6a)
\[
\frac{\partial^2 w_i}{\partial s_j \partial t_k} = \frac{\partial^2 w_i}{\partial d_j \partial d_k} \cdot \frac{s_j t_k}{d_j d_k} - \frac{\partial w_i}{\partial d_j} \cdot \frac{s_i t_k}{d_j^3} \delta_{jk} \quad (A.6b)
\]

\[
\frac{\partial^2 w_i}{\partial t_j \partial s_k} = \frac{\partial^2 w_i}{\partial d_j \partial d_k} \cdot \frac{s_j s_k}{d_j d_k} - \frac{\partial w_i}{\partial d_j} \cdot \frac{t_j s_k}{d_j^3} \delta_{jk} \quad (A.6c)
\]

\[
\frac{\partial^2 w_i}{\partial t_j \partial t_k} = \frac{\partial^2 w_i}{\partial d_j \partial d_k} \cdot \frac{t_j t_k}{d_j d_k} + \frac{\partial w_i}{\partial d_j} \cdot \frac{1}{d_j} \left(1 - \frac{t_j^2}{d_j^2}\right) \delta_{jk} \quad (A.6d)
\]

The second-degree partial derivatives of Eq. A.3 with respect to \(d_j\) are \((i \neq j \neq k)\)

\[
\frac{\partial^2 w_i}{\partial d_i^2} = \frac{2w_i(1-w_i)^2}{d_i^2} \quad (A.7a)
\]

\[
\frac{\partial^2 w_i}{\partial d_j^2} = -\frac{2w_i w_j(1-w_j)}{d_j^2} \quad (A.7b)
\]

\[
\frac{\partial^2 w_i}{\partial d_i \partial d_j} = \frac{w_i w_j(2w_i-1)}{d_i d_j} \quad (A.7c)
\]

\[
\frac{\partial^2 w_i}{\partial d_j \partial d_k} = \frac{2w_i w_j w_k}{d_j d_k} \quad (A.7d)
\]

Even though the second-degree partial derivatives in terms of \(s_i, t_i\) are not continuous at \(s_j = t_j = 0\), we can write Taylor’s expansions of \(w_i\) at \(d_i = 0\) and \(d_j = D_{ij} = d(u_i-u_j, v_i-v_j)\) as

\[
w_i = 1 - \left(\sum_{l \neq i} D_{il}^{-1}\right)d_i + 2\left(\sum_{l \neq i} D_{il}^{-1}\right)^2 d_i^2 + \frac{d_i(d_i-D_{ij})}{D_{ij}^2} + \cdots \quad (A.8)
\]
A.4.2. Asymptotic plane

Given Eq. A.4, the DIW-LV functions lead to simpler form of surface equation as shown in Equation A.9

\[ \Gamma(u, v) = \sum_{i=1}^{m} w_i [x_i + (u - u_i)a_i + (v - v_i)b_i] \]

\[ = \sum_{i=1}^{m} w_i (x_i - u_i a_i - v_i b_i) + u \sum_{i=1}^{m} w_i a_i + v \sum_{i=1}^{m} w_i b_i \]  \hspace{1cm} (A.9)

The global properties of this surface would be determined by Equation A.10

\[ A = \sum_{i=1}^{m} w_i a_i, \quad B = \sum_{i=1}^{m} w_i b_i, \quad C = \sum_{i=1}^{m} w_i (x_i - u_i a_i - v_i b_i) \]  \hspace{1cm} (A.10)

In the region where \( w_1 \approx w_2 \approx \cdots \approx w_m \), \( \Gamma \) can be approximated by the plane in Equation A.11

\[ \Gamma_0(u, v) = u \left( \frac{1}{m} \sum_{i=1}^{m} a_i \right) + v \left( \frac{1}{m} \sum_{i=1}^{m} b_i \right) + \frac{1}{m} \sum_{i=1}^{m} (x_i - u_i a_i - v_i b_i) \]

\[ = uA_0 + vB_0 + C_0 \]  \hspace{1cm} (A.11)

\( \Gamma_0 \) is called the asymptotic plane of \( \Gamma \). The vectors \( A_0, B_0 \) specify the direction of \( \Gamma_0 \).

Let \( \Omega = \Gamma - \Gamma_0 \). \( \Omega \) is the variation of \( \Gamma \) with respect to \( \Gamma_0 \), which can be used to show how much \( \Gamma \) suddenly changes from its asymptotic plane to the sample points.

A.4.3. Local shape nearby the sample points

Figure A.1 is a comparison of interpolation methods between DIW-LV method and B-Spline. A basis spline is a spline function that has minimal support with respect to a given degree, smoothness, and domain partition. It is often used in curve-fitting, a
process of constructing a curve or mathematical function that has the best fit to a series of data points.

**Figure A.1: A comparison of interpolation methods between (A) DIW-LV and (B) B-Spline.**

The sample points are taken from the surface \( z(x, y) = xe^{-x^2-y^2} \), over a 20 \( \times \) 20 grid in the region \([-1,1] \times [-1,1]\).

Suppose the sample points are distributed over the vertices of a square grid. Each element square's side length is \( l \). We hereby inspect Eq.A.8 at some vertices \( c_0 = (u_0, v_0) \). There are \( 3^2 - 1 \) points one square to \( c_0 \). Their distances to \( c_0 \) are within \([l, \sqrt{2}l]\). There are \( 5^2 - 3^2 \) points two squares to \( c_0 \). Their distances to \( c_0 \) are within \([2l, 2\sqrt{2}l]\). Follow this pattern, the coefficient of the linear term in Eq.A.8 is bounded by Equation A.12

\[
\frac{4\sqrt{2}r}{l} = \sum_{k=1}^{r} \frac{(2k+1)^2-(2k-1)^2}{\sqrt{2}kl} \leq -a \leq \sum_{k=1}^{r} \frac{(2k+1)^2-(2k-1)^2}{kl} = \frac{8r}{l}
\]  

(A.12)
The total number of points is characterized by \( m \sim (2r + 1)^2 \). Thus, \( a_1 \sim -\sqrt{\frac{m}{l}} \), and the coefficient of the quadratic term is \( a_2 = 2a_1^2 \sim \frac{m}{l^2} \). The region where \( c_0 \) gains control is characterized by \( a_1d + a_2d^2 < 0 \). Then the greatest distance controlled by \( c_0 \) is

\[
d_{\text{max}} = -\frac{1}{2a_1} = \frac{l}{8\sqrt{2r}} = \frac{l}{4\sqrt{2}(\sqrt{m-1})}
\]

(A.13)

Apparently, the region that \( c_0 \) can control is far less than one element square. The percentage of region it controls, defined by \( R = \frac{d_{\text{max}}^2}{l^2} \), is

\[
R = \frac{1}{32(\sqrt{m-1})^2}
\]

(A.14)

The percentage of region that \( c \) controls within one square decreases at a rate of the inverse of the number of samples. As a result, when there are a lot of sample points, \( \Gamma \) will suddenly go from the location of a sample point to the mean of the samples. In a global view, \( \Gamma \) is flatly expanded in the space with sharp bumps and pits at the sample points. When the sample points are many and dense, the surface looks like a thorny mat. Therefore, the major limitation in this approach is that DIW-LV would not be successful in providing a gently varying interpolation if there are too many experimental structures.
APPENDIX B CONSTRUCTION OF NEW BASIS

B.1 Introduction

Denote $C = (c_1, c_2, c_3, ..., c_n)$ and $O = (o_1, o_2, o_3, ..., o_n)$ as two orthonormal (Orthogonal and normalized) basis sets in $\mathbb{R}^n$, where $C$ and $O$ are matrices composed from all of the mode vectors for the closed form and the open forms, respectively. We can obtain an orthogonal transformation matrix $K = \{k_{ij}\}$, such that $C$ and $O$ have the following properties

$$O = CK \quad C = OK^T$$  \hspace{1cm} (B.1)

Where $K^T$ is the transpose of $K$. The entries of $K$ are direction cosines between the bases sets, or the inner product of $c_i$ and $o_j$, which is

$$(c_i, o_j) = k_{ij}$$  \hspace{1cm} (B.2)

where $(c_i, o_j)$ stands for the inner product of $c_i$ and $o_j$. Let $X = (x_1, x_2, x_3, ..., x_n)$ be a new orthonormal basis. By analogy to Eqs. B.1 and B.2, the inner products between $X$ and $C, O$ are given by

$$X = CL\quad X = OM$$  \hspace{1cm} (B.3)

$$(c_i, x_j) = l_{ij} \quad (o_i, x_j) = m_{ij}$$  \hspace{1cm} (B.4)

where $L = \{l_{ij}\}, M = \{m_{ij}\}$. The relationship between $K, L$ and $M$ is given by

$$K = LM^T$$  \hspace{1cm} (B.5)
We expect $X$ to be the best representation for both matrix $O$ and matrix $C$ together. To be specific, we expect that (a) The diagonal entries of $L$, $M$ are of the greatest absolute value within the column or row to which they belong; (b) The absolute values of their diagonal entries are as close to 1 as possible.

**B.2 Trivial solution**

There exists a trivial solution, i.e., either $X = C$ or $X = O$. In the first case, $L$ is the identity matrix and in the second case $M$ is the identity matrix. However, whether it is a good substitute for $O$ depends on $K$. If $K$ is really close to the identity matrix, i.e., all its diagonal entries are close to 1, then the trivial solution could be a good solution.

**B.3 A balanced solution**

A balanced solution is that $X$ resembles “equally” $C$ and $O$. We impose the following condition on $L$ and $M$

$$\left| (c_i, x_j) \right| = \left| (o_j, x_i) \right|$$  \hspace{1cm} (B.6)

When we take orthogonality into account, the solution is limited in Equation B.7, by using the matrix notation

$$M = SL^T, KS = L^2, X = CL$$  \hspace{1cm} (B.7)

where $S$ is a matrix with diagonal entries being ±1. Since we want $X$ to be a real number matrix, the diagonal entries shall not be complex numbers. It is worth
mentioning that we do not require \(|(c_i, x_j)| = |(o_j, x_i)|\), because this would lead to \(K = S\), which is generally false.

**B.4 A Preliminary Analysis: How good can the new basis be?**

The matrix \(K\) could be very far away from the identity matrix \(I\). Therefore, matrices \(C\) and \(O\) may not resemble each other at all. Our solution should be in between \(C\) and \(O\), i.e., \(X\) is somehow resembles \(C\), but also somehow resembles \(O\), which is a dilemma.

A rigorous analysis on what we can achieve is provided in the following. From Eq.B.5, we can obtain the following relation between \(k_{ii}, l_{ii}\) and \(m_{ii}\)

\[
k_{ii} = \sum_{k=1}^{n} l_{ik}m_{ik} = l_{ii}m_{ii} + \sum_{k \neq i} l_{ik}m_{ik}
\]

(B.8)

Utilizing Cauchy’s Inequality yields Equation B.9

\[
(k_{ii} - l_{ii}m_{ii})^2 \leq (1 - l_{ii}^2)(1 - m_{ii}^2)
\]

(B.9)

The equal sign holds if and only if \(\exists a, b \in \mathbb{R}\), such that \(al_{ik} = bm_{ik}, \forall k \neq i\). This is equivalent to

\[
l_{ii}^2 + m_{ii}^2 - 2k_{ii}l_{ii}m_{ii} \leq 1 - k_{ii}^2
\]

(B.10)

By diagonalizing the quadratic form on the left-hand side, we obtain

\[
\frac{(l_{ii}+m_{ii})^2}{2(1+k_{ii})} + \frac{(l_{ii}-m_{ii})^2}{2(1-k_{ii})} \leq 1
\]

(B.11)

And from the inequality in Eq.B.11, we can derive
\[ |l_{ii}| + |m_{ii}| \leq \sqrt{2(1 + |k_{ii}|)} \]  
\( \text{(B.12)} \)

The equal sign holds if and only if \( l_{ii} = \text{sgn}(k_{ii})m_{ii}, l_{ik} = -\text{sgn}(k_{ii})m_{ik}, \forall k \neq i. \)

**Figure B.1. Boundaries of diagonal entries of \( L \) and \( M \) for different \( k \) values.**

The x-axis indicates the diagonal values for matrix \( L \), and the y-axis indicates the diagonal values for matrix \( M \). Boundaries of diagonal entries of \( L \) and \( M \) corresponding to different \( k \) values are shown in different colors, with values of \( k \) marked aside. The red line denotes \( |l_{ii}| = |m_{ii}| \), and the intersection of the red line at the boundaries gives the values of \( |l_{ii}|/|m_{ii}| \) for a given \( |k_{ii}| \) value.

**Table B.1:** The optimal diagonal entries for \( |l_{ii}| \) and \( |m_{ii}| \) when \( |l_{ii}| = |m_{ii}| \) given the value of \( |k_{ii}| \).

| \( |k_{ii}| \) | 0.20 | 0.40 | 0.60 | 0.80 | 0.90 |
|----------------|------|------|------|------|------|
| \( |l_{ii}|/(|m_{ii}|) \) | 0.77  | 0.84 | 0.89 | 0.95 | 0.97 |

The inequality in Eq.B.12 addresses the dilemma. Generally, we cannot meet both ends simultaneously. Approaching to one basis set means deviating from the other.

Figure B.1 shows the boundary values of the diagonal entries of \( L \) and \( M \) for different \( k \) values. In the figure, the x-axis indicates the diagonal values \( l_{ii} \) for matrix \( L \), and the y-axis indicates the diagonal values \( m_{ii} \) for matrix \( M \). Boundaries of the diagonal
entries of $L$ and $M$ in response to $k$ values are color-coded, with corresponding $k$
values marked on the figure. We can assume that $L$ and $M$ have a pair of optimal
diagonal entries $l_{ii}$ and $m_{ii}$ at

$$|l_{ii}| = |m_{ii}| = \sqrt{\frac{(1+|k_{ii}|)}{2}}$$

(B.13)

Table B.1 above gives an idea about the best solution we may obtain by using
this approach based on the different values in $|k_{ii}|$ when $|l_{ii}| = |m_{ii}|$. It can be
concluded that the larger the value of $|k_{ii}|$, the larger value we can get for the diagonal
entries of $|l_{ii}|$ and $|m_{ii}|$. In addition, there don't exist such $i \neq j$ in general that Eq.B.12
takes equal sign for both $i$ and $j$. To put it simply, if we make one pair of diagonal
entries optimal, the other pairs cannot be optimal. Suppose that there exists such $i \neq j$,
then by using the orthogonality of $L$ and $M$, we can derive that

$$\sum_{k=1}^{n} m_{ik} m_{jk} = -\text{sgn}(k_{ii}) \text{sgn}(k_{jj})(l_{ii}l_{ji} + l_{ij}l_{jj}) + \text{sgn}(k_{ii}) \text{sgn}(k_{jj}) \sum_{k \neq i,j} l_{ik}l_{jk}$$

$$= -2\text{sgn}(k_{ii}) \text{sgn}(k_{jj})(l_{ii}l_{ji} + l_{ij}l_{jj}) = 0$$

$$k_{ij} = \sum_{k=1}^{n} l_{ik} m_{jk} = \text{sgn}(k_{jj}) l_{ij}l_{jj} - \text{sgn}(k_{jj}) \sum_{k \neq j} l_{ik} l_{jk} = 2\text{sgn}(k_{jj}) l_{ij}l_{jj}$$

which indicates that $\text{sgn}(k_{ii})k_{ji} + \text{sgn}(k_{jj})k_{ij} = 0$. This is generally not true for $K$.

In terms of the balanced solution, there are several questions: (1) how to
construct the balanced solution, since the matrix $S$ is still not determined yet; (2) how
to evaluate the balanced solution?
First of all, we need to determine the form of the matrix $S$. From linear algebra, $S$ needs to be chosen so that the solution is real. On one hand, the determinant of $KS$ must not be $-1$, since otherwise the determinant of $L$ would be $\pm i$, which means $L$ is imaginary. On the other hand, if the determinant of $KS$ equals 1, taking advantage of the properties of exponential mapping from Lie algebra, $so(n)$ to Lie group $SO(n)$, one can find that $L$ must be real. Therefore we need to make $KS$ an orthogonal matrix with the determinant of $KS$ equal to 1 and the determinant of $S$ equals the determinant of $K$.

To evaluate the performance of this approach, let $\{\lambda_i = e^{i\theta_i}, \theta_i \in [-\pi, \pi], i = 1, 2, ..., n \}$ be the eigenvalues of $KS$. The range restriction on $\theta_i$ ensures that the square root of the eigenvalues are closer to 1. From the property of the orthogonal matrix, a real coefficient polynomial equation has roots as complex conjugate pairs, i.e., a complex conjugate of one root must also be a root. Since the characteristic equation of an orthogonal matrix is a real coefficient polynomial, we know that $\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} Re\lambda_i$. Then we have

$$\sum_{i=1}^{n} Re\lambda_i = Tr(KS) \quad (B.14)$$

From Eq.B.7, the eigenvalues of $L$ are $\{\mu_i = e^{\frac{i\theta_i}{2}}, i = 1, 2, ..., n\}$. By the same logic, we can get Equation B.15 and Equation B.16:

$$\sum_{i=1}^{n} Re\mu_i = Tr(L) \quad (B.15)$$

$$2(Re\mu_i)^2 - 1 = Re\lambda_i, \forall i \quad (B.16)$$
It follows that

\[ \sqrt{\frac{\text{Tr}(KS) + n}{2}} \leq \text{Tr}(L) \leq \sqrt{\frac{n(\text{Tr}(KS) + n)}{2}} \]  

(B.17)

We may use \( \chi(A) = \frac{1}{\dim A} \text{Tr}(A) \) to roughly describe the closeness of an orthogonal matrix \( A \) to the identity matrix \( I \). By this means, we conclude that the average enhancement of the diagonals of matrix \( L \) is

\[ \sqrt{\frac{\chi(KS) + n}{2n}} \leq \chi(L) \leq \sqrt{\frac{\chi(KS) + 1}{2}} \]  

(B.18)

The right-hand side of Eq. B.18 shows the best result for the balanced solution, whereas the left-hand side of Eq. B.18 estimates the worst situation. The greater the \( \chi(KS) \), the better the result. We can perform transformation in \( K \) to increase its absolute values on its diagonal, and multiply it by \( S \), with the diagonals almost corresponding to the signs of \( K \)'s diagonals. The result might not exactly correspond to the signs of \( K \)'s diagonals because of the condition that the determinant of \( S \) equals the determinant of \( K \).

**B.5 Theoretical construction**

The steps of constructing a balanced solution are described in detail in this section. First, we need to obtain two eigenvector matrices from ENM on two conformational states of a protein structure. These two matrices are the normal mode vectors of motions that describes protein dynamics, in which the first 6 eigenvalues in the mode matrices are rigid body motions.
Second, we need to improve the absolute values of the diagonals in $K$, a step that could also be viewed as making matrices $C$ and $O$ more alike to each other. We exchange the orders of vectors in $C$ and $O$, i.e., shifting the mode indices in the normal mode space obtained from the closed state and those in the normal mode space from the open state. The two new basis after their indices shifted are denoted as $C'$ and $O'$. The order exchange in vectors of $C$ will exchange the order of rows in $K$, whereas the order exchange in vectors of $O$ will exchange the order of columns in $K$, as shown in Eq.B.1 and Eq.B.2. The relationship of $O'$ and $C'$ can be written as $O' = C'K'$. In our work, we simply keep the order of indices in the modes from the open form and shift the indices in the closed form, which ensures the x-axis of the overlap figure of two sets of vectors unaltered. By exchanging the rows in $K$, we can finally get a new overlap figure with the diagonal values in $K$ reach maximum. The way to exchange the orders of vectors in $C$ and $O$ depends on the specific form of $K$. In this study, we only shift the indices of the initial 10 normal modes. This can be justified by the fact that in ENM, the low frequency modes are usually of more importance. The high-indexed modes usually characterize local motions. They are trivial and do not account for much structural dynamics. In our study, we shift indices of the first 10 normal modes for the closed form (or the structure that has the minimal average RMSD among all the rest structures). The reasons are shown and explained in the result section in Chapter 5.4. Practically, we can also shift all modes in the mode vectors matrix of the closed form to match modes in the open form in order to produce highest sum up value over values on diagonal of $KS$ matrix. However, this step is not necessary and is time consuming.
Next, we need to determine matrix $S$. We create the matrix $S$ according to three principles: (a) the diagonals of $S$ are 1 or -1; (b) the determinant of $S$ equals the determinant of $K'$; (c) $\chi(K'S)$ is as great as possible. We create a diagonal matrix with values equal the signs of $K'$ diagonals. If this diagonal matrix has the same determinant with $K'$, then it is assigned as $S$. Otherwise in the diagonal matrix, flip the sign whose correspondent in $K'$ has the smallest absolute value along the diagonal. Denote $\{\lambda_i = e^{i\theta_i}, \ i = 1,2, ..., n \}$ as eigenvalues, and $\epsilon_1, \epsilon_n, ..., \epsilon_n$ as eigenvectors of $K'S$. We write all the eigenvectors as a matrix $P$, where $P = (\epsilon_1, \epsilon_n, ..., \epsilon_n)$. We can then obtain a new diagonal matrix

$$D = \begin{pmatrix} e^{i\theta_1} & 0 & \cdots & 0 \\ 0 & e^{i\theta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{i\theta_n} \end{pmatrix}$$

(B.19)

Now $L = P^H D P$, where $P^H$ is the Hermitian conjugate of $P$, i.e., the transpose and complex conjugate of $P$. Although $P$ and $D$ are complex, $L$ is real. The balanced solution is

$$X = C'L = O'SL^T$$

(B.20)
REFERENCES


stability, and free-energy landscapes: a molecular dynamics investigation.


158. Torrie, G.M., and J.P. Valleau. 1977. Nonphysical sampling distributions in
187–199.

Biol. 18: 134–139.

conformational sampling methods based on bivariate lag-distributions of


162. Juraszek, J., J. Vreede, and P.G. Bolhuis. 2012. Transition path sampling of

Combining statistical potentials with dynamics-based entropies improves
31.

164. Tanaka, S., and H. a Scheraga. 1976. Medium- and long-range interaction
parameters between amino acids for predicting three-dimensional structures of

contact energies from protein crystal structures: quasi-chemical approximation.
Macromolecules. 18: 534–552.


