PREDICTIONS OF PROTEIN CHEMICAL SHIFTS AND PROTEIN SLOW MOTIONS

BY SISHI TANG

A dissertation submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of
Dr. David A. Case
and approved by

______________________________
______________________________
______________________________
______________________________

New Brunswick, New Jersey
Oct, 2011
ABSTRACT OF THE DISSERTATION

Predictions of protein chemical shifts and protein slow motions

by Sishi Tang
Dissertation Director: Dr. David A. Case

Nuclear magnetic resonance spectroscopy (NMR) is one of the most powerful biophysical techniques for studying biomacromolecules. The advances of NMR techniques are often facilitated by the development of computational methods for the purpose of data interpretation and analyses. In this way, more information is extracted from the experimental measurements and complementary descriptions are given to details elusive to NMR probes, so that the structural and dynamical behavior of the biomolecule can be adequately described. Hence the two major goals of this thesis are:

i) To calculate chemical shift anisotropy (CSA) accurately and to understand how CSA is influenced by the local environment.

ii) To predict and characterize important metastable conformations of proteins probed in NMR relaxation experiments.

The first aspect is covered by three chapters (Chapter 2 - 4), where CSA calculations using QM and QM/MM models are described and compared. First, we used a small fragment (NMA$_3$) model to determine the effect of vibrational motion on the magnitude and orientation of CSAs. Next, we applied the same model to predict chemical shift tensors and achieve qualitative agreement with experimental measurements for the GB1 protein. Later we showed that a more expanded AF-QM/MM (automated
fragment quantum mechanical/molecular mechanical) model is able to provide better quantitative predictions to chemical shift tensors via an appropriate representation of environmental effects. Our study is expected to compensate for the lack of direct experimental measurements of CSAs, and help uncover the rich structural information hidden in CSA data.

The second aspect is covered by four chapters (Chapter 5 - 8), where we used the loop motion of triosephosphate isomerase (TIM) as our primary model to study protein conformational changes. To corroborate the “population shift” theory, conventional MD simulations was first performed to show that metastable states of TIM can be induced and stabilized. Then adaptively biased molecular dynamics (ABMD) simulations were used to predict and characterize the metastable conformations for monomeric TIM. In order to characterize the free energy landscape of this loop motion accurately and efficiently, an iterative approach combining ABMD and umbrella sampling was developed. Subsequently, this approach was applied to understand why TIM is only active as a dimer from energy and dynamics perspectives. Furthermore, we extended the ABMD method so that the metastable states of proteins can be predicted from their essential motions. The details of the methods used to predict and characterize protein minor conformations are described, providing insights into the energy and dynamics programmed in protein functions.
Acknowledgements

First and foremost, I’d like to thank my advisor, Dr. David A. Case, for his help and guidance throughout my thesis work. He is always more than willing to provide insights and suggestions to my thesis projects whenever I ask for his inputs. More importantly, he gives me the freedom to choose and explore my own research and academic interests. Furthermore, I’d like to thank him for all the moral support he kindly offered during my transition to Rutgers University. Given his kindness, patience and intellect, I could not have wished for a better mentor and friend for my Ph.D study.

I would also like to acknowledge the past and present members of the Case group. I’ve been blessed to work with a group of brilliant scientists who made our lab an enviably pleasant environment to work in. I truely appreciate all your generous help, as well as the inspiring discussions during work and lunch hours.

Finally, I’d like to thank my parents for respecting my choice of career path and for showering me with unconditional love and support throughout my entire life.
Dedication

To my parents, Fuyong Tang and Zhiping Huang.
Table of Contents

Abstract ................................................................. ii
Acknowledgements ...................................................... iv
Dedication ................................................................. v

List of Tables ............................................................. xii
List of Figures ............................................................ xiv
List of Abbreviations ...................................................... xviii

1. Introduction ........................................................... 1
   1.1. Chemical shift predictions ........................................ 2
       1.1.1. The origin of the chemical shift tensor .................. 2
       1.1.2. The significance and challenges of chemical shift tensor measurements ........................................... 4
       1.1.3. Quantum mechanical predictions of chemical shift tensors .................................................. 5
   1.2. Protein dynamics .................................................. 10
       1.2.1. Conformational selection and NMR relaxation theory 10
       1.2.2. Computational approaches to finding metastable conformations ........................................ 11
           1.2.2.1. Unbiased molecular dynamics and underlying force field ........................................... 12
           1.2.2.2. Umbrella sampling and potential of mean force calculation ........................................... 13
           1.2.2.3. Experimental & computational studies of TIM ..................................................... 15

2. Vibrational Averaging of Chemical Shift Anisotropies in Model Peptides .... 18
   2.1. Introduction ........................................................ 18
2.2. Theory and methods .............................................. 21
   2.2.1. Relaxation theory ........................................... 21
   2.2.2. Normal mode analysis .................................... 24
   2.2.3. Path integral molecular dynamics ....................... 25
2.3. Results .......................................................... 27
   2.3.1. $^{15}$N order parameters ................................. 28
   2.3.2. Effective CSA tensors .................................... 29
   2.3.3. $^{13}$C relaxation ........................................... 32
2.4. Discussion ....................................................... 33

3. Chemical Shift Anisotropy Calculations Using Small Fragment Models .. 40
   3.1. Introduction ..................................................... 40
   3.2. Methods ......................................................... 42
      3.2.1. Molecular dynamics simulations ........................ 42
      3.2.2. Energy minimization .................................... 43
      3.2.3. Chemical shift tensor calculations ..................... 43
      3.2.4. Calculation of hydrogen bonding strength ............ 44
   3.3. Results and discussions ..................................... 44
      3.3.1. Direct comparison with experimental data ............. 44
      3.3.2. Isotropy-anisotropy correlation ....................... 46
      3.3.3. Effect of minimization ................................... 46
      3.3.4. Effect of system size .................................... 48
      3.3.5. Effect of solvent hydrogen bonding .................... 48
      3.3.6. Effect of intramolecular hydrogen bond inclusion .... 49
      3.3.7. Effect of backbone dihedral angle $\omega$ ............... 49
      3.3.8. Effect of hydrogen bond geometry ...................... 51
         3.3.8.1. Hydrogen bond distance and angles ................ 51
         3.3.8.2. Hydrogen bond pattern ............................. 51
         3.3.8.3. Hydrogen bond strength ............................ 52
3.3.9. Limitations and scope ........................................ 52
3.4. Conclusions ....................................................... 53

4. Chemical shift tensor calculations using AF-QM/MM Model ............... 64
4.1. Introduction ..................................................... 64
4.2. Methods ......................................................... 66
  4.2.1. Energy minimizations ..................................... 66
  4.2.2. Chemical shift tensor predictions ......................... 67
  4.2.3. Deming regression ......................................... 69
4.3. Results and discussion .......................................... 70
  4.3.1. Comparison between AF-QM/MM model and NMA_{3} model .... 70
  4.3.2. $^{13}$C and $^{15}$N CSAs in GB1 protein .................... 71
  4.3.3. $^{15}$N CSA in GB3 protein .................................. 72
  4.3.4. $^1$H chemical shift tensors of GB3 protein ............... 73
  4.3.5. Relations of isotropic shifts and shielding tensors ............ 75
  4.3.6. Predictions of ubiquitin CSAs .............................. 76
  4.3.7. Computational improvements on CSA predictions ............. 77
4.4. Conclusions ..................................................... 79

5. Preliminary Characterization of Metastable States of TIM ................... 94
5.1. Introduction ..................................................... 94
  5.1.1. Methods ..................................................... 95
    5.1.1.1. Molecular dynamics simulations ........................ 95
    5.1.1.2. Trajectory stability analyses .......................... 96
    5.1.1.3. Screening for suitable order parameter ................. 97
    5.1.1.4. Essential dynamics analyses ........................... 97
  5.1.2. Results and Discussion .................................... 98
    5.1.2.1. Trajectory Stability analyses .......................... 98
    5.1.2.2. Screening for suitable order parameter ................. 98
    5.1.2.3. Essential dynamics analyses ........................... 99
5.1.3. Conclusions ........................................................................ 100

6. Exploring the Free Energy Landscape of TIM Conformational Changes .... 106
  6.1. Introduction ............................................................................. 106
  6.2. Theory ....................................................................................... 108
    6.2.1. Weighted histogram analysis method ................................. 108
    6.2.2. Adaptively biased molecular dynamics method ............... 109
    6.2.3. Combining ABMD with US ................................................. 111
  6.3. Methods .................................................................................... 112
    6.3.1. ABMD simulations ............................................................... 112
      6.3.1.1. Benchmarking study of NMA ....................................... 112
      6.3.1.2. ABMD simulations of TIM ........................................... 112
    6.3.2. Free energy evaluations ....................................................... 113
      6.3.2.1. Umbrella sampling ......................................................... 113
      6.3.2.2. Umbrella corrections to ABMD ..................................... 113
      6.3.2.3. Combining ABMD and US .............................................. 114
    6.3.3. Reaction rate approximations .............................................. 114
    6.3.4. Comparisons with predictions from PLOP ......................... 115
  6.4. Results and discussion ............................................................... 116
    6.4.1. Benchmarking study for ABMD simulations: cis-trans conversion of NMA ......................................................... 116
    6.4.2. Prediction of TIM excited state using ABMD simulations .... 117
    6.4.3. Comparison between four free energy methods for characterizing TIM loop motion .................................................. 118
    6.4.4. Comparison with experimental study and analogous computational studies ................................................................. 120
    6.4.5. Conclusions ....................................................................... 121

7. Why is TIM only active as a dimer? A study of energy and dynamics of TIM monomer vs. dimer ................................................................. 133
7.1. Introduction ................................................................. 133
7.2. Methods ................................................................. 136
   7.2.1. Unbiased MD simulations for TIM monomer, dimer and mono-
          TIM ................................................................. 136
   7.2.2. ABMD simulations for TIM monomer, dimer and monoTIM .... 136
   7.2.3. Trajectory analyses ................................................. 137
7.3. Results ................................................................. 137
   7.3.1. Monomer vs dimer based ABMD-US simulations ................. 137
   7.3.2. Free energy profile of TIM loop closure and predicted metastable
          state ................................................................. 138
   7.3.3. Effect of dimerization on B-factors ............................ 139
   7.3.4. Effect of dimerization on catalytic residues .................. 140
   7.3.5. Effect of dimerization on non-catalytic residues ............ 143
   7.3.6. Difference in dynamics between TIM monomer, dimer and mono-
          TIM ................................................................. 144
7.4. Conclusion ............................................................. 145

8. Implementation and Applications of Adaptively Biased Essential Molecular
   Dynamics ................................................................. 166
   8.1. Introduction .......................................................... 166
   8.2. Theory and implementation ......................................... 168
      8.2.1. Theory .......................................................... 168
      8.2.2. Implementation .................................................. 169
      8.2.3. Methods ........................................................ 171
         8.2.3.1. Benchmarking study: NMA cis-trans transformation .... 171
         8.2.3.2. TIM conformation change ................................. 172
   8.3. Results and discussion ............................................. 172
      8.3.1. NMA cis-trans conformation change ......................... 172
         8.3.1.1. PCA analysis ............................................... 172
8.3.2. ABEMD simulation of TIM conformation change .......................... 173
8.3.2.1. PCA and RMS analyses ................................................. 173
8.3.3. ABEMD simulation of TIM conformation change ...................... 174
8.3.3.1. TIM conformational space sampled by ABEMD simulations ... 175
8.3.4. Evaluation of ABEMD simulations ........................................ 177
8.3.4.1. Comparison between ABEMD and ABMD simulations ................ 177
8.3.4.2. Scope of ABEMD simulations ......................................... 178
8.4. Conclusion ........................................................................ 178

References ................................................................................. 191
Vita ............................................................................................. 205
# List of Tables

2.1. \(^{15}\text{N} \) CSA and N–H dipolar averages .............................................. 37
2.2. Effective \(^{15}\text{N} \) CSA values ............................................................... 37
2.3. \(^{13}\text{C} \) CSA and Ca–C dipolar averages .................................................. 38
2.4. Effective \(^{13}\text{C} \) CSA values ............................................................... 38
3.1. Correlation between chemical shift tensors and isotropic shifts for \(^{13}\text{C} \)
   and \(^{15}\text{N} \) nuclei ............................................................................. 54
3.2. RMSD of GB3 structures ................................................................. 54
3.3. Reduced CSA for helix and sheet of GB3 structures ......................... 55
3.4. Comparison of \(\alpha\)-helix and \(\beta\)-sheet CSA tensors .............................. 55
3.5. The effect of GB based minimization on backbone dihedrals ............. 55
4.1. RMSD of GB1 and GB3 Structures ......................................................... 81
4.2. Correlation between experimental and calculated isotropic shifts ....... 81
4.3. Reduced \(^{13}\text{C} \) and \(^{15}\text{N} \) CSAs of GB1 protein .................................................. 82
4.4. Pearson correlation and Deming regression statistics of reduced \(^{13}\text{C} \), \(^{15}\text{N} \)
   CSA for GB1 protein ............................................................................. 83
4.5. Regular \(^{15}\text{N} \) CSA of GB3 protein ......................................................... 84
4.6. Pearson correlation and Deming regression statistics of \(^{15}\text{N} \) regular CSA
   (calculated vs. experimental) for GB3 protein ........................................... 84
4.7. Pearson correlation and Deming regression statistics of of \(^{1}\text{H} \) Regular
   CSA of GB3 protein ............................................................................. 84
4.8. Pearson correlation and Deming regression statistics for \(^{1}\text{H} \) CSA of GB3
   protein .................................................................................................. 85
4.9. Pearson correlation between principal components of CSTs .......... 85
4.10. Pearson correlation and least squares regression statistics for \textsuperscript{1}H isotropic shift .......................................................... 86
4.11. Ubiquitin CSA ................................................................. 86
6.1. Summary of RMSD for plop predicted closed structures. ............... 123
7.1. His 95 - Gly 97 hydrogen Bond in TIM monomer, dimer and monoTIM. 146
7.2. Lys 13 - Glu 97 salt bridge in TIM monomer, dimer and monoTIM .... 146
8.1. Code modules for ABMD calculations .......................................... 180
# List of Figures

1.1. conformational selection theory ........................................ 17  
2.1. N-methylacetamide trimer .............................................. 38  
2.2. Typical CSA tensor principal directions for peptides .............. 39  
2.3. Histogram of $\Delta \sigma$ values for NMA ............................... 39  
3.1. Model systems for NMR calculations ................................... 56  
3.2. Overlapped secondary structure of GB1 and GB3 .................... 56  
3.3. Sequence alignment and secondary structure for GB1 vs. GB3 protein 57  
3.4. Comparison of $^{15}$N variations between experimental & calculated values .......................................................... 57  
3.5. Comparison of CSA variations between experimental & calculated values .......................................................... 58  
3.6. Correlations between CST principal components and $\delta_{iso}$ ........ 58  
3.7. Effect of minimization on CSA and CSA-isotropy correlation ....... 59  
3.8. Convergence of $^{13}$C CSA upon minimization ...................... 59  
3.9. Effect of system size ...................................................... 60  
3.10. Effect of water inclusion ................................................ 60  
3.11. Effect of hydrogen bond ............................................... 61  
3.12. Effect of minimization-$\omega$ angle ................................... 61  
3.13. Effect of $\omega$ angle .................................................... 61  
3.15. Effect of hydrogen bond strength ..................................... 62  
3.16. Effect of hydrogen bonding pattern .................................. 63  
4.1. AF-QM/MM model for CSA calculations. ............................ 87  
4.2. Overlapped GB1 and GB3 structures. .................................. 87
4.3. Correlations of isotropic chemical shifts. 88
4.4. Sequence alignment of GB1 and GB3 proteins. 88
4.5. Comparison of reduced $^{15}$N and $^{13}$C CSA for GB1 protein. 89
4.6. GB1 protein backbone $\omega$ angle changes due to full minimization. 90
4.7. The correlation between experimental and calculated chemical shifts for ubiquitin. 90
4.8. Relationship between $^1$H isotropic shift and CSA tensors for GB3 protein. 91
4.9. Relationship between $^{13}$C isotropic shift and CSA tensors for GB1 protein. 92
4.10. Relationship between $^{15}$N isotropic shift and CSA tensors for GB3 protein. 93
5.1. Loop motion of TIM is not ligand gated. 101
5.2. Induce minor conformers via MD simulation. 101
5.3. Potential energy profile of bound TIM in the open form. 102
5.4. Potential energy profile of apo TIM in the closed form. 103
5.5. RMS profile of apo TIM in the closed form. 104
5.6. Selection of reaction coordinate: exclusion of hydrogen bonds. 104
5.7. Pseudo-dihedral angle as a reaction coordinate. 105
5.8. PCA analyses for MD simulations of apo and bound TIMs. 105
6.1. cis-trans transformation of NMA. 123
6.2. Torsional angle sampling in ABMD simulations of NMA. 124
6.3. The evolution of ABMD potential for NMA. 124
6.4. The convergence of ABMD potential for simulations of NMA cis-trans conversion. 125
6.5. The umbrella correction to ABMD potential for NMA simulation. 126
6.6. Comparison between predicted loop and crystal structures of apo TIM. 127
6.7. Hydrogen bonding network in the predicted “closed” loop structure. 127
6.8. Time dependence of potential of mean force using standard umbrella sampling. 128
6.9. ABMD potential and umbrella corrections of apo TIM. 129
6.10. Time-dependence of PMF using umbrella sampling based on ABMD structures .......................... 130
6.11. Convergence of errors in the ABMD-US approach .......................... 130
6.12. Comparison between the free energy profiles calculated from different free energy methods .......................... 131
6.13. The diffusion coefficient of apo TIM monomer along the reaction coordinate .......................... 132

7.1. Loop interactions in TIM dimer interface .......................... 146
7.2. RMSD profile of monomer ABMD simulations .......................... 147
7.3. RMSD profile of dimer ABMD simulations .......................... 148
7.4. Convergence of errors in the ABMD-US approach .......................... 148
7.5. The evolution of ABMD potential of TIM monomer and dimer .......................... 149
7.6. Comparison between converged free energy profiles of TIM monomer and dimer .......................... 150
7.7. B-factor comparisons among TIM proteins .......................... 151
7.8. B-factor comparisons for short and accelerated time scale dynamics .......................... 152
7.9. Comparison of pseudo-dihedral angle space sampled from ABMD for monomer vs. dimer .......................... 153
7.10. Comparison of RMS space for monomer vs. dimer .......................... 154
7.11. Sampled backbone conformational space of Asn 11 .......................... 155
7.12. Sampled backbone conformational space of Lys 13 .......................... 156
7.13. Sampled backbone conformational space of His 95 .......................... 157
7.15. Sampled sidechain conformational space of Asn 11 .......................... 158
7.16. Sampled sidechain conformational space of Lys 13 .......................... 159
7.17. Sampled sidechain conformational space of His 95 .......................... 160
7.18. Sampled sidechain conformational space of Glu 167 .......................... 161
7.19. Sampled backbone conformational space of Gln 19 .......................... 162
7.20. Sampled backbone conformational space of Ala 73 .......................... 163
7.21. Sampled sidechain conformational space of Gly 135 ........................................ 164
7.22. Sampled backbone conformational space of Gly 175 ........................................ 165
8.1. Flow of ABMD code in SANDER ................................................................. 180
8.2. Sample input file ......................................................................................... 181
8.3. Principal component analysis of 10ns long NMA MD simulations ............... 182
8.4. Magnitude of projection in the cis and trans conformations in the 10 ns of unbiased MD ................................................................. 183
8.5. Evolution of biasing potential in 18 ns of ABEMD ....................................... 184
8.6. Evolution of projection and $\omega$ angle in 18 ns of ABEMD ......................... 185
8.7. RMS fluctuation of C-\(\alpha\) atoms along the selected eigenmode ................. 186
8.8. Evolution of magnitude of eigenmode projection for TIM conformation change ......................................................................................... 187
8.9. Evolution of biasing potential for TIM conformational change ................. 188
8.10. RMS fluctuation of TIM loop structure in the ABEMD simulation ............. 188
8.11. Comparison of RMS spaced by ABMD and ABEMD simulations ............. 189
8.12. Cluster analyses of RMS space sampled in ABEMD simulations .............. 190
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviations</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABMD</td>
<td>adaptively biased molecular dynamics</td>
</tr>
<tr>
<td>AF</td>
<td>automated fragment</td>
</tr>
<tr>
<td>CSA</td>
<td>chemical shift anisotropy</td>
</tr>
<tr>
<td>CST</td>
<td>chemical shift tensor</td>
</tr>
<tr>
<td>DFT</td>
<td>density functional theory</td>
</tr>
<tr>
<td>DHFR</td>
<td>dihydrofolate reductase</td>
</tr>
<tr>
<td>GIAO</td>
<td>gauge independent atomic orbital</td>
</tr>
<tr>
<td>IED</td>
<td>interactive essential dynamics</td>
</tr>
<tr>
<td>IGLO</td>
<td>individual gauge for localized orbital</td>
</tr>
<tr>
<td>MM</td>
<td>molecular mechanics</td>
</tr>
<tr>
<td>MD</td>
<td>molecular dynamics</td>
</tr>
<tr>
<td>NMA</td>
<td>N-methyl acetamide</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance</td>
</tr>
<tr>
<td>PB</td>
<td>Poisson Boltzmann</td>
</tr>
<tr>
<td>PC</td>
<td>principal component</td>
</tr>
<tr>
<td>PMF</td>
<td>potential of mean force</td>
</tr>
<tr>
<td>RMS</td>
<td>root mean square</td>
</tr>
<tr>
<td>QM</td>
<td>quantum mechanics</td>
</tr>
<tr>
<td>TIM</td>
<td>triosephosphate isomerase</td>
</tr>
<tr>
<td>US</td>
<td>umbrella sampling</td>
</tr>
<tr>
<td>WHAM</td>
<td>weighted histogram analysis method</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy plays an indispensable role in the study of structure and dynamics of biomacromolecules [1, 2]. Experimental spectra can often be successfully measured and interpreted empirically, but more difficult cases require predictions and analyses aided by computational methods. Over the past decade, the advances in quantum mechanical and molecular mechanical calculations have significantly expanded NMR theory and enhanced the utility of NMR techniques [3]. In this chapter, we lay the foundation for the original research presented in the subsequent chapters. First, we describe the basic theory behind the techniques of chemical shift measurements, as well as the challenges in the measurements and interpretations of chemical shift tensors. More importantly, the current quantum mechanical methods used for chemical shift tensor predictions are discussed, which forms the basis for chapter 2, 3 and 4. Second, we describe the conformational selection theory and its derivation from the NMR relaxation experiments, followed by an overview of computational methods used to explore and characterize the metastable conformations. In addition, we will include background knowledge of the loop closure motion of triosephosphate isomerase (TIM), which is further studied in chapter 5, 6, 7 and 8.
1.1 Chemical shift predictions

1.1.1 The origin of the chemical shift tensor

In NMR spectroscopy, the influence of an applied magnetic field to the designated molecule leads to precession of electrons around the nuclei [4]. This precession generates an induced magnetic field, which is antiparallel to the applied magnetic field for closed shell atoms. As a result, the magnetic field strength $B$ felt at the nucleus is slightly different from the applied field $B_0$ [5]:

$$B = (1 - \sigma)B_0 \quad (1.1)$$

where the chemical shielding $\sigma$ is the change in resonance frequency of the shield nucleus relative to the bare nucleus, and $\sigma B_0$ is the induced component of the magnetic field. The resonance frequency depends on the magnetic field strength at the nucleus, which is lowered when the nucleus is shielded by larger electron density. In this way, the difference in local electronic environment causes the resonance frequency to vary on submolecular distance scale for discrete nuclei. The chemical shielding can be analytically evaluated as the mixed second derivative of the interaction energy $E$ with respect to the magnetic dipole moment of the nucleus $\mu$ and the external magnetic field $B_0$ [6]:

$$\sigma - 1 = \frac{\partial^2 E}{\partial \mu \partial B_0} \quad (1.2)$$

The Chemical shielding tensor describes the magnitude and orientational dependence of the chemical shielding. For anisotropic electron distributions, the chemical shielding tensor can be transformed with respect to the principal axis system, where all the off-diagonal elements vanish:

$$\sigma = \begin{bmatrix} \sigma_{xx} & 0 & 0 \\ 0 & \sigma_{yy} & 0 \\ 0 & 0 & \sigma_{zz} \end{bmatrix} \quad (1.3)$$
The chemical shifts $\delta$ are in turn defined as the differences between chemical shielding of the reference nuclei $\sigma_{\text{ref}}$ (TMS for $^{13}$C nuclei, for instance) and the chemical shielding tensor $\sigma$:

$$\delta \equiv \sigma_{\text{ref}} - \sigma \quad (1.4)$$

In practice, the absolute shielding cannot be measured, because the magnetic field strength and the nuclear shielding are not know to parts per million (ppm) accuracy. Hence in experimental settings, the chemical shifts are measured in ppm with respect to the spectrometer frequency $v_0$, so that it is independent of the applied magnetic field strength:

$$\delta = \frac{v - v_{\text{ref}}}{v_0} \times 10^6 \quad (1.5)$$

When the molecule tumbles rapidly and isotropically in solution, the isotropic chemical shift can be measured as the average of the principal components:

$$\delta_{\text{iso}} = \frac{1}{3}(\delta_{xx} + \delta_{yy} + \delta_{zz}) \quad (1.6)$$

The magnitude of these principal components are ranked according to their deviation from the isotropic shift, $\delta_{\text{iso}}$:

$$|\delta_{zz} - \delta_{\text{iso}}| \geq |\delta_{xx} - \delta_{\text{iso}}| \geq |\delta_{yy} - \delta_{\text{iso}}| \quad (1.7)$$

The orientation-dependent part of chemical shift is known as the chemical shift anisotropy, which is oftentimes calculated as the difference between the largest tensor and the average of the other two tensors in solution NMR [5]:

$$\Delta \delta = \frac{1}{2}(\delta_{xx} + \delta_{yy}) - \delta_{zz} \quad (1.8)$$

Alternatively, solid state NMR studies often evaluates the reduced CSA as the difference between the largest tensor and the isotropic chemical shift according to the
Haeberlen conventions [7]:

\[ \Delta \delta_{\text{red}} = \delta_{zz} - \delta_{\text{iso}} \]  

(1.9)

1.1.2 The significance and challenges of chemical shift tensor measurements

Experimental and theoretical knowledge of CSA is required for both solution and solid state NMR experiments for the sake of interpreting structural and dynamics information [8]. In solution NMR experiments, the understanding of CSA is required for relaxation rates interpretation in relation to anisotropic internal motion [9]. More specifically, CSA-dipolar relaxation interference contains important angular information and forms the basis of TROSY experiments[10, 11]. In solid state NMR experiments, the orientation dependence of nuclear-spin interaction is not averaged by overall reorientation of molecules. As a result, the measurements of chemical shift tensor is a useful probe for the conformation of protein backbones [12].

For \(^{13}\text{C}\) and \(^{15}\text{N}\) nuclei of the amide backbone, accurate knowledge of both magnitude and orientation of the CSA tensor are important when quantitative interpretation of the effects of relaxation interference experiments are desired [9]. For \(^{1}\text{H}\) nuclei, knowledge of \(^{1}\text{H}\) CSA is important to predict what field strength can be applied to optimize line narrowing effects and for understanding the variations in \(^{1}\text{H}\) line width in TROSY [13]. Furthermore, \(^{1}\text{H}\) CSA info can also be important for the orientation of N-H bond and the use of potentially-powerful PISEMA experiments in solid state NMR [14, 15].

So far, progress in NMR instruments and methods have allowed for accurate measurements of relaxation rates of protein backbone nuclei [16]. Nevertheless, detailed information on the CSA tensor is required for quantitative analysis of relaxation rates, which is usually unavailable [17]. To date, the standard analysis of \(^{15}\text{N}\) relaxation rates assumes uniform \(^{15}\text{N}\) CSA across the backbone and axial symmetry, even though these assumptions have proven to be highly inaccurate [18]. While knowledge of CSA tensors is critical to the correct interpretation of protein dynamics, its accurate evaluation remains challenging. In solution NMR experiments, extraction of CSA information relies on \(^{15}\text{N}\) transverse relaxation and CSA-dipolar cross-correlation experiments [19].
These indirectly measured CSAs often result in large experimental errors and high sensitivity towards the experimental methods and analyses [20]. In the mean time, the solid state measurements are plagued by the difficulty to produce uniformly labeled proteins [21].

More recently, advances in experimental techniques, such as new cross-correlation relaxation methods and slow magic angle spinning have allowed for the removal of the effect of protein alignment from CSA measurements [22, 23]. As a result, chemical shift tensors have been accurately measured for small globular proteins in solid state NMR studies and in solution experiments of weakly oriented systems [23, 24, 25]. In the light of these experiments, quantum chemical predictions of chemical shift tensors can be made and directly compared with experimental data.

1.1.3 Quantum mechanical predictions of chemical shift tensors

Despite advances in experimental measurements of CSAs, currently there are not enough data to render empirical predictions of CSAs [26]. Hence accurate quantum mechanical calculations become critical in chemical shift anisotropy studies. In this section, we will give an overview of QM theory behind CSA calculations and a review of their application, both of which are restricted to closed shell systems.

To evaluate chemical shielding tensors quantum mechanically, the molecular Hamiltonian $\hat{H}$ in the presence of magnetic fields is obtained by replacing the operator of linear momentum $\vec{p}$ with $\vec{p} + \vec{A}$ [3]:

$$
\hat{H} = \hat{H}_0 + \vec{A} \cdot \vec{p} + \frac{1}{2} \vec{A}^2
$$

(1.10)

where $\vec{A}$ is the divergence free vector potential used to describe the external magnetic field $B_0$:

$$
\nabla \times \vec{A} = \vec{B}_0
$$

(1.11)

The vector potential $\vec{A}$ can be expressed in terms of the external field $\vec{B}_0$ and the
gauge origin $\vec{R}$:

$$\vec{A} = \frac{1}{2} \vec{B}_0 \times (\vec{r} - \vec{R})$$  \hspace{1cm} (1.12)

According to the Biot-Savart Law, a current density $\vec{j}(\vec{r})$ induces a magnetic field of $-\sigma B_0$ at position $\vec{s}$:

$$-\sigma B_0 = -\frac{1}{c^2} \int \frac{\vec{j}(\vec{r}) \times (\vec{r} - \vec{s})}{|\vec{r} - \vec{s}|^3} d\vec{r}$$  \hspace{1cm} (1.13)

The current density is then given by equation 1.14 for systems with $n$ doubly occupied orbital $\varphi_k$:

$$\vec{j} = i \sum_{k=1}^{n} \left\{ \varphi_k^* \vec{\nabla} \varphi_k - (\vec{\nabla} \varphi_k^*) \varphi_k \right\} - 2 \vec{A} \sum_{k=1}^{n} \varphi_k^* \varphi_k$$  \hspace{1cm} (1.14)

As an approximation, the linear response component of the induced current density, $\vec{j}^{(1)}$, is calculated as the sum of the contributions from diamagnetic (the first) and paramagnetic (the second) term:

$$\vec{j}^{(1)} = 2 \sum_{k=1}^{n} (\varphi_k^{(1)} \vec{\nabla} \varphi_k^{(0)} - \varphi_k^{(0)} \vec{\nabla} \varphi_k^{(1)}) - 2 \vec{A} \sum_{k=1}^{n} \varphi_k^{(0)} \varphi_k^{(0)}$$  \hspace{1cm} (1.15)

where $\varphi_k^{(0)}$ and $\varphi_k^{(1)}$ are the first order response of the orbitals $\varphi_k$, respectively. As a result, the paramagnetic component of the calculated current density $\vec{j}^{(1)}$ depends on the choice of gauge origin $\vec{R}$. When the Kohn-Sham equations are solved exactly, $\vec{j}^{(1)}$ is expected to be independent of gauge origin $\vec{R}$ [27]. But in actual QM calculations, only the atomic orbitals along the axis of the applied magnetic field can be accurately described. When an atom is situated far from the axis of rotation of the orbital, the magnetic field lead to perturbation in the momenta that twists the molecular orbitals, which can only be described by high angular momentum basis functions not included in the basis set [28]. Referring back to equation 1.15, $\varphi^{(1)}$ cannot be sufficiently expanded in terms of basis set functions without introducing a high cost of slow basis set convergence.
The IGLO method (individual gauges for localized orbitals) is the first solution to the gauge origin problem, where the local gauge origins were introduced by assigning individual gauge origins to localized orbitals $\varphi_k^{(0)}$ [29]:

$$\vec{j}^{(1)} = 2 \sum_{k=1}^{n} (\varphi_k^{(1)} \vec{\nabla} \varphi_k^{(0)} - \varphi_k^{(0)} \vec{\nabla} \varphi_k^{(1)}) - \sum_{k=1}^{n} (\vec{B}_0 \times (\vec{r} - \vec{R}_k)) \varphi_k^{(0)} \varphi_k^{(0)}$$ (1.16)

While the IGLO method is important and unprecedented, it also presents some disadvantages. In particular, the magnetic shielding tensors in a given calculation may depend on the choice of localized orbitals, as the orbital localizations is not always uniquely defined. In comparison, the GIAO (gauge-invariant atomic orbital) method resolves this problem by introducing the distributed gauge origins at the level of basis functions instead of molecular orbitals [30]. The following equation illustrates the GIAO method by expanding the orbitals into $N$ basis functions $\chi_{\mu}$:

$$\varphi_k^{(0)} = \sum_{\mu=1}^{N} C_{\mu k} \chi_{\mu}$$ (1.17)

$$\vec{j}^{(1)} = 2 \sum_{k=1}^{n} (\varphi_k^{(1)} \vec{\nabla} \varphi_k^{(0)} - \varphi_k^{(0)} \vec{\nabla} \varphi_k^{(1)}) - \sum_{k=1}^{n} \sum_{\mu,\nu=1}^{N} C_{\mu k} C_{\nu k} (\vec{B}_0 \times (\vec{r} - \frac{1}{2}(\vec{R}_{\mu} + \vec{R}_{\nu}))) \chi_{\mu} \chi_{\nu}$$ (1.18)

where $\vec{R}_\mu$ is the center of the basis function $\chi_{\mu}$.

When combined with the distributed gauge origin approaches such as IGLO and GIAO, density functional theory (DFT) methods have proven to be a promising for calculating CSTs. In comparison to Hartree Fock (HF) methods, DFT employs an exchange-correlation potential instead of ordinary exchange potentials, which leads to a better approximation of electron correlation effects [31]:

$$E = \sum_{k=1}^{n} \int d\vec{r} \varphi_k^* \left( \frac{\vec{p}_k^2}{2} + V_N \right) \varphi_k + \frac{1}{2} \int d\vec{r}_1 \int d\vec{r}_2 \frac{\rho(\vec{r}_1) \rho(\vec{r}_2)}{|\vec{r}_1 - \vec{r}_2|} + E_{xc}$$ (1.19)

where $\rho = \sum_{k=1}^{n} \varphi_k^* \varphi_k$ is the electron density of the system, $V_N$ is the nuclear potential
and \( \vec{p} \) is the linear momentum operator. In the mean time, it also allows for a significant increase in computational efficiency, in comparison to post-HF methods such as coupled cluster (CC) and Moller-Plesset perturbation theory (MP2, MP3, MP4, etc). The exchange correlation functional, \( E_{xc} \), is evaluated with three types of approximations, the local density approximation (LDA), the generalized gradient approximation (GGA) and hybrid functionals, which includes part of the exact HF exchange as well as LDA and/or GGA. For example, one of the most popular hybrid functionals, B3LYP has its exchange correlation potential calculated as follows [32]:

\[
E_{xc}^{B3LYP} = 0.20 \times (E_x^{HF} - E_x^{LDA}) + 0.72 \times (E_x^{GGA} - E_x^{LDA}) + 0.81 \times (E_c^{GGA} - E_c^{LDA}) + E_{xc}^{LDA}
\]

(1.20)

where \( E_x \) and \( E_c \) are exchange and correlation functionals.

Like the current density, the \( st \) tensor component chemical shift tensors in the DFT-GIAO formulation is obtained as a sum of the diamagnetic \( \sigma_{st}^d \) and paramagnetic components \( \sigma_{st}^p \) [31], where \( s \) and \( t \) are indices to a particular tensor component:

\[
\sigma_{st} = \sigma_{st}^d + \sigma_{st}^p
\]

(1.21)

\[
\sigma_{st}^d = \sum_i n_i \left\{ \frac{1}{\varepsilon^2} \varphi_i \sum_v \frac{2M}{\varepsilon} \frac{1}{2r_N} \left[ \vec{r}_N \cdot \vec{r}_v \delta_{st} - r_N r_v \right] \chi_v \right\} + \sum_i n_i \left\{ \frac{1}{\varepsilon} \sum_{\lambda,v} d_{\lambda i} d_{v i} \left( \vec{r}_v \times (\vec{R}_v - \vec{R}_\lambda) \right) h_{t}^{01} \chi_v \right\}
\]

(1.22)

\[
\sigma_{st}^p = \sum_i n_i \sum_{v} 2M \sum_{\lambda,\nu} d_{\lambda i} d_{v i} \left( \chi_\lambda \left\{ \frac{1}{\varepsilon} \left( \vec{R}_v - \vec{R}_\lambda \right) h_{t}^{01} \chi_v \right\} + \sigma_{st}^{p,oc-oc} + \sigma_{st}^{p,oc-vir}
\]

(1.23)

where \( \varphi_i \) with occupancy \( n_i \) is expanded into \( 2M \) atomic orbitals with expansion coefficients \( d_{\beta i} \), and the operator \( h_{t}^{01} \) is defined as:
And $\sigma_{st}^{p,oc-oc}$ and $\sigma_{st}^{p,oc-vir}$ are the contributions from the occupied and virtual zero-order molecular orbitals, respectively.

Given the advances in QM theory of chemical shielding, a lot of effort has been devoted to QM calculations of isotropic shifts of the backbone nuclei of proteins, including $^1$H, $^{13}$C' and $^{15}$N [33, 34]. In these studies, significant correlations were established between chemical shifts and a wide range of facts, including backbone and sidechain torsional angles, ring current effect, hydrogen bonding and local contacts [35]. As discussed in 1.1.2, chemical shift tensors carry more detailed information in dynamics and structures in comparison to their isotropic averages. From a computational perspective, the evaluation of individual tensor components is also a more rigorous test for the quality of predictions. However, the number of QM CSA studies cannot remotely compare with the ICS studies. This is largely due to the lack of accurate experimental observables as a reference point for the purpose of evaluating CSA predictions, as well as the high sensitivity of CSAs towards local environments[26].

To date, the characterization of CSAs have been mostly limited to di- & tri-peptide species. For example, Czinki and coworkers mapped the $^{15}$N and $^{13}$C CSA surface using L-Ala-NH$_2$ as a model for peptides and proteins. Their study, which does not take the nonbonded interactions into account, suggest that the magnitude of CSAs cannot be uniquely determined by the backbone conformation of peptides or secondary structure of proteins [36]. In another recent study, the $^{15}$N chemical shift tensors of the selected residues in GB3 protein were calculated using the N-formylalanyl model, but the accuracy of chemical shift anisotropy calculations were not addressed [37]. Hence so far there are very few systematic investigations for CSA on the complete protein backbone, and even fewer studies with adequate representation of the entire protein structure and solvent effect.

From Chapter 2 to Chapter 4 of this thesis, we focus on development, application and evaluation of new models used for CSA predictions. In Chapter 2, we apply a
small fragment QM model to study how CSA is influenced by vibrational motions of proteins. In Chapter 3, the same QM model is evaluated based on its ability to determine the helix/sheet propensity of proteins. In addition, we explore potential improvements to the QM model and the source of its sensitivity. In Chapter 4, we further expand this model to include effects of the complete protein environment and solvent exposure and show how the quality of CSA predictions can be significantly improved.

1.2 Protein dynamics

1.2.1 Conformational selection and NMR relaxation theory

The flexibility of proteins is associated with a variety of protein functions, including enzyme catalysis, molecular recognition, signal transduction and assisted protein folding [38]. According to the longstanding "induced-fit" model of protein motion, the presence of a ligand leads to accommodating motions of the target protein [39]. More recently, advances in NMR theory and techniques have led to the “conformational selection theory” which suggest that the multiple conformational states coexist in equilibrium, and perturbation to the system (e.g. introducing a ligand) induces a population shift towards the more favorable conformations (Figure 1.1).

According to NMR relaxation theory, the conformational change of proteins is reflected in the change in the electronic environment, which disrupts the normal nuclear Larmor precession. As a result, the slow motions of protein (μs-ms) can be probed by the exchange broadening of the resonance signal (Rex), which is indicated by an increase in transverse relaxation rate (R2) [40]. Two kinds of experiments are key to measuring Rex, including the CPMG-based relaxation dispersion experiment and R1ρ relaxation experiment [41, 40]. CPMG experiments have been used to measure R2 relaxation rates as a function of different interpulse delay at an interval τcp:

\[
R_2 \left( \frac{1}{\tau_{cp}} \right) = R_2^0 + R_{ex}
\]  

(1.25)
The resulting relaxation-dispersion data can be fitted to dispersion analysis, so that the rate of exchange \( (k_{ex}) \), magnitude of difference in chemical shifts \( (\Delta \omega) \), and relative population of the ground state \( (p_A) \) can be determined in the fast exchange limit \( (k_{ex} \leq \Delta \omega) \):

\[
R_{ex} = p_A p_B \frac{\Delta \omega^2}{k_{ex}} \left(1 - \frac{4v_{CPMG}}{k_{ex}} \tanh\left(\frac{k_{ex}}{4v_{CPMG}}\right)\right)
\] (1.26)

Alternatively, conformational exchanges can be observed by the off-resonance \( R_{1\rho} \) relaxation experiments by measuring signal decay in the presence of an off-resonance, continuous wave spin-locking RF pulse. The \( R_{1\rho} \) relaxation rate is measured as a function of longitudinal \( (R_1) \) and transverse relaxation \( (R_2) \) at various angles \( \theta \) and spin-locking fields \( \omega_e \).

\[
R_{1\rho} = R_1 \cos^2 \theta + R_0^1 \sin^2 \theta + R_{ex} \sin^2 \theta
\] (1.27)

Similar to CPMG experiments, the exchange parameters are then derived by fitting to \( R_{ex} \).

\[
R_{ex} = \frac{p_A p_B \Delta \omega^2 k_{ex}}{k_{ex}^2 + \omega_e^2}
\] (1.28)

These experiments have been applied to the studies/interpretation of \( \mu s-ms \) dynamics for adenylate kinase (ADKinase), dihydrofolate reductase (DHFR), T4 lysozyme, Ribonuclease A (RNase) and Triosephosphate Isomerase (TIM) [42, 43, 40]. The experimental evidences and theoretical studies of TIM loop motion will be discussed in details in 1.2.2.3.

### 1.2.2 Computational approaches to finding metastable conformationations

While CPMG and \( R_{1\rho} \) experiments have proven effective in detecting slow motions of proteins on \( \mu s-ms \) time scale, it remains a challenge to detect a minor population that goes below 0.5% of the entire conformational ensemble. In addition, a quantifiable
difference in chemical shift is required to characterize the exchange rate \( (k_{ex}) \) for individual residues. To complement experimental evidences, computational theory can be used to give a complete picture of the coordination of individual residues in the global dynamics of proteins, and even predict the functionally important conformations “invisible” to experiments. In the following sections (1.2.2.1 and 1.2.2.2), we provide an overview of molecular dynamic simulation and free energy methods, which can be adopted to capture the energy and dynamics of “invisible” conformations.

1.2.2.1 Unbiased molecular dynamics and underlying force field

Molecular dynamics (MD) simulation is one of the most widely used method for atomic-level models of macromolecules in condensed phase [44]. In MD simulations, the molecules are represented as collections of atom-centered interaction sites, where the potential energy terms are calculated according to classical molecular mechanics based force fields. The motions and new conformations of proteins are derived from the newtonian equation of motion for each atom:

\[
F_i = \frac{d}{dt} m_i \dot{v}_i = m_i a_i
\]  

(1.29)

where \( F_i \) is the force applied to the atom, and \( v_i, m_i \) and \( a_i \) are the velocity, mass and acceleration, respectively. A small time step is often used, after which the forces applied to all atoms in the molecule are updated simultaneously, and the new atomic positions and velocities are determined. The total energies and forces are calculated using the selected force field. For example, the basic potential function representing the AMBER force field has the following form [45]:

\[
V(r) = \sum_{bonds} K_b (b - b_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} \left( \frac{V_n}{2} \right) (1 + \cos(n\phi - \delta)) + \sum_{nonbond} \left\{ \epsilon_{ij} \left[ \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{12} - \left( \frac{R_{\text{min},ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon D r_{ij}} \right\}
\]

(1.30)
where $V(r)$ is the total potential energy, and the energy terms described are internal and external interactions, respectively. The bond and angle terms describe the deviation of the bond length and angles from their equilibrium position, $b_0$ and $\theta_0$, where $K_b$ and $K_{\theta}$ are the force constant of the harmonic potential. The dihedral term is a fourier series truncated to a single cosine term, where $V_n$, $n$ and $\delta$ are the rotational energy barrier, periodicity and phase, respectively. The bond, angle and dihedral terms are known collectively as the internal energy terms.

The nonbond energies, or the external energies, include the van der Waals term and the electrostatic term. The van der Waals term is represented by a Lennard Jones 6-12 potential, where the $\epsilon_{ij}$ and $R_{\text{min},ij}$ are the energy well depth and distance of minimal energy interaction, respectively. The electrostatic term describes the Coulombic interaction between two charged atomic centers, where $q$ and $\epsilon_D$ are the charge and appropriate dielectric constant between atoms.

The atomic level simulations derived from empirical force fields are capable of giving accurate predictions of structure and dynamics properties of macromolecules [46]. Given state-of-the-art computer hardware, MD simulations can be performed to explore protein structure and dynamics up to a few microseconds [47]. However, most current MD studies still rely on ps-ns time scale simulations due to limitations of computational resources. In order to study even slower motions of proteins (e.g. $\mu$s-ms) of functional importance, it is necessary to enhance sampling of protein conformations and accelerate slow conformational transitions.

### 1.2.2.2 Umbrella sampling and potential of mean force calculation

For regular MD simulations on ps-ns time scale, slow protein motions and the corresponding “invisible” conformations cannot be easily observed due to high energy barriers. As a result, it is hard to evaluate the relative population of the co-existing conformations as well as the potential of mean force of the transition. To resolve this problem, the umbrella sampling method was developed by introducing a bias potential along a predefined reaction coordinate [48].

In umbrella sampling, the Hamiltonian of the system is changed by introducing
an “umbrella” potential to bias the sampling into regions not adequately sampled in standard simulations. To define the reaction coordinate of umbrella sampling, a few degrees of freedom (DOF) are chosen to represent the structural changes involved in the transition.

\[ V'(r) = V(r) + w(r) \]  

(1.31)

where \( w(r) \) is the umbrella potential. In MD simulations, oftentimes \( w(r) \) takes the form of a harmonic potential with respect to a particular reference coordinate \( r_0 \).

\[ w(r) = \frac{1}{2} k (r - r_0)^2 \]  

(1.32)

As a result, umbrella sampling allows for relatively uniform sampling by driving the system along the reaction coordinate. The biased probability densities \( \rho'(r) \) can be obtained for \( w(r) \) centered at different \( r_0 \)'s:

\[ \rho'(r) = \frac{e^{-\beta V'(r)}}{\int e^{-\beta V'(r')} dr'} \]  

(1.33)

The original probability can be recovered from the biased probability density is related to density via the following equation:

\[ \rho(r) = \frac{\rho'(r) e^{\beta w(r)}}{<e^{\beta w(r)}>_{NVT}} \]  

(1.34)

The corresponding potential of mean force (PMF) can then be calculated as the free energy function of the chosen coordinate \( r \):

\[ F(r) = -kT ln(\rho(r)) + constant \]  

(1.35)

Numerically, the PMF of the collective coordinate transition is often calculated using the weighted histogram analysis method (WHAM) [49], the details of which are covered in 6.2.1.
1.2.2.3 Experimental & computational studies of TIM

We will use triosephosphate isomerase (TIM) as a model for our computational study of protein conformational changes, as its energy and dynamics of loop closure has been extensively studied in experimental investigations. TIM is a homodimeric enzyme with 249 residues in each monomer, catalyzing the reversible isomerization between dihydroxyacetone-phosphate (DHAP) and glyceraldehyde-3-phosphate (GAP) with diffusion-controlled efficiency [50]. Each monomer has an (α/β)$_8$ fold, which is commonly known as the TIM barrel fold. High resolution crystal structures of TIM shows that an 11-residue loop near the catalytic site remains open in the unligated form of TIM, whereas in the presence of a ligand the same loop moves more than 5 Å to close over the active site [51]. Site-directed mutagenesis studies proved that this loop conformation change is crucial to achieving its catalytic proficiency [52]. The dynamics of TIM has been characterized by solution and solid state NMR studies, which indicate that the loop opening is not ligand-gated but rate-limiting, or at least partially rate-limiting, for TIM catalysis [53]. The barrier of this loop motion has been estimated to be 10 - 12 kcal/mol, with the difference between the open and closed state in the range of 1.2 - 2.5 kcal/mol. Accordingly, the conformation exchange of the loop is expected to occur on the $10^{-14}$ s$^{-1}$ time scale.

So far, a number of theoretical investigation have been performed on the loop motion of TIM, but these studies failed to provide insight to the excited states of proteins or provide correct prediction of the energy barrier of the transition. For example, an earlier study based on coarse-grained (brownian motion) model suggested that the loop exhibit ligand-gated motions with nanosecond time scale transitions in the absence of the ligand, which is much faster than the $\mu$s-$ms$ time scale of motion observed from NMR experiments [54]. In another study, the enthalpic barrier of the loop motion was estimated to be 10 kcal/mol higher than what the experimental data suggested [55]. To this end, it is unclear where the discrepancy between theory and experiment comes from.

Using TIM as our model of conformational transitions, we seek to develop a more
robust method and protocol of molecular dynamics simulation and free energy calculations, so that these conformational transitions and metastable states can be characterized structurally and energetically. In Chapter 5, we use unbiased molecular dynamics simulations to characterize the stable and metastable conformations involved in TIM loop motions. The appropriate order parameter of TIM loop motion is subsequently defined, which is necessary for the biased MD simulations and free energy calculations. In Chapter 6, we introduce adaptively biased molecular dynamics (ABMD) simulations, which allows for continuous updates of suitable umbrella potentials so that the free energy landscape can be effectively traversed and the “metastable” conformations predicted. More importantly, we develop an iterative approach of free energy evaluation by combining ABMD biasing potential and umbrella sampling. The efficiency and accuracy of this method will be compared with other free energy approaches previously developed. In Chapter 7, we seek to understand why TIM is only active as a dimer with two independent active sites, by comparing the energy and structural changes of monomeric and dimer TIM when the loop motion is concerned. In Chapter 8, we address the problem that protein conformational changes often incorporate collective motions, which cannot be adequately characterized by a few simple reaction coordinates. Thus we develop an essential dynamics extension to ABMD, which is then applied to the cis-trans conformational change of NMA and the loop movement of TIM.
Figure 1.1: conformational selection theory
Chapter 2

Vibrational Averaging of Chemical Shift Anisotropies in Model Peptides

The effects of chemical shift anisotropy (CSA) are evident in line-shapes or side-band analysis in solid-state NMR, in the observed line positions in partially oriented samples, and in relaxation effects in liquid-state studies. In all of these cases, the effective shielding tensor is influenced by fast vibrational averaging in addition to larger-amplitude internal motions and to overall libration or rotation. Here we compute the contributions of vibrational averaging (including zero-point motions) to the CSA relaxation strengths for the nitrogen and carbonyl carbon in two simple peptide models, and for snapshots taken from a path-integral simulation of a small protein. Because the $^{15}$N shielding tensor is determined by all the atoms of the peptide group, it is less influenced by vibrational motion than (for example) the N-H dipolar interaction, which is more sensitive to the motion of the light hydrogen atom. Computed order parameters for CSA averaging are hence much closer to unity than are N-H dipolar order parameters. This leads to a reduction by about 9% in the magnitude of the amide nitrogen CSA that is needed to fit liquid-state relaxation data. Similar considerations apply to the carbonyl carbon shielding tensor, but in this case the differences between dipolar and CSA averaging are smaller. These considerations will be important for making comparisons between CSA tensors extracted from various NMR experiments, and for comparisons to quantum chemical calculations carried out on static conformers.

2.1 Introduction

Chemical shielding is a tensor quantity, and the effective field at the nucleus depends upon the orientation of this tensor with respect to the magnetic field. The way in
which this orientation is averaged can provide important information about macromolecular structure and dynamics. This averaging affects line-shapes in solid state NMR, residual chemical shifts in partially aligned samples, and relaxation behavior in all environments. The basic theory of how motions affect the observed parameters is formally analogous to that used for dipolar coupling [56], but things are more complicated in practice, since shielding is an electronic property that (unlike dipolar couplings) can depend in a very complex way on nuclear coordinates. The way in which isotropic shifts are averaged by vibrational motion has been well investigated [57, 58], but much less is known about CSA effects. Here we use quantum chemical calculations and normal mode theory to study vibrational averaging in model peptides, in a manner analogous to that used previously to examine dipolar couplings [59]. We also extract snapshots from a path-integral simulation of a small protein to compare with the simple peptide models.

Many standard treatments of relaxation theory implicitly assume that the principal components and directions of the CSA tensor are fixed in a local molecular frame, making the theory analogous to that used for dipolar couplings when bond distances and angles are assumed to be constant. This approach takes advantage of the fact that local vibrational averaging is invariably in the extreme narrowing limit, so that one can account for it through the use of an effective CSA tensor (or an effective bond length in the case of dipolar couplings). Since the components of the CSA tensor are (in practice) almost always treated as empirical parameters, using such an effective tensor as a fitting parameter makes good sense. Nevertheless, there are a number of reasons why it would be useful to understand the details of the vibrational averaging that creates the effective tensors. First, different experiments involve different sorts of averaging: qualitatively, “pure” CSA relaxation experiments depend upon the square of the shielding tensor, whereas CSA/dipole cross-correlation, residual shifts and line-shapes in solid-state experiments involve terms linear in the shielding anisotropy. One needs a model for vibrational motion and its effects on shielding in order to compare results, or to fit multiple data sets to a single atomic model. Second, electronic structure calculations typically consider only static (average) structures. If insight from
quantum chemistry is to become quantitatively useful for the interpretation of NMR
data, we need to understand how to connect static and vibrationally-averaged shielding tensors.

The classic treatment of effects of vibrational averaging on dipolar and quadrupolar
 coupling is that of Henry and Szabo [60]. The dipolar interaction depends only
upon nuclear positions, so that its behavior under vibrational averaging is straightforward, assuming that vibrational normal modes are available. Quadrupolar relaxation
depends upon the behavior of the electric field gradient (EFG) tensor, which is an electronic property; Henry and Szabo were able to make progress here by using a model
in which the EFG was assumed to depend in a certain way upon bond length (and not
upon any other geometric variables). Averaging of CSA interactions is a more complex
problem since the dependence of shielding on molecular geometry is not well understood. Recently, Zuiderweg and co-workers have used density functional calculations
to create a model for how carbonyl carbon CSA tensors in peptides depend upon the
local nuclear geometry [61]. Using this, they could average over vibrational motion
to obtain an effective tensor, *i.e.*, one that would show the same CSA relaxation in a
rigid molecule as the real tensor does when vibrational effects are included. Here we
use a simpler but more “brute force” approach, generating snapshots that sample local
vibrational motion, and carrying out quantum chemistry calculations on each such
snapshot. This eliminates the need to identify the most important geometric variables
and to provide a fit for the complex behavior that connects CSA tensors and geometries; on the other hand, it means that each new type of averaging requires a fresh
set of quantum calculations. As noted below, the two sorts of calculations give nearly
identical conclusions for carbonyl carbons.
2.2 Theory and methods

2.2.1 Relaxation theory

The general theory of NMR relaxation is covered in many places, and will not be repeated here [62, 63]. Under appropriate conditions, the influence of molecular motion on spin transition rates is governed by components of a spectral density function, \( j_m^{\lambda,\lambda'}(\omega) \), which is the Fourier transform of a time-correlation function:

\[
C_m^{\lambda,\lambda'}(\tau) = \left\langle \omega^\lambda(0)\omega^{\lambda'}(\tau)C_{2m}(u^\lambda(0))C_{2m}(u^{\lambda'}(\tau)) \right\rangle
\] (2.1)

Here \( \lambda \) and \( \lambda' \) denote the relaxation operators involved; we will be concerned here with dipolar relaxation, where \( \omega^{\lambda} = \gamma_i\gamma_j/\hbar r^3 \), and CSA relaxation, where \( \omega^{\lambda} = \gamma_i B \Delta \sigma \). Here the \( \gamma \) factors are nuclear magnetogyric ratios, \( r \) is the instantaneous distance between the spins whose dipolar coupling is being considered, \( B \) is the magnitude of the external magnetic field, and \( \Delta \sigma \) is the instantaneous value of the shielding anisotropy. For dipolar coupling, the unit vector \( u^\lambda \) lies along the vector connecting the two spins, whereas for CSA effects, it is the direction of unique principal component of the shielding tensor. The brackets in Eq. 2.1 indicate an average over all the molecules in the ensemble, and \( C_{2m} \equiv (4\pi/5)^{1/2}Y_{2m} \) is a modified spherical harmonic.

If there is no preferred direction in space, as in isotropic solution or in a randomly-oriented powder, the expression in Eq. 2.1 is independent of the subscript \( m \), and we can average over the five possible values using the spherical harmonic addition theorem. This gives

\[
C^{\lambda,\lambda'}(\tau) = \left\langle \omega^\lambda(0)\omega^{\lambda'}(\tau)P_2[u^\lambda(0) \cdot u^{\lambda'}(\tau)] \right\rangle
\] (2.2)

where \( P_2(x) \equiv (3x^2 - 1)/2 \) is a Legendre polynomial. For local vibrational relaxation, the decay of these correlation functions is in the sub-picosecond regime, and we are always in the extreme narrowing limit for NMR relaxation. This implies that we can set \( \tau \) to a value short compared to overall tumbling, and yet still large enough that the vectors at time \( \tau \) have become uncorrelated with their values at time 0. Then Eq. 2.1...
becomes

\[ C_{\lambda \lambda'}^m (\tau \to \infty) = \left\langle \omega^\lambda C_{2m}^* (\mathbf{u}^\lambda) \right\rangle \left\langle \omega^{\lambda'} C_{2m} (\mathbf{u}^{\lambda'}) \right\rangle \]  

\[(2.3)\]

where each average is over the equilibrium ensemble.

Since vibrational averaging takes place on a very rapid time scale, and has a small amplitude, we can effectively separate its correlation function from much slower processes such as overall rotational tumbling. In the Lipari-Szabo, or “model-free” approach [64], the spectral density for the combination of fast internal and slower overall rotation becomes:

\[ J_{LS} = \frac{2S^2 \tau_c}{1 + \omega^2 \tau_c^2} + \frac{2(1 - S^2) \tau}{1 + \omega^2 \tau^2} \]  

\[(2.4)\]

Here, we have assumed for simplicity that the overall motion is isotropic with a rotational relaxation time \( \tau_c \). The squared order parameter, \( S^2 \) represents the plateau value of Eq. 2.3 (where \( C_{\lambda \lambda'}^m \) is normalized to unity at time zero), and \( \tau^{-1} = \tau_c^{-1} + \tau_e^{-1} \), where \( \tau_e \) is the decay time for the internal motion. When \( \tau_e \ll \tau_c \), as is the case here, the second term of Eq. 2.4 can be neglected; then the effect of the rapid vibrational motion is just to scale the rotational spectral density by a factor of \( S^2 \). The purpose of this paper is to provide some estimates of these scaling factors for plausible models of vibrational motion in peptides.

Vibrational averaging also affects solid-state NMR spectra, including powder patterns (or the closely allied spinning side-band analysis) that have traditionally provided our primary knowledge about chemical shielding anisotropies. Vibrational effects on solid-state lineshapes have been considered in detail before [65, 60, 66], and we will not repeat this analysis here. The general result is that the vibrationally averaged interaction strength extracted from a powder pattern is the square root of the interaction defined in Eq. 2.2 (assuming identical vibrations in the solid and liquid states). This is just what one might expect from the fact that the interaction Hamiltonian appears to second order in relaxation analysis, but only to first order in lineshape analysis. The same sort of square root relationship also holds for motional effects.
on residual dipolar couplings, for example [67, 68]. This simple connection between motionally averaged frequencies and relaxation depends on an assumption of axial symmetry [60, 68], but is approximately true even for non-axial systems.

The dipolar interaction tensor is always axially symmetric about the (instantaneous) bond vector, so that the $u^λ$ direction in Eq. 2.1 can be taken as this unique direction. Shielding tensors, however, need not have any axial symmetry. As a general second rank tensor, $σ$ can be decomposed into a sum of tensors of rank 0, 1, and 2:

$$\sigma = \sigma^{(0)} + \sigma^{(1)} + \sigma^{(2)}$$

Here $\sigma^{(0)}$ is the unit matrix (tensor) multiplied by the isotropic shielding, which is $(σ_{xx} + σ_{yy} + σ_{zz})/3$; this is a scalar quantity, independent of orientation. The rank 1 component $\sigma^{(1)}$ is the antisymmetric component of the full tensor:

$$\sigma^{(1)} ≡ (σ - σ^T)/2 \quad (2.5)$$

and the rank 2 tensor is the (traceless) orientation-dependent part of the symmetric component:

$$\sigma^{(2)} ≡ (σ + σ^T)/2 - σ^{(0)} \quad (2.6)$$

If we rotate the coordinate system to a frame where $\sigma^{(2)}$ is diagonal (which can always be achieved for a real symmetric tensor), then we can write:

$$\sigma^{(2)} = \begin{bmatrix}
σ_{11} - σ_{iso} & 0 & 0 \\
0 & σ_{22} - σ_{iso} & 0 \\
0 & 0 & σ_{33} - σ_{iso}
\end{bmatrix}$$

$$= \frac{1}{3} Δσ_1 \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} + \frac{1}{3} Δσ_2 \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad (2.7)$$
Here \( \Delta \sigma_1 = \sigma_{11} - \sigma_{33} \), and \( \Delta \sigma_2 = \sigma_{22} - \sigma_{33} \). The final line decomposes a general non-axial symmetric tensor into two axially symmetric parts, which can always be done. Note that the two tensors in the final line of Eq. 2.7 have values of \( \sigma_{||} - \sigma_{\perp} \) of \( \Delta \sigma_1 \) and \( \Delta \sigma_2 \), respectively. We can treat the nitrogen CSA tensor as approximately axially symmetric, so that \( \Delta \sigma_1 \equiv \Delta \sigma \) and \( \Delta \sigma_2 = 0 \). For the carbonyl carbon, however, the tensor is very rhombic, and both tensors (plus the cross correlations between them) must be considered to understand the relaxation process.

### 2.2.2 Normal mode analysis

Normal mode vibrational analysis is a standard topic [69], and only an outline is given here. The basic idea is to expand the potential function \( V(x) \) in a Taylor series expansion about some point \( x_0 \):

\[
V(x) = V(x_0) + g \cdot (x - x_0) + \frac{1}{2}(x - x_0) \cdot F \cdot (x - x_0) \quad (2.8)
\]

If the gradient \( g \) of the potential vanishes at this point and one ignores third and higher-order derivatives, it is straightforward to show that the dynamics of the system can be described in terms of the normal mode directions and frequencies \( Q_k, \omega_k \), which satisfy:

\[
M^{-1/2}FM^{-1/2}Q_k = \omega_k^2 Q_k
\]

\[
Q_k \cdot Q_j = \delta_{kj} \quad (2.9)
\]

In Cartesian coordinates, the matrix \( M \) contains atomic masses on its diagonal, and the Hessian matrix \( F \) contains the second derivatives of the potential energy evaluated at \( x_0 \). The (classical) time evolution of the system is then:

\[
x_i(t) = x_i(0) + 2^{1/2} \sum_k Q_{ik} m_i^{-1/2} \sigma_k \cos(\omega_k t + \delta_k) \quad (2.10)
\]

where \( \sigma_k \) is an amplitude, \( \omega_k \) the angular frequency and \( \delta_k \) the phase of the \( k \)th normal
mode of motion. The phases and amplitudes depend upon the positions and velocities at time $t=0$. The thermal averages of the second moments $\sigma_k^2$ of the amplitude distributions can be calculated for both classical and quantum statistics:

$$
\sigma_{k,\text{class}}^2 = \frac{kT}{\omega_k^2}, \quad \sigma_{n_k,\text{qm}}^2 = \frac{\hbar}{4\pi \omega_k} \coth \frac{\hbar \omega_k}{4\pi kT}
$$

where $\hbar$ and $k$ are the Planck and Boltzmann constants. The two statistics coincide in the limits of low frequency or high temperature. For biomolecules, the most important difference is generally that higher frequency modes have little amplitude in classical statistics but have non-negligible zero-point motion in quantum statistics. Harmonic models thus provide one of the few practical ways for including quantum effects in biomolecular simulations.

Averages over the motion represented by normal modes are often carried out by Taylor-series expansions about $x_0$ for each normal mode [58], but for the present purposes, it is more convenient to randomly sample points from a thermal distribution:

$$
x^{(n)} = x_0 + \sum_k s_k^{(n)} M^{-1/2} Q_k
$$

where $s_k^{(n)}$ is a pseudo-random number drawn from a Gaussian distribution with mean zero and variance $\sigma_k^2$. A quantum chemistry calculation can then be done at each point $x^{(n)}$ and the results averaged (using Eq. 2.3) to yield the required averages over vibrational motion.

### 2.2.3 Path integral molecular dynamics

The equilibrium properties of a quantum system can also be approached by path-integral methods [70]. This approach exploits an isomorphism between the quantum Boltzmann distribution (which is what we are looking for) and a classical system with $P$ copies of each atom; in the classical system, the copies of each atom are connected by artificial spring forces, and the forces between atoms are reduced by a factor $1/P$. The classical system can then be simulated by conventional methods such as Monte
Carlo or molecular dynamics methods. As the number of copies grows, the partition function of the classical system approaches the quantum limit. We have recently implemented path-integral molecular dynamics (PIMD) methods into the Amber simulation programs [71, 72]. Each of the $P$ “beads” is simulated on a separate thread, and the springs are implemented via interprocess communication. A normal mode expansion in the bead coordinates is used to mitigate the stiffness of the dynamics between beads. Further details of the implementation are given by [72].

We carried out a PIMD simulation of fragment B3 of protein G (“GB3”) using the Amber ff99sb force field for the protein [73] and the q-SPCfw model for water [72]. (This is a preliminary simulation, since the water force field has been optimized for PIMD simulations but the protein force field has not.) After a classical equilibration, starting from the 1P7E pdb structure, the system was expanded to $P = 24$ beads, and 1 ns of normal-mode PIMD simulation was carried out with a time step of 0.5 fs. The temperature was regulated at 300 K with Nose-Hoover chains, and other details are as described by [72]. For this time period, the system was very stable, moving to structures with an RMS backbone deviation of about 0.7 Å from the starting structure, and maintaining all of the backbone secondary structure. Further analysis of this simulation will be given elsewhere.

Since we are primarily interested in local vibrational motion, we chose to analyze two residues, Phe30 and Phe52, that are in regular regions of the secondary structure of GB3: Phe30 is in the central helix, and Phe52 is part of the β-sheet; both have low backbone mobility as measured by $^{15}$N relaxation analysis [74]. In order to minimize the effects of overall tumbling, and of other slow internal motions that might be present, we sampled points over a short (32 fs) period, starting (arbitrarily) at 1.025 ns. The structures we used for analysis were each of the 24 beads at four (real) time points separated by 8 fs, for a total of 96 structures. Test calculations (not shown) indicated that going to longer time periods (out to 100 fs) did not significantly alter the spread of structures being sampled. The sampling period represents a compromise between having enough time (and beads) to sample local vibrational motions, without having contamination by slower processes such as overall rotational diffusion.
The 96 structures chosen in this fashion were then converted to model systems by extracting the Cα_{i−1}, C'_{i−1}, O_{i−1}, N_{i}, H_{i}, and Cα_{i} atoms, for i = 30 or 52, along with the corresponding atoms of two peptide groups to which Phe30 or Phe52 are hydrogen bonded. Hydrogens were added with standard geometries to the Cα atoms, so that the fragment system was (NMA)_{3} (see the right side of Fig. 2.1). These fragments were then subjected to the same analysis as described above for the normal mode analysis. That is, shielding calculations were carried out using B3LYP and the cc-pVTZ and cc-pVQZ basis sets, the results were extrapolated to the complete basis limit, and the results averaged using Eq. 2.3.

2.3 Results

The general behavior of shielding tensors for atoms in a peptide group is well-understood \[75\], and is illustrated in Fig. 2.2. For nitrogen, \(\sigma_{11}\) is roughly along the N–H bond, but displaced from it by about 20° as shown. The tensor is roughly axial about the \(\sigma_{11}\) direction, although rhombic components can be important for quantitative analysis. In our (NMA)_{3} calculations, the average shieldings in the principal axis frame were 0.6, 147.7 and 186.0 ppm, giving a rhombicity parameter \(\eta \equiv (\sigma_{22} - \sigma_{33}) / (\sigma_{11} - \sigma_{iso})\) of 0.42 and a value of \(\Delta\sigma \equiv \sigma_{11} - (\sigma_{22} + \sigma_{33})/2\) of -166.3 ppm. For isotropic motion, the effect of rhombicity is to increase \(\Delta\sigma\) by a factor of \((1 + \eta^2/3)^{1/2}\), which is an increase of 3%. For anisotropic motion, the results will be somewhat different, but quite high precision data would be required to detect the effects of non-axial symmetry in NMR relaxation.

The carbonyl carbon tensor, on the other hand, is very non-axial. Average values from the (NMA)_{3} model are -92.8, -17.7 and 80.3 ppm, so that any assumption of an axial tensor makes little sense. In our analysis, we will separate the total shielding tensor into two axial components as in Eq. 2.7. Overall CSA relaxation is then viewed as auto-relaxation of the two axial tensors, plus a cross-correlation between them.
2.3.1 \(^{15}\)N order parameters

For \(^{15}\)N relaxation, the principal results are given in Tables 2.1 and 2.2. The relative effects of vibrational averaging are shown in Table 2.1, where all of the correlation functions are normalized to unity at \(t = 0\). While there are some differences between the four models considered here, a number of general trends are evident, and likely to be reliable. First, the effects of librational motions of the N–H bond are significant, even when just zero-point vibrational motion is considered. This can be seen in the third column of Table 2.1, where values of \(\langle P_2[u(0) \cdot v(t)] \rangle\), which is the conventional \(S^2\) parameter in the Lipari-Szabo model free analysis, ranges from 0.85 to 0.90. In addition to this angular (or librational) dependence, fluctuations in the magnitude of the bond length have the effect of further reducing the effective dipolar coupling between the amide nitrogen and its attached proton. A more complete analysis of these zero-point motions, including estimates of anharmonic motions that are not included here, has been given earlier [59].

The dipolar order parameters shown here roughly match those measured for secondary structure residues in small proteins. For example, the \(S^2\) values fitted to \(^{15}\)N relaxation data for GB3 [74] are 0.861 and 0.827 for Phe30 and Phe52, respectively, if the N–H bond length is taken to be 1.01 Å, which is the actual equilibrium value from MP2 quantum chemistry calculations [59]. This indicates that the peptide groups of these and similar residues are actually quite rigidly held to the overall molecular frame, since a model that includes only zero-point librations (and no internal peptide group rotation) fits the observed data quite well. Since most N–H vectors in secondary structure regions in GB3 (and in many other small proteins) have relaxation rates close to those observed for Phe30 and Phe52 in GB3, it is reasonable to argue that only local vibrational motion is being seen in \(^{15}\)N relaxation measurements for these peptides. This view is in agreement with results from recent molecular dynamics simulations which also show peptide motions that can be interpreted in the same way [76, 77, 73].

A second key point from Table 2.1 is that the zero-point orientational averaging of
the CSA tensor is much less pronounced than for the N–H dipolar interaction: orientational $S^2$ values for the $\sigma_{11}$ direction of the nitrogen CSA tensor (third column of Table 2.1) range from 0.95 to 0.97, which are much closer to unity than are the corresponding dipolar averages. The shielding tensor is an electronic structure property, and its directionality appears to be more strongly influenced by the positions of the heavy atoms (and in particular, of the partial double bond between C' and N) than it is to the position of the amide proton. Since the zero-point vibrational amplitude of the light hydrogen atom is greater than that of the heavier carbon and nitrogen atoms, fluctuations in the direction of the CSA tensor are found to be much smaller than those of the N–H bond direction. This behavior is in contrast to the implicit assumption, found in most analyses of peptide relaxation, that the $S^2$ value for dipolar and CSA relaxation ought to be about the same. (The practical consequences of this error are mitigated by the fact the CSA tensor itself is often taken as a fitting parameter; we discuss this more below.)

As one might expect, the orientational averaging of the cross-correlated interaction between dipolar and CSA relaxation is intermediate between that of pure dipolar and pure CSA relaxation, as shown in the final four rows of Table 2.1. The cross-correlated values are always much closer to the CSA values than to the dipolar values.

All of these interactions involve averaging both of angles (the third column of Table 2.1) and magnitudes of the interactions. For dipolar interactions, it has been known for a long time that these fluctuations are almost uncorrelated, so that the final two columns are nearly equal [?]. This statistical independence is also very accurate for the CSA-related fluctuations, as can be seen in the Table. So, at least for these relatively small vibrational motions, one can consider the angle and magnitude fluctuations separately when analyzing CSA relaxation.

### 2.3.2 Effective CSA tensors

The vibrational averaging we are considering here occurs on a sub-picosecond time scale, and is invariably in the extreme narrowing regime for NMR relaxation. The exact time-dependence for relaxation then becomes irrelevant, and the effect of the
motion is to multiply the spectral density function expected for a rigid body by an order parameter, $S^2$. This order parameter can be, and often is, incorporated into an “effective” value of the bond length or shielding anisotropy. For dipolar interactions, we can write:

$$\left[ \frac{1}{r^6} \right]_{\text{eff}} \equiv \left\langle \frac{P_2[u(0) \cdot u(t)]}{r^3(0)r^3(t)} \right\rangle_{\text{vib}}$$

(2.13)

Here $u$ is a unit vector along the bond direction, and $t$ is a time long compared to a vibrational time scale. Hence, a rigid molecule with a bond length of $r_{\text{eff}} \equiv \left[ 1/r^6 \right]_{\text{eff}}^{-1/6}$ would have the same relaxation behavior as the actual system that includes the vibrational averaging. In a similar fashion, we can define an effective CSA tensor for pure CSA relaxation:

$$[\Delta \sigma]_{\text{eff}} \equiv \left[ (\Delta \sigma(0)\Delta \sigma(t)P_2[u(0) \cdot u(t)])_{\text{vib}} \right]^{1/2}$$

(2.14)

This state of affairs for dipolar relaxation has been recognized for many years [60, 59], and it is quite common for analyses of experimental relaxation behavior to approximately incorporate fast vibrational effects (including zero-point motion) into an effective bond length that is slightly longer than the actual equilibrium bond length. Estimates of these effective bond lengths can be derived from experiment [78], or from quantum chemistry calculations much like those considered here [59]. The two estimates are in quite good agreement, although both anharmonic effects and solvation corrections are required in a careful analysis. (For historical reasons, many analyses of $^{15}$N relaxation in the literature use an effective bond length for the amide N–H bond of 1.02 Å. This is intermediate between the equilibrium bond length, which is about 1.01 Å, and a value of 1.04 Å that reflects local vibrational motion [78, 59].)

Our analysis now allows us to estimate the effective CSA tensors as defined in Eq. 2.14; these are listed in Table 2.2. The key general point here is that the effective tensor differs from the average tensor (i.e. the average of the values computed for each snapshot) by only 1.2 to 2.5%. This is a consequence of the high order parameters seen in Table 2.1. (Remember that $[\Delta \sigma]_{\text{eff}}$ is squared in expressions for pure CSA relaxation.)
The actual values differ a lot in the four models we consider: NMA in vacuum has no hydrogen bond partners, and it is known that the effect of hydrogen bonding is to significantly increase $\Delta \sigma$. Hence we do not expect (or find) the NMA monomer results to be near to the tensors seen in proteins, which almost invariably have hydrogen bonding partners. The remaining three models have hydrogen bonding interactions, and consequently, more negative values of $\Delta \sigma$. The results from the PIMD simulation have the potential advantage of being extracted from a simulation of a real protein in water, but they have the disadvantage that the local geometries are determined by the molecular mechanics force field, which is known to have inaccuracies in both equilibrium geometries and in the extent of fluctuations about equilibrium [79]. The local geometries and fluctuations given by the DFT results on (NMA)$_3$ are probably the more accurate in this respect, but here the hydrogen bonds are energetically optimized in a way that might not be possible in the context of an entire polypeptide chain. Overall, this means that the differences of $\sim$10 ppm among the last three rows of Table 2.2 reflect real uncertainties in our models. One should also bear in mind that results for $\Delta \sigma$ in NMA vary by nearly 10 ppm depending on what density functional or correlation model is used, even at the CBS limit [80]. All of these factors limit the conclusions that can be drawn from the absolute CSA values (or effective values) reported here. Nevertheless, we expect our estimates of order parameters, which are a relative measure of motion, to be reliable.

It is worth noting that the individual tensors that go into the vibrational average have a lot of variability. Fig. 2.3 shows histograms for the CSA in the snapshots used for NMA$_3$ and for the GB3 simulation. The standard deviation about the mean is 27 ppm for NMA$_3$ and 23 ppm for residue 30 in GB3. This implies that small changes in the geometry can have a significant impact on the shielding tensor, and suggests that it may be very difficult to find a single structure that is representative of the ensemble. For example, for NMA and NMA$_3$ we know the equilibrium structure (about which the normal mode expansion was made). The $^{15}$N CSA for this average structure is -148.1 for NMA and -168.5 ppm for NMA$_3$; the latter is close to the vibrational average of -166.7 ppm (Table 2.2) but the former value is significantly different from the
vibrational average of -137.2 ppm. Hence, simply computing a shielding tensor for an optimized structure may not always be sufficient to represent even the local vibrational averaging that is taking place. Further investigation of this question is ongoing. Finally, one should note that the variations shown in Fig. 2.3 have only a minor effect on relaxation: values of \( \langle \omega_u(0) \cdot \omega_v(t) \rangle \) (fourth column of Table 2.1) vary between 0.96 and 0.98.

### 2.3.3 \(^{13}\text{C} \) relaxation

As mentioned above, CSA relaxation for the carbonyl carbon is more complex than for the amide nitrogen, since the CSA tensor is very rhombic. We can express this as a sum of two axial tensors (Eq. 2.7), and look at the effects of averaging on each tensor, as well as on their cross-correlations. These are given in Table 2.3. Again, there are some differences for the various models, but the \( S^2 \) values (given in the last two columns of the table) range from 0.92 to 0.98. Fluctuations in the magnitude of \( \sigma_1 = \sigma_{11} - \sigma_{33} \) are much smaller than those for \( \sigma_2 = \sigma_{22} - \sigma_{33} \), as seen in column 4 of Table 2.3. This is in accord with a model that \( \sigma_{22} \) varies more strongly with conformation than does either \( \sigma_{11} \) or \( \sigma_{33} \) [81, 82]. It should be noted, however, that the dependence of the carbonyl shielding tensor on environment in proteins is still under active study (see, e.g. [83]), and variations over vibrational motion need not necessarily reflect variations of the average from one residue to another.

As with the amide nitrogen tensor, we can express the effects of local vibrational averaging in terms of an effective CSA tensor, and these values are collected in Table 2.4. As one would expect from the high values of \( S^2 \) shown in Table 2.3, the difference between average and effective tensors is quite small; remember that tensor is scaled according to the square root of \( S^2 \), that is, by 1 to 3% for values of \( S^2 \) between 0.98 and 0.94, as found here. Hydrogen bonding is an important contributor to the carbonyl CSA tensor, especially for \( \Delta \sigma_2 \). In the snapshot we chose, Phe30 (in the central helix) has well-formed hydrogen bonds, but the carbonyl of the Phe52 peptide (actually in residue 51) has only a weak hydrogen bond (average H...O length of 2.1 Å), even though it is in a region of regular secondary structure. Hence, the Phe30 results are
close to that for the fully-hydrogen-bonded NMA$_3$ model, whereas those for Phe52 are close to those for NMA itself, which has no hydrogen bonds.

Recently, Jordan et al. [61] have carried out an analysis of motional effects on carbonyl carbon CSA values, using (classical) molecular dynamics simulations on calmodulin. They analyze internal fluctuations of 50 peptide planes (not including overall or local dynamical events involving the peptide plane as a whole,) and find a average value of $S^2$ of 0.93. This is comparable to the values found in Table 2.3, and the general conclusion is the same as ours: local distortions of the peptide group should have only a small (1-3%) effect on the effective CSA tensor for carbonyl relaxation, even though the shielding tensor is actually quite sensitive to such distortions. This sensitivity, however, means that considerable care must be taken in comparing quantum chemistry calculations to experiment: it may be necessary to explicitly carry out motional averaging in the quantum calculations in order to obtain reliable results.

The final four columns of Table 2.4 give information about the orientation of the carbonyl shielding tensor relative to the peptide bond direction. This information is especially relevant for the analysis of cross-correlated relaxation between the CSA tensor and the C’–N dipolar interaction. The angle $\beta_{11}$ is the effective angle between the $\sigma_{11}$ direction and the C’–N bond direction. It is an “effective” value in the sense that a completely rigid peptide with this angle would exhibit the same cross-correlated relaxation as the actual (vibrating) peptide group of our models. This angle is in the range of 35 to 39º, which is in excellent agreement with results from solid-state NMR and from liquid-state relaxation studies [75, 84, 82, 61].

### 2.4 Discussion

Although CSA relaxation effects become increasingly important at high magnetic fields, relatively little is known about how shielding tensors vary with molecular conformation, nor about how these variations affect NMR relaxation. Here we use modern methods of computational chemistry to address one piece of the puzzle: what are the likely effects of local vibrational motion on the $^{15}$N and $^{13}$C’ tensors in peptide groups?
Even this becomes a somewhat complex problem, since we need accurate calculations of electronic structures and reasonable models for nuclear motion that include zero-point vibrational effects. Uncertainties in how to best estimate these quantities leads to some variation in results, as seen in Tables 1 to 4. Nevertheless, some key conclusions can be drawn: fluctuations in the magnitude and directions of the $^{15}$N CSA tensor are quite small, with squared order parameters in the range 0.93 to 0.96; correspondingly, the effective shielding tensor (which would be used if vibrational averaging is ignored) is 2-4% smaller than the average tensor. The corrections for vibrational motion for CSA relaxation are much smaller than those needed for vibrational contributions to N–H dipolar relaxation. Squared order parameters for the carbonyl carbon CSA tensor are very similar (in the range of 0.92 to 0.95), with the $\sigma_{22}$ component (which points roughly along the C=O bond direction) showing larger fluctuations than $\sigma_{11}$ or $\sigma_{33}$.

Since the time scale for local vibrational motion is in the extreme narrowing limit for NMR relaxation, one can generally fold the effects of this motion into an effective tensor: the effective tensor shows the same relaxation behavior in a rigid system as the real tensor does for the vibrating system. Estimated values for these are given in Tables 2 and 4. These numbers can be compared to values extracted from various sorts of NMR experiments, but quantitative comparisons will require careful attention to the details of how the “experimental” tensors are determined. As an example, Kroenke et al. [85] estimated the mean value of the $^{15}$N chemical shift anisotropy in ribonuclease H to be $-172\pm13$ ppm, with a standard deviation in the site-to-site variation of 6 ppm. This analysis assumes an axially symmetric tensor, so is most directly comparable to $\Delta\sigma_{eff}(1 + \eta^2/3)^{1/2}$ in the nomenclature of the present paper. Furthermore, the data extraction assumes that the spectral density $J(\omega)$ is the same for dipolar and CSA relaxation. If we take into account the fact the the dipole-dipole squared order parameter ($S_{DD}^2$) is different from the CSA-CSA value ($S_{CSA}^2$), we arrive at the following relation (see the Appendix):

$$\Delta\sigma_{eff} \left[ 1 + \frac{\eta^2}{3} \right]^{1/2} = \Delta\sigma_{KRP} S_{DD}$$ (2.15)
Here $\Delta \sigma_{KRP}$ is the -172 ppm value extracted by Kroenke et al. [85] from their experimental data. Setting $S_{DD} \simeq 0.94$ and $\eta \simeq 0.42$, a value of -172 ppm for $\Delta \sigma_{KRP}$ corresponds to a value of $\Delta \sigma_{eff}$ of 157 ppm. This is close to the values given in Table 2.2 for hydrogen-bonded models (the final three columns). Hence the effect of ignoring the rhombicity of the $^{15}$N tensor (about a 3% reduction), and of assuming the same spectral density for dipolar and CSA relaxation (about a 6% reduction), combine to lower the fitted result by 15 ppm. This is roughly equal to the estimated standard deviation in the original paper. Hence, the elaborations given here are within the original statistical errors, but should become important as new measurements are made or as new types of data (such as from solid-state NMR) become available. We show in the Appendix that a similar re-interpretation can be made for other analyses of $^{15}$N CSA values in proteins [86, 87, 88]. Given the limitations of the models used here, it is not possible to provide a convincing estimate of the possible errors in our estimated value for $\Delta \sigma_{eff}$, but all of the models predict that local vibrational motion should affect N–H dipolar relaxation to a much greater extent that it affects CSA relaxation. Further calculations, and careful analysis of both liquid and solid-state data will be required to know for sure exactly how big the difference is in representative protein environments.

In the end, the question of the “best” value to use for effective shielding tensors has no unique answer, and is similar to the question of assigning an effective bond length in relaxation calculations. A parameter like $\Delta \sigma_{KRP}$ is probably the simplest way to summarize the observed field dependence of relaxation rates. As long as one is consistent, this value can be plugged into calculations that assume a fixed value for $\Delta \sigma$. On the other hand, to make comparisons to theoretical calculations, or to compare fits to experiment where different assumptions and parameters might be used, it is useful to be able to back out estimates of the CSA tensor itself, or its vibrationally averaged effective value, $\Delta \sigma_{eff}$ (Eq. 2.14). Differences between these various estimates amount to only a few percent, with larger effects for nitrogen than for the carbonyl carbon, but they can be important if we are to extract structural or dynamic information from quantitative analyses of CSA tensors in proteins.

It is important to note that the motional effects described here are restricted to
fast local vibrations. Peptide groups, particularly in loops and other mobile parts of a protein, can also undergo larger and slower internal motions, which we have not attempted to model here. For these larger motions, it is likely that the direction of the principal axes of the CSA tensor will follow the peptide plane, and the distinction between motion affecting dipolar relaxation and that affecting CSA relaxation will become less important. But these larger motions might also lead to fluctuations in the magnitude of $\Delta \sigma$ (as well as to its orientation); classical and path-integral dynamics simulations to investigate these larger motions are ongoing.
Table 2.1: $^{15}$N CSA and N–H dipolar averages
For each cell, the average is normalized by dividing by the same quantity but with $t = 0$. The final column gives the product of columns 3 and 4, that is $\langle \omega_u(0) \cdot \omega_v(t) \rangle \cdot \langle P_2[u(0) \cdot v(t)] \rangle$. For dipolar interactions, $\omega$ is $r^{-3}$, where $r$ is the N–H distance and $u$ or $v$ are unit vectors along the N-H bond direction. For the CSA interaction $\omega$ is $\Delta \sigma \equiv \sigma_{11} - (\sigma_{22} + \sigma_{33})/2$, and $u$ or $v$ are unit vectors along the $\sigma_{11}$ direction (see Fig. 2.2).

<table>
<thead>
<tr>
<th>interaction</th>
<th>model</th>
<th>$\langle P_2[u(0) \cdot v(t)] \rangle$</th>
<th>$\langle \omega_u(0) \cdot \omega_v(t) \rangle$</th>
<th>$\langle \omega_u(0) \cdot \omega_v(t) P_2[u(0) \cdot v(t)] \rangle$</th>
<th>$3 \times 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>dipolar</td>
<td>NMA</td>
<td>0.854</td>
<td>0.940</td>
<td>0.810</td>
<td>0.803</td>
</tr>
<tr>
<td></td>
<td>NMA$_3$</td>
<td>0.861</td>
<td>0.959</td>
<td>0.829</td>
<td>0.826</td>
</tr>
<tr>
<td></td>
<td>phe30</td>
<td>0.870</td>
<td>0.968</td>
<td>0.842</td>
<td>0.842</td>
</tr>
<tr>
<td></td>
<td>phe52</td>
<td>0.897</td>
<td>0.963</td>
<td>0.862</td>
<td>0.864</td>
</tr>
<tr>
<td>CSA</td>
<td>NMA</td>
<td>0.972</td>
<td>0.956</td>
<td>0.930</td>
<td>0.929</td>
</tr>
<tr>
<td></td>
<td>NMA$_3$</td>
<td>0.953</td>
<td>0.975</td>
<td>0.928</td>
<td>0.929</td>
</tr>
<tr>
<td></td>
<td>phe30</td>
<td>0.971</td>
<td>0.979</td>
<td>0.951</td>
<td>0.951</td>
</tr>
<tr>
<td></td>
<td>phe52</td>
<td>0.975</td>
<td>0.980</td>
<td>0.956</td>
<td>0.956</td>
</tr>
<tr>
<td>dipole-CSA</td>
<td>NMA</td>
<td>0.965</td>
<td>0.965</td>
<td>0.916</td>
<td>0.931</td>
</tr>
<tr>
<td></td>
<td>NMA$_3$</td>
<td>0.940</td>
<td>0.988</td>
<td>0.921</td>
<td>0.929</td>
</tr>
<tr>
<td></td>
<td>phe30</td>
<td>0.946</td>
<td>1.000</td>
<td>0.938</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>phe52</td>
<td>0.955</td>
<td>0.989</td>
<td>0.935</td>
<td>0.944</td>
</tr>
</tbody>
</table>

Table 2.2: Effective $^{15}$N CSA values
In the fourth column, $u$ is along the N–H bond and $v$ is along the $\sigma_{11}$ direction of the $^{15}$N CSA tensor. The angle $\beta$ is defined such that $P_2(\cos \beta) = \langle P_2[u(0) \cdot v(0)] \rangle$.

<table>
<thead>
<tr>
<th>model</th>
<th>$\Delta \sigma_{av}$</th>
<th>$\Delta \sigma_{eff}$</th>
<th>$\langle P_2[u(0) \cdot v(0)] \rangle$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMA</td>
<td>-137.2</td>
<td>-135.3</td>
<td>0.804</td>
<td>21.2°</td>
</tr>
<tr>
<td>NMA$_3$</td>
<td>-166.7</td>
<td>-162.6</td>
<td>0.797</td>
<td>21.6°</td>
</tr>
<tr>
<td>phe30</td>
<td>-160.3</td>
<td>-158.3</td>
<td>0.754</td>
<td>23.9°</td>
</tr>
<tr>
<td>phe52</td>
<td>-155.7</td>
<td>-153.7</td>
<td>0.779</td>
<td>22.6°</td>
</tr>
</tbody>
</table>
Table 2.3: $^{13}$C CSA and Cα–C dipolar averages

For each cell, the average is normalized by dividing by the same quantity but with $t = 0$. For CSA-11, both $u$ and $v$ point along the $\sigma_{11}$ direction and $\omega_u = \omega_v = \sigma_1$; for CSA-22, they both point along $\sigma_{22}$ and $\omega_u = \omega_v = \sigma_2$; for CSA-12, $u$ points along $\sigma_{11}$ and $v$ points along $\sigma_{22}$, with $\omega_u = \sigma_1$ and $\omega_v = \sigma_2$. The final column gives the product of columns 3 and 4, that is $\langle \omega_u(0) \cdot \omega_v(t) \rangle \cdot \langle P_2[u(0) \cdot v(t)] \rangle$.

Table 2.4: Effective $^{13}$C CSA values

In the fourth column, $u$ is along the N–C’ bond, $v_{11}$ is along the $\sigma_{11}$ direction and $v_{22}$ is along the $\sigma_{22}$ of the $^{13}$C CSA tensor. The angle $\beta$ is defined such that $P_2(\cos \beta) = \langle P_2[u(0) \cdot v(0)] \rangle$.

Figure 2.1: N-methylacetamide trimer

Left: DFT-optimized structure for normal mode calculations; right: one snapshot of the fragment constructed from Phe30 in the PIMD simulation.
Figure 2.2: Typical CSA tensor principal directions for peptides
The least-shielded component is denoted as $\sigma_{11}$, and the most-shielded component is $\sigma_{33}$. For each tensor, one of the components points in a direction very nearly perpendicular to the peptide plane.

Figure 2.3: Histogram of $\Delta\sigma$ values for NMA
Left: Phe30 in GB3; Right: Phe52 in GB3.
Chapter 3
Chemical Shift Anisotropy Calculations Using Small Fragment Models

The understanding of chemical shift anisotropy (CSA) of proteins is essential to deciphering structural and dynamics information from NMR measurements. However, the available experimental data on measured CSAs are still sparse and awaits more quantitative interpretations. Theoretical predictions of CSA serve to overcome this limitation by calculating arbitrary structures with systematic variations. In this study, the $^{13}\text{C}$ and $^{15}\text{N}$ CSA values were calculated for the entire GB3 protein using small fragment based DFT approaches. First, our calculations confirmed that both $^{13}\text{C}$ and $^{15}\text{N}$ CSA from the protein backbone are dependent on the secondary structure. Remarkable consistency with experimental observables was achieved when the protein structures were optimized using generalized born (GB) based implicit solvent model. Second, we examined the origin of $^{13}\text{C}$ and $^{15}\text{N}$ CSA variations. It was discovered that carbonyl $^{13}\text{C}$ CSAs are sensitive to backbone angles, and both $^{13}\text{C}$ and $^{15}\text{N}$ CSAs are highly dependent on their hydrogen bonding strength. These conclusions are expected to provide insight to CSA interpretation in NMR studies and improve the structural refinement procedure.

3.1 Introduction

NMR chemical shift is one of the most accurately measured spectroscopic parameters [89]. The accuracy of the chemical shift tensor (CST) measurement obtained from NMR studies significantly exceeds that of the nuclear overhauser effect (NOE), scalar coupling constants, or relaxation time. Chemical shift anisotropy (CSA) is the anisotropic component of chemical shift tensor, which makes it highly susceptible to the local
electronic and molecular environment [90, 91]. The amide nitrogen and carbon atoms lie in the peptide bond framework, which is a vital component of the protein structure. Previous studies showed that CSA tensors of $^{15}$N and $^{13}$C carry information on both the dynamics and the structures of the protein, which await further interpretation [9, 92]. Most of the NMR-resolved protein structures are determined based on scalar coupling and dipolar interactions. In comparison, the determination of chemical shift tensors is relatively difficult. In solution NMR, the direction and magnitude of chemical shift tensor cannot be directly measured due to isotropic tumbling. Solid state measurement requires site-specific labeling and multiple independent CSA measurements, which may be technically challenging. In addition, the interpretation of chemical shift tensors in proteins is particularly difficult due to their sensitivity to a large number environmental factors, including hydrogen bonding, neighboring contacts, solvent effect and long-range electrostatics [93]. As a result, most direct CSA measurements and interpretations have only been performed on small peptides [94, 95, 96, 97, 14].

Nevertheless, recent advances in solution and solid state NMR techniques have allowed CSA to be systematically measured for a few small, globular proteins, including GB1 and GB3 protein, binase and ubiquitin [25, 98, 99, 100, 101, 102]. For example, Fushman et al derived $^{15}$N chemical shift tensors of GB3 protein from $^{15}$N relaxation and CSA-dipolar cross correlation experiments [99, 102]. More recently, Rienstra et al directly measured the $^{13}$C and $^{15}$N CSA tensor components of GB1 protein via slow magic angle spinning (MAS) [25]. In comparison to the solution NMR results, the solid state investigations revealed more significant secondary structure dependency for both $^{13}$C and $^{15}$N CSA in the protein backbone, which shows prospects of CSA measurement as a means of improving accuracy in protein structure determination and refinement.

Over the past decade, advances in quantum mechanical methods have enabled CSA predictions for arbitrary model structures, which compensate for the limited dataset from experimental studies [103]. Some of these predictions have helped with NMR data refinement by establishing empirical relationships between structural features and isotropic chemical shifts [104, 105]. In addition, DFT based methods have
been used widely to evaluate chemical shielding tensors for small peptides [6, 97, 106, 107, 108]. While these studies have achieved considerable agreement with experimental measurements for short peptides, fewer systematic investigations have been performed on complete protein structures. Hence the precise effect of protein environment on chemical shift tensors remains elusive.

In the present study, we evaluated chemical shift tensors for every residue in the GB3 protein. The solvent effect was incorporated via minimization using generalized born (GB) implicit solvent model. Calculations of chemical shift tensor were performed on density functional theory (DFT) level. The $^{13}$C and $^{15}$N nuclei of interest were treated with locally dense basis set to enhance the accuracy of the shielding calculations [109, 110]. Our results deviate significantly from the solution NMR data, but show remarkable agreement with solid state measurements. More specifically, our study confirmed the strong dependency of CSA on protein secondary structure for both the carbonyl $^{13}$C and amide $^{15}$N nuclei. The correlation between chemical shift anisotropy (CSA) of proteins and their structural features, including dihedral angles, secondary structure and hydrogen bonding properties, were analyzed and elucidated.

3.2 Methods

3.2.1 Molecular dynamics simulations

Molecular dynamics simulations were performed on the NMR-resolved structure of GB3 (PDB code: 1P7E) using ff99SB force field from AMBER9 [111]. The protein was solvated in a 12.0 Å octahedral TIP4P-ew waterbox and minimized prior to MD. Briefly, the applied MD protocol included SHAKE constraints on bonds involving hydrogen, non-bonded cutoff of 8 Å, a time step of 2 fs, Particle Mesh Ewald (PME) for treatment of long-range electrostatics and an NVT ensemble with periodic boundary conditions. The system was first heated to 300 K in 60 ps with a 10.0 kcal/mol/Å² restraint on the protein, and then equilibrated with 1.0 kcal/mol/Å² restraint for 50 ps. The production run was performed with 0.1 kcal/mol/Å² restraint on the protein.
3.2.2 Energy minimization

The structures subjected to energy minimization included NMR and X-ray structures (PDB code: 1IGD, 1P7E, 2IGH, 2NMQ, 2OED). In order to understand the effect of molecular dynamics on CSA fluctuation, two MD snapshots were also extracted from the 160 ps and 203.660 ns of the MD trajectory of 1P7E. Their backbone RMSDs are 0.4 Å and 1.27 Å with respect to the 1P7E crystal structure. Two methods of minimization were tested. In the first method, the systems were minimized with 10 kcal/mol/Å² restraint on the protein, followed by a full minimization to an RMS gradient of $1.0 \times 10^{-4}$ kcal/mol/Å² by AMBER in vacuum. In the second method, the aqueous environment was approximated by the generalized born (GB) approach implemented in NAB [112]. The GB solvent model implicitly averages over solvent degrees of freedom and provides reliable estimates of solvation energies. The structures were minimized with LBFGS-TNCG preconditioned conjugate gradient algorithm to an energy gradient of $1.0 \times 10^{-7}$ kcal/mol/Å, followed by 3 iterations of Newton-Raphson minimization [113, 114].

3.2.3 Chemical shift tensor calculations

After minimization of the experimental structures and MD snapshots, GIAO calculations of the chemical shielding tensor were performed using the B3LYP functional [30, 32]. The locally dense basis set (cc-PVTZ) was used to treat the backbone atoms of the central fragment, while the rest of the system was treated with a 6-31G* basis set. Different model systems were used to represent the protein fragment (Figure 3.1). The first system included a single NMA fragment. The second model included a NMA fragment extracted along with its backbone hydrogen bonding partners of the amide group. Hydrogen bonds formed with water molecules or sidechains were excluded. The third model included all residues within the 2.5 Å sphere centered at the amide nitrogen atom of interest. The fourth model was based on the third model, but the hydrogen bonding partners of the amide group were removed. The fifth model included all components of the second model, plus any water molecules as the direct hydrogen...
3.2.4 Calculation of hydrogen bonding strength

The secondary structure of proteins is formally defined by their hydrogen bonding properties. The DSSP algorithm is a standard method used to assign secondary structure via geometric recognition of different hydrogen bonding patterns [115]. According to the DSSP method, the hydrogen bond strength is defined as the electrostatic interactions between the hydrogen bonding amide and carbonyl groups:

\[
E = q_1 q_2 \left( \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right) \times f
\]

where \( f \) is the dimensional factor (\( f = 332 \)), \( r \) is the distance between specified atoms, and \( q_1 \) and \( q_2 \) are predefined electron charges (\( q_1 = 0.42e \), \( q_2 = 0.2e \)). By this definition, a hydrogen bond is considered to have formed if \( E < 0.5 \text{ kcal/mol} \). The hydrogen bond strength of carbonyl and amide groups from GB3 will be estimated using the DSSP approach.

3.3 Results and discussions

3.3.1 Direct comparison with experimental data

Fushman et al extracted the \(^{15}\text{N}\) chemical shift tensors of GB3 protein from relaxation data using solution NMR techniques [99]. Their results showed a low to moderate dependency on secondary structure for the \(^{15}\text{N}\) nuclei from the backbone amide group. The anisotropic difference between the helical and sheet region of GB3 is determined to be 3.9 ppm for CSA, and 2.6 ppm for reduced CSA. In their solid state experiment, Rienstra et al performed direct measurement of CSA tensor components via slow magic angle spinning (MAS) on GB1 protein [25]. In comparison to Fushman’s study, Rienstra’s study revealed remarkable secondary structure dependence of CSAs for both the \(^{13}\text{C}\) and \(^{15}\text{N}\) nuclei. The helix-sheet difference they determined amounts to 11 ppm for CSA and 9 ppm for reduced CSA. Both studies showed linear correlations.
between chemical shift tensors and isotropic shift.

Hence the first goal of our study is to calculate the CSAs and evaluate their correlation to experimental results obtained from both solution and solid state investigations. GB1 (PDB code: 2QMT) and GB3 (PDB code: 1IGD) are both B domains of G protein. They exhibit exactly the same folds (one $\alpha$-helix and four $\beta$-sheets), with an RMSD of 0.30 Å for the protein backbone (Figure 3.2). Their sequences are comprised of 17 types of amino acids and differ by only 7 residues (Figure 3.3). While some differences may exist between their CSA values, similar helix-sheet dependence in CSA values is expected. Due to the sequence and structural similarity between GB1 and GB3, we chose GB3 as our model of study. The calculated CSAs were directly compared with both sets of experimental data described above.

The calculated trends of CSAs are much closer to the solid state measurements (Figure 3.4) than the data from solution NMR. This finding agrees with previous studies, which suggested that chemical shift tensor values are more likely to agree with solid state experiments rather than solution phase data [116]. While solid state experiments and quantum mechanical calculations can directly determine tensor orientation and magnitude, solution NMR measurements relies on CSA-dipolar cross correlation experiment, which may not effectively separate the effects of tensor orientation from tensor magnitude. As a result, the solid state data from Rienstra et al. will be used as the primary source for comparison with our calculations in the studies discussed below [25].

The calculated $^{13}$C reduced CSA values (83.8 ppm) for the $\alpha$-helix are consistent with experimental observables (83.5 ppm). But the calculated CSA of $\beta$-sheet is lower than experimental data by 2 ppm on average. Overall, the trend along the GB3 sequence was reproduced, and the $\alpha$-helical region can be easily identified (Figure 3.5a). The calculated range of $^{13}$C reduced CSA is 5.9 ppm, which is comparable to the 5 ppm range from Rienstra’s study.

In the case of $^{15}$N nuclei, the calculated CSA values are systematically lower than the experimental values by 14 ppm on average for the overall sequence (Figure 3.5b). Nevertheless, the mean CSA of the $\alpha$-helix is 9.5 ppm lower than that of the $\beta$-sheet,
which agrees with the experimental helix-sheet difference (9 ppm). The calculated spread of $^{15}$N CSA is 24 ppm, whereas the solid state and solution phase data give 28 ppm and 86 ppm range, respectively. The residues with larger than 10 ppm deviation from the experimental values are also examined. All these residues show either strong hydrogen bonding or hydrophobic interactions with neighboring sidechains. Since no residue sidechains were included in our system, large deviations were anticipated in these cases. The origin of the secondary structure dependence for both $^{13}$C and $^{15}$N nuclei was analyzed and discussed below.

### 3.3.2 Isotropy-anisotropy correlation

Previous studies showed that not all three principal components of CSA are correlated with the isotropic shift ($\delta_{\text{iso}}$) [25, 101]. More specifically, only the $\delta_{yy}$ component of $^{13}$C CSA exhibits dependence on $\delta_{\text{iso}}$, while both $\delta_{xx}$ and $\delta_{zz}$ component of $^{15}$N CSA are correlated to $\delta_{\text{iso}}$. The exact same correlations were observed from our predictions (Figure 3.6). For $^{13}$C nuclei, the $\delta_{xx}$ and $\delta_{zz}$ components of CSA are independent of the isotropic shift, resulting in very low $R^2$ values (Table 3.1) (0.08 and 0.007) from the linear least-square fit, giving averages of $96.89 \pm 2.36$ ppm and $264.49 \pm 3.84$ ppm, respectively. The $\delta_{yy}$ component of $^{13}$C CSA exhibits linear increase with respect to $\delta_{\text{iso}}$ at a slope of 2.72, which agrees well with the experimental slope of 2.82. For $^{15}$N nuclei, the $\delta_{yy}$ component of CSA shows poor linear dependence on $\delta_{\text{iso}}$ ($R^2 = 0.003$), yielding a mean of $100.97 \pm 5.63$ ppm. Both $\delta_{zz}$ and $\delta_{xx}$ components of $^{15}$N CSA increase with $\delta_{\text{iso}}$, giving slopes of 1.29 and 1.64, which are comparable to the experimental slopes of 1.24 and 1.04.

### 3.3.3 Effect of minimization

Since the structures acquired from PDB are obtained via alternative methods (X-ray or NMR) under different experimental conditions, energy minimization of protein structures is crucial to achieving uniform structures required for CSA calculation (Figure 3.7). Chemical shift tensors were obtained for GB3 protein based on a single PDB
structure (1IGD), where the NMA₃ fragments were extracted. The secondary structure dependence for ¹³C CSA manifests only after the structure was minimized with implicit solvent (GB) (Figure 3.8). Upon minimization, the reduced CSA of ¹³C for the α-helix (residues 23 - 37) became lower than the rest of the sequence. In addition, there is as much as a 10 ppm change in CSA on an individual residue after GB-based minimization.

To understand how the quality of experimental structures affects CSA, a comparison was made between the two structures acquired by X-ray crystallography and NMR. The crystallographic structure (1IGD) was obtained at 1.1 Å resolution without hydrogen atom coordinates, whereas the NMR structure (1P7E) was based on 1IGD and refined with residual dipolar coupling [117, 118]. The difference between their reduced CSAs is as large as 7 ppm (Figure 3.8a). While the 1P7E-based CSA values can be used to identify the helical structure, the 1IGD-based values appears random and uncorrelated to secondary structure prior to structural optimization. After minimization, the CSA values for 1IGD and 1P7E converge to a mean absolute deviation of 1 ppm, and both showed significant secondary structure dependence (Figure 3.8b). Therefore the isotropy-anisotropy dependence can be used to determine whether the corresponding protein structure has been properly and sufficiently optimized.

Hence the same minimization protocol was applied to five experimental GB3 structures, as well as the two snapshots taken at 160 ps and 203.66 ns of an MD simulation of the solvated 1P7E structure. These snapshots were chosen to compare the effect of the length of MD on the structures and their associated CSAs. The RMSD differences of these structures are summarized in Table 3.2. The experimental and MD-based GB3 structures differ significantly from the 1IGD structure, with RMSD varying from 0.7 Å to 2.4 Å for the backbone, and 0.60 Å to 3.7 Å for all atoms. The GB-based minimization affects the GB3 structures to different extents, as the all-atom RMSD varies from 1.0 Å to 3.2 Å when their original structures were referenced.

The reduced CSA values were calculated for these seven GB3 structures (Table 3.3). Consistency in CSA was observed for both ¹³C and ¹⁵N nuclei. The largest difference in ¹³C CSA exists between 1IGD and 2NMQ structures, which amount to 2.26 ppm
in the α-helical region and 0.99 ppm in the β-sheet regions. For $^{15}$N nuclei, the most significant discrepancy is 2.13 ppm for the β-helix residues (1IGD vs. 2NMQ) and 2.71 ppm for the β-sheet residues (2OED vs. 2NMQ). There is no obvious correlation between the deviation in CSAs and the RMSDs between minimized and unminimized structures. Therefore, structural minimization using Generalized Born approach leads to convergent structures and uniform CSA values, regardless of the quality of the starting structure.

3.3.4 Effect of system size

Previous studies showed that CSA is dependent on non-bonded interactions [6]. To examine the effect of the model system size, the CSA values were evaluated for the standard fragment (up to 24 atoms) and the system comprised of all residues within the 2.5 Å sphere of the amide nitrogen (up to 90 atoms) (Table 3.4 and Figure 3.9). The increased system size adds more noise to the CSA profile for both the $^{13}$C and $^{15}$N nuclei. The CSA comparison between the two model compounds gives a mean absolute difference of 3.6 ppm/residue ($^{13}$C) and 8.1 ppm/residue ($^{15}$N), respectively. The expansion of system size also leads to an increase in the helix-sheet CSA difference for $^{13}$C (from 2.3 ppm to 5.2 ppm), but a decrease in $^{15}$N (from 9.5 to 6.3 ppm). It remains unclear whether the inclusion of three to four times more atoms guarantees more accurate CSA values. Thus we conclude that the NMA$_3$ fragment is sufficient for our present study.

3.3.5 Effect of solvent hydrogen bonding

In order to understand the effect of water hydrogen bonding, CSA calculations were performed for model systems including directly hydrogen bonded water molecules (Figure 3.10). A cutoff distance of 3.5 Å was employed as the hydrogen bonding criteria between the donor and acceptor atoms. The hydrogen bonding geometry was derived from the minimized MD snapshot of the 203.066 ns, where the GB3 protein was solvated in a truncated octahedral waterbox. Inclusion of solvent water does not
significantly change the CSA profile of $^{13}\text{C}$ or $^{15}\text{N}$. Nevertheless, the helix-sheet difference of $^{13}\text{C}$ and $^{15}\text{N}$ are reduced by 0.4 ppm and 2.5 ppm, respectively. Hence there is no significant advantage for the inclusion of immediate hydrogen bonding solvent molecules.

3.3.6 Effect of intramolecular hydrogen bond inclusion

The effect of intramolecular hydrogen bonding on CSA values was investigated (Figure 3.11). For the $^{13}\text{C}$ nuclei, both the average CSAs for helix and sheet structures are increased by only 1 ppm. The $^{13}\text{C}$ reduced CSA values are not significantly affected by the removal of hydrogen bonds on both sides of the central NMA fragment. The small effect of hydrogen bonding observed on $^{13}\text{C}$ chemical shift tensors is corroborated by the experimental studies of ubiquitin [98]. However, the removal of hydrogen bonding partners affected $^{15}\text{N}$ CSAs considerably, reducing the helix-sheet CSA difference from 6 ppm to 1 ppm. This sensitivity of $^{15}\text{N}$ CSA tensor to hydrogen bonding is supported by both experimental and theoretical studies [91, 93]. Therefore, the inclusion of intramolecular hydrogen bonds is critical to accurate $^{15}\text{N}$ CSA calculations but not to $^{13}\text{C}$ CSA.

3.3.7 Effect of backbone dihedral angle $\omega$

Energy minimization with GB solvent model improves the quality of protein structures and accuracy of $^{13}\text{C}$ CSA values as described above. In order to understand the effect of minimization on individual geometric parameters, the backbone conformations prior to and after optimization were compared. Neither the $\phi$ or $\psi$ angle, nor the $\omega$ angle of the coil and $\beta$-sheet region were significantly affected (Table 3.5). However, the mean $\omega$ angle of the helical region was lowered by 4.5° on average upon minimization (Figure 3.12). As a result, the standard deviation of $\omega$ angle was increased considerably from 2.9° to 6.3°. The latter value is consistent with former statistical analysis based on high-quality structure databases (6.2° from Priestle et al and 5.5° from Thornton et al) [119, 120]. Both structures were then examined by PROCHECK, which
suggested that the $\omega$ angles are overconstrained in the unminimized structure [121]. Previous studies also showed that in X-ray or NMR based structural refinement, the $\omega$ angles are often tightly constrained to planarity, which may lower the quality of resulting structures [122]. The extent of peptide planarity was compared for GB structures subjected to varying minimization constraints from 0 kcal/mol/Å to 500 kcal/mol/Å (Figure 3.12b). The $\omega$ angles from the $\alpha$-helical component of GB3 deviate further from planarity when the system is progressively relaxed. Taken together, the GB-based minimization gave rise to more accurate structures by relaxing the tight $\omega$-angles restraints.

The range of $\beta$-sheet angles ($\sigma = 8.1^\circ$) is much wider compared to that of the $\alpha$-helix ($\sigma = 2.3^\circ$). Consequently, this implies that $\omega$ angles are less variable in $\alpha$-helix, while exhibiting significant fluctuations in the $\beta$-sheet. This finding is consistent with the moderate correlation between the helical propensity and $\omega$ angle observed in previous statistical analyses [120]. A similar trend was observed for $^{13}$C CSA: the range for $\alpha$-helix is much smaller (4.01 ppm) compared to that of $\beta$-sheet (7.24 ppm), which is substantiated by solid state NMR experiments [116]. Taken together, lower $\omega$ angles of $\alpha$-helix residues may lead to smaller $^{13}$C CSAs in the $\alpha$-helical region of GB3 overall. Thus $^{13}$C CSA is more dependent on through-bond ($C'-N-CA-C'$) effects represented by the $\omega$ angle. In comparison, $^{15}$N CSA is not significantly affected by $\omega$ angle changes.

In order to understand the direct correlation between reduced CSA and $\omega$ angle values, chemical shift tensors were evaluated as a function of $\omega$ angle (Figure 3.13). The single NMA structure was extracted from Phe30 of GB3 structure (1IGD), and fully optimized using the locally dense basis set previously described. Partial optimizations were then performed with the $\omega$ angle varied from $160^\circ$ to $200^\circ$ at $0.5^\circ$ intervals. The corresponding chemical shift tensors were calculated for each of the optimized structures. The reduced CSAs for $^{13}$C and $^{15}$N nuclei reach maximum when the $\omega$ angle is $179.5^\circ$ and $179.0^\circ$, respectively. As the peptide bond deviates further from planarity, both the $^{13}$C and $^{15}$N CSAs decrease. The ranges of $^{13}$C and $^{15}$N CSA are 2.8 and 1.4 ppm, respectively, over a $20^\circ$ $\omega$ angle change centered at the maximum CSA. Overall,
$^{15}\text{N}$ shows a very small change, as it only varies by 0.5 ppm in the range of $179.5 \pm 10^\circ$. Our calculations show that while $^{13}\text{C}$ chemical shift tensor responds to $\omega$ angle variation, $^{15}\text{N}$ chemical shift tensor is not as sensitive to the $\omega$ angle change.

### 3.3.8 Effect of hydrogen bond geometry

#### 3.3.8.1 Hydrogen bond distance and angles

Hydrogen bonding distances were defined as N-H...O and C=O...H distances, where the N-H and C=O groups belong to the central NMA fragment. Hydrogen bonding angles were defined as N-CO...H and C-N-H...O dihedrals. Former experimental and theoretical studies showed strong correlation between H(N) CSA and hydrogen bonding distance [6, 123]. A moderate correlation was observed between hydrogen bonding dihedral angle C-N-H...O and the $^{15}\text{N}$ CSA ($R^2 = 0.37$). The C-N-H...O angle is localized at $-96.0 \pm 10^\circ$ for the helical component (excluding residue 23 as an outlier), but varies from $+97.2^\circ$ to $-73.1^\circ$ for the $\beta$-sheet region. Similarly, the N-C-O...H angle is clustered at $36.04 \pm 11.8^\circ$ for the $\alpha$-helical structure but diverges from $+165.5$ to $-161.7$ degrees for the $\beta$-sheet region. However, no similar correlation was derived between hydrogen bonding distance and $^{13}\text{C}$ or $^{15}\text{N}$ CSA. The variation of hydrogen bonding distance appears to be insensitive to protein secondary structure. Our results suggest that GB3 CSA values are more dependent on the hydrogen bond dihedrals than the hydrogen bond distances (Figure 3.14).

#### 3.3.8.2 Hydrogen bond pattern

The calculated CSA values exhibit periodicity in both $\beta$-sheet and $\alpha$-helix (Figure 3.16). In the $\beta$-sheet region, reduced CSA increases and decreases alternatingly along the sequence. In the $\alpha$-helical region, the pattern is repeated every three or four residues. As a standard $\alpha$-helix contains 3.6 residues per turn, those periodicity and secondary structure dependence are likely results of the hydrogen bonding pattern, as indicated by earlier studies of amide $^1\text{H}$ chemical shift [124]. The alternating patterns in $\beta$-sheets are likely caused by alternating hydrogen bond partners of the amide backbone. Apart
from the α-helical region, double-sided hydrogen bonding patterns also occurred in
the C and N terminals, where large deviations with respect to experimental data oc-
cur. The inherent conformational flexibility and solvent exposure may have added
complexity to the accurate experimental measurement in the terminal regions.

3.3.8.3 Hydrogen bond strength

The hydrogen bonding strength was calculated for the NH...O bond and C=O...H
bond for the entire GB3 backbone (Figure 3.15). There appears to be a strong sec-
ondary structure dependence on the hydrogen bonding strength. When the C=O...H
hydrogen bond strength (measured in kcal/mol) was shifted by an arbitrary unit of 80
kcal/mol, significant overlap was observed between the hydrogen bonding strength
and $^{13}$C CSA profiles, especially for the two β-sheets (residue 13-19, residue 42-46).
Overall, the average hydrogen bonding strength from C=O...H bond is stronger in
the α-helical region (-2.26 kcal/mol) compared to the β-sheet regions (-1.58 kcal/mol),
which agrees with previous theoretical studies predicting a difference of 0.45 kcal/mol
[125]. A similar correlation was observed between the NH...O hydrogen bonding en-
ergy and $^{15}$N reduced CSA.

3.3.9 Limitations and scope

One limitation to our study was the exclusion of solvent and sidechain interactions.
Nevertheless, our system adequately describes the secondary structure dependence of
CSA, whereas explicit treatment of additional interactions did not yield better results
in comparison. In order to capture the effect of internal motion, it may be interesting to
sample different protein conformations and solvent configurations via molecular dy-
namics simulations in the future. A few previous studies indicated linear correlation
between hydrogen bond length and the magnitude of chemical shift tensors, which
is not observed by our calculations or the solid state NMR data. It is likely that this
direct correlation is obscured by the heterogeneity of amino acid residues. In order to
test or confirm these correlations, a larger sampling size containing the same residue
type is necessary. Nevertheless, our methodology is generalizable and can be applied to globular proteins other than GB3.

3.4 Conclusions

In this study, the theoretical calculations of CSAs were extended from peptide to proteins, which are more biologically relevant and thus more desirable as targets of NMR studies. The quantitative predictions of protein backbone CSA values proved to be feasible using contemporary QM methods. Two main observations were made in terms of structural dependency of CSAs:

First, $^{13}$C CSA is highly dependent on the backbone $\omega$ angle. Our results also suggest that even with the high-resolution experimental structures, the $\omega$ angles are likely to be inaccurate due to over-constrained refinement procedures. This opens up the possibility that experimentally acquired CSAs can be adopted to enhance backbone refinement for the corresponding proteins. Meanwhile, theoretically predicted CSA values can be used to examine the quality of refined protein structures. Second, both $^{13}$C and $^{15}$N CSA are highly correlated to the hydrogen bonding strength in the $\beta$-sheet regions.

In terms of improving the approaches to CSA predictions, the incorporation of solvent effect using GB-based implicit solvent model is shown to significantly increase the consistency and accuracy of CSA calculations. In comparison, the inclusion of larger fragments or explicit water molecules does not appear to make a significant difference. Thus it is hoped that better representations of the environment, such as implicit solvent models, may improve the quality of the predictions.
Table 3.1: Correlation between chemical shift tensors and isotropic shifts for $^{13}$C and $^{15}$N nuclei

<table>
<thead>
<tr>
<th>Nuclei</th>
<th>Plot</th>
<th>LS fit, exp</th>
<th>R</th>
<th>LS fit, calc</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^{13}$C</td>
<td>$\delta_{xx} \text{ vs. } \delta_{iso}$</td>
<td>$y = 197.980 + 0.245 x$</td>
<td>0.27</td>
<td>$y = 303.540 - 0.219 x$</td>
<td>0.08</td>
</tr>
<tr>
<td>$^{13}$C</td>
<td>$\delta_{yy} \text{ vs. } \delta_{iso}$</td>
<td>$y = -304.910 + 2.815 x$</td>
<td>0.91</td>
<td>$y = -318.710 + 2.761 x$</td>
<td>0.59</td>
</tr>
<tr>
<td>$^{13}$C</td>
<td>$\delta_{zz} \text{ vs. } \delta_{iso}$</td>
<td>$y = 106.910 - 0.0598 x$</td>
<td>-0.07</td>
<td>$y = 15.148 + 0.458 x$</td>
<td>0.28</td>
</tr>
<tr>
<td>$^{15}$N</td>
<td>$\delta_{xx} \text{ vs. } \delta_{iso}$</td>
<td>$y = -69.637 + 1.041 x$</td>
<td>0.63</td>
<td>$y = -110.340 + 1.287 x$</td>
<td>0.81</td>
</tr>
<tr>
<td>$^{15}$N</td>
<td>$\delta_{yy} \text{ vs. } \delta_{iso}$</td>
<td>$y = -10.542 + 0.715 x$</td>
<td>0.47</td>
<td>$y = 92.106 + 0.733 x$</td>
<td>0.06</td>
</tr>
<tr>
<td>$^{15}$N</td>
<td>$\delta_{zz} \text{ vs. } \delta_{iso}$</td>
<td>$y = 80.515 + 1.241 x$</td>
<td>0.84</td>
<td>$y = 18.237 + 1.640 x$</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 3.2: RMSD of GB3 structures
(RMSDs were calculated by fitting all structures to residue 3-54 of GB-minimized 1IGD structure. The RMSD measurements are in Å.)

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>ref: minimized</th>
<th>1IGD</th>
<th>ref: X-ray</th>
<th>1IGD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Backbone</td>
<td>All atom</td>
<td>Backbone</td>
<td>All atom</td>
</tr>
<tr>
<td>1IGD</td>
<td>0.66</td>
<td>1.08</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1P7E</td>
<td>0.76</td>
<td>1.13</td>
<td>0.32</td>
<td>0.37</td>
</tr>
<tr>
<td>2IGH</td>
<td>2.38</td>
<td>3.63</td>
<td>2.14</td>
<td>3.53</td>
</tr>
<tr>
<td>2NMQ</td>
<td>0.85</td>
<td>2.61</td>
<td>0.56</td>
<td>2.54</td>
</tr>
<tr>
<td>2OED</td>
<td>0.75</td>
<td>1.07</td>
<td>0.32</td>
<td>0.71</td>
</tr>
<tr>
<td>MD of 1P7E, 160ps</td>
<td>0.72</td>
<td>1.13</td>
<td>0.48</td>
<td>0.92</td>
</tr>
<tr>
<td>MD of 1P7E, 203.66ps</td>
<td>1.64</td>
<td>2.28</td>
<td>1.51</td>
<td>2.33</td>
</tr>
<tr>
<td>1IGD_gbmin</td>
<td>0</td>
<td>0</td>
<td>0.66</td>
<td>1.08</td>
</tr>
<tr>
<td>1P7E_gbmin</td>
<td>0.11</td>
<td>0.90</td>
<td>0.67</td>
<td>1.22</td>
</tr>
<tr>
<td>2IGH_gbmin</td>
<td>2.30</td>
<td>3.65</td>
<td>2.23</td>
<td>3.66</td>
</tr>
<tr>
<td>2NMQ_gbmin</td>
<td>1.80</td>
<td>3.32</td>
<td>1.61</td>
<td>3.25</td>
</tr>
<tr>
<td>2OED_gbmin</td>
<td>0.14</td>
<td>0.61</td>
<td>0.65</td>
<td>1.03</td>
</tr>
<tr>
<td>MD of 1P7E, 160ps</td>
<td>0.28</td>
<td>0.81</td>
<td>0.76</td>
<td>1.16</td>
</tr>
<tr>
<td>MD of 1P7E, 203.66ps</td>
<td>1.52</td>
<td>2.13</td>
<td>1.38</td>
<td>2.09</td>
</tr>
</tbody>
</table>
Table 3.3: Reduced CSA for helix and sheet of GB3 structures
(Only CSAs available from Rienstra et al were included in calculating averages)

<table>
<thead>
<tr>
<th>Origin</th>
<th>Helix, $^{13}$C</th>
<th>Sheet, $^{13}$C</th>
<th>Helix, $^{15}$N</th>
<th>Sheet, $^{15}$N</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp data</td>
<td>-83.50 ± 2.04</td>
<td>-77.7 ± 2.12</td>
<td>114.94 ± 2.58</td>
<td>106.95 ± 2.50</td>
</tr>
<tr>
<td>1IGD</td>
<td>-83.83 ± 1.19</td>
<td>-81.87 ± 1.62</td>
<td>103.98 ± 1.93</td>
<td>93.46 ± 3.52</td>
</tr>
<tr>
<td>1P7E</td>
<td>-83.02 ± 1.31</td>
<td>-81.79 ± 1.31</td>
<td>104.49 ± 1.66</td>
<td>93.94 ± 3.04</td>
</tr>
<tr>
<td>2IGH</td>
<td>-83.28 ± 0.84</td>
<td>-81.37 ± 2.47</td>
<td>105.05 ± 1.98</td>
<td>93.57 ± 3.67</td>
</tr>
<tr>
<td>2OED</td>
<td>-82.87 ± 0.86</td>
<td>-81.75 ± 1.28</td>
<td>104.41 ± 1.76</td>
<td>93.98 ± 3.07</td>
</tr>
<tr>
<td>2NMQ</td>
<td>-81.55 ± 1.80</td>
<td>-80.88 ± 1.65</td>
<td>104.17 ± 2.21</td>
<td>91.27 ± 3.40</td>
</tr>
<tr>
<td>MD of 1P7E, 160 ps</td>
<td>-82.90 ± 1.31</td>
<td>-81.79 ± 2.18</td>
<td>103.81 ± 2.60</td>
<td>93.30 ± 3.86</td>
</tr>
<tr>
<td>MD of 1P7E, 203.66ns</td>
<td>-83.19 ± 0.96</td>
<td>-81.84 ± 2.82</td>
<td>104.68 ± 1.23</td>
<td>94.97 ± 2.52</td>
</tr>
</tbody>
</table>

Table 3.4: Comparison of $\alpha$-helix and $\beta$-sheet CSA tensors
The CSAs are measured in ppm.

<table>
<thead>
<tr>
<th>Angle</th>
<th>Before Mean</th>
<th>Minimization Stdev</th>
<th>After Mean</th>
<th>Minimization Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$ angle - coil</td>
<td>230.0</td>
<td>70.3</td>
<td>229.8</td>
<td>69.0</td>
</tr>
<tr>
<td>$\phi$ angle - sheet</td>
<td>232.6</td>
<td>24.9</td>
<td>232.0</td>
<td>25.7</td>
</tr>
<tr>
<td>$\phi$ angle - helix</td>
<td>294.5</td>
<td>9.25</td>
<td>293.0</td>
<td>10.0</td>
</tr>
<tr>
<td>$\psi$ angle - coil</td>
<td>181.7</td>
<td>113.6</td>
<td>180.3</td>
<td>117.4</td>
</tr>
<tr>
<td>$\psi$ angle - sheet</td>
<td>141.3</td>
<td>25.9</td>
<td>142.1</td>
<td>28.7</td>
</tr>
<tr>
<td>$\psi$ angle - helix</td>
<td>298.0</td>
<td>77.3</td>
<td>299.7</td>
<td>77.0</td>
</tr>
<tr>
<td>$\omega$ angle - coil</td>
<td>180.6</td>
<td>1.6</td>
<td>178.5</td>
<td>6.4</td>
</tr>
<tr>
<td>$\omega$ angle - sheet</td>
<td>177.7</td>
<td>3.7</td>
<td>176.8</td>
<td>7.8</td>
</tr>
<tr>
<td>$\omega$ angle - helix</td>
<td>179.6</td>
<td>1.2</td>
<td>175.2</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Table 3.5: The effect of GB based minimization on backbone dihedrals
The angles are measured in degrees.
Figure 3.1: Model systems for NMR calculations
Five different pure QM models were employed to calculated the chemical shift tensors, including: a) single NMA b) NMA$_3$ with direct hydrogen bonding partners c) includes all residues within the 2.5 Å radius of amide nitrogen. d) includes all residues in 2.5 Å radius but without hydrogen bonding partners. e) include water as direct hydrogen bonding partner. All geometries are based on GB3 structure (1P7E) centered at Thr 25.

Figure 3.2: Overlapped secondary structure of GB1 and GB3
The backbone structure overlaps with an RMS of 1.1 Å. The 1P7E and 1IGD structures were visualized in VMD.
Figure 3.3: Sequence alignment and secondary structure for GB1 vs. GB3 protein
The aligned sequences of GB1 and GB3 proteins shows that their sequences only differ by 7 residues. The upper and lower sequences are taken from PDB structures of 1PGA and 1P7E.

Figure 3.4: Comparison of $^{15}\text{N}$ variations between experimental & calculated values
Figure 3.5: Comparison of CSA variations between experimental & calculated values
The figure on the left hand side shows the CSA variations of the $^{13}\text{C}$ CSAs, whereas the figure on the right shows that of the $^{15}\text{N}$ CSAs. The red lines indicate the experimental CSAs from Rienstra et al, and the black lines are from our quantum mechanical predictions.

Figure 3.6: Correlations between CST principal components and $\delta_{iso}$
- a) $^{13}\text{C}$ CST, experimentally measured.
- b) $^{15}\text{N}$ CST, experimentally measured.
- c) $^{13}\text{C}$ CST, calculated.
- d) $^{15}\text{N}$ CST, calculated.
Figure 3.7: Effect of minimization on CSA and CSA-isotropy correlation

Figure 3.8: Convergence of $^{13}$C CSA upon minimization
a) $^{13}$C reduced CSA prior to minimization for GB3 structures (1IGD & 1P7E). b) $^{13}$C reduced CSA after minimization for GB3 structures (1IGD & 1P7E).
Figure 3.9: Effect of system size
a) $^{13}$C CSA of minimized structures. b) $^{15}$N CSA of minimized structures. The black curves are derived from CSA calculations using (NMA)$_3$ as the small fragment model. The red curves are derived from CSA calculations where the fragment model includes all residues within the 2.5 Å radius of central amide nitrogen.

Figure 3.10: Effect of water inclusion
a) $^{13}$C CSA of minimized structures b) $^{15}$N CSA of minimized structures
Figure 3.11: Effect of hydrogen bond
a) $^{13}$C CSA of minimized structures b) $^{15}$N CSA of minimized structures

Figure 3.12: Effect of minimization-$\omega$ angle
a) $^{13}$C CSA of minimized structures b) $^{15}$N CSA of minimized structures

Figure 3.13: Effect of $\omega$ angle
a) $^{13}$C CSA of minimized structures b) $^{15}$N CSA of minimized structures
Figure 3.14: Effect of hydrogen bond geometry
a) $^{13}$C CSA of minimized structures b) $^{15}$N CSA of minimized structures

Figure 3.15: Effect of hydrogen bond strength
a) Plot of the C-O...H hydrogen bonding strength (red) and the $^{13}$C CSA of GB3 residues (black). The hydrogen bonding energy was shifted by -80 kcal/mol for comparison with $^{13}$C CSA magnitude.
b) Plot of the N-H...O hydrogen bonding strength (red) and the $^{15}$N CSA of GB3 residues (black). The hydrogen bonding energy was shifted by +90 kcal/mol for comparison with $^{15}$N CSA magnitude.
Figure 3.16: Effect of hydrogen bonding pattern

a) Experimental $^{13}$C CSA with hydrogen bonding pattern. b) Experimental $^{15}$N CSA with hydrogen bonding pattern. c) Calculated $^{13}$C CSA with hydrogen bonding pattern. d) Calculated $^{15}$N CSA with hydrogen bonding pattern. The hydrogen pattern is determined according to the (NMA)$_3$ model.
Chapter 4

Chemical shift tensor calculations using AF-QM/MM Model

Individual peptide groups in proteins must exhibit some variation in the chemical shift anisotropy (CSA) of their constituent atoms, but not much is known about the extent or origins of this dispersion. Direct spectroscopic measurement of CSA remains technically challenging, and theoretical methods can help to overcome these limitations by predicting CSAs for arbitrary structures. An automated fragmentation quantum mechanics/molecular mechanics (AF-QM/MM) approach has recently been developed and shown promising results in chemical shift calculations. Here we present the calculated chemical shift tensors for ubiquitin and GB1/GB3 proteins using both the AF-QM/MM approach. The AF-QM/MM approach appears to give better results compared to the NMA$_3$ model (described in Chapter 3), largely due to a more correct treatment of solvent effects and long-range electrostatics, but still has significant uncertainties. Correlation coefficients between solid-state experimental and calculated CSAs were 0.7 and 0.8 for $^{13}$C and $^{15}$N nuclei, respectively. Both $^{13}$C and $^{15}$N CSAs depend on the local secondary structure, consistent with solid state NMR studies.

4.1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy provides valuable information for the three dimensional structures of macromolecules including proteins and DNA [126]. Chemical shifts can be easily measured and used as restraints in protein structure determination and refinement. Chemical shift anisotropy (CSA) can also be an excellent indicator of the local electronic and molecular environment [6, 90, 91], but much less is known about its connection to structure, especially for N, C and H atoms in the peptide group itself. [9, 92, 6]. The interpretation of chemical shift tensors in proteins is
expected to be complicated by their susceptibility to a large number of environmental factors, such as the conformation of neighboring residues, hydrogen bonding and long-range electrostatics [6, 93].

In solution NMR, the direction and magnitude of chemical shift tensors cannot be directly measured due to isotropic tumbling. Instead, anisotropies can be estimated from relaxation and CSA-dipolar cross-correlation experiments at multiple spectrometer fields [19, 127], or from shifts in peaks upon partial alignment [128]. Solid state NMR experiments provide more direct information, but often require site-specific labeling and multiple independent CSA measurements, which can be technically challenging. As a result, most direct CSA measurements and interpretations have been performed on small peptides [94, 95, 129, 96, 97, 14]. Nevertheless, recent advances in solution and solid state NMR techniques have allowed CSA to be systematically measured for a few small, globular proteins, including GB1 and GB3 protein, binase and ubiquitin [98, 99, 101, 100, 23, 24, 25]. These investigations of CSAs revealed a significant dependence of the CSAs of amide nuclei on the protein backbone conformation, which shows prospects of using additional information from CSA for the purpose of protein structure determination and refinement.

Theoretical predictions enable CSA calculations for experimental and model structures, which compensate for the limited dataset from experimental studies and help elucidate the structural information implicated in experimental measurements [103]. Over the past decade, quantum mechanical calculations have facilitated NMR structure refinement by establishing empirical relationships between structural features and isotropic chemical shifts [104, 105]. Density functional theory (DFT) based methods have been used extensively to characterize chemical shielding tensors for di- and tri-peptide species [97, 106, 107, 6, 108]. More recently, Czinki and coworkers mapped the $^{15}$N and $^{13}$C CSA surface using L-Ala-NH$_2$ as a model for peptides and proteins [36]. Cai and Fushman also calculated the $^{15}$N chemical shift tensors of the selected residues in GB3 protein using a variety of peptide models [130, 37]. While these recent studies provide some understanding to the influence of protein geometry on chemical shift tensors, the effects of the complete protein environment remains to be assessed.
In the present study, we adapted and extended the automated fragment-quantum mechanics/molecular mechanics (AF-QM/MM) model developed by He and coworkers [131]. In this model, the central protein fragment is treated with quantum mechanics and the rest of the protein and solvent environment are represented by point charges. It was originally applied to Trp Cage mini-protein to predict proton isotropic shieldings and achieved considerable agreement with experimental measurements [131]. We report here the CSA tensors of $^{15}$N, $^{13}$C and $^1$H nuclei for GB1 and GB3 proteins using a number of different experimental structures, where the quality of the CSA predictions are evaluated and the environmental effects are reassessed.

4.2 Methods

4.2.1 Energy minimizations

A number of high-resolution protein structures solved by NMR spectroscopy or X-ray crystallography were adopted for our study of GB1 and GB3 proteins. As discussed below, some “regularization” of the experimental structures via molecular mechanics based energy minimization appears to improve the results. The structures subjected to energy minimization included NMR and X-ray structures for both GB1 (PDB codes 2QMT and 1PGA) and GB3 (PDB codes 1IGD, 1P7E, and 2OED) proteins [117, 118, 132, 133]. During energy minimization, the aqueous environment was approximated by the Hawkins, Cramer and Truhlar (HCT) form of pairwise generalized Born (GB) model approach implemented in the Nucleic Acid Builder (NAB) program [134, 112]. The HCT form of GB solvent model averages over solvent degrees of freedom and provides reliable estimates of solvation energies. To analyze the effect of minimization on protein structure and chemical shifts, both partial minimization and full minimization was performed for GB1 protein. Only partial minimization was performed for GB3 protein. For partial minimization, each of the structures were relaxed for 10 steps using the PRCG Polak-Ribiere conjugate gradient algorithm. This mainly serves to bring bond lengths and bond angles close to the “ideal” values specified in
the force field, while changing little else in the structure. As shown in Table 4.1, the backbone change is 0.1 to 0.2 Å. For full minimization, each of the structures was relaxed with the same algorithm, until a gradient RMS of $10^{-6}$ was reached; these structures differ significantly (backbone RMSD of 0.6 to 0.8 Å) from their starting structures.

4.2.2 Chemical shift tensor predictions

After minimization of the experimental structures, Gauge Independent Atomic Orbital (GIAO) calculations of the chemical shielding tensor were performed using the B3LYP functional from density functional theory (DFT)[32, 30, 135]. Two quantum mechanical program, Gaussian03 and Demon2k were used to calculate the chemical shift tensors and their results were compared [135, 136]. The AF-QM/MM model takes the entire protein and solvent effects into consideration, where the protein was partitioned into the core and buffer region to be treated with quantum mechanics, whereas the rest of the protein and solvent effects are represented by partial atomic charges from the AMBER94 force field[45]. The core region includes the CA-N-CA segment of the backbone of the $n^{th}$ amino acid, together with directly attached sidechains. Within the core region, the central fragment CA-C(O)-N(H)-CA was treated with a locally dense basis set, whereas the rest of the core region and buffer region were treated with a smaller basis set [137]. In Gaussian and Demon calculations, the locally dense basis set scheme are cc-pvTZ/6-31G** and iglo-iii/dzvp, respectively. The buffer region is defined to include residues described by the following criteria:

i) The $(n-2)^{th}$, $(n-1)^{th}$, $(n+1)^{th}$ and $(n+2)^{th}$ residues in the protein.

ii) The residues within 4 Å of the core region, where at least one of the contacting atoms is non-hydrogen.

iii) The residues within 3 Å of the core region, where both contacting atoms are hydrogens.

iv) The residues within 5 Å of the core region, where the contacting atoms include a heavy atom from an aromatic ring from the buffer region, and any atom from the core region.
The rest of the protein environment is represented by partial atom charges defined in the AMBER94 force field[45]. In addition, the solvent effect is approximated by Poisson Boltzmann based grid charges (Figure 4.1).

In order to generate these surface charges, the self consistent reaction field was calculated iteratively by solving the Poisson Boltzmann (PB) equation iteratively using divcon program [138]:

$$\nabla \cdot [\epsilon(r) \nabla \phi(r)] - \kappa^2(r)\phi(r) = -4\pi\rho(r)$$

where $\kappa$ is the modified debye-Huckel parameter reflecting the salt concentration and temperature. $\epsilon(r)$ is the dielectric constant distribution in space, which were set to 1.0 and 80.0 for solute and solvent, respectively. $\phi(r)$ is the electrostatic potential to be calculated. $\rho(r)$ is the solute charge distribution, described by CM2 point charges calculated using PM3 Hamiltonian. According to PB theory, the solute produces an electrostatic field in the solute region and solvent region, including the solvent reaction field and the Coulumbic field. The Coulumbic field is produced by the atomic charges within the solute boundary as a uniform medium. The reaction field is generated by the polarization in the solvent at the position of the solute. The PB equation was solved for each SCF cycle, until the solute wave function is consistent with the solvent reaction field [139, 140]. After mapping the solvent accessible surface onto a grid with 1.8 grid/Å spacing, the dielectric boundary is defined by assigning the grid points farther than 1.4 Å(water probe radius) away from the solvent accessible surface to the interior of the molecule [141]. At the end of the PB calculation, the converged electrostatic potential was then used to calculate the induced polarization charge at each of the grid points:

$$q^{\text{ind}}(i,j,k) = -q^{\text{real}}(i,j,k) + \frac{3\hbar}{2\pi} (\phi(0) - \frac{1}{6} \sum_{l=1}^{6} \phi(l))$$

where $q^{\text{ind}}(i,j,k)$ and $q^{\text{real}}(i,j,k)$ are the induced and real charge on the (i,j,k) grid cube, respectively. $\phi(0)$ and $\phi(l)$ are the potential at the (i,j,k) grid points and $\phi(l)$ is the potential at neighboring grid points.
4.2.3 Deming regression

In most cases, a simple linear regression of calculated values vs. experiment is sufficient to characterize the level of agreement. Such a comparison assumes that the uncertainties in the measured values are smaller than the errors arising in the model (e.g. from the use of a single structure, from a density functional calculation with a limited basis set, or from the approximate treatment of solvation effects.) For experimental CSA estimates, however, this may not be true, and significant uncertainties exist in the measured values, both from noise in the measurement itself and from limitations in the models used to extract CSA estimates from the raw data. The Deming regression method can be used to test the functional relationship between the calculated CSAs and the experimental observables, where uncertainties in both are taken into account [142]. The measurement errors of the calculated CSAs are assumed to be proportional to the magnitude of the CSAs as follows:

\[
\frac{s_{calc}}{\Delta \delta_{calc}} = \frac{s_{exp}}{\Delta \delta_{exp}} \tag{4.1}
\]

where \(s_{calc}\) and \(s_{exp}\) are the errors from calculation and experimental measurement, respectively, and \(\Delta \delta_{calc}\) and \(\Delta \delta_{exp}\) are the calculated and measured reduced CSA.

According to the Deming regression model, the loss function \(Q\) is described as:

\[
Q = \sum (x_i - u_i)^2 / k_i + (y_i - \beta_0 - \beta_1 u_i)^2 / \lambda_i \tag{4.2}
\]

where \(k_i\) and \(\lambda_i\) are the variance of \(x_i\) and \(y_i\), respectively, and \(u_i\) is the observed value of \(x_i\). The maximum likelihood estimate of the intercept \(\beta_0\) and the slope \(\beta_1\) were obtained from consecutive minimization of \(Q\) over \(u_i\) and \(\hat{\beta}_1\):

\[
\hat{u}_i = w_i [\lambda_i x_i + k_i \beta_1 (y_i - \alpha)] \tag{4.3}
\]

\[
Q_m(\alpha, \beta) = \min_{u_i} Q = \sum w_i [(y_i - \beta_0 - \beta_1 x_i)^2] \tag{4.4}
\]
where \( w_i = \frac{1}{(\lambda_i + \beta_i^2 x_i)} \).

4.3 Results and discussion

GB1 and GB3 proteins are the B1 and B3 domains of immunoglobulin G, respectively. GB1 (PDB code: 2QMT, 1PGA) and GB3 (PDB code: 1IGD, 1P7E, 2OED) protein exhibit exactly the same fold, including one \( \alpha \)-helix and four antiparallel \( \beta \)-sheets (Figure 4.2). Their sequences are comprised of 17 types of amino acids and differ by only 7 residues from each other (Figure 4.4). These two proteins have been used as model systems for the study of chemical shift tensors in various NMR experiments mentioned above [99, 25, 23]. The earlier solution NMR study of Hall and Fushman showed low to moderate dependency on the secondary structure for the \( ^{15}N \) nuclei [99]. In comparison, the latter two studies showed significant secondary structure dependence of CSAs for both the \( ^{13}C \) and \( ^{15}N \) nuclei [25, 24].

4.3.1 Comparison between AF-QM/MM model and NMA\(_3\) model

In agreement with solid state measurements, the general secondary structure dependency of \( ^{13}C \) and \( ^{15}N \) CSA of the GB3 protein was reproduced using both the NMA\(_3\) model and the AF-QM/MM model. Yet more importantly, we need to evaluate the accuracy of these two models on predicting CSAs and isotropic chemical shifts for individual residues. First, we examined the correlation between the experimentally determined CSA and calculated CSA. The NMA\(_3\) model gave relatively low quality prediction for both the \( ^{13}C \) and \( ^{15}N \) nuclei, with correlation coefficient of 0.39 and 0.40, respectively. The quality of CSA prediction was significantly improved when the AF-QM/MM model was applied. The correlation with solid state data increased to 0.63 for \( ^{13}C \) CSA and to 0.82 for \( ^{15}N \) CSA. Second, we compared the experimental isotropic chemical shift to the theoretical predictions (Table 4.2). NMA\(_3\) model failed to estimate the isotropic chemical shifts effectively, giving correlations of 0.13 and 0.31 for the \( ^{13}C \) and \( ^{15}N \) nuclei. In contrast, high correlations were observed between the experimental chemical shieldings and calculated chemical shieldings from AF-QM/MM model,
which equal 0.83 and 0.79 for the $^{13}\text{C}$ and $^{15}\text{N}$ nuclei, respectively. Hence AF-QM/MM approach is a significant improvement from the NMA$_3$ model in terms of predicting individual chemical shift tensors and isotropic chemical shifts (Figure 4.3).

4.3.2 $^{13}\text{C}$ and $^{15}\text{N}$ CSAs in GB1 protein

The predictions of reduced CSAs of GB1 protein were based on two different PDB structures (PDB code: 1PGA and 2QMT). We evaluated the reduced CSAs of the $^{13}\text{C}$ and $^{15}\text{N}$ nuclei and compared them with the experimental CSAs measured by Rienstra and co-workers (Figure 4.5).

For the $^{15}\text{N}$ nuclei, the calculated CSAs are systematically lower than that of the experimental observables (Table 4.3). This systematic difference amounts to -13 ppm from Demon predictions and -6 ppm from Gaussian predictions. The helix and sheet dependence of reduced CSAs were also examined: For partially minimized structures, the average $^{15}\text{N}$ CSAs of the $\alpha$-helical region are calculated to be 5.03 - 6.19 ppm higher than that of the $\beta$-sheet region, which is slightly lower than that of the experimental data (7.31 ppm). Full minimized structures give rise to larger helix-sheet CSA differences, and smaller standard deviations with respect to the mean CSAs. In the case of Gaussian-calculated 2QMT CSAs, the difference of 7.39 ppm is directly comparable with the experimental derived difference. The standard deviation of the reduced CSAs of the $\alpha$-helix and $\beta$-sheet are determined to be in the range of 0.91-1.47 and 0.76-0.92 respectively, which are similar to the variations found in experimental CSAs (1.27 and 0.66, respectively). The Pearson correlations between the measured and calculated reduced CSAs are between 0.58 and 0.69, which are slightly higher than the correlations derived for GB3 proteins (Table 4.4). Interestingly, the $^{15}\text{N}$ CSAs derived from fully minimized structures do not show a significant improvement in the strength of correlations. In terms of Deming regression, the slopes of all regression models are close to unity, with the full minimized structures giving slightly larger slopes. The Deming intercepts suggest a wide range of shift with respect to experimental values. Intercept aside, there likely is an excellent one-to-one correspondence between individual $^{15}\text{N}$ predictions and experimental measurements. While the minimization protocol affects
the helix-sheet difference of $^{15}$N CSA, its effect on correlations for individual CSA and magnitude of Deming slopes are rather ambiguous. Hence the extent of minimization is deemed to have a small effect on $^{15}$N CSAs.

For the $^{13}$C nuclei, the experimental CSAs are systematically higher than that of Demon predictions by around 13 ppm, but systematically lower than that of the Gaussian predictions by 6 ppm (Table 4.3). For partially minimized structures, the helix-sheet difference between reduced CSAs were evaluated to be between -2.49 and -1.79 ppm. In comparison, the full minimized structures give a trend much closer to the experiment measurements (-4.81 ppm), in the range of -2.67 to -4.67 ppm. It is worth noting that the Gaussian-calculated 1PGA CSAs are the closest to experimentally determined helix-sheet CSA differences, as the rest of the predictions gives differences at least 1.33 ppm lower than expected. The Pearson correlation between predicted and experimental values are significantly boosted by minimization as well, giving correlations between 0.71 and 0.75, in comparison to correlations between 0.58 and 0.64 from partial minimization (Table 4.4). The Deming regression models result in slopes in the range of 0.61 - 0.80 and large intercepts between -10.90 and -33.65. Given the significantly underestimated helix-sheet difference of CSAs and relatively low correlation due to partial minimization, it is likely that the quality of $^{13}$C CSA predictions are considerably affected by the extent of minimization. The effect of minimization on structures will be further explored in section 4.3.7.

### 4.3.3 $^{15}$N CSA in GB3 protein

Three different PDB structures (PDB code: 1IGD, 1P7E and 2OED) were employed to calculate the chemical shift tensors of GB3 protein. The resulting anisotropies are compared with the experimental data in Table 4.5. For $^{15}$N nuclei, the calculated CSA values from both models are systematically smaller in absolute magnitude than the experimental values by 10 - 20 ppm. The difference between the averages of $^{15}$N CSA for the $\alpha$-helical versus the $\beta$-sheet region is between 9.9 and 10.2 ppm, depending on the PDB structure and method used. These calculated differences are much closer to the measurements of Yao and coworkers (9.9 ppm) than that of Hall and coworker
(3.9 ppm). The calculated spread of $^{15}$N CSA is 33.2 ppm, which is comparable to the spread of the measurements of Yao et al (39.4 ppm). The study of Hall and Fushman on the other hand, shows significantly larger variations in comparison, giving a spread of 129 ppm. Hence the calculated regular $^{15}$N CSAs are consistent with the study of Yao and co-workers in terms of both helix-sheet dependence and magnitude of the CSA spread.

Furthermore, we examined the accuracy of the individual regular CSAs with respect to the measurements of Yao and co-workers (Table 4.6). The highest Pearson correlation (0.61) was given by the Gaussian calculation of the 1P7E structure, whereas the lowest correlation (0.53) was given by the Demon calculation of the 2OED structure. We also performed Deming regression on the calculated CSAs against the experimental CSAs. The estimated slopes are given in a range of 0.80 - 0.88, with a standard deviation between 0.17 and 0.23. The largest slope again is given by the Gaussian-based calculation of the 1P7E structure (0.88 ± 0.17).

The RMS deviation of the protein backbone were calculated and examined in the context of CSA values (Table 4.1). No obvious correlation was observed between CSA values and the calculated RMSD or experimental structure resolution. These observations suggest that our minimization protocol is robust and will generate CSA predictions of consistent quality, regardless of the resolution and origin of the starting structure.

4.3.4 $^1$H chemical shift tensors of GB3 protein

Yao and co-workers measured the regular $^1$H chemical shift anisotropy with cross-correlation relaxation experiments in liquid crystalline state [23]. Two models were used to produce fitted CSA values: The three-parameter model uses three cross-correlated relaxation parameters to determine the CSAs, while assuming tensor symmetry relative to the peptide plane. In comparison, the five parameter model takes advantage of the precise RCSA measurements and does not require the assumption of symmetry. Yao and co-workers also performed density functional theory (DFT) based calculations on NMA-acetamide models, where only the residues with backbone hydrogen
bonds were accounted for. In addition, they derived empirical models trained on the results of DFT calculations for $\delta_{yy}$ and $\delta_{zz}$. Hence we seek to assess the quality of our predictions compared to the experimental data from liquid crystalline measurements. We are also interested in how the AF-QM/MM model perform comparing to the NMA-acetamide model and the derived empirical model mentioned above.

We examined the difference in regular $^1$H CSA between $\alpha$-helical and $\beta$-sheet regions (Table 4.7). On average, the $^1$H CSA of the $\alpha$-helical region is lower than that of the $\beta$-sheet region, by 1.7-1.8 ppm. In comparison, the difference predicted by the AF-QM/MM model is slightly larger, which is in the range of 2.00 - 2.82 ppm. Correlations between the experimental and calculated CSAs were also examined, where we discovered that the correlations are close to 0.6 when solvent-exposed residues are excluded. The Pearson correlation and functional relationship between the calculated CSAs and fitted CSAs from the two models are similar, which suggest that the predicted CSAs does not significantly favor one model over the other.

Then we compared our calculated CSAs with the theoretical predictions of the NMA-acetamide model applied by Yao and co-workers (Table 4.8). We found that the Pearson correlation for the NMA-acetamide model are lower (0.34 - 0.43) than that of AF-QM/MM model (0.51 - 0.61). A more significant difference was observed in terms of functional relationship: When regressing the calculated data against the fitted values from the five-parameter model, the Deming regression model results in a slope of $1.37 \pm 0.68$ (PDB structure: 2OED) with an intercept of $-2.56 \pm 6.26$. In comparison, the calculated CSAs from the NMA-acetamide give a considerable smaller slope of $0.1 \pm 0.22$ with an intercept of $11.56 \pm 2.3$, suggesting that the functional relationship is nonexistent. We also compared the results from AF-QM/MM model with the empirical models of Yao and co-workers, which shows that our calculations are highly consistent with the empirical predictions ($R = 0.90$), whereas the NMA-acetamide model only gives moderately similar values ($R = 0.67$) (Table 4.10). Therefore the AF-QM/MM model is significantly more effective in predicting $^1$H CSAs qualitatively than small peptide QM models.
4.3.5 Relations of isotropic shifts and shielding tensors

The three principal components of the chemical shifts tensors were also examined individually (Table 4.9, Figure 4.8). The best correlations between experimental and calculated CSAs is given by $\delta_{zz}$ ($R = 0.76$), followed by $\delta_{yy}$ ($R = 0.69$) and $\delta_{xz}$ ($R = 0.59$), which are very close to the correlations derived from the empirical model with hydrogen bond geometric parameters. However, it is worth noting that all these comparisons of $^1$H chemical shift tensor and anisotropy were performed with exclusion of the solvent exposed residues. The predictions for $^{13}$C and $^{15}$N nuclei on the other hand, are not as sensitive to solvent exposure. Thus if further improvements were to be made on the AF-QM/MM model for predictions of $^1$H CSAs in particular, a more careful treatment of the solvent effect is required.

In previous experimental studies, correlations between individual tensor components and isotropic chemical shifts were observed. Hence we would like to see whether these properties can be predicted from our calculations. For GB1 protein, the $^{13}$C chemical shift tensor calculated from the 2QMT structure was examined and compared with the results of Wylie and co-workers (Figure 4.9). The experimental data suggest that $\delta_{yy}$ is the only principal component significantly correlated to the isotropic shift ($R = 0.84$), giving a regression slope of 2.82. This is very similar to our calculated results, where the Pearson correlation is given as 0.81 with a regression slope of 2.18. For GB3 protein, the $^{15}$N and $^1$H chemical shift tensors were obtained from the 1IGD structure and compared with the experimental measurements of Yao and co-workers (Figure 4.10). Our calculated results suggest that both $\delta_{zz}$ and $\delta_{xx}$ of the $^{15}$N tensors are highly correlated to the isotropic shift, giving regression slopes of 1.29 ($R = 0.89$) and 0.92 ($R = 0.78$), respectively. The data obtained by Yao and co-workers leads to a least squares linear regression model with regression slopes of 1.03 ($R = 0.73$) and 0.63 ($R = 0.45$). In comparison, the experimental results from Wylie and co-workers give regression slopes of 1.24 ($R = 0.84$) and 1.04 ($R = 0.63$). Hence the strength of correlations and regression model derived from our calculations, while qualitatively comparable to both datasets, are closer to the solid state measurements for GB1 protein. For $^1$H tensors,
both $\delta_{zz}$ and $\delta_{yy}$ are highly correlated to the isotropic shift according to the GB1 protein data, giving regression slopes of 1.79 ($R = 0.89$) and 1.22 ($R = 0.82$), which are in agreement with the calculated slopes of 1.58 ($R = 0.89$) and 1.14 ($R = 0.85$).

The magnitude of the $^{13}\text{C}$, $^{15}\text{N}$ and $^1\text{H}$ tensor components were also compared with experimental data (Table 4.9). It is noted that for all three nuclei, the tensor components most correlated to the isotropic chemical shift are the most consistent with the experimental measured counterparts (Figure 4.8, 4.9, 4.10). For example, the $\delta_{zz}$ component of $^{15}\text{N}$ tensor is most correlated to its corresponding isotropic shift, and the Pearson correlation with respect to GB1 data is as larger as 0.82. Similarly, the $\delta_{yy}$ component of $^{13}\text{C}$ tensor and $\delta_{zz}$ components of $^1\text{H}$ tensor are the most accurately predicted. For the purpose of comparison, the individual correlations from fully-minimized 1PGA structures are also included in the Table 4.9, it is noted that the effect of minimization is ambiguous when it comes to individual tensor components.

4.3.6 Predictions of ubiquitin CSAs

Both the AF-QM/MM method and the NMA$_3$ method confirmed the secondary structure dependence of CSAs in GB1 and GB3 proteins. In principle, these general methods can be used to predict the trends in other globular proteins. Therefore we selected ubiquitin as a second model for the prediction of CSAs (Table 4.11). The backbone chemical shift tensors of ubiquitin have been extensively studied in solution NMR experiments [98, 19, 143, 144, 145]. However, direct measurements of these chemical shift tensors in solid state experiments are yet to become available. Thus our study on ubiquitin will serve as predictions to chemical shift tensors to be measured in solid state. In addition, it is of great interest to compare the solution phase chemical shift tensors to the predicted values. In terms of isotropic chemical shift, the predictions show relatively good agreement with respect to experimental measurements [146]. More specifically, the correlations between calculated and experimental shifts are 0.82 and 0.83 for $^{13}\text{C}$ and $^{15}\text{N}$ nuclei, respectively (Figure 4.7). In terms of CSAs, The AF-QM/MM method correctly predict that $^{13}\text{C}$ CSAs in the $\alpha$-helix regions (residue 22-32,
residue 51-55) is lower than that of the β-sheet regions by 2.5 ppm, which is very similar to the results from the solution NMR experiments, giving differences in the range of 1.7 - 3.8 ppm. In the case of $^{15}$N nuclei, the experimental measurements by Loth and coworkers from cross-correlation relaxation showed that sheet CSA is higher by 3.6 - 4.9 ppm than the helix CSA [101]. This is in agreement with our predictions, which gives a helix-sheet difference of 2.5 ppm (AF-QM/MM) and 4.8 ppm (NMA$_3$). Since systematic measurements of chemical shift tensors are yet to be performed on the ubiquitin in solid state experiments, our predictions may serve as a starting point for the understanding of the structure-dependent behavior of the ubiquitin chemical shift tensors.

4.3.7 Computational improvements on CSA predictions

In previous quantum mechanical studies of chemical shielding tensors, the proteins are often fragmented and modeled as NMA or small peptides in vacuum. While these models provide qualitative understanding to the impact of electronic environment on chemical shieldings, a lot of important information in the protein structures are often neglected. The AF-QM/MM model is by far the most complete model in terms of inclusion of environmental effects. Hence it is of interest to evaluate how and why AF-QM/MM model outperforms small fragment models.

In our modified AF-QM/MM model, implicit-solvent based structural refinement of proteins was performed prior to fragmentation and quantum mechanical calculations. Here we’d like to discuss the two stages of minimization and their effects on CSTs.

The first ten cycles of GB based minimization effectively removes sidechain steric clashes, while preserving the backbone geometry (including $\phi$, $\psi$ and $\omega$ angles) with respect to the experimental structure. This stage proved to be crucial in achieving relatively consistency, given different starting PDB structures acquired under different experimental conditions. For example, prior to minimization, the $^{13}$C CSA values calculated from 1P7E and 1IGD structures deviate by as much as 9.34 ppm for individual nuclei, but this difference was reduced to 3.32 ppm after minimization. Furthermore,
partial minimization leads to a much better trend in helix and sheet $^{15}$N CSAs. Take the 1IGD structure for example, the helix-sheet difference for the $^{15}$N CSA increased from 3.87 ppm to 9.75 ppm after the refinement, which is much closer to the experimental difference of 9.93 ppm.

The full minimization performed after the first 10 cycles does not affect the quality of $^{15}$N CSA significantly. However, it shows a remarkable influence on the $^{13}$C CSA, in terms of the general helix-sheet trend as well as individual CSAs. While GB based full optimization does not affect $\phi, \psi$ angles significantly, it results in significant deviation of the $\omega$ angle from planarity. The effect of full minimization on protein structures is illustrated in Figure 4.6. Prior to minimization, the average $\omega$ angle is $179.242 \pm 2.12$, whereas the average $\omega$ angle is $177.067$, stddev = 6.49 post minimization. The latter value is consistent with former statistical analyses based on high-quality structure databases (6.2° from Priestle et al and 5.5° from Thornton et al)[119, 120]. Previous studies also showed that in X-ray or NMR based structural refinement, the $\omega$ angles are often tightly constrained to planarity, which may lower the quality of resulting structures [122]. Our calculations show that while $^{13}$C chemical shift tensor responds to $\omega$ angle variation, $^{15}$N chemical shift tensor is not as sensitive to the $\omega$ angle change. Taken together, the full GB minimization gives rise to more accurate $^{13}$C CSA values by relaxing the $\omega$ angles restraints.

The other significant advantage of AF-QM/MM model is the inclusion of solvent effect in the quantum mechanical calculations. The solvent environment were represented by Poisson-Boltzmann surface charges on the solvent exposed surface of proteins. When these charges are not accounted for, the correlation to experimental data are 0.48 ($^{13}$C) and 0.45 ($^{15}$N) for the GB1 protein (pdb code: 2QMT), which is similar to small peptide predictions despite the inclusion of a much larger quantum mechanical buffer region and the point-charges representing the entire protein environment. Once the solvation effect was included in the form of surface charges, these correlations were increased to 0.61 ($^{13}$C) and 0.66 ($^{15}$N).
4.4 Conclusions

In this study, we extended the theoretical calculations of CSAs from peptide to proteins, and used our method to evaluate chemical shift tensors of the GB1 and GB3 proteins. The quantitative predictions of protein backbone CSA values proved to be feasible with the AF-QM/MM model, which significantly outperforms other peptide models in reproducing chemical shift tensors for individual residues. The calculated Pearson correlation are close to 0.7 for CSAs and 0.8 for individual tensors of the $^{13}\text{C}$ and $^{15}\text{N}$ nuclei, respectively. The magnitude of calculated $^{15}\text{N}$ CSA tensors show remarkable agreement with solid state measurements of Wylie et al and with the solution/liquid crystal NMR measurements from Yao et al [23, 25]. However, all these experimental and theoretical values of $^{15}\text{N}$ CSAs are considerably different from the $^{15}\text{N}$ relaxation measurements of earlier studies in solution NMR, which showed high variability of regular CSAs and low secondary structure dependency [99]. It is likely that the earlier studies relied on relaxation and CSA-dipolar cross correlation experiment, which may contain large experimental errors [116, 127].

The AF-QM/MM approach achieves accuracy by correct approximation of close contacts and long range electrostatics, and can be used when high quality CSA predictions for individual residues are desired. There are two important advances in the AF-QM/MM approach in comparison to earlier models: First, an additional GB based optimization prior to CSA calculation was applied to achieve consistent and accurate estimations of CSAs, which effectively relaxes over-restrained geometries and provide better refined structures for further assessment. A full minimization is particularly beneficial to $^{13}\text{C}$ CSA calculations, given its dependency on the magnitude of $\omega$ angle. Second, the Poisson-Boltzmann treatment of solvent effects can significantly enhance the accuracy of CSA predictions in comparison to calculations of the entire protein system in vacuo. It is also noted that our chemical shift tensor predictions are most accurate for the principal components highly correlated to the isotropic chemical shift. If further improvements were to be made to the current method, it will be crucial to understand how the predictions can be improved for the tensors which are only
weakly correlated to the isotropic shift. In addition, a better treatment of solvent effect is expected to improve the quality of $^1$H CSA predictions.

More importantly, we have developed a protocol for relatively accurate prediction of chemical shift tensors of proteins. Our theoretical evaluations, combined with solid state data as they become available, are expected to provide a better understanding to CSA and their dependence on local and global environment.
Table 4.1: RMSD of GB1 and GB3 Structures

For all RMSDs calculated, the unminimized crystal structures are used as the reference structure. The fitted atoms include all backbone atoms between residue 3 and residue 54. The RMS are measured in Å.

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB ID</th>
<th>Backbone RMS after partial min</th>
<th>Backbone RMS after full min</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB3</td>
<td>1P7E</td>
<td>0.15</td>
<td>0.79</td>
</tr>
<tr>
<td>GB3</td>
<td>1IGD</td>
<td>0.21</td>
<td>N/A</td>
</tr>
<tr>
<td>GB3</td>
<td>2OED</td>
<td>0.14</td>
<td>N/A</td>
</tr>
<tr>
<td>GB1</td>
<td>1PGA</td>
<td>0.20</td>
<td>0.66</td>
</tr>
<tr>
<td>GB1</td>
<td>2QMT</td>
<td>0.11</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 4.2: Correlation between experimental and calculated isotropic shifts

The first row contains results where the NMA₃ model was used. For the rest of the table, the model of calculation was AF-QM/MM. 1P7E a, b and c are the three lowest energy snapshots from a 203.66 ns long MD trajectory.
<table>
<thead>
<tr>
<th>PDB</th>
<th>Nuclei</th>
<th>Helix</th>
<th>Error</th>
<th>Sheet</th>
<th>Error</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2qmt - Gaussian</td>
<td>$^{13}$C</td>
<td>-86.3</td>
<td>0.8</td>
<td>-84.5</td>
<td>0.50</td>
<td>-1.8</td>
</tr>
<tr>
<td>2qmt - Demon</td>
<td>$^{13}$C</td>
<td>-76.2</td>
<td>0.8</td>
<td>-74.9</td>
<td>0.51</td>
<td>-1.4</td>
</tr>
<tr>
<td>2qmt - Gaussian - Full min</td>
<td>$^{13}$C</td>
<td>-86.5</td>
<td>0.6</td>
<td>-83.3</td>
<td>0.50</td>
<td>-3.2</td>
</tr>
<tr>
<td>2qmt - Demon - Full min</td>
<td>$^{13}$C</td>
<td>-76.2</td>
<td>0.5</td>
<td>-73.5</td>
<td>0.47</td>
<td>-2.7</td>
</tr>
<tr>
<td>1pga - Gaussian</td>
<td>$^{13}$C</td>
<td>-87.6</td>
<td>0.5</td>
<td>-84.1</td>
<td>0.48</td>
<td>-3.5</td>
</tr>
<tr>
<td>1pga - Demon</td>
<td>$^{13}$C</td>
<td>-76.8</td>
<td>0.8</td>
<td>-74.3</td>
<td>0.51</td>
<td>-2.5</td>
</tr>
<tr>
<td>1pga - Gaussian - Full min</td>
<td>$^{13}$C</td>
<td>-87.3</td>
<td>0.4</td>
<td>-82.6</td>
<td>0.60</td>
<td>-4.7</td>
</tr>
<tr>
<td>1pga - Demon - Full min</td>
<td>$^{13}$C</td>
<td>-76.7</td>
<td>0.6</td>
<td>-73.4</td>
<td>0.53</td>
<td>-3.3</td>
</tr>
<tr>
<td>experimental</td>
<td>$^{13}$C</td>
<td>-83.0</td>
<td>0.8</td>
<td>-78.2</td>
<td>0.52</td>
<td>-4.8</td>
</tr>
<tr>
<td>2qmt - Gaussian</td>
<td>$^{15}$N</td>
<td>108.2</td>
<td>1.5</td>
<td>102.1</td>
<td>0.84</td>
<td>6.1</td>
</tr>
<tr>
<td>2qmt - Demon</td>
<td>$^{15}$N</td>
<td>101.8</td>
<td>1.4</td>
<td>95.6</td>
<td>0.83</td>
<td>6.2</td>
</tr>
<tr>
<td>2qmt - Gaussian - Full min</td>
<td>$^{15}$N</td>
<td>108.3</td>
<td>0.9</td>
<td>100.9</td>
<td>0.81</td>
<td>7.4</td>
</tr>
<tr>
<td>2qmt - Demon - Full min</td>
<td>$^{15}$N</td>
<td>101.3</td>
<td>0.9</td>
<td>94.2</td>
<td>0.80</td>
<td>7.2</td>
</tr>
<tr>
<td>1pga - Gaussian</td>
<td>$^{15}$N</td>
<td>107.9</td>
<td>0.9</td>
<td>103.6</td>
<td>0.92</td>
<td>4.3</td>
</tr>
<tr>
<td>1pga - Demon</td>
<td>$^{15}$N</td>
<td>101.6</td>
<td>1.3</td>
<td>96.6</td>
<td>0.80</td>
<td>5.0</td>
</tr>
<tr>
<td>1pga - Gaussian - Full min</td>
<td>$^{15}$N</td>
<td>107.7</td>
<td>0.9</td>
<td>102.2</td>
<td>0.82</td>
<td>5.5</td>
</tr>
<tr>
<td>1pga - Demon - Full min</td>
<td>$^{15}$N</td>
<td>100.8</td>
<td>0.9</td>
<td>95.3</td>
<td>0.76</td>
<td>5.5</td>
</tr>
<tr>
<td>experimental</td>
<td>$^{15}$N</td>
<td>114.0</td>
<td>1.3</td>
<td>106.7</td>
<td>0.66</td>
<td>7.3</td>
</tr>
</tbody>
</table>

Table 4.3: Reduced $^{13}$C and $^{15}$N CSAs of GB1 protein. The average CSAs of α-helical and β-sheet regions are described and their differences calculated in ppm. The errors are the standard errors of the mean.
<table>
<thead>
<tr>
<th>PDB</th>
<th>Nuclei</th>
<th>Correlation</th>
<th>Slope</th>
<th>Error</th>
<th>Intercept</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1pga - Demon</td>
<td>$^{13}$C</td>
<td>0.64</td>
<td>0.74</td>
<td>0.13</td>
<td>-15.52</td>
<td>11.00</td>
</tr>
<tr>
<td>1pga - Gaussian</td>
<td>$^{13}$C</td>
<td>0.58</td>
<td>0.65</td>
<td>0.14</td>
<td>-33.65</td>
<td>12.10</td>
</tr>
<tr>
<td>1pga - Demon - Full min</td>
<td>$^{13}$C</td>
<td>0.75</td>
<td>0.80</td>
<td>0.10</td>
<td>-10.90</td>
<td>7.72</td>
</tr>
<tr>
<td>1pga - Gaussian - Full min</td>
<td>$^{13}$C</td>
<td>0.74</td>
<td>0.61</td>
<td>0.24</td>
<td>-36.38</td>
<td>19.25</td>
</tr>
<tr>
<td>2qmt - Demon</td>
<td>$^{13}$C</td>
<td>0.60</td>
<td>0.63</td>
<td>0.12</td>
<td>-24.93</td>
<td>9.57</td>
</tr>
<tr>
<td>2qmt - Gaussian</td>
<td>$^{13}$C</td>
<td>0.61</td>
<td>0.69</td>
<td>0.14</td>
<td>-29.41</td>
<td>11.05</td>
</tr>
<tr>
<td>2qmt - Demon - Full min</td>
<td>$^{13}$C</td>
<td>0.74</td>
<td>0.68</td>
<td>0.09</td>
<td>-19.94</td>
<td>7.03</td>
</tr>
<tr>
<td>2qmt - Gaussian - Full min</td>
<td>$^{13}$C</td>
<td>0.71</td>
<td>0.70</td>
<td>0.10</td>
<td>-28.43</td>
<td>8.47</td>
</tr>
<tr>
<td>1pga - Demon</td>
<td>$^{15}$N</td>
<td>0.65</td>
<td>0.86</td>
<td>0.22</td>
<td>4.85</td>
<td>24.37</td>
</tr>
<tr>
<td>1pga - Gaussian</td>
<td>$^{15}$N</td>
<td>0.58</td>
<td>0.88</td>
<td>0.26</td>
<td>8.81</td>
<td>28.78</td>
</tr>
<tr>
<td>1pga - Demon - Full min</td>
<td>$^{15}$N</td>
<td>0.69</td>
<td>0.96</td>
<td>0.14</td>
<td>-0.44</td>
<td>15.53</td>
</tr>
<tr>
<td>1pga - Gaussian - Full min</td>
<td>$^{15}$N</td>
<td>0.69</td>
<td>0.95</td>
<td>0.14</td>
<td>-6.44</td>
<td>15.52</td>
</tr>
<tr>
<td>2qmt - Demon</td>
<td>$^{15}$N</td>
<td>0.66</td>
<td>0.99</td>
<td>0.19</td>
<td>-9.87</td>
<td>21.06</td>
</tr>
<tr>
<td>2qmt - Gaussian</td>
<td>$^{15}$N</td>
<td>0.66</td>
<td>1.04</td>
<td>0.20</td>
<td>-8.05</td>
<td>21.60</td>
</tr>
<tr>
<td>2qmt - Demon - Full min</td>
<td>$^{15}$N</td>
<td>0.64</td>
<td>1.05</td>
<td>0.23</td>
<td>-18.25</td>
<td>25.57</td>
</tr>
<tr>
<td>2qmt - Gaussian - Full min</td>
<td>$^{15}$N</td>
<td>0.66</td>
<td>1.12</td>
<td>0.26</td>
<td>-19.06</td>
<td>27.67</td>
</tr>
</tbody>
</table>

Table 4.4: Pearson correlation and Deming regression statistics of reduced $^{13}$C, $^{15}$N CSA for GB1 protein.

The reference values are from Rienstra et al. One data point (residue #40) was removed as an extreme outlier for $^{13}$C nuclei prior to analysis. The reference values are from Rienstra et al. One data point (residue #41) was removed as an extreme outlier for $^{15}$N nuclei prior to analysis. For full minimizations, the last residue (residue #56) for $^{15}$N is also removed prior to analysis.
Table 4.5: Regular $^{15}$N CSA of GB3 protein

The average CSAs of $\alpha$-helical and $\beta$-sheet regions are described and their differences calculated in ppm. The errors in the average CSAs are the standard errors in the mean.

<table>
<thead>
<tr>
<th>PDB</th>
<th>Helix Error</th>
<th>Sheet Error</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1igd</td>
<td>-164.1</td>
<td>2.0</td>
<td>-9.8</td>
</tr>
<tr>
<td>1igd_demon</td>
<td>-154.42</td>
<td>1.8</td>
<td>-9.9</td>
</tr>
<tr>
<td>1p7e</td>
<td>-163.8</td>
<td>1.9</td>
<td>-9.5</td>
</tr>
<tr>
<td>1p7e_demon</td>
<td>-153.8</td>
<td>1.8</td>
<td>-9.5</td>
</tr>
<tr>
<td>2oed</td>
<td>-167.3</td>
<td>1.9</td>
<td>-9.3</td>
</tr>
<tr>
<td>2oed_demon</td>
<td>-153.8</td>
<td>1.8</td>
<td>-10.22</td>
</tr>
<tr>
<td>exp</td>
<td>-172.3</td>
<td>1.9</td>
<td>-9.9</td>
</tr>
</tbody>
</table>

Table 4.6: Pearson correlation and Deming regression statistics of $^{15}$N regular CSA (calculated vs. experimental) for GB3 protein

<table>
<thead>
<tr>
<th>PDB</th>
<th>Pearson correlation</th>
<th>Slope</th>
<th>Error</th>
<th>Intercept</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1igd - Gaussian</td>
<td>0.56</td>
<td>0.85</td>
<td>0.20</td>
<td>-16.87</td>
<td>33.96</td>
</tr>
<tr>
<td>1igd - Demon</td>
<td>0.56</td>
<td>0.80</td>
<td>0.21</td>
<td>-15.10</td>
<td>35.63</td>
</tr>
<tr>
<td>1p7e - Gaussian</td>
<td>0.61</td>
<td>0.88</td>
<td>0.17</td>
<td>-11.46</td>
<td>28.11</td>
</tr>
<tr>
<td>1p7e - Demon</td>
<td>0.59</td>
<td>0.82</td>
<td>0.17</td>
<td>-12.52</td>
<td>28.39</td>
</tr>
<tr>
<td>2oed - Demon</td>
<td>0.54</td>
<td>0.82</td>
<td>0.19</td>
<td>-25.39</td>
<td>32.49</td>
</tr>
<tr>
<td>2oed - Demon</td>
<td>0.53</td>
<td>0.86</td>
<td>0.23</td>
<td>-5.07</td>
<td>39.27</td>
</tr>
</tbody>
</table>

Table 4.7: Pearson correlation and Deming regression statistics of $^1$H Regular CSA of GB3 protein

<table>
<thead>
<tr>
<th>PDB</th>
<th>Nuclei</th>
<th>Helix</th>
<th>Error</th>
<th>Sheet</th>
<th>Error</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1IGD - Demon</td>
<td>$^1$H</td>
<td>8.55</td>
<td>0.30</td>
<td>10.55</td>
<td>0.47</td>
<td>-2.00</td>
</tr>
<tr>
<td>1P7E - Demon</td>
<td>$^1$H</td>
<td>8.44</td>
<td>0.29</td>
<td>11.00</td>
<td>0.44</td>
<td>-2.56</td>
</tr>
<tr>
<td>2OED - Demon</td>
<td>$^1$H</td>
<td>8.37</td>
<td>0.30</td>
<td>11.18</td>
<td>0.47</td>
<td>-2.82</td>
</tr>
<tr>
<td>exp, parm3</td>
<td>$^1$H</td>
<td>8.07</td>
<td>0.42</td>
<td>9.76</td>
<td>0.33</td>
<td>-1.70</td>
</tr>
<tr>
<td>exp parm5</td>
<td>$^1$H</td>
<td>8.39</td>
<td>0.40</td>
<td>10.19</td>
<td>0.32</td>
<td>-1.80</td>
</tr>
<tr>
<td>Src</td>
<td>Exp</td>
<td>Correlation</td>
<td>Slope</td>
<td>Error</td>
<td>Intercept</td>
<td>Error</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----</td>
<td>-------------</td>
<td>-------</td>
<td>-------</td>
<td>-----------</td>
<td>-------</td>
</tr>
<tr>
<td>1igd - Demon</td>
<td>parm3</td>
<td>0.58</td>
<td>1.47</td>
<td>0.59</td>
<td>-3.15</td>
<td>5.17</td>
</tr>
<tr>
<td>1igd - Demon</td>
<td>parm5</td>
<td>0.51</td>
<td>1.49</td>
<td>0.73</td>
<td>-3.87</td>
<td>6.76</td>
</tr>
<tr>
<td>1p7e - Demon</td>
<td>parm3</td>
<td>0.61</td>
<td>1.41</td>
<td>0.53</td>
<td>-2.40</td>
<td>4.72</td>
</tr>
<tr>
<td>1p7e - Demon</td>
<td>parm5</td>
<td>0.51</td>
<td>1.37</td>
<td>0.68</td>
<td>-2.56</td>
<td>6.26</td>
</tr>
<tr>
<td>2oed - Demon</td>
<td>parm3</td>
<td>0.52</td>
<td>1.52</td>
<td>0.59</td>
<td>-3.45</td>
<td>5.22</td>
</tr>
<tr>
<td>2oed - Demon</td>
<td>parm5</td>
<td>0.60</td>
<td>1.51</td>
<td>0.76</td>
<td>-3.86</td>
<td>6.98</td>
</tr>
<tr>
<td>NMA-acetamide</td>
<td>parm3</td>
<td>0.43</td>
<td>0.14</td>
<td>0.10</td>
<td>11.10</td>
<td>0.95</td>
</tr>
<tr>
<td>NMA-acetamide</td>
<td>parm5</td>
<td>0.34</td>
<td>0.10</td>
<td>0.23</td>
<td>11.56</td>
<td>2.30</td>
</tr>
</tbody>
</table>

Table 4.8: Pearson correlation and Deming regression statistics for $^1$H CSA of GB3 protein

The “NMA-acetamide” rows indicates the values were taken from the results of Yao and co-workers using NMA-acetamide as the model of their calculation ([24]). The second column describes the experimental sources that we are comparing with. “Parm3” and “Parm5” indicates the three parameter and five parameter models, respectively. Only the 34 residues with intramolecular backbone hydrogens were included for comparisons.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Nuclei</th>
<th>PC</th>
<th>Correlation</th>
<th>PC</th>
<th>Correlation</th>
<th>PC</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB1</td>
<td>$^{13}$C</td>
<td>$\delta_{xx}$</td>
<td>0.29</td>
<td>$\delta_{yy}$</td>
<td>0.78</td>
<td>$\delta_{zz}$</td>
<td>0.23</td>
</tr>
<tr>
<td>GB1</td>
<td>$^{15}$N</td>
<td>$\delta_{xx}$</td>
<td>0.60</td>
<td>$\delta_{yy}$</td>
<td>0.58</td>
<td>$\delta_{zz}$</td>
<td>0.82</td>
</tr>
<tr>
<td>GB1-full min</td>
<td>$^{13}$C</td>
<td>$\delta_{xx}$</td>
<td>0.17</td>
<td>$\delta_{yy}$</td>
<td>0.70</td>
<td>$\delta_{zz}$</td>
<td>0.39</td>
</tr>
<tr>
<td>GB1-full min</td>
<td>$^{15}$N</td>
<td>$\delta_{xx}$</td>
<td>0.62</td>
<td>$\delta_{yy}$</td>
<td>0.56</td>
<td>$\delta_{zz}$</td>
<td>0.77</td>
</tr>
<tr>
<td>GB3</td>
<td>$^1$H</td>
<td>$\delta_{xx}$</td>
<td>0.59</td>
<td>$\delta_{yy}$</td>
<td>0.69</td>
<td>$\delta_{zz}$</td>
<td>0.76</td>
</tr>
<tr>
<td>GB3</td>
<td>$^{15}$N</td>
<td>$\delta_{xx}$</td>
<td>0.45</td>
<td>$\delta_{yy}$</td>
<td>0.67</td>
<td>$\delta_{zz}$</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 4.9: Pearson correlation between principal components of CSTs

The GB1 data were calculated from 1PGA structure, while the GB3 data were calculated from the 1P7E structure.
Table 4.10: Pearson correlation and least squares regression statistics for $^1$H isotropic shift
The correlations were calculations between our theoretical calculations and the experimental data for Yao and co-workers [24]. The “NMA-acetamide” rows indicates the values were taken from the results of Yao and co-workers using NMA-acetamide as the model of their calculations. Only the 34 residues with intramolecular backbone hydrogens were included for comparisons.

<table>
<thead>
<tr>
<th>PDB/Source</th>
<th>Correlation</th>
<th>LS Slope</th>
<th>Error</th>
<th>LS Intercept</th>
<th>Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>1IGD - Demon</td>
<td>0.86</td>
<td>1.14</td>
<td>0.12</td>
<td>0.12</td>
<td>1.04</td>
</tr>
<tr>
<td>1P7E - Demon</td>
<td>0.90</td>
<td>1.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.82</td>
</tr>
<tr>
<td>2OED - Demon</td>
<td>0.90</td>
<td>1.09</td>
<td>0.10</td>
<td>0.10</td>
<td>0.78</td>
</tr>
<tr>
<td>NMA-acetamide</td>
<td>0.67</td>
<td>0.79</td>
<td>0.16</td>
<td>0.16</td>
<td>1.33</td>
</tr>
</tbody>
</table>

Table 4.11: Ubiquitin CSA
The NMR and X-ray method refers to the bond length and angles taken from either NMR or X-ray structures as described by Cisnetti and coworkers [98]. The details and assumptions of Model1 to Model5 were defined in the work of Loth and coworkers [143]. The average CSAs of $\alpha$-helical and $\beta$-sheet regions are measured in ppm.

<table>
<thead>
<tr>
<th>Method</th>
<th>Src</th>
<th>Helix</th>
<th>Sheet</th>
<th>Difference</th>
<th>Src</th>
<th>Helix</th>
<th>Sheet</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMR</td>
<td>$^{13}$C</td>
<td>-89.76</td>
<td>-88.12</td>
<td>-1.64</td>
<td>$^{15}$N</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>X-ray</td>
<td>$^{13}$C</td>
<td>-89.42</td>
<td>-87.71</td>
<td>-1.71</td>
<td>$^{15}$N</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Model1</td>
<td>$^{13}$C</td>
<td>-95.63</td>
<td>-91.86</td>
<td>-3.77</td>
<td>$^{15}$N</td>
<td>113.57</td>
<td>109.85</td>
<td>3.72</td>
</tr>
<tr>
<td>Model2</td>
<td>$^{13}$C</td>
<td>-94.17</td>
<td>-90.48</td>
<td>-3.69</td>
<td>$^{15}$N</td>
<td>107.48</td>
<td>103.86</td>
<td>3.62</td>
</tr>
<tr>
<td>Model3</td>
<td>$^{13}$C</td>
<td>-98.41</td>
<td>-95.17</td>
<td>-3.24</td>
<td>$^{15}$N</td>
<td>102.19</td>
<td>97.30</td>
<td>4.89</td>
</tr>
<tr>
<td>Model4</td>
<td>$^{13}$C</td>
<td>-85.10</td>
<td>-81.67</td>
<td>-3.43</td>
<td>$^{15}$N</td>
<td>100.66</td>
<td>96.18</td>
<td>4.48</td>
</tr>
<tr>
<td>Model5</td>
<td>$^{13}$C</td>
<td>-85.19</td>
<td>-81.69</td>
<td>-3.50</td>
<td>$^{15}$N</td>
<td>109.98</td>
<td>105.19</td>
<td>4.79</td>
</tr>
<tr>
<td>AF-QM/MM</td>
<td>$^{13}$C</td>
<td>-78.08</td>
<td>-75.57</td>
<td>-2.51</td>
<td>$^{15}$N</td>
<td>97.82</td>
<td>95.29</td>
<td>2.53</td>
</tr>
</tbody>
</table>
Figure 4.1: AF-QM/MM model for CSA calculations. The central QM region, represented by ball-and-stick model, is calculated on B3LYP/6-31G* level. The rest of the protein, represented in ribbons, are treated as AMBER charges. The entire protein is embedded in surface charges rendered as small blue dots –make these black and bigger.

Figure 4.2: Overlapped GB1 and GB3 structures. The α-helix is rendered in purple and the β-sheets are rendered in yellow.
Figure 4.3: Correlations of isotropic chemical shifts. Left: the red line is the linear regression line: $y = 0.69 \times x + 19.62$, $R = 0.83$. Right: the red line is the linear regression line: $y = 0.84 \times x + 16.2$, $R = 0.78$.

Figure 4.4: Sequence alignment of GB1 and GB3 proteins

MTYKLINGK TLKGETTEA VDAATAKVF KQYANDNGVD GETYDDATK TFTVTE
MQYKLINGK TLKGETTTKA VDAETAHKF KQYANDNGVD GETYDDATK TFTVTE
Figure 4.5: Comparison of reduced $^{15}$N and $^{13}$C CSA for GB1 protein
The calculation of $^{13}$C nuclei is based on the 1PGA structure and the calculation of $^{15}$N nuclei is based on the 2QMT structure.
Figure 4.6: GB1 protein backbone $\omega$ angle changes due to full minimization

Figure 4.7: The correlation between experimental and calculated chemical shifts for ubiquitin.
The chemical shift data are taken from the work of Wand and coworkers[146]. The predictions are based on the human ubiquitin structure(pdb code: 1UBQ).
Figure 4.8: Relationship between $^1$H isotropic shift and CSA tensors for GB3 protein. The predicted $^1$H chemical shift tensors and isotropic shifts are calculated from the 1PGA structure. The Pearson correlation for $\delta_{zz}$, $\delta_{yy}$, $\delta_{xx}$ are 0.89, 0.85 and 0.32, respectively.
Figure 4.9: Relationship between $^{13}$C isotropic shift and CSA tensors for GB1 protein

The results presented in this plot is based on the calculation of 2QMT structure. The Pearson correlation between $\delta_{yy}$ and $\delta_{iso}$ is 0.81.
Figure 4.10: Relationship between $^{15}$N isotropic shift and CSA tensors for GB3 protein. The results presented in this plot is based on the calculation of 1IGD structure. The Pearson correlation for $\delta_{zz}$, $\delta_{yy}$, $\delta_{xx}$ are 0.89, 0.59 and 0.78, respectively.
Chapter 5

Preliminary Characterization of Metastable States of TIM

Conformational changes of proteins are closely associated with a wide range of protein functions. The "conformational selection" theory derived from NMR relaxation studies suggests that the ground state of a protein can coexist with one or more lowly-populated excited states. These excited states will become ground states in the ensuing steps of protein function. While NMR studies suggest that the minor conformers resemble stable states in other parts of the catalytic cycle, there are no direct descriptions of these metastable states from experimental studies. Here we propose a novel strategy to induce and characterize the metastable states of proteins using molecular dynamics (MD) simulations. The model of our study is triosephosphate isomerase (TIM), which shows a localized loop motion tightly coupled to its catalysis. Suitable order parameters of this conformational transition are also determined in our investigation, which is critical to describing the dynamic energy landscape programmed into TIM function (Chapter 6 and Chapter 7).

5.1 Introduction

NMR spectroscopy provides a dynamical view of protein structures, using chemical shift differences as a probe for the minor conformational states with functional significance [40]. In the case of triosephosphate isomerase (TIM), NMR experiments suggest that its catalytic competence is associated with a slow loop motion on \( \mu \)s-ms time scale, where the closed and open loop conformations are exchanged regardless of the presence of ligand [53] (Figure 5.1). If as the NMR data suggested, both conformations of TIM active site loop have \( \mu \)s-ms time scale lifetimes, it is likely that a large-scale conformational transition will not occur on a short (ps-ns) time scale starting from the
excited state. Hence if we make the initial guess that the minor conformer of unligated TIM takes a closed loop conformation, we may start from this structure and stabilize it with a nanosecond long MD simulation, which may result in the desired metastable form not far from its initial conformation. To achieve this, the crystal structure of TIM bound to a substrate mimic (glycerol-3-phosphate, G3P) is adopted [147]. An instantaneous removal of G3P followed by MD-based stabilization at room temperature, is likely to yield the minor conformers of the apo enzyme. Similarly, docking the ligand into the open enzyme may lead to the minor conformer of the bound enzyme (Figure 5.2).

Thus in the present study, we propose a novel strategy to induce the metastable states of proteins via instant removal of ligand to the apo conformation or addition of ligand to the bound conformations. Then we used MD simulations to further compare and study the dynamics behavior of these conformations. More importantly, we are aiming to find suitable order parameters of the TIM conformational transition, which are used in later chapters for the study of energy and dynamics of the entire transition pathway of TIM loop conformational change.

5.1.1 Methods

In this study, we performed MD simulations for four different systems of monomeric TIM, including TIM in their native conformations (apo TIM, G3P-bound TIM), as well as the presumed “metastable” states (“apo” TIM docked with G3P and “bound TIM” with G3P removed). All systems went through MD equilibration, which were then analyzed in terms of structural stability and large-scale, coordinated motions. Furthermore, the structures from respective trajectories were compared to identify one or more appropriate reaction coordinate(s) of TIM loop conformation change.

5.1.1.1 Molecular dynamics simulations

MD simulations were perform to gain insight into the structure and dynamic behavior of metastable states of TIM. In addition, they were expected to remove packing
artifacts, and quantify the fast time scale flexibility independent of the slow conformational transition. The model of our study includes four systems. Two of those, the apo and bound forms of TIM monomer were directly adopted from the dimeric crystal structure (PDB code: 6TIM) [147]. The crystal structure contains two subunits, which takes the open and closed loop conformation, respectively. The starting structure for apo TIM in its presumed metastable state is derived by removing the ligand from the bound TIM. The bound TIM in its metastable state is constructed by re-docking the G3P ligand into the active site. To simulate TIM in solution, each of these systems was solvated in a truncated octahedral box of TIP4P water molecules, where each atom is at least 10 Å away from the solvent boundary. The calculation protocol included the periodic boundary conditions, a time step of 2 fs, non-bonded cut-off at 10 Å, and Particle Mesh Ewald (PME) for treatment of long range electrostatics. Initially, the minimized model systems were heated to 300 K and equilibrated for 50 ps using the canonical (NVT) ensemble, followed by 2.95 ns long density equilibration using the isothermal-isobaric (NPT) ensemble. The production run includes a 30 ns constant energy (NVE) simulation, with MD snapshots collected and energies monitored at 5 ps intervals.

5.1.1.2 Trajectory stability analyses

At the end of the simulation, the potential energies along the MD trajectory were extracted and their time dependent behavior analyzed. In order to quantify the structural stability of TIM, we also performed RMSD analyses using the PTRAJ module of AMBER program [111]. For the trajectory of each model system, the atomic coordinates were RMS fitted to the backbone of their respective starting structures, excluding that of the flexible loop (Tyr 165 - Val 176). The structural fluctuation was then evaluated as the deviation of the loop structures in these fitted trajectories. Only the portion of trajectory that are structurally and energetically stabilized will be used for further analyses.
5.1.1.3 Screening for suitable order parameter

Once the systems are equilibrated, we were able to examine the dynamic behavior of the ligated and unligated TIM species. The conformational changes reflected in individual residues were also examined. We were looking to extract one or more order parameters, indicated by a significant difference in a few degrees of freedom between the apo and bound form of TIM. These order parameters will be adopted as reaction coordinates to evaluate the activation energy of the TIM loop conformational change.

While the comparison between static structures of the ligand-free and ligand-bound forms of TIM reveals a number of sidechain and backbone movements, many of these differences may be caused by crystal packing or fast time scale fluctuations. In order to find true order parameters indicative of long time scale motions, a large number of different geometric parameters in the active site, including dihedral angles, pseudo-dihedral angles and distances were screened. More specifically, the screened parameters include all the C-α distances between Tyr 209 and the central loop residues (Val 169-Gly 175), all the backbone dihedrals (φ and ψ) from Tyr 166 to Lys 176, as well as all the pseudo-dihedral angles from Pro 167 to Pro 179.

5.1.1.4 Essential dynamics analyses

The essential dynamics analyses were employed to distinguish between large-scale, concerted motions from random thermal fluctuations [148]. In this method, the covariance matrix of atomic displacement $C$ are first calculated from the snapshots of the MD trajectory. The principal component modes are then obtained by the diagonalization of $C$ as follows:

$$C = \langle (x - \langle x \rangle)(x - \langle x \rangle)^T \rangle = T \Lambda T^T \tag{5.1}$$

where $\Lambda$ is the diagonal matrix of eigenvalues, and $T$ is a matrix of column eigenvectors of an orthonormal basis set. The resulting eigenvectors and eigenvalues represent the direction and magnitude of these motions, respectively. The eigenvectors are sorted in decreasing order, so that the first eigenvector gives the largest positional
fluctuation. Earlier studies showed that the first few principal modes represent 90% of the important internal motions [149].

Thus we analyzed the combined trajectory of the four systems to understand the essential motions of TIM in absence and presence of the ligand. The average coordinate of all snapshots were used as the reference structure to construct the covariance matrix, which was then diagonalized to obtain the principal component modes. The resulting eigenvectors and eigenvalues representing the direction and magnitude of these motions were then analyzed.

5.1.2 Results and Discussion

5.1.2.1 Trajectory Stability analyses

The subsequent analyses of the MD trajectory revealed that energetic stabilization was achieved within 1 ns (Figure 5.3, 5.4). Overall, the RMSD profile revealed that all systems achieved structural stability after 5 ns (Figure 5.5). The ligand-bound form of TIM shows the smallest loop fluctuation on the order of 0.8 Å. When the G3P ligand is removed from the active site, the “closed loop” shows more flexibility with RMSD on the order of 1.8 Å. The loop fluctuation is comparable (1.6 Å) for the ligated form of TIM with an open loop conformation. The largest RMS flexibility comes from the apo form of TIM, where the loop structures deviate by 2.1 Å from the starting structure.

5.1.2.2 Screening for suitable order parameter

A number of different geometric parameters, including dihedral angles, pseudo-dihedral angles and hydrogen bonding distances from the active site loop, were extracted from the respective trajectories and compared. For example, the tip of the loop residues is observed to move 7 Å in the X-ray structures, yet MD simulation reveals that some of these loop residues have high intrinsic flexibility on picosecond time scale.

Among these candidates, a hydrogen bond distance (Tyr 209 (OH) - Ala 177 (NH)) was considered promising, as it was used as the reaction coordinate in the previous study of Schlick et al [55]. However, after examining its time-dependent profile, it
appears that this hydrogen bond can be easily broken on nanosecond time scale even when TIM is ligated (Figure 5.6). Therefore this hydrogen bond distance is deemed unsuitable as an order parameter for characterizing the loop conformational change.

Nevertheless, we were able to identify the loop hinge movement, represented by a pseudo-dihedral angle as a suitable order parameter. This angle is formed between four consecutive C-α carbon atoms (Glu165-Pro166-Val167-Trp168), which differs significantly for the open and closed form of TIM (Figure 5.7). For the ligated vs. unligated forms, the pseudo-dihedral angle is stabilized at -125° ± 6° vs. -189° ± 6°, with 60° of separation between them without any overlap. Therefore, this angle can be used as the reaction coordinate for our potential of mean force calculations (in chapter 6 & 7), which will be progressively changed from one conformational state to the other.

As a result, the pseudo-dihedral angle formed by four consecutive C-α carbon atoms from the hinge residues was chosen as the reaction coordinate.

5.1.2.3 Essential dynamics analyses

Essential dynamics analyses showed that the amplitude and direction of motion is significantly different for the systems derived from the unligated and the ligated forms (Figure 5.8). More importantly, the essential motions of TIM did not change significantly after instantaneous removal or docking of the ligand, which suggested that the metastable states have been induced and stabilized. Visualization of the essential dynamics of these states showed dynamic behavior resembling the ground state of TIM in the other functional form [150]. Hence, it is likely that the flexibility of TIM in its essential subspace is predisposed to subsequent functional transition. Furthermore, TIM structures with the open and closed loop can be distinguished by the direction and amplitude of the movement along the principal component modes, suggesting that the principal component may also be used as an order parameter for further studies of TIM conformational changes, which will be discussed in chapter 8.
5.1.3 Conclusions

In agreement with the “conformational selection theory”, the preliminary results showed that the quasi-stable minor conformational state of the unligated state exist on the time scale of MD simulation, and bears resemblance to the major conformation of the ligated state. In particular, the structural and dynamics analyses showed that excited states of TIM exist for both apo and bound forms of TIM. More importantly, a pseudo-dihedral angle has been identified as a desirable order parameter for TIM loop motion, which will be used to calculate the activation barrier of TIM conformation transition in chapter 6 & 7.
Figure 5.1: Loop motion of TIM is not ligand gated
The conformational exchange rate were measured by Rozovsky et al [53].
Figure 5.3: Potential energy profile of bound TIM in the open form
The trajectory was taken for the first 12 ns.
Figure 5.4: Potential energy profile of apo TIM in the closed form.
The trajectory was taken for the first 12 ns.
Figure 5.5: RMS profile of apo TIM in the closed form
The trajectory was taken for the first 20 ns.

Figure 5.6: Selection of reaction coordinate: exclusion of hydrogen bonds
Left: hydrogen bonding profile for Ala177(NH)-Tyr209(OH) during the MD simulations. Right: the ball-and stick representation of the hydrogen bonding residues.
Figure 5.7: Pseudo-dihedral angle as a reaction coordinate
Left: pseudo-dihedral angle profile during the MD simulations. Right: the pseudo-dihedral angle is formed by four consecutive C-α atoms, represented by green and red spheres for the open and closed form of TIM.

Figure 5.8: PCA analyses for MD simulations of apo and bound TIMs
Chapter 6
Exploring the Free Energy Landscape of TIM Conformational Changes

NMR relaxation experiments provide a dynamical view of protein structures, using chemical shift differences as a probe for minor conformers of proteins with functional importance. Nevertheless, a quantifiable difference in chemical shift is required to characterize these minor conformers, and populations smaller than 0.5% cannot be easily detected. In principle, computational methods are able to compensate for the limitations in NMR techniques and can be used to give a complete picture of concerted, global dynamics for a wide range of proteins. More importantly, it is hoped that theoretical investigations can be used to predict the functionally important conformations “invisible” to or cannot be completely characterized by experiments, yet so far it remains challenging. In this study, we show how the metastable conformations can be predicted using adaptively biased molecular dynamics (ABMD) simulations. Furthermore, we develop a novel approach, termed the iterative ABMD-US approach, which combines umbrella sampling (US) and ABMD simulations to describe the energy profile of the conformational transition. We show that this method leads to faster convergence in comparison to other free energy methods, and therefore can be adapted to characterize the energy landscape of a variety of protein conformational changes.

6.1 Introduction

The flexibility of proteins are associated with a wide variety of protein functions, including enzyme catalysis, molecular recognition, signal transduction and assisted protein folding [38]. According to the longstanding "induced-fit" model of protein motion, the presence of a ligand leads to accommodating motions of the target protein [39].
Nevertheless, recent NMR spectroscopy studies support the "conformational selection" theory, which suggest that multiple conformational states coexist in equilibrium, and that the introduction of a ligand simply biases the protein population towards the conformational state favored for binding [41]. The CPMG-based relaxation dispersion experiment is one of the most commonly used NMR techniques to probe the thermodynamics and kinetics of protein population shifts [40]. While CPMG experiments have proven effective in detecting slow motions of proteins on \( \mu s \)-ms time scale, a detailed description of the metastable state and corresponding pathways of transitions remain elusive.

Computational methods are expected to complement experiment and give a complete picture of the energy and dynamics of conformational transitions on the atomic level. However, so far molecular simulations have generally fallen short in providing insights to the excited states of proteins. For example, contrary to experimental evidence, computational studies of the A/B-DNA conversion and adenylate kinase showed no sizable energy barrier for the conversion between the ground state and the excited state in the absence of the ligand [42, 151]. To this end, it is unclear where the discrepancy between theory and experiment comes from. This inconsistency may be due to intrinsic limitations of molecular mechanics based force fields. Additionally, it may indicate that more robust methods and protocols of free energy simulations need to be developed for the characterization of these excited states.

In this study, we seek to answer the following questions: First, can we predict the excited states of TIM using a molecular mechanics based force field? Second, how accurate are the current free energy methods and can they be further improved? To answer these questions, we first performed adaptively biased dynamics simulations to explore the free energy landscape of NMA cis-trans conversion and TIM loop motion and predict the "invisible" conformations and compare them with the crystallographic structures. Then the AMBD-US approach and other free energy methods were used to evaluated the TIM energy landscape. By comparing the ABMD-US approach and other free energy methods, we show that ABMD-US is a robust method for free energy
calculation which quickly achieves convergence. Furthermore, our approach is gen-
eralizable and is expected to provide insight into the slow conformational changes for a
wide range of large biomolecules.

6.2 Theory

In this section, we describe four free energy approaches used to evaluate the energy
landscape of TIM loop motion. First we briefly discuss weighted histogram analysis in
the context of umbrella sampling. Next, we introduce the ABMD method, which uses a
history-dependent potential to improve sampling of the system. Then we discuss how
umbrella sampling and ABMD simulations can be combined in two different ways to
further improve accuracy and convergence of the free energy calculations.

6.2.1 Weighted histogram analysis method

In our study of TIM loop motion, umbrella sampling simulations was first applied to
evaluate the free energy profile of TIM conformational transition. For each sampling
window $w_j$, the constraint on each window is described as a harmonic potential:

$$ w_j(\xi) = K_{rmsd}(\xi - \xi_{min})^2 $$

(6.1)

where $\xi$ is the reaction coordinate, namely the pseudo-dihedral angle characteristic
of the loop hinge movement; $\xi_{min}$ is the minimum value for the harmonic potential and
$K_{rmsd}$ is the harmonic force constant. The weight histogram analysis method (WHAM)
was used to estimate the unbiased probability distribution $\rho(\xi)$ as the weighted sum
of biased distributions $\rho'(\xi)$ using the following equations[49]:

$$ < \rho(\xi) > = \frac{\sum_i n_i < \rho'(\xi) >_i}{\sum_i n_i e^{(F_j - w_j(\xi)) / k_B T}} $$

(6.2)

$$ F_j = -k_B T ln \left( \int e^{-w_j(\xi)/k_B T} < \rho(\xi) > d\xi \right) $$

(6.3)
where \( N_i \) is the number of umbrella sampling windows and \( n_i \) is the number of points in each window, and \( F_j \) is the free energy shift for each window. The WHAM equations were solved iteratively using Alan Grossfield’s WHAM code [152].

The potential of mean force can then be calculated from the recovered probability distribution \(<\rho(\xi)>\):

\[
W(\xi) = -k_B T \ln <\rho(\xi)> + \text{constant} \tag{6.4}
\]

6.2.2 Adaptively biased molecular dynamics method

As a variation of the conformational flooding and metadynamics methods, the ABMD method is based on the Newtonian equation of motion with an additional force from the biasing potential \( U(t|\xi) \) [153]:

\[
m_i \frac{d^2 r_i}{dt^2} = F_i - \frac{\partial}{\partial r_i} U[t|\sigma] \tag{6.5}
\]

\[
\frac{\partial U(t|\xi)}{\partial t} = \frac{k_B T}{\tau_F} G[\xi - \sigma] \tag{6.6}
\]

where \( \tau_F \) is the flooding time scale, \( \sigma \) is the collective reaction coordinate, and \( \xi \) is the variable used to calculate the biweight kernel \( G(\xi) \):

\[
G(\xi) = \begin{cases} 
\frac{48}{41} \left(1 - \frac{\xi^2}{4}\right)^2, & -2 \leq \xi \leq 2 \\
0, & \text{otherwise}
\end{cases}
\]

In Amber 11, the ABMD method is implemented so that \( U(t|\xi) \) is discretized using a cubic B-spline \( B(\xi) \):

\[
U(t|\xi) = \sum_m U_m(t) B(\xi / \Delta \xi - m) \tag{6.7}
\]
The time-dependent, discretized component, $U_m(t)$, is calculated by adding many small Gaussians along the course of the simulation:

$$U_m(t + \Delta t) = U_m(t) + \Delta t \frac{k_B T}{\tau_F} G[\sigma / \Delta \xi - m]$$  \hspace{1cm} (6.9)$$

In this way, the system is discouraged from revisiting the previously sampled low energy states, and eventually the biasing potential outgrows the size of the original potential energy basin. Then the system will be able to cross the lowest transition state and move into the neighboring local minimum, which may correspond to the excited state. When both potential wells of the ground and excited states are flooded, the biasing potential along the reaction coordinate is complementary to the free energy surface. In practice, the simulation is stopped right before the system recrosses to the ground state, and the instantaneous adaptive biasing potential provides an initial guess to the underlying free energy landscape [154].

Ideally, if the ABMD simulation were performed with an infinitely large $(\tau_F)$ and given sufficient time to sample all conformations, eventually the biasing potential will be exactly complementary to the intrinsic free energy surface, and the biased probability density will be uniform regardless of the reaction coordinate. In practice, the biased probability may vary with respect to the reaction coordinate due to slow convergence of the repulsive potential. Hence an "umbrella correction" term, $k_B T ln p^B(\xi)$, will be added to the biasing potential to obtain the corrected free energy profile, $f(\xi)$, as follows:

$$f(\xi) = -V_h(\xi) - k_B T ln p^B(\xi)$$  \hspace{1cm} (6.10)$$
6.2.3 Combining ABMD with US

While umbrella sampling is a powerful method widely used to improve sampling of protein conformations, starting multiple windows at a single conformation may lead to significantly slower convergence. When large restraints are used inappropriately, the sample conformational space may become overly restricted, and never have the opportunity to relax to the true minimum. In the mean time, while ABMD simulation can be used to explore the transition pathway and sample a large conformational space along the reaction coordinate, the resulting biasing potential is oftentimes not accurate enough to be considered complementary to the intrinsic free energy surface.

Hence it is a good idea to combine ABMD with US. A first combined approach entails ABMD simulations followed by umbrella sampling. More specifically, a number of structures along the reaction coordinate are extracted from the ABMD simulations, and used as starting points for further umbrella sampling. In this way, better starting structures are employed for umbrella sampling, yet any prior knowledge of the biasing potential from ABMD simulations is ignored.

Alternatively, a second approach, termed the iterative ABMD-US approach, takes advantage of the aforementioned biasing potential as an initial energy estimate. First, an ABMD simulation is performed to generate the intermediate structures along the transition path and to give an initial estimate to the appropriate biasing potential complementary to the free energy surface. Next, umbrella sampling simulation is performed with the ABMD-simulated structures together with the underlying bias. The umbrella-corrected free energy profile will then be updated as the new biasing potential and a second iteration of umbrella sampling was performed with the new bias. In this way, multiple iterations was employed until desired convergence was achieved. This method effectively takes advantage of both structures and energetic estimates from ABMD simulations.
6.3 Methods

6.3.1 ABMD simulations

6.3.1.1 Benchmarking study of NMA

In this study, we used N-methyl acetamide (NMA) as a minimal model to understand the evolution of flooding potential and the extent of conformational sampling in ABMD simulations. More specifically, the accuracy and convergence of the predicted free energy profile were examined and discussed.

The NMA molecule was embedded in a 10 Å TIP4P truncated octahedral water box and modeled using the ff99SB force field[73]. The system was then minimized, heated over 20 ps and equilibrated for 100 ps using Langevin dynamics, where the collision frequency was set to be 1 ps$^{-1}$. Subsequently, ABMD simulation was performed for 15 ns, with a flooding time scale of 25 ps. The reaction coordinate was defined as the torsional angle formed by the backbone atoms of NMA (Figure 6.1).

Furthermore, we performed “umbrella correction” on the converged free energy estimate, where umbrella sampling was performed in addition to the biasing potential taken at the 15th ns of the ABMD simulation. Umbrella correction was performed with 36 windows at 10 degree intervals on the backbone torsional angle of NMA. The MD simulation for each window was performed for 10 ns, with a harmonic restraint of 50 kcal/mol/rad$^2$. The combined trajectories are 360 ns long, where the torsional angle statistics were collected every 0.2 ps. The WHAM program was then used to analyze the collected pseudo-dihedral angle statistics.

6.3.1.2 ABMD simulations of TIM

Adaptively biased molecular dynamics (ABMD) simulations were applied to simulate the transition between the ground state and excited state of TIM, using the aforementioned pseudo-dihedral angle as the reaction coordinate. The starting points of ABMD simulations was the equilibrated structures generated after 20 ns of constant energy MD. ABMD simulations were performed in three stages. The first stage was
performed for 3 ns with a 50 ps flooding time scale ($\tau_F$) and $20^\circ$ resolution ($4\Delta \xi$). A second stage includes a 5 ns run with $\tau_F = 50\,\text{ps}$ and $4\Delta \xi = 4^\circ$. In the last stage, 15 ns simulation was performed with $\tau_F = 100\,\text{ps}$ and $4\Delta \xi = 4^\circ$.

6.3.2 Free energy evaluations

6.3.2.1 Umbrella sampling

Umbrella sampling simulations was first performed to evaluate the free energy profile. A total of 16 windows were established: the windows were separated by $10^\circ$ intervals from $-190^\circ$ to $-180^\circ$ and $-120^\circ$ to $-110^\circ$. The starting point for all windows is the MD-equilibrated structure of the monomeric TIM in its open form. Each window was simulated for 5 ns in Langevin dynamics under constant pressure, where the friction coefficient was set at $5\,\text{ps}^{-1}$. The harmonic force constant, $K_{r_{\text{rmsd}}}$ was set to be $50\,\text{kcal/mol/rad}^2$, except for the windows centered at $-175^\circ$ and $-135^\circ$, where a $100\,\text{kcal/mol/rad}^2$ force constant was applied. The unbiased probability distribution is obtained by running Alan Grossfield’s WHAM program, where the analyses were performed from $-200^\circ$ to $-110^\circ$ with 90 bins and a convergence tolerance of $0.0001$ kcal/mol [152]. The Monte Carlo bootstrap errors were analyzed accordingly with 100 random trials.

6.3.2.2 Umbrella corrections to ABMD

Following the ABMD simulations, the ABMD potentials at the 10th ns were used as the initial guess of the biasing potential. To enhance the sampling of different conformations, an umbrella correction term was calculated by running 10 different windows with the same bias, where the starting structures vary from $-200^\circ$ to $-110^\circ$ at $10^\circ$ intervals. Each window was simulated for 6 ns, where data points of the reaction coordinate were recorded at $2\,\text{ps}$ intervals. Thus a total trajectory of 60 ns was collected. The correction term $-k_B T \ln p^B(\xi)$ was then calculated with a 95% confidence interval.
6.3.2.3 Combining ABMD and US

For the first combined approach previously described, 10 structures close to the center of the windows were extracted from the trajectory of initial ABMD simulations, from -200° to -110° at 10° intervals. Then umbrella sampling was performed for each window for 6 ns, with the reaction coordinate monitored at 2 ps intervals. WHAM analyses were then performed on the data points for the six consecutive nanoseconds.

For the second method, termed the iterative ABMD-US approach, the ABMD flooding potentials at the 10th ns were used as an initial estimate to the underlying free energy landscape. Then umbrella sampling was performed with the ABMD-generated structures together with the underlying bias. The umbrella-corrected free energy profile was then updated as the new biasing potential and a second iteration of umbrella sampling was performed with the new bias. Thus the free energy calculations were performed for five iterations, until which the mean umbrella correction is less than 1 kcal/mol.

6.3.3 Reaction rate approximations

According to Kramer-Smoluchowski reaction rate theory, the diffusion coefficient and free energy changes can be used to estimate the loop opening and closing rate [155]:

\[ k \approx D_{\text{max}} \frac{|W''_{\text{min}} W''_{\text{max}}|^{1/2}}{2 \pi k_B T} e^{(\frac{\Delta G^*}{k_B T})} \]  \hspace{1cm} (6.11)

\[ D_{\text{max}} = D(Q_i) = \frac{<\delta Q^2>}{\tau_i} \]  \hspace{1cm} (6.12)

where the decorrelation time, \( \tau_i \) was approximated by the following equation [156]:

\[ \tau_i = \frac{\int_0^\infty <\delta Q(t)\delta Q(0)> dt}{<\delta Q^2>} \]  \hspace{1cm} (6.13)

where \( D_{\text{max}} \) is the diffusion coefficient at the transition state; \( <\delta Q^2> \) is the variance of the reaction coordinate for the \( i^{th} \) window, and \( W''_{\text{min}} \) and \( W''_{\text{max}} \) are the second derivative of the potential energy surface at the bottom and top of the barrier, respectively.
The diffusion coefficient $D_{max}$ was calculated as the median diffusion coefficient from thirteen umbrella sampling trajectories, each of which is 500 ps in length.

The second derivative of energies were generated by fitting quadratic functions in the forms of $y = \frac{1}{2} kx^2 + bx + c$ to the potential energy surface, centered at the ground states and the transition state, respectively. The resulting $k$ was used as an estimate to the second derivative.

### 6.3.4 Comparisons with predictions from PLOP

Protein Local Optimization Program (PLOP) is known as an effective program for the prediction of loop conformations [157]. Previous studies on TIM loop conformation suggest that PLOP can be used to predict the closed loop conformations in absence of the ligand. To evaluate the quality of our ABMD-predicted structures in comparison to other computational methods, we also performed loop predictions using PLOP.

The open form of TIM in absence of the ligand was used as the initial conformation. Three iterations of loop predictions were performed both with and without the crystal contacts. The analytical generalized Born plus nonpolar (AGBNP) implicit solvent model was used to approximate the solution environment, whereas the SGBNP solvent was used to model the solvent conditions in the crystal packing environment [158]. Initial predictions were produced with 3 different overlap factors (0.55, 0.65, 0.75) and subjected to constrained refinement. The eight lowest energy structures were subject to further sampling using a Cartesian constraint of 4 Å on each of the loop backbone C-α atoms. These structures from the initial and refinement predictions were subject to the last iteration of refinement, using a tighter Cartesian constraint of 2 Å. The resulting structures were then examined and compared to the crystallographic closed loop conformation.
6.4 Results and discussion

6.4.1 Benchmarking study for ABMD simulations: cis-trans conversion of NMA

In this benchmarking study, we seek to answer two questions that are critical to the evaluation of the ABMD method in predicting free energy profiles. First, how accurate can the current force field predict the simple rotational energy landscape of NMA with respect to experimental measurements? Second, what kind of consistency, or error margin can we expect from the energy predictions given by free energy simulations (such as ABMD and umbrella sampling)?

To answer these questions, the ABMD potential recorded at 1 ns intervals and the corresponding torsional angle changes was examined for the 15 ns long simulation (Figure 6.2, 6.3). We observe that the flooding potential accumulated in the ground state for the first 3 ns. Next, the cis-conformation of NMA was extensively sampled between 3 to 6 ns. After 6 ns of flooding, the biasing potential started to converge. Additional insights for the convergence of ABMD potential were gained from Figure 6.4, where we calculated the error in the flooding potential for the last 7 ns of the ABMD simulations. Overall, the deviation between PMF estimates is bound between 1 and -1 kcal/mol. The absolute mean error fluctuates between 0.1 kcal/mol and 0.6 kcal/mol, whereas the maximum error along the reaction coordinate ranges from 0.2 kcal/mol and 0.8 kcal/mol. Notably, the error calculated from the last two ns of simulation still shows maximum error around 0.5 kcal/mol. Furthermore, we performed “umbrella correction” on the converged free energy estimate, and the calculated PMF is shown in Figure 6.5. Overall, the maximum correction is calculated to be around 0.4 kcal/mol, whereas the average correction is 0.2 kcal/mol. Taken together, the ABMD potential approaches convergence after 6 ns of flooding. More importantly, the margin of error for the “converged” ABMD potential is on the order of 0.5 kcal/mol.

We estimated the energy profile of NMA cis-trans conversion by averaging the energy profile from the last 6 ns of the simulations. The calculated rotational barrier and energy difference between trans and cis conformations are estimated to be 17.5
kcal/mol and 4.5 kcal/mol, respectively. In comparison, previous experimental study showed that the barrier of rotation is 21.2 kcal/mol starting from the trans conformation, and that the cis conformation is 2.4 kcal/mol less stable comparing to the trans conformation (water, 30°C) [159]. While ABMD flooding potential correctly reproduces the shape of the rotational energy landscape of NMA, the energy barrier was underestimated by 4.7 kcal/mol, whereas the energy difference between the two conformations was overestimated by 2.1 kcal/mol.

Taken together, our benchmarking study of NMA as a minimal model of peptide and proteins suggest that ABMD simulations are capable of giving qualitative predictions for minor conformations observed in experiments. The converged ABMD flooding potential suggests that the energy barrier is considerably underestimated, whereas the relative energy difference is considerably overestimated. Given the trivial size of the NMA system and the extensive conformational sampling performed in explicit solvent, it is likely that these discrepancies reflect the intrinsic inaccuracies of the force field.

6.4.2 Prediction of TIM excited state using ABMD simulations

The time-dependent evolution of ABMD potential was examined (Figure 6.3). For the apo form of TIM monomer, the biasing potential was observed to accumulate in the ground state (-190°) for the first 2 ns. After 4 ns of flooding the system diffused into the potential energy well corresponding to the closed conformation (-120°). At 10 ns the closed conformations were fully explored and the intermediate conformations were reflooded. Similarly for the bound form of TIM, the energy minimum corresponding to the open form can be observed after 10 ns. Hence our ABMD simulation suggest that the unligated TIM samples the closed conformations, whereas the ligated TIM samples the open conformations.

ABMD simulations generated an ensemble of quasi-equilibrated structures along the reaction coordinate. In order to evaluate the structures of minor conformers predicted by ABMD, RMSD analyses was performed on the ABMD trajectories. For apo TIM monomer, the loop RMSD with respect to the closed form of TIM crystal structure
decreased from 7.6 Å at 0 ns to 1.9 Å at 10 ns (Figure 6.6). In the mean time, the loop RMSD comparing to the closed TIM structure increased to over 6 Å. Similarly for the TIM dimer, the loop RMSD with respect to the closed loop crystal structure decreased from 6.0 Å to 2.5 Å within 12 ns of conformational flooding. However, the hydrogen bonds observed in the crystal structure, namely the hydrogen bonds between Ser 212 and Gly 174, as well as Tyr 209 and Gly 177 are broken in our predicted structures. This finding is not surprising, considering that these hydrogen bonds are frequently seen disrupted in our MD simulation of the ligated form of TIM, which are likely to be even weaker in the apo TIM dynamics. Alternatively, both hydrogen bonds appear to be solvent mediated (Figure 6.7).

To evaluate our predicted loop conformations, we also compared our structures to the structures generated from PLOP predictions, where the open form of TIM was used as the starting structure (Table 6.1) [160]. The PLOP based loop predictions were performed for the dimeric TIM both in solution and in the crystallographic environment, giving RMSDs around 2.8-3.0 Å, which are higher than our predictions (1.8 Å). Like the ABMD predicted structures, no strong hydrogen bonds characterized in the crystal structure were observed in the PLOP predictions. Thus the structures from ABMD predictions are comparable, or slightly better than that of the PLOP predictions. One weakness in the PLOP methodology is that it is hard to identify the true local minima amongst a number of predicted structures base on the relative energies, in case that the true structure is not known a priori. In comparison, ABMD overcomes this problem by identifying the minor conformation based on the flooding potential.

6.4.3 Comparison between four free energy methods for characterizing TIM loop motion

When we applied the standard umbrella sampling approach, we observed that umbrella sampling (US) alone is inefficient in achieving convergence. Our first US simulation, where all windows used the open structure as the starting conformation, showed time-dependent changes in the free energy profile for every 1 ns long simulation, and could not properly identify the minor conformations even after 8 ns of simulations.
(Figure 6.8). WHAM analyses gave the statistical error on the order of 0.01 - 0.02 kcal/mol using the Monte Carlo bootstrap estimates, but the magnitude of statistical error may not reflect the true error on the free energy estimates.

ABMD simulations can be used to generate an ensemble of structures along the reaction coordinate, as well as predict the minor conformations. However, the biasing potential of ABMD cannot provide an accurate error estimate for the free energy profile. When we applied "umbrella correction" approach to the monomeric TIM, we observed that due to the inaccuracy of the initial bias, TIM is often trapped in the local minima (Figure 6.9a). As a result, 60 ns of sampling statistics leads to large umbrella corrections on the order of 2 - 4 kcal/mol, which are comparable to the magnitude of the biasing potentials (Figure 6.9b). In addition, comparison between the ABMD potentials and corresponding umbrella corrections suggest that the potential energy wells may have been significantly over-flooded. In addition, previous studies showed that for a small peptide, the complete convergence of ABMD requires hundreds of nanoseconds of replica exchange based ABMD [161]. Therefore this approach is undesirable for the simulations of proteins and other biomacromolecules, which needs significantly more replicas and much longer simulations.

Our third method employs the flooded structures generated from ABMD to perform umbrella sampling. While the free energy profiles showed much less time-dependent behavior than the previous US simulation, there were still large deviations between consecutive PMF profiles (Figure 6.10). The size of the barrier varied from 3.19 kcal/mol to 5.38 kcal/mol, whereas the free energy difference between the ground state and excited state varied from 1.77 kcal/mol to 4.4 kcal/mol in twelve 1 ns long simulations performed sequentially. Averaging of all trajectories is feasible, but it is unclear whether the average serves as a better estimate.

Last but not least, the iterative ABMD-US approach was used to further drive the free energy towards convergence. For TIM monomer, the average umbrella corrections decreased from 1.76 kcal/mol to the 0.35 kcal/mol (Figure 6.11), and the maximum correction dropped from 3.3 kcal/mol to 0.78 kcal/mol after 5 iterations. Overlap of the four energy profiles suggest that both the "umbrella correction" method and the
non-iterative ABMD-US approach are not as effective as the iterative ABMD-US approach (Figure 6.12). In comparison, the profile generated from the standard umbrella sampling approach overlaps well with that of the iterative ABMD-US approach, only deviating in the metastable conformations (-130° - -110°) by 1.5 kcal/mol on average. This is likely due to the slow convergence of umbrella sampling shown in Figure 6.8. In other words, if umbrella sampling were carried out for an additional 10 - 20 ns, we are likely to see reasonable convergence between standard US and ABMD-US. More importantly, this observation raises questions to the quality of potential of mean force studies commonly seen in literature, where only a few hundred picoseconds worth of conformations were sampled.

6.4.4 Comparison with experimental study and analogous computational studies

Solid state NMR experiments were used to derive the population ratio between the open and close conformations of TIM in exchange with each other. Subsequently, the expected difference between the open and closed state was estimated to be between 1.2 and 2.5 kcal/mol [162]. In addition, the Arrhenius kinetics combined with an exchange rate of $3 \times 10^4$ s$^{-1}$ leads to an energy barrier on the order of 10 kcal/mol.

Hence the estimated free energy difference between the ground state and excited state (2 - 3 kcal/mol) is in excellent agreement with experiments. However, the energy barrier from our approximation is much lower than the experimental measurement. In order to correct for the distance-dependent diffusion rate, we estimated the reaction rate using the Kramer-Smoluchowski reaction rate theory described earlier. The transition state diffusion rate is estimated to be 0.0013 rad$^2$/ps, at -132.5° (Figure 6.13). The second derivative of the potential with respect to the reaction coordinate was approximated to be 72.7 kcal/mol/rad$^2$, 47.5 kcal/mol/rad$^2$ and 172.8 kcal/mol/rad$^2$ for the ground state, transition state and the excited state, respectively. Given an energy barrier of 5 kcal/mol, we estimated the reaction rate to be on the order of $10^7$ s$^{-1}$ for $K_{open}$ and $10^8$ s$^{-1}$ for $K_{close}$. Thus the calculated reaction rate is four orders faster than the experimentally measured rates ($10^3$ - $10^4$ s$^{-1}$). Nevertheless, this is consistent with an
independent computational study on DHFR, where the predicted transition appears to be much faster than the experimentally determined rate [163].

In our study of TIM loop conformational change, the energy barrier of TIM loop closing is estimated to be 5 - 6 kcal/mol, whereas the energy difference is 2 - 3 kcal/mol. Comparing to experimental measurements, our calculations overestimated the barrier by about 5 - 6 kcal/mol, and the energy difference is overestimated by 1 - 2 kcal/mol, which is very similar to the discrepancy we observed in the benchmarking study of cis-trans conversion of NMA. Interestingly, despite the huge difference in system size, the deviations in TIM loop energy profile and NMA torsional energy profile from experimental studies are remarkably similar, in terms of both sign and magnitude. Hence the inaccuracy of the molecular mechanics based force field is likely responsible for the inaccuracy in the predicted energetics. Furthermore, we may expect an error margin of 5 - 6 kcal/mol in energy barrier and 1 - 2 kcal/mol from our free energy calculations of small and large biological systems.

6.4.5 Conclusions

We showed that the minor conformations of TIM can be predicted and characterized via computer simulations. In good agreement with experimental dynamics studies, our studies showed that TIM samples functionally important conformations in its apo and bound forms. Structural analyses suggest that the ABMD-predicted metastable conformations are in good agreement with the crystal structures. The observed differences likely reflect the difference between crystal packing and solution environment. Thus our result serves as further evidence to the “conformational selection” theory of protein dynamics. In terms of thermodynamics, our calculations are in good agreement with the experimental estimates. The comparison of kinetics shows that the predicted transition rate is orders of magnitude faster than the observed rate[163]. Yet the same trend was observed in other protein conformational studies using a different force field, which suggest that molecular mechanics method may have systematically underestimated the energy barrier.

Our study also shows a new approach of free energy calculation by combining
ABMD simulations and umbrella sampling calculations. ABMD offers the unique advantage of efficiently exploring unknown transition pathways and corresponding metastable states, which may have been “invisible” to experimental probes, whereas many theoretical studies of protein conformational change rely on the prior knowledge of all stable conformations and a predetermined transition pathways. ABMD is not commendable when it comes rigorous energy estimates, as it tends to overestimate barrier size and underestimate free energy differences between local minimum for large biological systems like TIM. However, the iterative ABMD-US approach can help us converge errors to around 0.5 kcal/mol within a few iterations. In comparison to other free energy methods discussed here, ABMD-US method offers the unique advantage of faster convergence and accurate estimates.

In summary, our study show that novel computational methods such as ABMD can be used to gain insights into protein conformational changes beyond experimental limitations. These methods may evolve into popular strategies for predictions of “invisible” proteins conformation, provided that suitable order parameters can be obtained via normal mode analysis or principal component analysis a priori. In the future, our methods could potentially be used to derive of functionally important metastable state in collaboration with experimental studies (NMR and single molecular fluorescence spectroscopy). It is quite promising that predictions made from our strategy may guide experimentalist to some interesting discoveries in functionally important protein dynamics.
Table 6.1: Summary of RMSD for plop predicted closed structures.
The loop RMSDs are measured in Å. The relative energies are measured in kcal/mol. The energy of the open structure is referenced as the zero point. The 1YPI data (with asterisks) are provided by Dr. Victor Guallar.

<table>
<thead>
<tr>
<th>Src</th>
<th>loop RMSD from closed</th>
<th>loop RMSD from open</th>
<th>ofac</th>
<th>energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>6TIM soln</td>
<td>2.95</td>
<td>5.01</td>
<td>0.75</td>
<td>1.05</td>
</tr>
<tr>
<td>6TIM xtal</td>
<td>2.74</td>
<td>4.04</td>
<td>0.55</td>
<td>11.76</td>
</tr>
<tr>
<td>1YPI soln*</td>
<td>2.88</td>
<td>1.05</td>
<td>0.70</td>
<td>-1.58</td>
</tr>
<tr>
<td>1YPI xtal*</td>
<td>similar</td>
<td>similar</td>
<td>0.70</td>
<td>~8</td>
</tr>
</tbody>
</table>

Figure 6.1: cis-trans transformation of NMA
The trans- (left) and cis- (right) conformations of NMA are rendered in ball-and-stick forms.
Figure 6.2: Torsional angle sampling in ABMD simulations of NMA
The torsional angle of NMA backbone sampled in 15 ns is shown in radians.

Figure 6.3: The evolution of ABMD potential for NMA
The ABMD potentials for the first 10 ns are shown in 1 ns intervals. To make all potentials directly comparable, the energy of the lowest potential in each profile is set to 0.
Figure 6.4: The convergence of ABMD potential for simulations of NMA cis-trans conversion

The deviation of the ABMD potentials in the last 7 ns of the simulation was evaluated. More specifically, the deviation was calculated as the difference in energy between flooding potentials that are taken 1 ns apart from each other.
Figure 6.5: The umbrella correction to ABMD potential for NMA simulation

The biasing potential was taken from the 15th ns of the ABMD flooding simulation. The potential of mean force was calculated from 36 windows evenly spaced along the torsional angle of NMA. Each window includes a 10ns long trajectory.
Figure 6.6: Comparison between predicted loop and crystal structures of apo TIM. The left hand side shows the RMSD between 8th and 12th ns of ABMD simulation with respect to the open (black) and closed (red) form of TIM crystal structures. The right hand side shows the ABMD snapshot at the 10th ns overlapping with the crystallographic loop structures. The RMSD of the predicted closed structure is 1.8 Å away from the crystal structure.

Figure 6.7: Hydrogen bonding network in the predicted “closed” loop structure. In this particular snapshot, the OH (Tyr 209)...O(Water) distance is 2.24 Å; The NH (Ala 174)...O(water) is 1.93 Å; The OH (Ser 212) forms a 1.93 Å hydrogen bond with a water molecule, which also interacts with the carbonyl group of Gly 174.
Figure 6.8: Time dependence of potential of mean force using standard umbrella sampling.
The 8 lines shows the PMF derived from eight consecutive, 1 ns long umbrella sampling simulations.
Figure 6.9: ABMD potential and umbrella corrections of apo TIM
The black line and the green line are the negative of ABMD potential and the corresponding umbrella corrections, respectively. The error bars on the green line show the 95% confidence interval of the umbrella corrections. The red line shows the sum of ABMD potential and the umbrella correction terms.
Figure 6.10: Time-dependence of PMF using umbrella sampling based on ABMD structures.
The 6 lines shows the PMF derived from eight consecutive, 1 ns long umbrella sampling simulations.

Figure 6.11: Convergence of errors in the ABMD-US approach
The top figures show the error convergence of the TIM monomer. The bottom figure shows how the average correction decreases over the five iterations.
Figure 6.12: Comparison between the free energy profiles calculated from different free energy methods
The red line gives the corrected ABMD potential, whereas the black line represents the potential of mean force obtained from WHAM analysis of umbrella sampling simulations.
Figure 6.13: The diffusion coefficient of apo TIM monomer along the reaction coordinate.
The transition state is estimated to be between -140° and -130° of the pseudo-dihedral angle.
Chapter 7  
Why is TIM only active as a dimer? A study of energy and dynamics of TIM monomer vs. dimer

Protein oligomerization is critical for the proper function of a wide range of proteins both in vitro and in vivo. Triosephosphate isomerase is known as an enzyme with high catalytic proficiency existing only in the dimeric form. While the two subunits of TIM show no cooperativity, their dissociation lead to total loss of catalytic activity. A number of biophysical and biochemical studies have aimed to understand this phenomena, but a comprehensive explanation was never given. In this study, we are able to give a complete description to the energy and dynamics of monomeric and dimeric TIM via ABMD simulations and conformational analyses. By comparing the behavior of monomeric and dimeric TIM, we showed that the lack of activity of TIM monomer can be mostly attributed to the increase in allosteric flexibility in the interface loops, but not the trivial changes in the active site dynamics.

7.1 Introduction

In the cellular environment, the majority of the proteins exist as protein oligomers [164]. Protein oligomerization is indispensable for the functional control of proteins, providing structural support, allosteric regulation and an additional level of functional flexibility [165]. The study of mechanism and role of protein polymerization are expected to give insights into protein-protein interactions and thereby elucidate the association between protein dynamics and function.

A good example of protein oligomerization is given by triosephosphate isomerase (TIM), which is a dimeric glycolytic enzyme with two identical subunits [50]. Each subunit of TIM consists of a protein scaffold called the TIM barrel, where eight parallel
β-sheets are succeeded by eight α-helices. There are eight loops (loop-1 to loop-8), each of which is preceded by a β-sheet. The interface loops include loop-1, loop-2, loop-3, loop-4, which directly interact with the neighboring unit at the dimeric interface. Among these four interface loops, loop-3 is the longest interface loop (15 residues) and protrudes from the interface to dock into a groove between loop-1 and loop-4 of the other monomeric subunit (Figure 7.1). Loop-6 is the active site loop, which extends into the solution in its open form and closes over the active site when ligated.

Previous studies showed that the TIM is only active in its dimeric form [166]. Nevertheless, the catalytic site of one monomer is observed to be fully active even when it’s bound to another monomer with a mutated, non-functional active site [167, 168]. This finding suggests that the active sites of the two respective subunits are completely independent and have no cooperativity. To better understand why TIM has to function as a dimer while the subunits are catalytically independent, a number of experimental and theoretical studies have been carried out to evaluate the effect of dimerization on TIM structure and dynamics. First, a number of site-directed mutagenesis studies have been performed to induce activity in monomeric TIM, or abolish activities of dimeric TIM [169, 170]. In a group of studies by Wierenga and co-workers, they were able to construct a moderately active form of TIM monomers by a 7-residue deletion of loop-3 [171, 170, 172, 173, 174, 147, 175]. In the mean time, an independent investigation also showed that the perturbation of the interaction between loop-1 (Cys15) of one subunit and loop-3 of the other subunit can reduce the activity of dimeric TIM by 50% [169]. All these experimental evidences point to an important role of allosteric dynamics in TIM catalysis. Nevertheless, these biochemical studies rely on point mutations and static crystallographic structures as probes of protein structure and dynamics, without offering a comprehensive interpretation for the dynamic effect of dimerization.

Theoretical studies are expected to complement these experimental evidences, when molecular simulations are appropriately performed and correctly interpreted. A recent computational study by Cansu and Doruker seeks to understand the difference between monomeric and dimeric TIM via nanosecond time scale MD simulations [176]. In this study, they suggested that the difference in catalytic activity between
monomeric and dimeric TIM is largely due to the difference in the first essential mode obtained from principal component analyses. Furthermore, they observed that almost all helices and loops show higher fluctuation in dimeric form, except for loop-1 and a slight broadening in mobility in loop-3. There are at least two caveats in this theoretical study: First, it is doubtful whether the first essential mode from a nanosecond time scale MD can faithfully reflect behavior in a much longer time scale (µs-ms) computational study. The study of DNA base flipping as well as our PCA analysis on NMA simulation (see Chapter 8) have both suggested that the long-term behavior are usually not reflected in the first ten eigenmodes from a short MD simulations [177]. Second, increase in loop flexibility in their dimeric TIM simulation is counterintuitive and goes against all experimental evidences.

In this study, we performed systematic investigations for the effect of dimerization on TIM, by assessing the free energy landscape and residue-level dynamics of TIM conformational change on regular and accelerated time scales. To characterize the free energy profile of TIM monomer and dimer, we combined ABMD simulations and umbrella sampling (see Chapter 6) and mapped out the converged free energies along the designated reaction coordinate. To determine whether certain residues are critically affected by TIM dimerization, we compared the backbone and sidechain flexibility observed in three ABMD simulations for wild type monomeric TIM, dimeric TIM and the active mutant monoTIM, respectively [170]. If certain residues exhibit high similarity between dimeric TIM and monoTIM but significant difference from monomeric TIM, they are deemed as major contributing factors to the lack of activity of TIM following dimer dissociation. Similarly, when residues in monoTIM shows contrasting behavior to wildtype TIM monomer and dimer, they are considered irrelevant to TIM dimerization even if they may have direct contributions to TIM catalysis. Overall, we seek to establish a causal relationship between energy, dynamics and function of different TIM variants, instead of relying solely on the static structure from crystallographic sources.
7.2 Methods

To explore how dimerization affect TIM energy and dynamics, we studied three systems of TIM variants, including monomeric TIM (pdb code: 6TIM), dimeric TIM (pdb code: 6TIM), and monoTIM (pdb code: 1ML1) [170]. The molecular flexibility on nanosecond and accelerated time scales were characterized by MD and ABMD simulations, respectively. The potential of mean force of TIM loop closure was calculated using the ABMD-US approach as illustrated in Chapter 6. In addition, B-factor and geometric analysis were performed for a number of residues using PTRAJ program of AMBER11 package [111].

7.2.1 Unbiased MD simulations for TIM monomer, dimer and monoTIM

Each of the three systems was solvated in a truncated octahedral box of TIP4P water molecules, where each atom is at least 10 Å away from the solvent boundary. The calculation protocol included the periodic boundary conditions, a time step of 2 fs, non-bonded cut-off at 10 Å, and Particle Mesh Ewald (PME) for treatment of long range electrostatics. Initially, the minimized model systems were heated to 300 K and equilibrated for 50 ps using the canonical (NVT) ensemble, followed by 2.95 ns long density equilibration using the isothermal-isobaric (NPT) ensemble. For each MD simulation, the trajectory were recorded with snapshots at 5 ps intervals for further analyses.

7.2.2 ABMD simulations for TIM monomer, dimer and monoTIM

The starting points of ABMD simulations were the equilibrated structures from the 3 ns long regular MD simulation described above. Langevin dynamics were applied with a collision frequency of 1.0 s\(^{-1}\), with snapshots taken at 5 ps intervals and the sampled reaction coordinate (pseudo-dihedral angle) monitored and recorded every 0.2 ps. All ABMD simulations were performed with the flooding time scale \(\tau_F = 100\) ps and the flooding resolution \(4\Delta \xi = 4^\circ\). The resulting trajectories of TIM monomer, dimer and monoTIM are 26 ns, 33 ns and 29 ns long, respectively.

The ABMD-US approach were then applied to TIM monomer and dimer. The
ABMD flooding potentials at the 10th ns and 21st ns were used as initial estimates to the underlying free energy landscape for TIM monomer and dimer, respectively. Then umbrella sampling was performed with the ABMD-generated structures together with the underlying bias. The umbrella-corrected free energy profile will then be updated as the new biasing potential and a second iteration of umbrella sampling was performed with the new bias. In this way, multiple iterations were employed until the mean umbrella correction is less than 1 kcal/mol. Five and four iterations of ABMD-US simulations were performed for TIM monomer and dimer, respectively.

7.2.3 Trajectory analyses

For each trajectory, an RMS fit to the first structure was performed for all backbone atoms except the three residues from the N- and C- terminals. In the case of dimeric TIM, RMS fit was only performed for the first subunit where the conformational change took place. The deviation between the loop structures were calculated from Tyr 165 to Val 176. To understand the extent of fluctuation for individual residues, the average mass-weighted B-factor were then calculated for all residues of TIM. Apart from the critical catalytic residues, the additional residues of interest were identified as the residues showing the most significant differences between monomeric and dimeric TIM in the B-factors. The backbone and sidechain movements of these residues were described by mapping their $\phi - \psi$ and $\chi_1 - \chi_2$ space sampled during ABMD simulations.

7.3 Results

7.3.1 Monomer vs dimer based ABMD-US simulations

ABMD simulations generated an ensemble of quasi-equilibrated structures along the reaction coordinate. In order to evaluate the structures of minor conformers predicted by ABMD, RMS analyses was performed on the ABMD trajectories. For apo TIM monomer, the loop RMSD with respect to the closed form of TIM crystal structure decreased from 4.9 Å at 0 ns to 1.4 Å in 5 ns (Figure 7.2), In the mean time, the loop
RMSD comparing to the closed TIM structure increased to over 4 Å. Similarly for the TIM dimer, the loop RMSD with respect to the closed loop crystal structure decreased from 6.0 Å to 2.5 Å within 10 ns of conformational flooding (Figure 7.3). However, the hydrogen bonds seen in the crystal structure, namely the hydrogen bonds between Ser 212 and Gly 174, as well as Tyr 209 and Gly 177 cannot be observed from our predicted structures. This finding is not surprising, considering that these hydrogen bonds are frequently seen disrupted in our MD simulation of the ligated form of TIM, which are likely to be even weaker in the apo TIM dynamics. Alternatively, both hydrogen bonds appear to be solvent mediated (Figure 6.7).

We then used the iterative ABMD-US approach to drive the free energy to convergence (Figure 7.4). For TIM monomer, the average umbrella corrections decreased from 1.76 kcal/mol to the 0.35 kcal/mol, and the maximum correction dropped from 3.3 kcal/mol to 0.78 kcal/mol after 5 iterations. Similarly, for TIM dimer, the average and maximum corrections were reduced to 0.4 and 0.73 kcal/mol, respectively, after 4 iterations.

7.3.2 Free energy profile of TIM loop closure and predicted metastable state

The converged free profiles show that the monomeric form of TIM visits two conformations (pseudo-dihedral angle = -190° and -125°) in the open state, separated by a small barrier (Figure 7.6). The energy minimum corresponding to the closed form is centered at -125°, with 4.3 kcal/mol higher energy than the ground state. The transition state emerges at -133°, which is only 0.48 kcal/mol higher than the metastable closed form. More importantly, the crystal structure of the ligated TIM showed a pseudo-dihedral angle of -132.9°, which corresponds to the least stable point out of the entire free energy landscape. This suggests that the lack of loop rigidity in monomeric TIM may have shifted the energy landscape, so that the catalytically important conformation becomes unfavorable. The alternative loop conformational minimum (-125°), which is 9° away, may be clamped too tight or too close to the rest of the protein to accommodate the desired substrates or to perform catalysis. In comparison, the dimeric TIM shows a different free energy landscape. Not only is the minor conformation
characterized by a broader well, but also both the transition state and the minor conformation have shifted. The predicted minor conformation is 0.7 kcal/mol lower in energy compared to the monomeric TIM. Furthermore, the closed conformations is centered at -135°, which is very close to the -133° angle seen in the crystal structure. The corresponding transition state was shifted to -150°, which is 18° larger than the TIM monomer. In addition, the energy well corresponding to the ground state becomes narrower in the dimerized TIM. It is also noted that the center of distribution for the major conformation has also shifted by 5 degrees (from -190° to -195°). This loop conformation with a wider cleft may also promote the chance of ligand binding.

7.3.3 Effect of dimerization on B-factors

The wildtype TIM functions as a homodimeric enzyme, where the two subunit are tightly bound to each other via the protruding interface loops. The major difference between TIM monomer and dimer is not in their ground state conformations, but the size of buried surface, which could be linked to protein stability [178]. In order to fully understand the effect of dimerization on the catalytic activity of TIM, we examined the dynamics of TIM monomer and dimer.

Evaluation of B-factors from our ABMD simulations highlights these key differences. Three regions in the dimer sequence shows significantly lower flexibility in comparison to the monomeric TIM. The three regions are loop-1 (residue 11-16), loop-3 (residue 64-77) and helix-6 (residue 129-135) (Figure 7.7). A close inspection of the these regions showed that helix-6 is adjacent to the active site loop-6, whereas loop-1 and loop-3 are 15 to 30 Å away from loop-6. The average B-factor for loop-3 is 8.9 Å² on average from the dimer simulation, but was increased to 167.3 Å² in the monomer simulation. Similarly, the average B-factor from loop-1 was decreased from 29.6 Å² to 12.1 Å² when dimerized. In addition, visualization based on our principal component analyses suggested that the allosteric loop movements are tightly coupled to the loop-6 motion in TIM monomer, but not for TIM dimer. In addition, we examined the structural flexibility of monoTIM (PDB code: 1ML1), which is engineered to be a stable, catalytically active monomer [175]. We observed remarkable similarity in the
residue-dependent fluctuations between the dimeric TIM and the mutant monoTIM: Like the dimeric TIM, the high loop flexibility in TIM monomer from loop-1, loop-3 and loop-6 is quenched in the monoTIM dynamics. Our observations may also serve as further explanation to previous studies where TIM was engineered in loop-1 or loop-3 to produce stable, active monomers [174, 175].

In order to evaluate the difference in dynamics on short versus long time scale, we also compared the loop fluctuations from regular MD to that from the ABMD simulations (Figure 7.8). In general, the accelerated time scale fluctuations are higher in magnitude compared to the short time scale ones. One prominent difference occurred in helix-6 of TIM monomer, where high flexibility only manifested itself on the accelerated time scale. In comparison, this behavior is not observed in TIM dimer. These observations suggest that dimerization of TIM affects its dynamics on both ns and longer time scales. Taken together, our study suggest that increase in allosteric loop rigidity (especially in loop-1 and loop-3), whether through dimerization or site-directed mutagenesis, is critical to the catalytic competence of TIM. Therefore, our observations strongly disagrees with Cansu and Drouker’s study on the dimerization effect of TIM [176]. While our study showed reduced flexibility in the overall protein both on nanosecond and longer time scales for TIM dimer, their study suggested more fluctuations in the dimeric form overall. Furthermore, their study showed comparable loop movements for monomeric and dimeric TIM in both the loop-1 and loop-3 regions, which could be viewed as a considerable aberration from experimental observations.

### 7.3.4 Effect of dimerization on catalytic residues

There are four critical catalytic residues located in the active site, which are directly involved in the catalytic activities of TIM, including Asn 11 (loop-1), Lys 13 (loop-1), His 95 (loop-4) and Glu 167 (loop-6). Among these residues, the carboxylate group of Glu 167 functions as a base for proton abstraction of DHAP from C-1. Similarly, His 95 acts as an acid for proton donation to the carbonyl group attached to C-2 of GAP. In the reverse reaction, Glu 167 abstracts the proton from C-2 and His 95 donates the proton
to C-1 carbonyl group. Furthermore, the proton shuffling of substrate is facilitated by two oxyanion holes. The first oxyanion hole is formed by NZ of Lys 13 and NE of His 95 around the keto-oxygen O2 and facilitates proton exchange from C-1. The second oxyanion hole is formed by NE2 of His 95 and ND2 of Asn 11, where a proton between C-2 is exchanged. In addition, Lys 13 is known to be important for substrate binding due to its positively charged sidechain [179, 180, 181]. In order to understand why dimerization of TIM is critical to catalytic competency of TIM, we analyzed the backbone and sidechain dynamics of these four individual residues during ABMD simulations for TIM monomer, dimer and the monoTIM mutant.

We first examined and compared the backbone dynamics of the four residues of interest (Figure 7.11, 7.12, 7.13, 7.14). The $\phi-\psi$ space of Asn 11 sampled in the monomer and dimer simulations are both roughly centered around the ($\phi, \psi$) angle of (-90°, 120°) with considerable overlap, suggesting that the backbone dynamics of Asn 11 are analogous for TIM monomer and dimer. In comparison, the monoTIM simulations is centered around (-90°, 90°), while showing a different conformational space from the two systems with a downward drift in the $\psi$ angle. For Lys 13, the $\phi-\psi$ map of TIM monomer and dimer are also comparable, with the $\phi$ angles in the range of (15°, 80°), and $\psi$ angles between (-180°, -50°). The Lys 13 from monoTIM on the other hand, samples a discrete conformational space with zero overlap. In this case, the $\phi$ and $\psi$ angles sampled are in the ranges of (-163°, -43°) and (-75°, -43°), respectively. His 95 of monoTIM shows considerably larger backbone flexibility, sampling two conformations centered at (-55°, 130°) and (-130°, 120°), whereas TIM monomer and dimer sampled nearly identical spaces centered at (-60°, 150°). Similar to Asn 11 and Lys 13, the backbone dihedral spaces of Glu 167 show great resemblance for TIM monomer and dimer, but differ significantly from that of monoTIM.

The sidechain dynamics of the four catalytic residues were also evaluated by mapping the $\chi_1 - \chi_2$ conformational spaces (Figure 7.15, 7.16, 7.17, 7.18). For Asn 11, the $\chi_1 - \chi_2$ space of all three systems are very similar. In the case of Lys 13, it is observed that the TIM dimer samples only one conformations of $\chi_1$, whereas TIM monomer and monoTIM samples two and three conformations, respectively. This observation
suggest that TIM dimer has the most rigid sidechain, while the same sidechain of monoTIM is the most flexible. His 95 sidechains show similar behavior to its backbone, where monomer and dimer TIM samples almost identical spaces in contrast to the much wider sampling range of monoTIM. For Glu 167, four accessible sidechain conformations were observed: TIM dimer and monoTIM samples three conformations each, two of which are very similar. TIM monomer samples all four conformations.

In addition, two stabilizing interactions conserved in the dimerized TIM active site were also analyzed. The first interaction is described as a salt bridge formed between Lys 13 and Glu 97, the latter of which is a completely conserved residue among more than thirty TIM sequences. We calculated the average length of this salt bridge along with their standard errors from the ABMD simulations, as well as the occupancy with a 4 Å cutoff (Table 7.2). TIM dimer shows a strong salt bridge (3.12 Å) between Lys 13 and Glu97 throughout the simulation as expected (99.7% occupancy). In the meantime, TIM monomer shows a salt bridge of comparable strength with slightly longer distance (3.32 Å) and lower occupancy (98.0% occupancy). In contrast, this salt bridge is not well conserved in monoTIM, as the salt bridge is only present 21% of the time and the average distance is as large as 7.66 Å. The other important interaction entails a hydrogen bond between His 95 and Glu 97, which properly positions the histidine side chain for its catalytic function (Table 7.1). This hydrogen bond is present in all three simulations, even though the monoTIM hydrogen bond is still relatively the weakest of all three, given a 65.9% occupancy compared to over 98% occupancy for both TIM monomer and dimer.

Hence the backbone conformational space of TIM monomer and TIM dimer are comparable for all the four catalytic residues, whereas that of monoTIM shows little or no overlap with either system. In terms of sidechain conformations, oftentimes the dimer sidechains are the most rigid while the monoTIM sidechains are the most flexible. As for the conserved salt bridge and hydrogen bond, the interactions observed in monoTIM are significantly weaker compared to that of TIM monomer and dimer. As we recall, TIM monomer is completely inactive due to the dissociations of subunits, whereas monoTIM retains 0.1% of the catalytic activity of TIM dimer [170]. Given the
similarity in backbone dynamics between TIM monomer and dimer, it is most likely that dimerization alone does not have a significant impact on the dynamics within the active site. Considering that TIM monomer shows intermediate sidechain flexibility with respect to TIM dimer and monoTIM, we also suggest that the sidechain rigidity in the active site is not an absolute requirement for catalytic competence. Similarly, the destabilization of the aforementioned conserved interaction does not abolish catalytic activities of TIM. Taken together, we conclude that dimerization, while critical for the catalytic activity of TIM, does not have a significant impact on the backbone and sidechain dynamics and certain conserved interactions in the TIM active site.

7.3.5 Effect of dimerization on non-catalytic residues

In the previous section, we observed that the dissociation of TIM subunits does not significantly affect the dynamics of TIM active site. Hence the changes in allosteric dynamics are most likely responsible for the catalytic proficiency of TIM upon dimerization. Consequently, we evaluated and compared the allosteric displacement and conformational dynamics of TIM from the ABMD simulations.

In section 7.3.3, the B-factor analyses showed considerable difference in backbone flexibility between TIM monomer and dimer, which suggest that the dimerization affects a number of allosteric regions, including residue in loop-1 (residue 16-18), helix-1 (residue 19, 27-29), the turn succeeding helix-1 (residue 30-31), loop-3 (residue 68-77), loop-5 (residue 129-137), loop-6 (residue 171-176) and loop-7 (residue 214-216). Among these regions, the most significant dimerization-induced rigidity are reflected in loop-1, loop-3, loop-5 and loop-6. For each of these regions we select one representative residue to illustrate how dimerization affects allosteric dynamics of TIM as well as the active site loop (loop-6) motion. The residues selected for our dynamics studies are: Gln 19 (loop-1/helix-1), Ala 73 (loop-3), Glu 135 (loop-5), and Gly175 (loop-6).

For Gln 19, we observe that a much larger $\phi - \psi$ space was sampled for TIM monomer compared to the other two systems, with the $\phi$ angle in the range of (-35°, -161°) and $\psi$ angle in the range of (-72°, -115°) (Figure 7.19). In the case of Ala 73 backbone, both monoTIM and TIM dimer adopts a single conformation, centered at
(-68°, -34°) and (-76°, 75°), respectively (Figure 7.20). In comparison, TIM monomer samples a much wider φ − ψ space, where the backbone adopts two to three different conformations, with the φ angle varying from -175° and -50° and ψ angles adopting two states centered around -150° and -50°. The behavior of Gly 175 backbone and sidechain of Glu 135 are very similar: The TIM dimer and monoTIM samples similar conformational space, while TIM monomer shows much larger conformational flexibility (Figure 7.21, 7.22). Taken together, we observed that the TIM monomer tend to sample multiple backbone conformations, while TIM dimer and monoTIM sample much more restricted φ − ψ spaces. Thus it is most likely that dimerization enhances the conformational rigidity of backbone and sidechain of allosteric regions, as well as a hinge residue (Gly 175) of the active site loop.

7.3.6 Difference in dynamics between TIM monomer, dimer and monoTIM

In the study of Borchert et al, they were able to characterize three new crystal structures of monoTIM variants and compare their structure to that of dimeric TIM [172]. They further concluded that the residual activity of monoTIM is due to the ability of catalytic residues to adopt similar conformations as wtTIM, despite the significantly higher conformational flexibility in the loop-1 and loop-4 regions. Their study, while insightful, did not include the wildtype TIM in the monomeric form for which no crystallographic structure exists. Our computational study, on the other hand, overcomes this challenge by including the simulations of monomeric wildtype TIM. In this way, we were able to distinguish the essential and non-essential residues affected by TIM dimerization. Our comparative analyses of residue-level dynamics leads to the following observations: First, our study suggests that the reason that monoTIM shows residual activity is not its ability to adopt similar conformations as wtTIM. This argument is supported by our observation that all the catalytic residues in monomeric TIM adopt conformations very similar to that of dimeric TIM (more so than monoTIM), despite its complete loss of activity. Furthermore, the conformational space sampled by dimeric TIM in in loop-3 and loop-5 are all accessible to monomeric TIM. Instead, the main reason for the complete loss of catalytic activity upon subunit dissociation is the
enhancement of allosteric flexibility in loop-3 and loop-5 regions. In other words, the residual activity of monoTIM is due to the retention of a certain rigidity in these regions. In contrast, while there is an observable decrease in backbone rigidity for loop-1 of monoTIM with respect to dimeric TIM, this change is still relatively small compared to that of monomeric TIM. For residues in loop-4, the backbone flexibility of monoTIM is even higher than that of monomeric TIM, suggesting that this moderate increase in loop flexibility is not the most relevant to TIM activity.

7.4 Conclusion

In this study, we seek to understand why triosephosphate isomerase is only active as a dimer, even though the active sites in the respective subunits show no cooperativity. While previous studies suggested a link between loop flexibility and TIM activity, we were able to evaluate the direct impact of dimerization on the free energy landscape of TIM. We showed that the effect of dimerization on the catalytic residues are not significant enough to abolish catalytic activities. Instead, the residues in loop-3 and loop-5 segments of TIM have a significant allosteric effect on TIM dynamics upon dimerization. Increasing the rigidity of these regions, whether through dimerization or site-directed mutagenesis, can help shape the free energy landscape for the desired catalytic functions. The dimerization of TIM leads to a more stable “closed” conformation in the excited state and a more “open” structure in the ground state. Thus, our study points to the global, concerted dynamics of proteins, which could be significantly modified with the slightestest perturbation at some remote allosteric sites.

Overall, our investigation suggests oligomerization as an important mechanism for regulating protein dynamics and functions. Additionally, we showed that the key differences in protein dynamics can be uncovered via sequence-dependent B-factor analyses combined with ABMD simulations, which may have applications in knowledge-based protein engineering. All this knowledge could potentially lead to abundant opportunities for protein engineering and structure-based drug design.
<table>
<thead>
<tr>
<th></th>
<th>His 95 ND1.. GLY97NH</th>
<th>Std Err</th>
<th>% Occupancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIM monomer</td>
<td>2.38 Å</td>
<td>0.004 Å</td>
<td>98.0%</td>
</tr>
<tr>
<td>TIM dimer</td>
<td>2.23 Å</td>
<td>0.002 Å</td>
<td>99.7%</td>
</tr>
<tr>
<td>monoTIM</td>
<td>2.86 Å</td>
<td>0.011 Å</td>
<td>65.9%</td>
</tr>
</tbody>
</table>

Table 7.1: His 95 - Gly 97 hydrogen Bond in TIM monomer, dimer and monoTIM
hbond cutoff 3.0 Å

<table>
<thead>
<tr>
<th></th>
<th>Lys13 N-HZ ... Glu 97(OE)</th>
<th>Std Err</th>
<th>% Occupancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIM monomer</td>
<td>3.32 Å</td>
<td>0.006 Å</td>
<td>80.4%</td>
</tr>
<tr>
<td>TIM dimer</td>
<td>3.12 Å</td>
<td>0.004 Å</td>
<td>85.9%</td>
</tr>
<tr>
<td>monoTIM</td>
<td>7.66 Å</td>
<td>0.017 Å</td>
<td>21.5%</td>
</tr>
</tbody>
</table>

Table 7.2: Lys 13 - Glu 97 salt bridge in TIM monomer, dimer and monoTIM
salt bridge cutoff 4.0 Å

Figure 7.1: Loop interactions in TIM dimer interface
The loop-1 and loop-4 of the first subunit of TIM are colored in green and orange, respectively. The loop-3 of the second subunit of TIM is colored in red. Loop-3 is positioned for direct interaction with loop-1 and loop-4.
Figure 7.2: RMSD profile of monomer ABMD simulations
Figure 7.3: RMSD profile of dimer ABMD simulations

Figure 7.4: Convergence of errors in the ABMD-US approach
Top: the umbrella correction converges over four iterations. Bottom: the average and maximum corrections decreases over the iterations.
Figure 7.5: The evolution of ABMD potential of TIM monomer and dimer
The left plot shows the accumulation of ABMD potentials of the apo monomer TIM from 2-18 ns at 2 ns intervals. The right plot shows the ABMD potential of TIM dimer from 2-26 ns in 2 ns intervals.
Figure 7.6: Comparison between converged free energy profiles of TIM monomer and dimer.
Figure 7.7: B-factor comparisons among TIM proteins
All set of b factors were derived from 10 ns long trajectories of ABMD simulations. Prior to calculating atomic fluctuations, the protein were fitted to the first snapshot using all the backbone atoms except for the terminal residues.
Figure 7.8: B-factor comparisons for short and accelerated time scale dynamics. The figures on top and bottom describes TIM dynamics for monomer and dimer, respectively.
Figure 7.9: Comparison of pseudo-dihedral angle space sampled from ABMD for monomer vs. dimer.
Figure 7.10: Comparison of RMS space for monomer vs. dimer.
Figure 7.11: Sampled backbone conformational space of Asn 11
Figure 7.12: Sampled backbone conformational space of Lys 13
Figure 7.13: Sampled backbone conformational space of His 95
Figure 7.14: Sampled backbone conformational space of Glu 167

Figure 7.15: Sampled sidechain conformational space of Asn 11
Figure 7.16: Sampled sidechain conformational space of Lys 13
Figure 7.17: Sampled sidechain conformational space of His 95
Figure 7.18: Sampled sidechain conformational space of Glu 167
Figure 7.19: Sampled backbone conformational space of Gln 19
Figure 7.20: Sampled backbone conformational space of Ala 73
Figure 7.21: Sampled sidechain conformational space of Gly 135
Figure 7.22: Sampled backbone conformational space of Gly 175
Regular ABMD simulations can be used to predict the metastable states of proteins when the order parameters can be fully characterized by a few degrees of freedom in the protein structures. Nevertheless, most protein conformational changes incorporate large scale, coordinated motions, the details of which may not have been known. Hence we have implemented the adaptively biased essential molecular dynamics (ABEMD), which combines ABMD simulations with the use of essential coordinates derived from principal component analyses of fast time scale MD simulations. We successfully applied ABEMD simulations to the cis-trans conformational change of NMA and the TIM loop conformational change. We show that ABEMD is comparable to the regular ABMD in terms of efficiency and accuracy, and may become a powerful approach for studying protein slow motions when the reaction coordinates are not known \textit{a priori} or cannot be easily specified.

### 8.1 Introduction

Previous studies have shown that the essential dynamics analysis of short MD trajectories can be used to predict slow structural transitions crucial to protein functions on a longer time scale [182]. The other reason that makes essential dynamics appealing is that it captures large-scale, collective motions of functional importance, which are distinguished from random thermal motions. This idea was taken advantage of in earlier studies of Bouvier and Grubmuller, where the conformational flooding method was used to study the slow base flipping in DNA [177, 183, 184]. Nevertheless, the conformational flooding method uses a single Gaussian biasing potential to escape from the
free energy minimum, which may not allow alternative conformations to be explored sufficiently.

Later the conformational flooding method was superceded by the metadynamics method, which builds a history-dependent biasing potential based on conformations previously sampled [154]. The metadynamics methods, while powerful, comes at a numerical cost of $O(t^2)$, as the number of Gaussians added to the biasing potential increases linearly with time ($t$). In addition, the metadynamics requires a number of convoluted control parameters, where the exact balance and effects are unknown. Subsequently, the adaptively biased molecular dynamics (ABMD) method has been developed as a further improvement to metadynamics, and its standard implementation has been incorporated into AMBER10 [161, 153]. Like metadynamics, ABMD method can be used to accelerate the conformational transition of biophysical systems, which cannot be observed in unbiased MD on nanosecond time scales. The advantage of ABMD is that its numerical cost scales linearly with simulation time ($O(t)$), and the number of control parameters has been effectively reduced to include only flooding time scale $\tau_F$ and kernel width $4\Delta\zeta$.

The transition pathway and alternative states sampled by ABMD is critically dependent on the prechosen reaction coordinate. The current implementation of ABMD allows for the choice of a number of different reaction coordinates, including angles, distances, dihedrals, RMSD, radius of gyration and handedness. As a result, the application of ABMD assumes that one or two reaction coordinates are sufficient to describe the desired conformational change accurately. Therefore the choice of appropriate reaction coordinates requires the detailed understanding of both end structures of the conformational change, as well as the adequate representation of the conformational change by the selected reaction coordinates. Hence the current implementation of ABMD has limited application in terms of predictions of an unknown protein conformation. Furthermore, the important conformational changes often entails global changes in protein structures, which cannot be completely characterized by a few degrees of freedom.
Hence in this study, we seek to combine ABMD simulations with flooding along essential coordinates for the first time. Briefly, we extend the current ABMD method, so that it incorporates many degrees of freedom required to describe the protein conformational change, yet does not rely on the knowledge of the alternative conformation. Hence we implemented adaptively biased essential molecular dynamics (ABEMD), which constructs the biasing potential based on the eigenvectors of essential protein motions, which are derived from nanosecond long MD trajectories. The relevant theory and implementation of ABEMD are described in the next section. Once ABEMD was implemented, we were able to use it to perform a benchmarking study, where the cis-trans conformational change of N-methyl acetamide (NMA) was predicted, without pre-specifying a simple geometric reaction coordinate or an alternative end structure.

The ultimate goal of ABEMD development is to predict slow motions in proteins without specifying a simple order parameter \textit{a priori}. While we were able to identify a pseudo-dihedral angle as a single order parameter for TIM conformational change in Chapter 4, it remains unclear as to whether this order parameter is adequate in capturing all details of TIM loop motion. Thus we hope that the direct comparison between ABMD and ABEMD simulations will not only establish the validity and efficiency of ABEMD method, but also serve as a cross validation to the earlier studies of TIM relying on a single order parameter.

\section{Theory and implementation}

\subsection{Theory}

Performing the ABEMD simulation includes two stages. In the first stage, the Cartesian coordinates of a protein from an MD trajectory were fitted to that of a reference structure to remove the translational and rotational degrees of freedom. A covariance matrix was then constructed based on the fitted coordinates according to equation 8.1:

\begin{equation}
C = \langle (x_{\text{fitted}} - \langle x \rangle)(x_{\text{fitted}} - \langle x \rangle)^T \rangle = \langle XX^T \rangle = T^T \Lambda T
\end{equation}
where \( x_{\text{fitted}} \) and \( <x> \) are 3N-dimensional vectors representing the fitted coordinates and the average of the fitted coordinates, respectively. \( \Lambda \) is the matrix of eigenvalues describing the magnitude of essential motions, \( T \) is the matrix of corresponding column eigenvectors describing the direction of essential motions, and \( X \) is the zero-mean trajectory matrix.

The matrix of projection from the trajectory matrix onto the eigenvectors can then be calculated as (equation 8.2):

\[
P = T (x_{\text{fitted}} - <x>)
\]

(8.2)

In the second stage of ABEMD simulation, the selected eigenvectors will be used to construct the flooding potential, where the force of flooding for each atom will be added onto the original force calculated from the force field. In this case, we apply the chain rule and calculate the additional forces using equation 8.3:

\[
F_{fl} = \frac{\partial U}{\partial x} = \frac{\partial U}{\partial P} \frac{\partial P}{\partial x} = \frac{\partial U}{\partial P} * TR
\]

(8.3)

where \( R \) is the rotational matrix used to perform RMS fit according to the quaternion method [185], and \( x \) is the matrix of the original coordinates related to the fitted coordinates \( x_{\text{fitted}} \) by the following equation:

\[
x_{\text{fitted}} - <x> = R(x - x_{\text{cm}}) + r - <x>
\]

(8.4)

where \( x_{\text{cm}} \) is the center of mass and \( r \) is the translational transformation.

### 8.2.2 Implementation

In the existing ABMD code, ncsu-sander-hooks.f is the main “hook” between the main sander program and the ABMD extensions. The functional calls made in the “hook” are parsed to ncsu-colvar.f, which is the core program for dispatching the function calls to different reaction coordinate modules depending on the information given in the input file. The functions declared in ncsu-colvar.f include subroutines for order
parameter and force calculation, memory allocation and deallocation, as well as file output. The names and utilities of these functions are described in Table 8.1. The detailed implementation of all these functions are given in respective reaction coordinate modules: For example, ncsu-cv-angle.f, which describes the subroutines when an angle is chosen as the reaction coordinate. Thus to follow the convention of the existing ABMD code, we add another module “ncsu-cv-PCA.f”, where the magnitude of one or more principal components will be used as the reaction coordinate. The flow of ABMD program, including the PCA extension, is illustrated in Figure 8.1.

In addition to ncsu-cv-PCA.f, a number of helper functions were implemented, so that they can be directly called from ncsu-cv-PCA.f. An additional function in ncsu-rmsd.f, “rmsd_q3u” was implemented to calculate the rotational matrix instead of the transpose. This function will be called in the force calculation routine, colvar_force. A number of I/O functions were also implemented in ncsu-read-pca.f, including functions for reading the eigenvector, the average coordinates and the reference coordinates. The desired format of the eigenvector and average coordinates follows the output format from the ptraj program after the principal component analysis of the trajectory matrix. Each of the eigenvector and average coordinate files have two header lines followed by numerical data displayed in 7 columns. When reading the reference coordinate, the AMBER restart file format is supported, where numerical data are displayed in 6 columns after two header lines. In addition, a number of constants were added for the unit opening of eigenvector file, average coordinate file and reference coordinate file. Two additional arrays associated with the average coordinates and eigenvector were declared in ncsu-colvar-type.f.

To illustrate how ABEMD simulation can be performed, the AMBER input script for the cis-trans conformation transition of NMA is included in Figure 8.2.
8.2.3 Methods

8.2.3.1 Benchmarking study: NMA cis-trans transformation

As a benchmarking study, we explored the pathway of NMA cis-trans transformation using ABEMD. As discussed in the previous paragraph, this conformational change has an energy barrier on the order of 21.2 kcal/mol, and therefore is unlikely to be observed in a regular unbiased MD simulation on nanosecond time scale.

First, the trans-NMA molecule was embedded in a 10 Å TIP4P truncated octahedral water box and modeled using the AMBER ff99SB forcefield. The system were then minimized, heated over 20 ps and equilibrated for 100 ps using Langevin dynamics, where the collision frequency was set to be 1 ps$^{-1}$. Then the regular MD simulation was performed for 10 ns, where the MD snapshots were recorded every ps. The MD trajectory was RMS fitted to the first set of equilibrated coordinates of the solute molecule, and the principal component analyses were then performed on the trajectory matrix, where the mass weighted covariance matrix was calculated and the 20 eigenvectors with the largest eigenvalues were selected following matrix diagonalization. As a result, twenty eigenvectors and the average coordinates were acquired from the analyses. These twenty eigenmodes were visualized using the IED program [150]. The 11th eigenmode, which is the one that has the largest eigenvalue among the eigenmodes reflecting significant cis-trans conformational change was selected for the flooding simulations. Furthermore, a 10 ns long MD simulation was performed on the solvated cis-NMA. We then analyzed the projection of the 11th eigenmode over both sets of MD trajectories of the cis and trans NMA systems.

The last snapshot of the 10 ns production run was used as the starting point of our ABEMD simulation. The additional inputs required are the selected eigenvector, the reference coordinates, which is the first set of coordinate of the production run, and the average fitted coordinates calculated from the entire MD trajectory. Langevin dynamics was employed to perform our ABEMD simulation, with a collision frequency of 1 ps$^{-1}$. The flooding protocol includes a flooding time scale of 50.0 ps, a resolution of 0.1 Å, and a monitoring frequency of 0.1 ps for the sampled reaction coordinates.
and flooding potential. For the purpose of our study, all of the atoms in NMA, but none of the solvent molecules were included for calculating the eigenmode projection. An 18 ns long ABEMD simulation was then performed and the corresponding conformational changes were monitored.

8.2.3.2 TIM conformation change

For TIM conformational change, the eigenvector of essential motion was first generated by overlapping all residues of the closed and open form of TIM monomer, with the closed form as the reference structure. The eigenvector with the largest magnitude was selected as the eigenvector most representative of the TIM loop motion and was adopted for further ABEMD simulations.

For the purpose of ABEMD simulation, a flooding potential were only applied to the atoms in the loop residues (Glu 166-Thr 178). The closed form of TIM was used as the reference coordinates and the average fitted coordinates was then calculated as the average of the two initial structures in the open and closed forms. The flooding resolution and flooding time scale were set at 4 and 50 ps, respectively. Then Langevin dynamics were performed with a collision frequency of 1 ps$^{-1}$, where a restraint weight of 1.0 kcal/mol/Å$^2$ was applied to all allosteric regions. These restraints were meant to restrict the movement of allosteric loops at the dimer interface, so that the dynamic behavior of TIM monomer would resemble that of a dimer. For the 35 ns of ABEMD simulations, the open form of TIM was adopted as the starting structure and the snapshots were collected at 5 ps intervals.

8.3 Results and discussion

8.3.1 NMA cis-trans conformation change

8.3.1.1 PCA analysis

Following the unbiased MD simulation and PCA analysis, we observed that the 11th eigenmode is the first mode that reflects the cis-trans conversion of NMA (Figure 8.3),
where an $\omega$ change was interpolated to be over 50 degrees according to IED visualization. Interestingly, while the first five modes from PCA analyses are often considered to incorporate 90% of the large-scale motions, these modes are dominated by methyl group rotations and vibrations, as well as N-H and C-O bond stretches. In comparison, these modes include a minimal component of $\omega$ angle rotations. Nevertheless, this observation is in agreement with previous studies [182], which suggested that the long-term motions of interest are often not included in the first 10 modes of essential movements derived from short MD simulations. We also observed that the magnitude of projection are significantly different between the cis and trans NMA for the 11th eigenmode. More specifically, the value of the eigenmode projection varies between +0.5 Å to -0.5 Å for trans-NMA, whereas that of the cis-NMA fluctuates on a much large magnitude, between +1.8 Å to -1.5 Å. Thus we hope that this difference in projection can be used to identify cis-trans transition in the ABEMD simulation described below (Figure 8.4).

8.3.1.2 ABEMD simulation

After 18 ns of ABEMD simulation, the conformational transition of the NMA molecule was analyzed in terms of transitions in the eigenmode projections and the $\omega$ angle (Figure 8.4). The evolution of the biasing potential is described in Figure 8.5 accordingly. We observed that the evolution of the eigenmode projection is synchronized with the changes in $\omega$ angle (Figure 8.6). In the case of eigenmode projection, a large change in magnitude was observed around 6.3 ns, which coincides with the first $\omega$ angle flip, reflected by the $\omega$ angle change. During the last 10 ns of the ABEMD simulation, frequent transitions were observed in both projection magnitude and the $\omega$ angle.

8.3.2 ABEMD simulation of TIM conformation change

8.3.2.1 PCA and RMS analyses

To generate the eigenmode to be used for ABEMD simulation, PCA analysis was first performed on the closed and open structures of TIM. First, an all-atom fitting was
performed, resulting in two well overlapped structures with a large RMS (3.9 Å) in the loop region and a small RMS in the rest of the backbone (0.5 Å). Then the mass-weighted covariance matrix was generated and the first principal component reflects the loop motion primarily. As we analyzed the rms fluctuation of Cartesian coordinates along the eigenmode direction, the loop motion was predominant in comparison to the rest of the backbone; some N- and C-terminal motions were also observed but were deemed irrelevant to our study due to their tendency to be more disordered (Figure 8.7). The PCA analysis shows that the magnitude of projection for the closed and open forms of TIM are -45.84 Å and 45.84 Å, respectively. Thus in the ensuing ABEMD simulation, the changes in the projection magnitude are expected to reflect the conformational changes in the TIM active site loop region.

8.3.3 ABEMD simulation of TIM conformation change

The magnitude of projection in the 30 ns long ABEMD simulation was monitored and analyzed (Figure 8.8). From 0 ns to 2.8 ns, the magnitude of projection were decreased from -35 Å to -57 Å, indicating further sampling of the open loop conformation. From 2.8 ns to 8 ns, the magnitude of projection was increased from -35 Å to 32 Å, suggesting that the intermediate conformations were being progressively sampled, until the metastable “closed” conformation was reached. A sharp decrease in projection magnitude followed, indicating that the initial transition path had been explored, and more intermediate conformations between the two states were being sampled. Further down the trajectory, the “closed” conformations were sampled a second time between 16 and 30 ns.

Some interesting observations were also made on the evolution of the biasing potential (Figure 8.9). Initially, the biasing potential was centered at around -35 Å in terms of PC magnitude. By 6 ns, the biasing potential has been extended to include repulsive components for positive magnitude of projections up to 12 Å. This component was increased to include large, repulsive components for PC magnitudes up to 34 Å by 8 ns. By 16 ns, the biasing potential has been further expanded, this time to include a repulsive potential centered at 35 Å, which continued to increase until the end of
the simulation. The evolution of biasing potential can be seen as directly correlated to the magnitude of PC projection sampled: For example, the buildup of biasing potential towards more positive PC projections between 3 ns and 8 ns is responsible for the progressively larger projections being sampled in the same time frame. Similarly, the accumulation of biasing potential in the negative PC projection region between 8 ns and 12 ns is indicative of the resampling of monotonously decreasing PC projections.

Given that the ABEMD simulation was started from the open conformation, the structures from the entire trajectory were compared with the crystal structure with the closed loop conformation to better understand the extent of sampling in the metastable states. For all the snapshots in the trajectory, only the loop RMSD was measured while the rest of the backbone was fitted with the reference structure. We observed that from 2.8 ns and onward, the sampled structure slowly approached the “closed” conformation. At 7.7 ns, the sampled loop conformation showed the smallest difference (1.8 Å) with respect to the crystal structure (Figure 8.10). This observation naturally raises the question of whether 1.8 Å is “close enough” to the experimentally determined structure. Therefore we will compare the quality of this structure to what was derived in earlier studies in Chapter 5 & Chapter 6. In Chapter 5, we performed unbiased MD of the “induced” metastable state with closed loop structure. In this study, the loop RMS was stabilized at 1.8 Å in comparison to the crystal structure. In Chapter 6, the ABMD simulation was performed with the pseudo-dihedral angle as the order parameter, where the structure of best quality also gave an 1.8 Å deviation from the crystal structure. Taken together, the quality of prediction given by ABEMD is as good as that of regular ABMD; both of these methods lead to predicted structures within the range of RMS fluctuations from unbiased MD simulations.

8.3.3.1 TIM conformational space sampled by ABEMD simulations

In order to further evaluate the effectiveness of ABEMD sampling, we compared the conformational space sampled by ABEMD simulations with that of ABMD simulations. Considering that the allosteric regions of TIM monomer were restrained in the ABEMD simulation to mimic the dimeric TIM environment, the conformational space
sampled in the dimeric ABMD simulation was used as the reference space of this comparison. A two dimensional conformational space was constructed, where the RMS deviations from the open and closed loop structures were used as respective coordinates (Figure 8.11). We observed that the two conformational spaces sampled are quite similar, while some differences exist in the ground and excited states sampled: For the “open” state, the ABEMD simulation sampled a large number of conformations with RMS on the order of 1.2 Å, whereas the ABMD simulation of TIM dimer sampled conformations with a larger RMS, on the order of 1.8 Å. Similarly, for the “closed” state, the ABMD predicted conformational cluster appear to be 0.5 Å further from the ABMD predictions. This is most likely due to the 1 kcal/mol/Å² restraints imposed on ABEMD simulations. As an attempt to mimic the dimeric environment, some regions of the monomer protein may have been over-restrained, leading to smaller deviations from the ground state, and larger deviations from the excited states. Another interesting observation is that the ABEMD simulations appear to have sampled more intermediate conformations centered at (4 Å, 4 Å) and also at (6 Å, 3 Å), which suggests that ABEMD has the ability to sample a wider conformational space than regular ABMD.

We also performed cluster analyses to understand the conformational subspace sampled in ABEMD simulations. The partition around medoid (PAM) algorithm was used to describe the conformational space as eight discrete clusters in Figure 8.12. Two clusters colored in black and green, respectively, were identified as the ground and excited state. The structures corresponding to their medoid are also displayed to show their remarkable semblance to the open and closed conformations. In addition, two clusters colored in red and blue, centered at (3.1 Å, 2.4 Å) and (2.4 Å, 3.8 Å) could be seen as intermediates directly connecting these two states.
8.3.4 Evaluation of ABEMD simulations

8.3.4.1 Comparison between ABEMD and ABMD simulations

In order to evaluate the performance of ABEMD simulations with respect to the regular ABMD simulations, we compared the two studies we performed on NMA cis-trans conversion and TIM loop motion.

First, we examined the profile of sampled conformations in the NMA simulations. For regular ABMD simulation with $\omega$ angle as the reaction coordinate, a flooding time scale of 25 ps was used, with a resolution of $5^\circ$ (Chapter 7). The first transition from trans- to cis-NMA was observed at 3.2 ns, whereas the first inversion (back to trans-NMA) was observed at 5.3 ns. For ABEMD simulations, a flooding time scale of 50.0 ps was applied with a resolution of 0.1 Å. At 6.3 ns the first transition to cis-NMA was achieved, and the system then inverted back to trans-NMA spontaneously. When comparing the last 10 ns of both simulations, we observed that ABEMD simulations shows progressively more even distribution between cis- and trans-conformations sampled, whereas the same behavior was not observed in regular ABMD. It is also noticed that ABEMD simulation leads to more spontaneous transitions on ps intervals, whereas the regular ABMD simulation exhibits transitions on the order of 0.1 ns intervals. This can be an indicator that the biasing potential constructed by ABEMD is a better complement to the original free energy landscape, which leads to a smoother biased energy surface, enabling more spontaneous transitions.

Second, we compared the respective trajectories from TIM simulations, where the flooding time scale for both simulations were set at 50 ps and the flooding resolution $(4\Delta \xi)$ were set at $4^\circ$ and 4 Å, respectively. In regular ABMD simulation, the metastable “closed” state closest to the experimental structure was reached after 10 ns, whereas that of the ABEMD simulation was achieved after 7.8 ns. When comparing the RMS space of the sampled conformations, we observed that many more intermediate conformations were sampled from ABEMD simulations.
8.3.4.2 Scope of ABEMD simulations

While ABEMD has the potential of becoming a powerful tool in protein conformation predictions, a basic understanding of the desired motion is required prior to its application. In other words, the most prominent modes uncovered by PCA analysis are more likely to be the modes characteristic of ps-ns time scale motions, which are different from the \( \mu s-\text{ms} \) time scale motions we are interested in. Hence randomly choosing a couple of modes or constructing a linear combination out of the first 10 modes is unlikely to give rise to the mode of our interest. This is where theoretical prediction of protein motions should be combined with experimental observables, such as NMR relaxation data. Then we can choose the appropriate eigenmodes based on their capacity to incur large magnitude of motions in the highlighted regions from experimental studies. In any case, careful identification of modes of essential motions is needed prior to ABEMD simulations.

The other caveat of ABEMD method lies in the free energy calculation of the transition pathway post-simulation. Our study suggest that for the motion that is characteristic of cis-trans conformational change of NMA, the magnitude of the motion can not be used to uniquely identify the conformation of the NMA molecule. This could lead to difficulty in umbrella sampling calculations if the magnitude of the projection was to used as the reaction coordinate. Nevertheless, we could analyze the the trajectory of ABEMD simulation to derive alternative reaction coordinates. More importantly, we could use these intermediate snapshots as the starting points of subsequent umbrella sampling simulations.

8.4 Conclusion

To summarize, we have performed ABEMD simulations to predict the conformational transitions of small peptide-like molecule (NMA) and globular protein (TIM). our study showed that ABEMD is capable of predicting end structures and intermediates of functionally important conformational transitions on long time scales. This method has the potential of being a powerful prediction tool for the study of conformational
changes, when the exact order parameter is not known \textit{a priori}, or when the transition evolves large scale, concerted motions. In addition, there is a great prospect that the ABEMD method could become very useful when combined with experimental data, given that some understanding of the specific conformational transition problem is required prior to applying this method.
### Functions Description

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>colvar_value</code></td>
<td>Calculates the value of the reaction coordinate</td>
</tr>
<tr>
<td><code>colvar_force</code></td>
<td>Calculates the additional force needed for flooding</td>
</tr>
<tr>
<td><code>colvar_bootstrap</code></td>
<td>Allocates memory for mass array, reference &amp; COM coordinates</td>
</tr>
<tr>
<td><code>print_details</code></td>
<td>Print out the first/last atoms selected for flooding</td>
</tr>
<tr>
<td><code>colvar_cleanup</code></td>
<td>Memory cleanup</td>
</tr>
</tbody>
</table>

Table 8.1: Code modules for ABMD calculations

![Diagram of code modules](image)

**Figure 8.1: Flow of ABMD code in SANDER**

Explanation of how the ABMD modules are related and how the new module is added
Figure 8.2: Sample input file
Explanation of how the ABMD modules are related and how the new module is added.
Figure 8.3: Principal component analysis of 10ns long NMA MD simulations
The 11th eigenmode was employed for our ABEMD simulation as the representative
eigenmode of cis-trans transformation of NMA.
Mode #1: methyl group rotation & symmetric stretch; Mode #2: methyl group rotation
& asymmetric stretch; Mode #3: methyl group rotation & symmetric stretch; Mode #4:
C=O, N-H symmetric stretch; Mode #5: asymmetric methyl group stretch
Figure 8.4: Magnitude of projection in the cis and trans conformations in the 10 ns of unbiased MD
Figure 8.5: Evolution of biasing potential in 18 ns of ABEMD
The biasing potential from 2 to 18 ns at 2 ns intervals was monitored and shown here.
Figure 8.6: Evolution of projection and $\omega$ angle in 18 ns of ABEMD
The plot on the top shows how the projection of eigenmode changes in the ABEMD trajectory. The plot on the bottom shows how the $\omega$ angle, reflecting cis-trans conformational changes, vary in the 18 ns long simulations.
Figure 8.7: RMS fluctuation of C-α atoms along the selected eigenmode.
The residues in the active site loop region are highlighted by the red line.
Figure 8.8: Evolution of magnitude of eigenmode projection for TIM conformation change
Figure 8.9: Evolution of biasing potential for TIM conformational change

Figure 8.10: RMS fluctuation of TIM loop structure in the ABEMD simulation
Left: RMS profile of ABEMD simulations. Right: Loop prediction at 7.73 ns compared to the open/closed loop structures.
Figure 8.11: Comparison of RMS spaced by ABMD and ABEMD simulations.
Figure 8.12: Cluster analyses of RMS space sampled in ABEMD simulations
References


Vita

Sishi Tang

2011  Ph. D. in Computational Biology and Molecular Biophysics, Rutgers University

2011  M.S. in Statistics, Rutgers University

2006-08  Ph.D Candidate in Biophysics, The Scripps Research Institute

2004-2006  M.S. in Pharmaceutical Sciences, University of Toronto

2001-2004  B.S. in Chemistry and Computer Science, Carleton University


