Computer simulations exploring conformational preferences of short peptides and developing a bacterial chromosome model

Shuxiang Li

University of Iowa

Copyright © 2017 Shuxiang Li

This dissertation is available at Iowa Research Online: https://ir.uiowa.edu/etd/5960

Recommended Citation
https://doi.org/10.17077/etd.iw0pnk9x

Follow this and additional works at: https://ir.uiowa.edu/etd

Part of the Biochemistry Commons
COMPUTER SIMULATIONS EXPLORING CONFORMATIONAL
PREFERENCES OF SHORT PEPTIDES AND DEVELOPING A BACTERIAL
CHROMOSOME MODEL

by
Shuxiang Li

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Biochemistry
in the Graduate College of
The University of Iowa

December 2017

Thesis Supervisor: Professor Adrian H. Elcock
This is to certify that the Ph.D. thesis of

Shuxiang Li

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biochemistry at the December 2017 graduation.

Thesis Supervisor: Adrian H. Elcock

Thesis Committee: Madeline A. Shea, Thesis Chair

Ernesto J. Fuentes

Michael J. Schnieders

Maria Spies

Claudio J. Margulis
To my beloved wife and daughters
ACKNOWLEDGMENTS

The work in this thesis was carried out with the contributions of others. I will endeavor to acknowledge them because of their meaningful efforts during the past five years.

First and foremost, I want to thank my advisor Dr. Adrian Elcock for introducing me to the beautiful world of computational biology. I want to thank him for not only giving me the knowledge but also showing me how to discover it. I also want to thank him for constantly reminding me to be ambitious. I am thankful for his tremendous effort in helping me to publish my scientific papers. There are no words that can express my gratitude for his guidance and help in those years. I thank him for everything that he has done for me.

Second, I wish to express my gratitude to past and current members of Elcock Lab. I want to thank Dr. Casey Andrews, who provided excellent technical support and generous help for my research since the day I started in the lab. In particular, I want to thank William Hacker, who worked together with me on a project for developing the E.coli chromosome models and provided knowledge, help and friendship for which I am very grateful. I also want to thank Dr. Tam Frembgen Kesner, Wesley Lay, Mark Miller, Timothy Collingsworth, Robert McDonnell and Venkata Sanaboyana, who gave me helpful suggestions and discussions on both science and life.

I want to give special thanks to my thesis committee: Dr. Madeline Shea, Dr. Ernesto Fuentes, Dr. Michael Schnieders, Dr. Maria Spies and Dr. Claudio Margulis. I thank them for their time, patience, and their helpful comments and constructive suggestions, which helped moving my research project forward. I would like to thank the Department of Biochemistry at the University of Iowa for
accepting me into their graduate program and allowing me to pursue this degree.

Lastly, I would like to thank my family: my parents, for their faith in me and allowing me to pursue this degree, and, most importantly, my wife Xiaomin who supported and encouraged me at all times.
ABSTRACT

Computer simulations provide a potentially powerful complement to conventional experimental techniques in elucidating the structures, dynamics and interactions of macromolecules. In this thesis, I present three applications of computer simulations to investigate important biomolecules with sizes ranging from two-residue peptides, to proteins, and to whole chromosome structures.

First, I describe the results of 441 independent explicit-solvent molecular dynamics (MD) simulations of all possible two-residue peptides that contain the 20 standard amino acids with neutral and protonated histidine. $^3J_{HNHa}$ coupling constants and $\delta_{Ha}$ chemical shifts calculated from the MD simulations correlated quite well with recently published experimental measurements for a corresponding set of two-residue peptides. Neighboring residue effects (NREs) on the average $^3J_{HNHa}$ and $\delta_{Ha}$ values of adjacent residues were also reasonably well reproduced. The intrinsic conformational preferences of each residue, and their NREs on the conformational preferences of adjacent residues, were analyzed. Finally, these NREs were compared with corresponding effects observed in a coil library and the average $\beta$-turn preferences of all residue types were determined.

Second, I compare the abilities of three derivatives of the Amber ff99SB force field to reproduce a recent report of $^3J_{HNHa}$ scalar coupling constants for hundreds of two-residue peptides. All-atom MD simulations of 256 two-residue peptides were performed and the results showed that a recently-developed force field (RSFF2) produced a dramatic improvement in the agreement with experimental $^3J_{HNHa}$ coupling constants. I further show that RSFF2 also improved modestly agreement with experimental $^3J_{HNHa}$ coupling constants of five model proteins. However, an analysis of NREs on the $^3J_{HNHa}$ coupling constants of the
two-residue peptides indicated little difference between the force fields’ abilities to reproduce experimental NREs. I speculate that this might indicate limitations in the force fields’ descriptions of nonbonded interactions between adjacent side chains or with terminal capping groups.

Finally, coarse-grained (CG) models and multi-scale modeling methods are used to develop structural models of entire E. coli chromosomes confined within the experimentally-determined volume of the nucleoid. The final resolution of the chromosome structures built here was one-nucleotide-per-bead (1 NTB), which represents a significant increase in resolution relative to previously published CG chromosome models, in which one bead corresponds to hundreds or even thousands of basepairs. Based on the high-resolution final 1 NTB structures, important physical properties such as major and minor groove widths, distributions of local DNA bending angles, and topological parameters (Linking Number (Lk), Twist (Tw) and Writhe (Wr)) were accurately computed and compared with experimental measurements or predictions from a worm-like chain (WLC) model. All these analyses indicated that the chromosome models built in this study are reasonable at a microscopic level. This chromosome model provides a significant step toward the goal of building a whole-cell model of a bacterial cell.
The use of computer simulation to study the dynamics and interactions of biomolecules has become an important tool in the field of structural biology. The first part of this thesis describes 441 independent all-atom computer simulations of two-residue peptides to validate current force fields and explore the intrinsic conformational preferences of amino acids and the extent to which they are modulated by neighboring residues. NMR measurements calculated from the simulations were shown to correlate quite well with recently published experimental measurements for a corresponding set of two-residue peptides. In particular, a recently developed force field produced a dramatic improvement in the agreement with experimental $J$ coupling constants. Due to the limitations in computational expense, all-atom computer simulation is not always practical for simulating large scale biomolecular systems on long timescales. An alternative is to use coarse-grained models, which allows for a tremendous increase in computational efficiency. The second part of this thesis describes the use of coarse-grained models and multi-scale simulation techniques to build structural models of the whole $E. coli$ chromosome at a resolution of one-bead-per-nucleotide. The higher-resolution chromosome models developed in the thesis can contribute to constructing structural models of entire bacterial cells and the study of more complex cellular processes.
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................ x
LIST OF FIGURES ....................................................................................................... xi
LIST OF ABBREVIATIONS ....................................................................................... xv
LIST OF SYMBOLS .................................................................................................... xvii

CHAPTER

I. INTRODUCTION ................................................................................................. 1

II. EXPLORING THE CONFORMATIONAL PREFERENCES AND NEIGHBORING RESIDUE EFFECTS OF TWO-RESIDUE PEPTIDES BY MOLECULAR DYNAMICS SIMULATION ........................................................................... 8

   Introduction ............................................................................................................ 8
   Methods .................................................................................................................. 11
     Systems studied .................................................................................................. 11
     Molecular dynamics simulations ....................................................................... 12
     Convergence of molecular dynamics simulations .......................................... 13
     Analysis of conformational distributions .......................................................... 14
     Analysis of β-turn distributions ....................................................................... 16
     Calculation of $^3J_{HN\alpha}$ coupling constants .............................................. 16
     Calculation of chemical shifts ......................................................................... 17
   Results ................................................................................................................... 18
     Convergence of molecular dynamics simulations ............................................ 18
     Comparison of experimental and simulated $^3J_{HN\alpha}$ coupling constants .... 20
     Investigation of discrepancies between experimental and simulated $^3J_{HN\alpha}$
     coupling constants ........................................................................................... 23
     Comparison of experimental and simulated chemical shifts ......................... 25
     Average intrinsic backbone conformational preferences ................................ 27
     Neighboring residue effects on average backbone conformational
     distributions ........................................................................................................ 29
     Analysis of β-turn populations .......................................................................... 31
   Discussion .............................................................................................................. 33

III. VALIDATION OF THREE AMBER FF99SB DERIVED FORCE FIELDS WITH EXPERIMENTAL NMR MEASUREMENTS ON PEPTIDE AND PROTEIN SYSTEMS ........................................................................ 67

   Introduction ............................................................................................................ 67
   Methods .................................................................................................................. 69
     Systems studied .................................................................................................. 69
     Molecular dynamics simulation setup ............................................................... 70
     Calculation of $^3J_{HN\alpha}$ coupling constants .................................................. 71
     Calculation of neighboring residue effects ....................................................... 72
   Simulation differences for ff99SB-ILDN-NMR force field ............................ 73
LIST OF TABLES

Table

2.1. Residue-specific Karplus parameters derived by optimizing agreement between simulation and experiment for two-residue peptides .....................65

4.1. The calculated writhe, twist and linking number for the four test models......................................................................................................................140

4.2. Tw, Wr and Lk calculations for a complete E.coli model.............................141

4.3. Topological properties of the final 20 oriC@pole models ..........................142

4.4. Topological properties of the final 20 oriC@midcell models ........................143
LIST OF FIGURES

Figure

1.1. Application ranges for computer simulations at different resolutions ........7

2.1. Representation of an example two-residue peptide (Asp-Asp)........39

2.2. Representation of the phi and psi dihedral angles and a Ramachandran map with the most preferred conformations ..........40

2.3. Convergence of average Ramachandran maps in MD simulations ....41

2.4. Ramachandran maps obtained from 3 replicate MD simulations of selected peptides .................................................................42

2.5. Convergence of Ramachandran maps in MD simulations for termini ....43

2.6. Comparison of computed and experimental $^3J_{HNHa}$ coupling constants for two-residue peptides .....................................................44

2.7. Comparison of simulation and experimental $^3J_{HNHa}$ coupling constants for residues in histidine-containing peptides ......................45

2.8. Comparison of computed and experimental $^3J_{HNHa}$ coupling constants averaged by type of amino acid .................................................................46

2.9. Factors influencing agreement between computed and experimental $^3J_{HNHa}$ coupling constants ..................................................................................47

2.10. Comparison of computed and experimental $^3J_{HNHa}$ coupling constants for peptides grouped according to their constituent amino acids ..........48

2.11. Residue-specific Karplus parameters improve agreement with experiment .................................................................................................49

2.12. Comparison of computed and experimental $\delta_{Ha}$ chemical shifts .......50

2.13. Comparison of computed and experimental $\delta_{Ha}$ chemical shifts with different backbone conformations .........................................................51

2.14. Comparison of computed and experimental $\delta_{Ha}$ chemical shifts with neighboring residue effects .................................................................52

2.15. Ramachandran maps for the N-terminal residues sampled in MD simulations of 441 two-residue peptides .........................................................53

2.16. Ramachandran maps for the C-terminal residues sampled in MD simulations of 441 two-residue peptides .........................................................54

2.17. Average Ramachandran maps at the N-terminus ................................55
2.18. Average Ramachandran maps at the C-terminus ........................................... 56

2.19. Fractional populations of four major backbone conformations ..................... 57

2.20. Comparison of average secondary structure populations observed in this study for two-residue peptides with those obtained for single-residue peptides .......................................................... 58

2.21. Neighboring residue effects on the average fractional populations of four major backbone conformations ................................................................. 59

2.22. Comparison of simulated effect fractional populations of four major backbone conformations with that obtained from a PDB-derived coil library ................................................................. 60

2.23. Neighboring residue effects on α and β fractional populations ..................... 61

2.24. Populations of β-turn conformations obtained from MD simulations .......... 62

2.25. Propensity of each type of amino acid to be found for each type of β-turn ................................................................. 63

2.26. Comparison of computed β-turn conformations with that obtained from an analysis of PDB structures ................................................................. 64

3.1. Representation of an example two-residue peptide (Ala-Ala) ...................... 81

3.2. Cartoon images of the five proteins simulated ........................................... 82

3.3. Comparison of computed \(^3J_{HNHa}\) coupling constants with the ff99SB-ILDN-NMR force field using two different water models for 256 two-residue peptides ................................................................. 83

3.4. Comparison of computed \(^3J_{HNHa}\) coupling constants with experimentally measured values for 256 two-residue peptides using the Karplus parameterization due to Vögeli et al. ................................................................. 84

3.5. Comparison of computed \(^3J_{HNHa}\) coupling constants with the ff99SB-ILDN and ff99SB-ILDN-NMR force fields for 256 two-residue peptides ................................................................. 86

3.6. Comparison of computed \(^3J_{HNHa}\) coupling constants with experimentally measured values for 256 two-residue peptides using the Karplus parameterization due to Hu & Bax ................................................................. 87

3.7. Comparison of computed \(^3J_{HNHa}\) coupling constants using two different Karplus equation parameterizations for 256 two-residue peptides ................................................................. 89

3.8. Comparison of root mean square (RMS) errors in the computed \(^3J_{HNHa}\) coupling constants obtained using two different Karplus equation parameterizations for 256 two-residue peptides ................................................................. 91

3.9. Comparison of computed neighboring residue effects on \(^3J_{HNHa}\) coupling constants with experimentally measured values for 256 two-
residue peptides using the Karplus parameterization due to Vögeli et al. ................................................................. 92

3.10. Comparison of computed neighboring residue effects on $^3J_{HNHa}$ coupling constants averaged by type of amino acid using the Karplus parameterization due to Vögeli et al. ................................................................. 94

3.11. Comparison of computed neighboring residue effects on $^3J_{HNHa}$ coupling constants averaged by type of amino acid using the Karplus parameterization due to Hu & Bax ................................................................... 96

3.12. Linear regression $r^2$ values for comparisons of computed and experimental $^3J_{HNHa}$ coupling constants for 256 peptide systems ................... 98

3.13. Scatter plots comparing computed and experimental $^3J_{HNHa}$ coupling constants for the model proteins ................................................................. 99

3.14. Linear regression $r^2$ values for comparisons of computed and experimental $^3J_{HNHa}$ coupling constants for five protein systems using the Karplus parameterization due to Vögeli et al............................................ 101

3.15. Linear regression $r^2$ values for comparisons of computed and experimental $^3J_{HNHa}$ coupling constants for five protein systems using the Karplus parameterization due to Hu & Bax.................................................. 102

3.16. RMS deviations of the simulated protein structures from their initial structures as a function of simulation time. .................................................. 103

3.17. Average RMS deviations of the simulated protein structures from their initial structures. ................................................................. 104

4.1. Several available bacterial chromosome models with limited resolutions or incomplete coverage of the genome ........................................ 125

4.2. Comparison of all-atom model with different resolutions of coarse-grained DNA models................................................................. 126

4.3. General properties of persistence length (P) and contour length (L) ........ 127

4.4. An illustration of the major and minor grooves........................................ 128

4.5. An illustration of the DNA bending angle ............................................. 129

4.6. Parameterization of the 5 BPB model to match persistence length of DNA at 150 mM salt concentration .................................................. 130

4.7. Ionic strength dependence of the persistence length for the 5BPB model ......................................................................................... 131

4.8. Overview of the protocol for constructing an E.coli chromosome model at 1 NTB resolution ................................................................. 132

4.9. Comparison of the 500 BPB model (red spheres) and the 5 BPB model (green spheres) ......................................................................................... 133
4.10. Overview of one final *E. coli* chromosome structure at 1 NTB resolution ................................................................................................................................. 134

4.11. The 20 oriC@pole structures ................................................................................................................................. 135

4.12. The 20 oriC@midcell structures ........................................................................................................................ 136

4.13. Groove width distributions of the 1 NTB model ................................................................................................. 137


4.15. Plasmid models used to validate the calculations of writhe, twist and linking number ........................................................................................................... 139

5.1. Distributions of void sizes in the chromosome models ......................................................................................... 150
LIST OF ABBREVIATIONS

1NTB: one nucleotide per bead
5 BPB: 5 base pairs per bead
500 BPB: 500 base pairs per bead
Amber: Assisted Model Building with Energy Refinement
BD: Brownian dynamics
CHARMM 27: chemistry at Harvard molecular mechanics version 27
CG: coarse-grained
CD: circular dichroism
GROMACS: Gröningen machine for chemical simulations
DH: Debye-Hückel
ff99SB-ILDN: force field 1999 Stony Brook with corrections for isoleucine, leucine, aspartate and asparagine sidechains
HIP: protonated histidine
IDP: intrinsically disordered protein
IR: infrared spectroscopy
LINCS: linear constraint solver
MD: molecular dynamics
NMR: nuclear magnetic resonance
NRE: neighboring residue effects
DNA: deoxyribonucleic acid
OPLS: optimized potential for liquid simulations
PDB: protein data bank
PME: particle mesh ewald
PMF: potential of mean force
QM: quantum mechanical
REMD: replica exchange molecular dynamics
RMSD: root mean square deviation
RMS: root mean square
RSFF2: residue-specific force field
TIP3P: transferable intermolecular potential 3 points
TIP4P: transferable intermolecular potential 4 points
TIP4P-Ew: transferable intermolecular potential 4 points for use with Ewald summation
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>angstrom ($10^{-10}$ meter)</td>
</tr>
<tr>
<td>°C</td>
<td>degrees Celsius</td>
</tr>
<tr>
<td>δ</td>
<td>chemical shift</td>
</tr>
<tr>
<td>θ</td>
<td>bond angle</td>
</tr>
<tr>
<td>σ</td>
<td>supercoil density</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>alpha carbon</td>
</tr>
<tr>
<td>$f_{max}$</td>
<td>maximum frequency</td>
</tr>
<tr>
<td>ΔG</td>
<td>free energy change</td>
</tr>
<tr>
<td>J</td>
<td>couplings constant</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Debye-Hückel screening parameter</td>
</tr>
<tr>
<td>L</td>
<td>contour length</td>
</tr>
<tr>
<td>Lk</td>
<td>linking number</td>
</tr>
<tr>
<td>P</td>
<td>persistence length</td>
</tr>
<tr>
<td>q</td>
<td>atomic charge</td>
</tr>
<tr>
<td>$r_{ij}$</td>
<td>inter-atomic distance between atom i and atom j</td>
</tr>
<tr>
<td>R</td>
<td>end-to-end distance</td>
</tr>
<tr>
<td>T</td>
<td>temperature</td>
</tr>
<tr>
<td>Tw</td>
<td>twist</td>
</tr>
<tr>
<td>V</td>
<td>potential energy</td>
</tr>
<tr>
<td>Wr</td>
<td>writhe</td>
</tr>
<tr>
<td>$&lt;x&gt;$</td>
<td>ensemble average of variable $x$</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

Computer simulations have made very significant improvements over the last few decades. Within the fields of biology and chemistry, they are carried out in the hope of characterizing biomolecules in terms of their structures, motions and the microscopic interactions between them (Klepeis et al., 2009). Early computer simulations studied simple systems such as noble gases (Verlet 1967). Similar simulations were later carried out to examine the dynamics of protein structures by the Karplus group in 1977 (McCammon et al., 1977). Over the years, improvements in computer simulations have allowed them to model accurately complicated biomolecular processes such as protein-folding (Best et al., 2013; Freddolino et al., 2008; Voelz et al., 2010; Voelz et al., 2012) and protein crystallization (Martonak et al., 2005; Martonak et al., 2003). More recently, an all-atom representation of the human immunodeficiency virus type 1 (HIV-1) capsid was simulated by the Schulten group for over 1 µs in explicit solvent, a system containing 64 million atoms (Perilla and Schulten 2017). In these applications, computer simulation acts as an in silico microscope and has proven to be an exceptionally powerful tool for elucidating details currently inaccessible with conventional experimental techniques such as X-ray crystallography and NMR. The utility of computer simulation of molecular systems was formally acknowledged with the award of the 2013 Nobel Prize to Martin Karplus, Michael Levitt, and Arieh Warshel “for the development of multiscale models for complex chemical systems”.

Many recent computer simulations have employed all-atom, explicit-solvent molecular dynamics (MD). Molecular dynamics is a numerical technique for studying the equilibrium and physical movements of N-body atomic systems.
(Allen and Tildesley 1987). It was first introduced by Alder and Wainwright in the late 1950's to examine the interactions of hard spheres (Alder and Wainwright 1957; 1959; 1960). In molecular dynamics, each atom is treated as a point mass, and a potential energy function is utilized to describe the interactions between atoms. Given the potential and position of each atom, the force acting on each atom and its acceleration can be determined. Integration of Newton's equations of motion yields a trajectory that describes the positions, velocities, and accelerations of the atoms in the system as they vary with time. Thus, the state of the system can be predicted at any time. From the trajectory, the macroscopic thermodynamic properties of the system can be determined. Molecular dynamics can be interpreted as a virtual experiment that acts as a bridge between laboratory experiments and theoretical computations. Now molecular dynamics is widely used to investigate the structure, dynamics, and thermodynamics of biological molecules and their complexes. Broadly used molecular dynamics packages include GROMACS (Berendsen et al., 1995; Pronk et al., 2013), NAMD (Phillips et al., 2005), AMBER (Case et al., 2005) and CHARMM (Brooks et al., 1983).

A molecular mechanics force field or potential energy function describes the covalent connectivity and interactions of the molecular system. In molecular dynamics, it is assumed that the force field depends only on the atomic positions. Empirical force fields consist of a summation of (a) bonded forces associated with chemical bonds, bond angles, bond dihedrals, and (b) non-bonded interactions including van der Waals forces and electrostatic interactions (Allen and Tildesley 1987). The general functional form for a non-polarizable force field is often written:
where $V(r^N)$ denotes the potential energy of the system where $r^N$ represents the spatial coordinates of the $N$ atoms. The first three summations define the potential energies of the bonds, angles and dihedrals. The last summation defines non-bonded interactions. The potential energy equation also contains parameters such as equilibrium bond length ($l_0$), bond angle ($\theta_0$), dihedral angle ($\omega$), atomic charge ($q_i$ and $q_j$) and van der Waals parameters ($\varepsilon_{ij}$ and $\sigma_{ij}$). These parameters are obtained by fitting against detailed quantum mechanical (QM) calculations or experimental physical properties such as elastic constants, viscosity, density and NMR $J$ coupling constants, etc. Different force fields have been specifically developed for the simulations of proteins, nucleic acids, and other biomolecules and small organic molecules. Some commonly-used empirical force fields include AMBER (Hornak et al., 2006), CHARMM (Brooks et al., 1983) and GROMOS (Oostenbrink et al., 2004).

All-atom molecular dynamics simulation is a very powerful method, yet it also has limitations due to computational resources. Without specialized computational hardware (Shaw et al., 2008) or distributed computing approaches Folding@Home (Larson et al., 2009), MD simulation is limited to studying relatively small systems, often at time-scales much times shorter than the biologically-relevant time scale. To overcome this limitation, approximations or enhanced sampling techniques such as replica exchange molecular dynamics (REMD) (Sugita and Okamoto 1999), metadynamics (Laio and Gervasio 2008),
umbrella sampling (Kastner 2011) and transition path sampling (Escobedo et al., 2009) have been developed to decrease computational time. Even with such enhanced techniques, all-atom MD simulations cannot realistically access the length- and time-scales relevant to biological processes (Chung et al., 2015; Shaw et al., 2010). These limitations in scale have motivated the use of coarse-grained models of biomolecular systems. In coarse-grained (CG) models, the all-atom model of the system is mapped to a lower dimensional space. The reduction in complexity of the system minimizes the atomistic degrees of freedom and enlarges the envelope accessible to simulation (Figure 1.1). For example, an N-fold reduction in the number of sites in the system leads to a corresponding N-fold increase in accessible time-scales. More significantly, coarse-graining decreases the largest characteristic frequency of the system, allowing for larger integration time steps and faster sampling in simulations (Ingolfsson et al., 2014). Therefore, CG models typically result in significant computational acceleration of several orders of magnitude, allowing these models to overcome both the spatial and temporal limitations of all-atom models (Bond et al., 2007; Ingolfsson et al., 2014).

Generally, there are two main routes to develop CG models: top-down and bottom-up approaches (Hinckley et al., 2013). In the top-down approach, the parameterization of CG models is based on reproducing important experimental measurements, especially thermodynamic properties. In the bottom-up approach, higher resolution data such as all-atom simulations are used to parametrize interaction potentials for CG models. To achieve this, iterative parameterization methods like iterative Boltzmann inversion (IBI) (Reith et al., 2003), inverse Monte Carlo (IMC) (Lyubartsev and Laaksonen 1995) and force matching (FM) (Izvekov and Voth 2005; Noid et al., 2008) are often utilized.
My goal in this dissertation was to use computer simulations for studying important biomolecules such as proteins and DNA. In Chapter II, I describe the use of all-atom MD simulations to investigate the intrinsic conformational preferences of amino acids and the extent to which they are modulated by neighboring residues. 441 two-residue peptides containing all twenty standard amino acids, with histidine modeled in both its neutral and protonated states, were simulated. \(^3J_{HNHa}\) coupling constants and chemical shifts calculated from the MD simulations were calculated and compared with recently published experimental measurements for a corresponding set of two-residue peptides. Neighboring residue effects on the average \(^3J_{HNHa}\) and \(\delta_{Ha}\) values of adjacent residues were also examined. NREs on the secondary structure preferences of adjacent amino acids were computed and compared with corresponding effects observed in a coil library and the average \(\beta\)-turn preferences of all amino acid types were determined.

In Chapter III, I describe comparisons of the abilities of three derivatives of the Amber ff99SB force field to reproduce experimental \(^3J_{HNHa}\) scalar coupling constants for small peptides and proteins. All-atom explicit-solvent MD simulations of 256 two-residue peptides and five model proteins ranging in size from 67 to 149 residues were performed. \(^3J_{HNHa}\) coupling constants were computed for all the systems and compared with available experimental measurements. For the two-residue peptide systems, analyses of neighboring residue effects on the \(^3J_{HNHa}\) coupling constants were also carried out. I show that a recently derived residue-specific force field (RSFF2) produces excellent agreements with experimental \(^3J_{HNHa}\) coupling constants of peptides.

In Chapter IV, I describe the application of CG models and multi-scale modeling methods for constructing structural models of the whole *E. coli* chromosome at a resolution of one-bead-per-nucleotide. A combination of
‘morphing’ and Brownian dynamics (BD) simulations were used to build two sets of chromosome models, differing in their positioning of the origin of replication. Further analyses were performed to calculate the physical properties of the models, which were compared with experimental data. The results of Chapters II, III and IV have already been published. Finally, in Chapter V, I summarize the major conclusions of the work described in Chapters II, III, and IV and discuss some possible future directions.
Figure 1.1. Application ranges for computer simulations at different resolutions. The time scales and system sizes are at approximate ranges (Kmiecik et al., 2016).
CHAPTER II
EXPLORING THE CONFORMATIONAL PREFERENCES AND NEIGHBORING RESIDUE EFFECTS OF TWO-RESIDUE PEPTIDES BY MOLECULAR DYNAMICS SIMULATION

Most of the contents of this chapter were published in 2015 in the article “Molecular dynamics simulations of 441 two-residue peptides in aqueous solution: conformational preferences and neighboring residue effects with the Amber ff99SB-ildn-NMR force field” by Shuxiang Li et al. in the Journal of Chemical Theory and Computation 11, 1315-1329. The introduction presented here, however, differs significantly from that in the publication.

Introduction

The native conformational preferences of amino acids in unfolded or denatured proteins have attracted considerable attention in recent years (Avbelj et al., 2006; Hagarman et al., 2010; Kragelj et al., 2013; Oh et al., 2012; Sugase et al., 2007). A thorough understanding of this question could help to interpret the behavior of intrinsically disordered proteins (IDPs). A variety of experimental and theoretical techniques, including spectroscopic methods (IR, NMR, CD) and protein database statistical analyses, have been utilized to explore the intrinsic conformational preferences of the 20 common amino acids in short peptides (Brown and Zondlo 2012; Shi et al., 2005). These studies have been focused on the extent to which the conformational propensities depend not only on the identity of individual residues but also on the identities of neighboring residues. There have been numerous studies from different research groups presenting contradictory evidence regarding the preferred conformations adopted by amino
acid residues in unfolded proteins and peptides (Makowska et al., 2006; Shi et al., 2002; Zagrovic et al., 2005).

Because of the complications associated with interpreting results on large proteins in which the influences of the environment can be complex, short and simple host peptides, into which an amino acid of interest in introduced as a guest, are widely used to explore the conformational propensities of amino acids and effects due to nearest neighbor residues (Avbelj et al., 2006; Jiang et al., 2010; Makowska et al., 2006; Oh et al., 2012). Despite their simplicity, most short peptides are able to adopt the same conformational angles observed for polyproline II (PPII), α helix and β strand motifs in proteins (Jha et al., 2005).

Ever since Tiffany et al. proposed that poly-L-lysine adopts a locally ordered conformation rather than a random coil by comparing its far-UV circular dichroism spectra with that of trans-poly-L-proline (Tiffany and Krimm 1968), a number of groups have published data supporting the notion that amino acid residues in unfolded peptides display more restricted conformations than predicted by the random coil model (Avbelj et al., 2006; Ding et al., 2003; Eker et al., 2004; Garcia 2004; Hagarman et al., 2010; Kelly et al., 2001; Rucker et al., 2003; Shi et al., 2005; Shi et al., 2002). Furthermore, the conformational distributions of amino acids have been shown to be highly influenced by neighboring residues (Avbelj et al., 2006; Hagarman et al., 2006). Shi and coworkers analyzed the conformational distributions of all natural amino acids besides proline and glycine and found that the PPII conformation is dominant for residue X in an AcGGGXGNNH2 host-guest system (Shi et al., 2005). Hagarman and coworkers used GXG peptide systems to investigate the conformational distributions of amino acids and found that except for alanine, both the PPII and β-strand conformations are populated (Hagarman et al., 2011; Hagarman et al., 2010). In addition, a high propensity of turn structures was also observed for short polar
residues in their studies. Grdadolnik and coworkers employed IR, Raman and NMR spectra and reported the relative populations of three major backbone conformations in 19 amino acid dipeptides (Grdadolnik et al., 2011). A similar study was also carried out by Brown et al. who combined CD and NMR spectra to develop a propensity scale for the PPII conformation using a proline-rich host–guest system (Brown and Zondlo 2012). In addition to experimental studies, Jha and coworkers undertook a quite detailed analysis of the backbone conformations based on a protein coil library and concluded that the conformational preferences of amino acids are affected by both the chemical identity and the conformation of the nearest neighboring residues (Jha et al., 2005). Protein coil libraries classify residues outside the α-helices and β-sheets of known protein structures in an attempt to identify amino acids’ intrinsic propensities to adopt particular secondary-structure conformations (Jiang et al., 2010). Furthermore, Jiang et al used a restricted coil library derived from high-resolution protein structures to obtain the local conformational features of amino acids free from the influence of side-chain conformations (Jiang et al., 2010).

As a complementary approach to experimental and coil library studies for investigating the conformational properties of unfolded proteins and peptides, molecular dynamics (MD) simulation using all-atom representations is increasingly utilized to study the effects of sequence and conformational context on backbone conformations. Over the last decade, a number of MD studies have been performed investigating peptide conformations (Beck et al., 2008; Cruz et al., 2012; Feig 2008; Garcia 2004; Graf et al., 2007; Ioannou et al., 2011; Toal et al., 2013). Although the results obtained from MD simulations depend significantly on the force fields as well as the solvent model (Schweitzer-Stenner 2012), recent MD studies that investigated the intrinsic conformational propensities of peptides have been found to be in quite good agreement with experimental $^3J_{HNHa}$
coupling constants (Beauchamp et al., 2012; Nerenberg and Head-Gordon 2011; Toal et al., 2013). Moreover, modified MD force fields have been developed in order to obtain a dominant PPII/β conformation for short peptides, which are the preferred conformations observed in most experimental studies (Adzhubei et al., 2013; Brown and Zondlo 2012; Nerenberg and Head-Gordon 2011).

In this chapter, I used a combination of the Amber ff99SB-ildn-NMR force field (Hornak et al., 2006; Li and Brüschweiler 2012; Lindorff-Larsen et al., 2010) and the TIP4P-Ew water model (Horn et al., 2004) to simulate a large number of two-residue peptides. I used this combination of force field and water model due to its previous success in reproducing NMR observables for peptide systems (Beauchamp et al., 2012). I employed the same procedures to compute $^3J_{\text{HNIH}}$ coupling constants (Hu and Bax 1997) and $\delta_{\text{H}}$ chemical shifts (Shen and Bax 2010) as used in that previous study (Beauchamp et al., 2012), and I compared my results to the corresponding experimental data for as many two-residue peptides as possible (Jung et al., 2014). The comparisons indicate a reasonably high level of correspondence between the simulations and experiment and provide the first truly comprehensive description obtained from explicit-solvent MD simulations of the effects of neighboring residues on the conformational preferences of amino acids.

**Methods**

**Systems studied**

The systems in this work consist of all possible two-residue peptides immersed in $35 \times 35 \times 35$ Å cubic box containing explicit solvent molecules. Since one of my principal objectives was to sample the conformational behavior that peptides would exhibit when embedded within protein sequences, each two-
residue peptide was constructed in its fully extended conformation and then capped (blocked) with an acetyl group (Ace) at the N-terminus and an N-methyl group (Nme) at the C-terminus in order to mimic the adjacent residues that would be present in proteins. Therefore, every two-residue peptide that was simulated had the form Ace-X-Y-Nme, where X and Y represent one of the 20 standard amino acids. Figure 2.1 shows an example of the two-residue peptide (Asp-Asp). For neutral histidine, one proton was added on HE2 position; for protonated histidine, two protons were added on ND1 and HE2 positions, respectively. It should be noted that in this work I use the term “two-residue peptide” to describe these molecules instead of “dipeptide” as the latter is used in multiple ways in the literature (i.e. sometimes it is meant to describe a single amino acid with peptide bonds at both the N- and C-termini – as in the alanine dipeptide – and sometimes it is used to mean two amino acids connected by a single peptide bond). Since histidine is modeled in both its neutral and protonated states, a total of $21 \times 21 = 441$ possible two-residue peptides have been simulated.

Molecular dynamics simulations

All simulations were performed using the MD software package GROMACS version 4.5.1 (Hess et al., 2008; Van der Spoel et al., 2005). Simulations were performed by me and the following members of the Elcock lab: Dr. Casey T. Andrews, Dr. Tamara Frembgen-Kesner, Mark S. Miller, Stephen L. Siemonsma, Timothy D. Collingsworth, Isaac T. Rockafellow, Nguyet Anh Ngo, Brady A. Campbell, Reid F. Brown, Chengxuan Guo, Michael Schrodt and Yu-Tsan Liu. While a number of people performed the simulations, all of the analyses presented here were performed by me. As noted above, the Amber ff99SB-ILDN-NMR force field (Hornak et al., 2006; Li and Brüschweiler 2010;
Lindorff-Larsen et al., 2010) together with the TIP4P-Ew water model (Horn et al., 2004) was used because the combination of this force field and water model has previously been shown to perform very well in reproducing the NMR observables of a wide variety of peptide systems (Beauchamp et al., 2012; Nerenberg and Head-Gordon 2011). All systems were first energy minimized using steepest descent minimization for 1000 steps, gradually heated to 298 K over the course of 350 ps, and then equilibrated for a period of 1 ns. Then, production simulations were carried out in the NPT ensemble for 300 ns, with the temperature maintained at 298 K using the Nosé-Hoover thermostat (Hoover 1985; Nose 1984), and the pressure maintained at 1 atm using the Parrinello-Rahman barostat (Parrinello and Rahman 1981). A cutoff of 10 Å was applied to short-range nonbonded interactions and the PME method (Essmann et al., 1995) was used to calculate all long-range electrostatic interactions. All covalent bonds were constrained to their equilibrium lengths using the LINCS algorithm (Hess et al., 1997), allowing a 2.5 fs time step to be employed. Coordinates of the solutes were collected every 0.1 ps, giving a total of 3 million structures for analysis for each two-residue peptide.

Convergence of molecular dynamics simulations

I calculated the standard deviations of the population distribution for the backbone dihedral angles $\phi$ and $\psi$ to evaluate the completeness of sampling for the MD simulations. For each two-residue peptide, the whole 300 ns simulation period was divided into three time intervals (0-100, 100-200, 200-300ns). The GROMACS utility g_rama was used to calculate the backbone dihedral angles $\phi$ and $\psi$ of each residue during the course of each of the three time intervals. Then, 2D histograms of the sampled $\phi$ and $\psi$ values were constructed using a bin size of 5° and the standard deviation for each bin was calculated over the three 100 ns
intervals. The standard deviations of all bins were summed to obtain total standard deviations and this procedure was used to calculate the standard deviations for both the N-terminal and C-terminal residues of each two-residue peptide. The results were averaged to give a single number for each two-residue peptide. For the peptides that exhibited the largest average standard deviations (Ile-Pro, Ile-Tyr, and Trp-Arg), two independent additional MD simulations, also of 300 ns duration, were performed in order to demonstrate that sampling is likely to be sufficient.

Analysis of conformational distributions

To aid comparison of the conformational preferences of different peptides, each 2D histogram of φ and ψ values was converted into free energy form using the relation:

\[ \Delta G = -RT \ln (f/f_{\text{max}}) \]  

(2-1)

where \( f \) is the frequency of sampling a particular bin in the 2D φ, ψ histogram and \( f_{\text{max}} \) is the maximum frequency found in any of the bins. Since each of the 441 two-residue peptides has two residues, a total of 882 2D free energy distributions were constructed in this way. The data in these histograms were also expressed in an alternative, more concise form as fractional populations of the four major backbone conformations (right-handed helix α, near right-handed helix α’, β-strand, and PPII) computed by summing the populations in the relevant bins. Definitions for these secondary structure conformations were based on Cruz’s work (Figure 2.2) (Cruz et al., 2012), which use the following φ and ψ combinations:
\[ \alpha: -100^\circ \leq \phi \leq 0^\circ, -80^\circ \leq \psi \leq 40^\circ \]
\[ \alpha': -160^\circ \leq \phi < -100^\circ, -80^\circ \leq \psi \leq 60^\circ \]
\[ \text{PPII: } -100^\circ \leq \phi \leq 0^\circ, 70^\circ \leq \psi \leq 180^\circ \]
\[ \beta: -180^\circ \leq \phi < -100^\circ, 70^\circ \leq \psi \leq 180^\circ \]

To obtain average 2D free energy maps and average secondary structure populations that describe the intrinsic conformational preferences of each type of amino acid when at the N- and C-terminal positions, I summed all frequency distributions for peptides that contained the residue of interest, excluding those that had Gly or Pro as neighbors. For example, to calculate the average fractional populations of the different secondary structures for an Ala residue at the N-terminal position, the fractional populations of the N-terminal residue in the peptides Ala-Ala, Ala-Cys, Ala-Gln, …, Ala-Val (excluding Ala-Gly and Ala-Pro) were all combined.

To describe the average effects exerted by each type of amino acid on the properties of neighboring amino acids (i.e. neighboring residue effects: NREs), I use the same procedure used by the Cho group to compute NREs on \( ^3J_{HNH_a} \) and \( \delta_{H_a} \) values, which can be illustrated as follows. To compute the average NRE exerted by a C-terminal Trp on the \( \alpha \) population at the N-terminal position, for example, I consider in turn each of the peptides Ala-Trp, Cys-Trp, Gln-Trp, …, Val-Trp (excluding Gly-Trp and Pro-Trp). For Ala-Trp, I calculated the difference between: (a) the \( \alpha \) population at the N-terminal Ala in Ala-Trp, and (b) the mean \( \alpha \) population at the same position in all peptides Ala-Ala, Ala-Cys, Ala-Gln, …, Ala-Val. I carried out similar calculations for the Cys in Cys-Trp, the Gln in Gln-Trp, etc. and averaged the results to obtain the mean change in \( \alpha \) population at the N-terminal position induced by the presence of a C-terminal Trp. Similar calculations can be performed for all other types of secondary structure populations and for \( ^3J_{HNH_a} \) and \( \delta_{H_a} \) values.
Analysis of β-turn distributions

To calculate populations of β-turn conformations, which are defined in terms of the backbone conformations of four consecutive residues i, i+1, i+2, and i+3, I identify i with the acetyl capping group, i+1 with the N-terminal amino acid, i+2 with the C-terminal amino acid, and i+3 with the N-methyl capping group. I consider all possible β-turns that can be formed with peptide bonds in trans configurations using definitions provided in the literature (Hutchinson and Thornton 1994). It should be noted that these definitions depend only on φ and ψ of residues i+1 and i+2 and so can be meaningfully applied to the capped two-residue peptides studied here even though they do not possess complete amino acids at the i and i+3 positions.

Calculation of $^3J_{HNHa}$ coupling constants

Since the work of Karplus, $^3J$ coupling constants have been interpreted in terms of the dihedral angle connecting the two coupled nuclei (Karplus 1959). To compute $^3J_{HNHa}$ coupling constants from the simulations, the φ backbone dihedral angles of each amino acid were calculated for every simulation snapshot and converted using an empirical Karplus relation of the following formula:

$$^3J(\phi) = A \cos^2(\phi + \theta) + B \cos(\phi + \theta) + C$$ (2-2)

where θ denotes a phase shift value, and A, B, C represent parameters. In this chapter, I used the parameterization of Hu and Bax (Hu and Bax 1997), for which A = 7.09, B = −1.42, C = 1.55, and θ = −60°. While a large number of alternative parameterizations are available (Graf et al., 2007; Hu and Bax 1997; Vuister and Bax 1993), including at least one that has been derived from MD simulations of proteins (Markwick et al., 2009), the particular parameterization used here has
been selected for consistency with the earlier study of the Pande group (Beauchamp et al., 2012) and with other studies (Best et al., 2008; Graf et al., 2007). Those studies have shown that correlation coefficients between simulation and experimental $^3J_{HNHa}$ values are insensitive to the specific choice of Karplus parameters (Beauchamp et al., 2012; Jiang et al., 2014). As is usual, it can be assumed that for most calculations that a single parameterization of the Karplus equation applies equally well to all residue types.

In comparing with the experimental data reported by the Cho group (Jung et al., 2014), all peptides containing either Asp or Glu were omitted. The reason for this is that the experiments for these peptides were performed at pH values between 4 and 5, where sidechains are expected to be partially deprotonated, while our simulations were all performed in a situation where the Asp and Glu sidechains are in fully deprotonated forms (i.e. more appropriate to a pH of 7). For comparison with experiment, I also omit $^3J_{HNHa}$ coupling constants of Gly residues since the experimental data do not identify which of the two Hα protons is probed: experimentally, it has been shown that the $^3J_{HNHa}$ values for the two Hα atoms can differ enormously (e.g. by 5.1 Hz for Gly79 in staphylococcal nuclease (Vuister and Bax 1993)).

Calculation of chemical shifts

Calculations of $\delta_{H\alpha}$ chemical shifts for both residues in each two-residue peptide were performed using the SPARTA+ program (Shen and Bax 2010); owing to the greater computational expense of these calculations they were carried out only on simulation snapshots sampled at 1 ps intervals (a total of 30000 snapshots for each peptide). SPARTA+, in common with other chemical shift prediction algorithms (Han et al., 2011), uses information on the $\phi$, $\psi$ angles at neighboring residues i-1 and i+1 when calculating the chemical shifts at
position i. This makes use of the program to calculate chemical shifts in our peptides challenging, since the first residue is capped with an acetyl group and so does not have a complete neighbor at the i-1 position, and the second residue is capped with a N-methyl group and so does not have a complete neighbor at the i+1 position. This problem can be mitigated by renaming the two capping groups as Ala residues: this is done by renaming the appropriate atoms and residues in the structure (pdb) file of each snapshot. This simple approach allows me to obtain quite good results for the δHα chemical shifts, although it works less well for δHN chemical shifts, which are in any case predicted more poorly by SPARTA+; I therefore report only the former here. It should be also note that, while a chemical shift prediction program has recently been reported that has been specifically parametrized using MD data on proteins (Li and Brüssweiler 2012), the peptides studied here are apparently too short to be used with that program.

Results

Convergence of molecular dynamics simulations

In an attempt to obtain reasonable conformational sampling of the 441 two-residue peptides studied here, comparatively long simulation times of 300 ns were employed. While this is much longer than has been used in other MD studies that have attempted to examine a very large number of peptide systems (Beauchamp et al., 2012; Oh et al., 2012), it is still essential to consider the extent to which such a simulation length allows sampling of the conformational behavior. To determine this, I divided each simulation trajectory into three 100 ns time intervals and calculated the standard deviations of the φ and ψ backbone dihedral angle distributions over these three time intervals. The results from N-
terminal and C-terminal positions were averaged and are shown in Figure 2.3 as a 2D matrix. In this matrix, the x-axis indicates the identity of the N-terminal residue (i.e. residue X in Ace-X-Y-Nme) and the y-axis indicates the identity of the C-terminal residue (i.e. residue Y in Ace-X-Y-Nme). Separate plots for the N-terminal and C-terminal residues are shown in Figure 2.4A and 2.4B, respectively. In these three plots, red squares indicate those two-residue peptides for which standard deviations of the φ and ψ distributions are particularly high.

The largest standard deviations - and therefore the poorest degrees of sampling - are generally associated with two-residue peptides that contain Ile (column I or row I in Figure 2.3) and/or Trp (column W or row W in Figure 2.3); the effects of the former can be understood in terms of the effects of its β-branched sidechain limiting backbone flexibility, the latter may be a consequence of restrictions due to the steric effects of the large side-chain and/or the presence of comparatively long-lived stacking interactions. To further examine sampling issues with the ‘worst case scenario’ peptides, I performed additional 300 ns MD simulations for the Ile-Pro, Ile-Tyr and Trp-Arg peptides. As shown in Figures 2.4, the Ramachandran maps obtained from the 3 independent repeats of each simulation are quite similar to each other, suggesting that reasonable sampling of backbone conformations can probably be obtained in a 300 ns MD simulation even for those two-residue peptides for which conformational sampling is poorest. Interestingly, slightly poorer sampling at the N-terminal position appears to be a common result: a separate examination of the standard deviations for the N-terminal residue (Figure 2.5A) shows that they are, on average, higher than those for the C-terminal residue (Figure 2.5B). In general, however, the Ramachandran maps for most peptides appear to be quite well converged. For example, Figure 2.4D shows the Ramachandran maps obtained from three independent 300 ns MD simulations of the Asp-Val peptide (chosen
because it has the median standard deviation of those displayed in Figure 2.3). For this peptide, which is representative of the majority of peptides, the maps obtained from the simulations are effectively identical in the regions encompassing the \( \alpha, \alpha', \beta \) and PPII conformations, indicating that 300 ns is likely to allow reliable Ramachandran maps to be obtained for most of the peptides studied here.

**Comparison of experimental and simulated \( 3J_{\text{HINH}a} \) coupling constants**

After reasonable conformational sampling of the two-residue peptides were obtained from the simulations, it was first important to determine the extent to which they are consistent with experimental measurements (Jung et al., 2014) before further conformational analyses were carried out. Figure 2.6 compares the calculated and experimental \( 3J_{\text{HINH}a} \) coupling constants for all two-residue peptides that do not contain an Asp or Glu residue. Because of uncertainties in experimental assignments, the coupling constants for Gly residues were omitted. It shows that the overall level of agreement between the 544 computed and experimental \( 3J_{\text{HINH}a} \) coupling constants is quite reasonable with \( r^2 = 0.58 \). When this figure is replotted including data for Gly residues the \( r^2 \) actually improves significantly to 0.68; however, when data for peptides containing Asp or Glu are included, the \( r^2 \) worsens again to 0.64 (data not shown). The degree of agreement with experiment obtained here is similar to that reported recently (Jiang et al., 2014) when the same force field and Karplus parameterization was applied to \( 3J_{\text{HINH}a} \) coupling constants of 19 dipeptides (i.e. capped single amino acids) (Avbelj et al., 2006). The Pearson correlation coefficient here is 0.76, while in that work, the coefficient for the dipeptides is 0.62 (Avbelj et al., 2006). In addition, however, the simulations reported here also reproduce the experimental observation (Jiang et al., 2014) that, within any given
peptide, the $^3J_{HNH\alpha}$ coupling constants of residues at the C-terminal position are generally significantly higher than those at the N-terminal position: in Figure 2.6, C-terminal $^3J_{HNH\alpha}$ coupling constants (red triangles) are consistently shifted upward relative to N-terminal $^3J_{HNH\alpha}$ coupling constants (blue circles).

In experiments, all of the measurements were performed at pH 2 with the exception of the Asp/Glu-containing peptides (Jung et al., 2014). We expect that all histidine sidechains are in a fully protonated state at this pH value. Therefore, we expected better agreement with experiment in those simulations that modeled the protonated form of the residue (Hip) than in those that modeled the neutral form (His). Figure 2.7A and Figure 2.7B show that this expectation is satisfied. Figure 2.7A compares the computed and experimental $^3J_{HNH\alpha}$ coupling constants for those peptides that contain histidines using data from simulations that employed the neutral His residue type; the agreement is clearly very poor ($r^2 = 0.15$). Figure 2.7B on the other hand, shows a corresponding comparison using data obtained from simulations that employed the charged Hip residue type; in this case, the agreement is much better ($r^2 = 0.67$), indicating that the simulations are correctly sensing the overall effects of the side chain’s charge state on the conformational properties of the peptide backbone.

The above results indicate that the simulations perform quite well in reproducing the experimental $^3J_{HNH\alpha}$ coupling constants. As might be expected, however, when the data are examined in more detail, significant discrepancies between simulation and experiment emerge. Figure 2.8A compares the average computed and experimental $^3J_{HNH\alpha}$ coupling constants of each type of amino acid when at the N-terminal position of a two-residue peptide. The agreement between the two sets of data is far from perfect ($r^2 = 0.33$): in particular, relative to their experimental values, the simulation-average $^3J_{HNH\alpha}$ coupling constants for Ala, Leu, Trp and Val are too high while those of Asn, Cys and Thr are too low.
Again, these results are in line with those reported recently for simulations of capped single amino acids using the same force field (Jiang et al., 2014). In their work, Ala and Val were again found to be too high, while Asn, Cys and Ser were too low. Similar discrepancies between the average computed and experimental \(3^{1}J_{HNH_{\alpha}}\) coupling constants are also apparent at the C-terminal position of the two-residue peptides (Figure 2.8B), although here the agreement between simulation and experiment is somewhat higher \((r^2 = 0.45)\).

As is considered further below, discrepancies between the computed and experimental \(3^{1}J_{HNH_{\alpha}}\) coupling constants could be caused by a number of different factors, the most likely of which may be the accuracy of the force field. A comparison between simulation and experiment that we anticipate might be less dependent upon potential errors in the force field is to compare the effect of a given residue type on the average \(3^{1}J_{HNH_{\alpha}}\) coupling constant at a neighboring position. Figure 2.8C shows the effects that each type of amino acid, when present at the N-terminal position, has on the average \(3^{1}J_{HNH_{\alpha}}\) coupling constant at the C-terminal position of the same peptide: red bars show the experimental values calculated as recently described (Jung et al., 2014), while blue bars show the corresponding simulation values. As noted by the Cho group, the most noticeable effect in the experimental data is that aromatic residues at the N-terminal position increase the average \(3^{1}J_{HNH_{\alpha}}\) coupling constants of C-terminal residues by \(\sim 0.3\) Hz. Importantly, this effect is nicely reproduced by the simulations: we find that Phe, Trp, and Tyr at the N-terminus cause upward shifts of the average \(3^{1}J_{HNH_{\alpha}}\) coupling constant at the C-terminus of 0.32, 0.51, and 0.50 Hz, respectively. In fact, the overall level of agreement between the computed and experimental neighboring residue effects (NREs) of N-terminal amino acids is quite high: \(r^2 = 0.80\), and for 14 out of the 17 residue types studied, the sign of the NRE is correctly reproduced.
For the NREs caused by amino acids at the C-terminal position (Figure 2.8D), the agreement is considerably less good but still reasonable ($r^2 = 0.39$). The largest effects seen in the experimental data are provided by the aromatic amino acids Phe and Tyr, which increase the average $^3J_{\text{HNH}_\alpha}$ coupling constants of N-terminal residues by ~0.3 Hz, and the positively charged amino acids Arg and Lys, which decrease them by ~0.2 Hz. Encouragingly, both of these effects are correctly captured by the simulations. Nevertheless, other significant effects that are predicted by the simulations are not apparent in the experimental data. Gly and Trp, for example, are both predicted to exert positive NREs on the $^3J_{\text{HNH}_\alpha}$ values at the N-terminus, but these are not supported by the experimental data; Hip, on the other hand, is predicted to cause a large negative NRE, and while this is qualitatively supported by the experimental data, the actual effect is significantly smaller in magnitude.

Investigation of discrepancies between experimental and simulated $^3J_{\text{HNH}_\alpha}$ coupling constants

Although the overall level of agreement between the computed and experimental $^3J_{\text{HNH}_\alpha}$ coupling constants is quite good, it is evident from the scatter in Figure 2.6 and from the more detailed comparisons in Figure 2.8 that there are notable discrepancies. While the obvious and most likely potential source of disagreement is inadequacies of the simulation force field, there are a number of other factors that might play contributing roles. One possible factor is the degree of convergence of the simulations: it might easily be imagined, for example, that simulations that are less completely sampled might produce worse agreement with experimental $^3J_{\text{HNH}_\alpha}$ coupling constants than simulations that are well sampled. To examine this issue I first arranged the peptides studied experimentally in order of increasing standard deviation of their simulated
Ramachandran maps. I then computed correlation coefficients between the computed and experimental $^3J_{HNH\alpha}$ coupling constants for each of the 10 different groups of peptides with ranks in the range 0-10%, 10-20%, … 90-100%. I also computed the mean unsigned difference between the computed and experimental $^3J_{HNH\alpha}$ coupling constants for each ranked group of peptides. The resulting correlation coefficients and mean unsigned errors are plotted versus the average rank in each group in Figure 2.9A. There is no significant relationship between the extent of conformational sampling and the degree of correlation between the computed and experimental $^3J_{HNH\alpha}$ coupling constants. In fact, surprisingly, those peptides for which sampling of the Ramachandran maps is probably the best (left-most datapoints in Figure 2.9A) produce a somewhat worse correlation with experiment (blue symbols), and a higher error (red symbols) than those for which sampling is likely to be worse (right-hand side of Figure 2.9A).

A second possible factor that might contribute to differences between the computed and experimental $^3J_{HNH\alpha}$ coupling constants is the identity of the capping group employed at the C-terminus of the peptides. In the experiments of the Cho group (Jung et al., 2014) a simple amide group (NH$_2$) was used at the C-terminus, whereas in the simulations an N-methyl group (NHCH$_3$) was used; at the N-terminus, both the experiments and the simulations used an acetyl group. To test the potential effects of the difference at the C-terminus, I randomly selected 20 peptides and performed replicate 300 ns MD simulations using the same amide C-terminal capping group used in the experiments. As shown in Figure 2.9B, the $^3J_{HNH\alpha}$ coupling constants calculated from simulations using the NH$_2$ capping group (residue-type NHE) are effectively identical with those calculated from simulations using the NHCH$_3$ capping group (residue-type NME): the difference between the two corresponding $^3J_{HNH\alpha}$ coupling constants
averages $0.09 \pm 0.09$ Hz, with the largest discrepancy (0.42 Hz) being for Met-Trp which is likely to be poorly sampled. This suggests that differences in the termini are not likely to be a significant cause of discrepancies with experiment.

With the above two issues dealt with, we can consider whether inaccuracies in the force field are likely to be responsible for differences between the simulated and experimental $^3J_{HN\alpha}^\text{values}$. One very real possibility is that differences in the intrinsic conformational preferences of different amino acids might be incorrectly described by the force field (Jiang et al., 2014; Zhou et al., 2015). An indication that this might be the case is provided by Figure 2.10, which shows that quite different regression lines are obtained when the computed and experimental $^3J_{HN\alpha}$ values are compared separately for aliphatic (Ile, Leu, and Val), aromatic (Phe, Trp, and Tyr), “polar” (Cys, Ser, and Thr), and positively charged amino acids (Arg, Hip, and Lys); in particular, the regression line for “polar” amino acids (yellow) is shifted down by ~0.5 Hz relative to that for aliphatic amino acids (blue). We note in passing that an alternative explanation that might be considered for this result is that different parametrizations of the Karplus equation might be required for different types of amino acids. Figure 2.11 shows that the agreement between computed and experimental $^3J_{HN\alpha}$ couplings can be significantly improved if Karplus parameters are optimized separately for individual residue types; a full listing of the resulting residue-specific Karplus parameters is provided in Table 2.1.

**Comparison of experimental and simulated chemical shifts**

In addition to $^3J_{HN\alpha}$ coupling constants, the Cho group also reported $\delta_{H\alpha}$ and $\delta_{HN}$ chemical shifts for both residues in 361 two-residue peptides (Jung et al., 2014). To compare with their data, I used the SPARTA+ program to compute the chemical shifts for all two-residue peptides that do not contain an Asp or
Glu residue and the results are shown in Figure 2.12. The chemical shift values for Gly residues were again omitted owing to assignment ambiguities. The correlation between simulation and experiment is good although the simulated values are clearly shifted upwards relative to experiment. While the generally high level agreement is an encouraging result, it is important to determine the extent to which this results from correct sampling of conformations in the simulations: the chemical shift calculations use a neural network that takes as input the identity of the amino acids involved (Shen and Bax 2010), so it is possible that this information is instead primarily responsible for the good agreement. To explore this issue, I repeated the chemical shift calculations of all peptides with each built in the following, idealized backbone conformations: α-helix, β-strand and PPII. If correct sampling of the backbone distributions is responsible for the good agreement shown in Figure 2.12, then we should expect to obtain poorer agreement with experiment when a single backbone conformation is used in the calculations. Figure 2.13 compares the computed δHα values of all peptides with the corresponding experimental values. All three secondary structures produce good correlations with experiment, with that of the α-helical conformation, surprisingly, being the highest (r² = 0.77). Importantly, however, the correlation coefficient obtained from the MD simulations is considerably higher (r² = 0.84), thereby suggesting that MD’s reasonably realistic sampling of different backbone conformations plays a significant role in determining the agreement with experiment.

In the similar way that NREs on 3JHNHα coupling constants can be compared with experiment, it is possible to compare NREs on the average δHα chemical shifts. Figure 2.14A shows the effects that each type of amino acid, when present at the N-terminal position, has on the average δHα coupling constant at the C-terminal position of the same peptide: red bars show the
experimental values calculated as recently described (Jung et al., 2014), while blue bars show the corresponding simulation values. Experimentally, the largest NREs are again due to the aromatic residues, each of which causes a decrease in the average $\delta_{\alpha}$ chemical shift at a C-terminal position of $\sim$0.1 ppm. Qualitatively, these effects are correctly reproduced by the simulations – the three largest NREs are all due to the aromatic residues – but there are other effects predicted by the simulations that are not mirrored in the experiments: e.g., Ile and Val are both predicted to cause large increases in the average $\delta_{\alpha}$ chemical shifts at the C-terminal position, but much smaller effects are seen experimentally. Despite these discrepancies, the overall level of agreement between the computed and experimental neighboring residue effects (NREs) of N-terminal amino acids is quite high: $r^2 = 0.70$, with the sign of the NRE being correct for 14 out of the 17 residue types studied. Similarly, Figure 2.14B shows the effects that each type of amino acid, when present at the C-terminal position, has on the average $\delta_{\alpha}$ coupling constant at the N-terminal position of the same peptide. As was the case with the $^3J_{\text{HNH}_\alpha}$ coupling constants, the agreement for NREs exerted by C-terminal residues on the average $\delta_{\alpha}$ chemical shifts at the N-terminal position is much less: $r^2 = 0.40$, even though the sign of the NREs is correctly predicted for 13 out of 17 residue types.

Average intrinsic backbone conformational preferences

I proceed now with an analysis of the conformational distributions of all 441 two-residue peptides predicted by the simulations. The individual Ramachandran maps ($\phi/\psi$ distributions) of all peptides are plotted separately for residues at the N-terminal and C-terminal positions in Figures 2.15 and 2.16, respectively. Composite maps that describe the intrinsic backbone conformational preferences of each type of amino acid averaged over all possible
types of neighboring residues are shown in Figures 2.17 and 2.18 for N-terminal and C-terminal positions, respectively. Since in all cases we average only over those neighboring residues that are not Gly or Pro, each of the results shown in Figures 2.17 and 2.18 represents an average of $19 \times 300 \text{ ns} = 5.7 \mu\text{s}$ of simulation data; I therefore expect them to be more reliable than Ramachandran maps of any individual system. Figures 2.19A and 2.19B show the corresponding average fractional populations of the four major backbone conformations that I consider here ($\alpha, \alpha', \beta, \text{ PPII}$) for each residue type at the N-terminal and C-terminal positions, respectively. As was the case with the $3J_{HN\alpha}$ coupling constants described earlier, the average fractional populations of $(\alpha + \alpha'), \alpha$ and PPII obtained from the present simulations are very similar to those reported recently from replica-exchange simulations of 19 dipeptides (capped single residues) performed using the same force field (Zhou et al., 2015) with $r^2 = 0.91$ (Figure 2.20).

Generally speaking, for all non-Pro and non-Gly residue types, the PPII conformation (red squares) is the most preferred conformation, with $\beta$-strand (yellow downward triangles) being the next most preferred conformation (Figure 2.19). Amino acids with less clear preferences for PPII include the negatively charged Asp and Glu, protonated His (Hip) and the small polar amino acids Thr and Ser. Asp and Glu show an increased preference for $\alpha$-helical conformations (blue circles) at both the N- and C-terminal positions, and when found at the N-terminal position (Figure 2.19A), their $\alpha$-helical populations are competitive with those of the PPII conformation. For Hip, on the other hand, the preferred conformation in the simulations is the $\beta$-strand, regardless of whether it is at the N- or C-terminal position. Other residues showing comparatively high preferences for the $\beta$-strand conformation when at the C-terminal position include the three aromatic amino acids (Phe, Trp, and Tyr), Arg and Val. One
final notable effect is that while the α conformation is clearly preferred over α' for residues at the N-terminal position (compare blue circles with green upward triangles), this preference larger disappears for residues at the C-terminal position (Figure 2.19B).

Neighboring residue effects on average backbone conformational distributions

In addition to examining the intrinsic conformational preferences of each type of amino acid, it is also of interest to compute the average effects of each type of amino acid on the conformational preferences of their neighboring residues. Figure 2.21A shows the effects that each type of amino acid, when present at the N-terminal position, has on the average fractional population of each type of secondary structure at the C-terminal position of the same peptide. Perhaps the most obvious effect is that the presence of an aromatic residue (His, Phe, Trp, Tyr) at the N-terminal position strongly decreases the chances of finding a PPII conformation at the C-terminal position (red squares) and, with the exception of His, also has the surprising added effect of increasing the preference for the α' conformation (green triangles). Importantly, it is this increase in the α' population at the C-terminal position that is most responsible for the increases in the average computed $^{3}J_{HNH_{α}}$ coupling constants caused by these residue types in Figure 2.10C, not an increase in the β population. A second clear effect is that the presence of the positively charged amino acids Arg and Lys at the N-terminal position increases the average PPII population at the C-terminal position while suppressing both the α and α' populations (Figure 2.21A); the former effect is responsible for the decrease in the average computed $^{3}J_{HNH_{α}}$ coupling constants caused by these residue types in Figure 2.10C. A third significant observation is that very different NREs are exerted by the different protonation states of the histidine sidechain: the neutral form, His, increases the
α and α’ populations (blue circles and green upward triangles, respectively) at the expense of β and PPII, while the protonated form, Hip, does the opposite.

Figure 2.21B is a corresponding plot showing the effects that each amino acid type, when present at the C-terminal position, has on secondary structure populations at the N-terminal position. The presence of a Pro at the C-terminal position strongly decreases the chances of finding an N-terminal residue in either the α or α’ conformations and accentuates its preference for a PPII conformation; this is in marked contrast to the very small effects exerted by a N-terminal Pro on the conformational preferences at the C-terminal position (Figure 2.21A). Again, the presence of a positively charged amino acid (Arg, Hip, Lys) at the C-terminal position increases the population of the PPII conformation at the N-terminal position, and a corresponding but smaller effect results from the presence of a β-branched amino acid at the C-terminal position (Ile, Val). Again, the two protonation states of histidine produce quite different NREs on the conformational properties of N-terminal residues: a C-terminal Hip increases the average PPII population at an N-terminal position by $\sim$8%, while a C-terminal His decreases it by $\sim$5%.

I can compare all of the above changes in populations with corresponding results obtained from a Bayesian statistical analysis of the “TCB” coil library carried out by the Dunbrack group (Ting et al., 2010). Figure 2.22A compares the change in average PPII populations at the C-terminus accompanying with each type of N-terminal amino acid, while Figure 2.22B shows corresponding effects exerted by C-terminal amino acids on populations at the N-terminus. Analogous plots for NREs exerted on α-helical and β-strand populations are shown in Figure 2.23. For NREs exerted by N-terminal residues (Figure 2.22A), correspondence between the simulations and the PDB analysis is quite poor: a scatter plot of the two datasets has $r^2 = 0.28$ and for only 13 out of 20 residue
types are the same qualitative effects predicted. For NREs exerted by C-terminal residues (Figure 2.22B), on the other hand, agreement is much better: $r^2 = 0.78$, and qualitative agreement is obtained for 15 out of 20 residue types. In particular, the very strong effect exerted by a C-terminal Pro seen in the PDB analysis is reproduced by the simulations, as is the tendency for aromatic residues at the C-terminal position to decrease the populations of PPII at N-terminal residues, and for the β-branched residues Ile and Val to increase the populations of PPII. The tendency for Arg and Lys to increase the PPII conformation at neighboring residues is also a consistent feature of the simulations and the PDB analysis.

**Analysis of β-turn populations**

Since the two-residue peptides contain acetyl groups at the N-terminus and N-methyl groups at the C-terminus they can be considered minimal model systems capable of adopting β-turn conformations. The average populations of each of the various types of β-turns that can form with all-trans peptide bonds are shown in Figure 2.24. In the simulations, the populations of β-I and β-VIII are in general much higher than those of the remaining types β-I', β-II and β-II', although β-II' is common in peptides that contain a Gly at the N-terminal position (i.e. the “i+1” position of a β-turn involving residues i, i+1, i+2 and i+3) and β-II is common in peptides containing a Gly at the C-terminal position (i.e. the “i+2” position). For almost all residue types the population of β-I turns is higher than that of β-VIII; exceptions involve peptides containing N-terminal Arg and Lys residues, for which the populations are very similar (Figure 2.24A) and peptides containing C-terminal Ile and Val residues, for which the β-VIII population is clearly higher (Figure 2.24B).
The preferences of residue types for forming different β-turn types are expressed in the form of β-turn potentials (Hutchinson and Thornton 1994) in Figures 2.25A and 2.25B for the N-terminal and C-terminal positions, respectively. Each set of symbols refers to a different type of β-turn; the β-turn ‘potential’ then indicates the fraction of occurrences of that turn type found in all simulations that involved a residue of a given type. For example, Figure 2.25A shows that in the combined set of (infrequently sampled) β-I’ conformations (cyan upward triangles), the residue most often found at the N-terminal (i+1) position is Hip, with Gly being a close second; Figure 2.24B, on the other hand, shows that in β-II conformations (green downward triangles), which are also infrequently sampled, the residue at the C-terminal (i+2) position is almost always a Gly.

For the more frequently sampled turn types (β-I, β-II and β-VIII), the simulation data can again be compared with statistics compiled from the PDB (Hutchinson and Thornton 1994). Given that 60% of all β-turns formed in the simulations are β-I type turns I show results only for this type of turn; for comparison, 57% of the turns identified by Hutchinson and Thornton in their PDB analysis were of this type (Hutchinson and Thornton 1994). Figures 2.26A and 2.26B compares the β-turn potentials of β-I type turns reported by Hutchinson and Thornton with those calculated here. Figure 2.26A shows results for residues at the i+1 (N-terminal) position, Figure 2.26B shows results for the i+2 (C-terminal) position. For the i+1 position, the correspondence is only modest ($r^2 = 0.40$), but for the i+2 position it is more convincing ($r^2 = 0.63$): in particular, the simulations reproduce the finding that polar sidechains have high turn potentials while Pro and the β-branched Ile and Val have low turn potentials. No closer correspondence is obtained when I compare with an updated β-turn database (Guruprasad and Rajkumar 2000).
Discussion

The overall goal of this chapter has been to provide a reasonably comprehensive view of the conformational properties of two-residue peptides simulated with a very widely used force field and water model combination. The comparisons with experimental $^{3}J_{\text{HNH}_{\alpha}}$ coupling constants and $\delta_{\text{H}_{\alpha}}$ chemical shifts show that the simulations achieve an encouraging level of correspondence with experiment. In particular, neighboring residue effects on the $^{3}J_{\text{HNH}_{\alpha}}$ and $\delta_{\text{H}_{\alpha}}$ values are quite well reproduced, especially so for the NREs exerted by C-terminal residues (Figures 2.8C, 2.7D, 2.13C and 2.13D). Discrepancies with experiment are likely to be primarily a function of inadequacies in both the force field since we have found that neither the convergence of sampling (Figure 2.9A) nor the type of capping group (Figure 2.9B) appear to play significant roles in dictating the level of correspondence with experiment. The results have shown that for $^{3}J_{\text{HNH}_{\alpha}}$ values it is also possible that residue-specific parameterizations of the Karplus equation may be required for full correspondence (Figure 2.13). Agreement between the simulations and previous analyses of conformational distributions in the PDB is less good, but there are obvious structural differences between two-residue peptides and proteins that would place limits on the correspondence expected.

Although the force field and water model combination used in this study has already been shown to be one of the best currently available for reproducing a variety of NMR observables in both peptide (Beauchamp et al., 2012) and protein systems (Long et al., 2011), the simulations reported in this chapter provide an opportunity to explore its ability to reproduce a much larger body of NMR data for two-residue peptides. The generally good reproduction of the experimental $^{3}J_{\text{HNH}_{\alpha}}$ values for two-residue peptides appears notable since the force field was optimized to reproduce NMR data for proteins, not peptide
systems (Li and Brüschweiler 2010), and since the original parameterization did not consider scalar coupling constants but focused instead on chemical shifts. This, together with the observation that the NREs on $J_{\text{HaN}}$ and $\delta_{\text{H}}$ values can also be reasonably well captured reiterates the point made by others that the force field selected here performs quite well (Beauchamp et al., 2012); it is to be noted, however, that it has recently been shown to be incapable of producing a stable structure for the GB1 hairpin (Zhou et al., 2015).

In other words, the agreement between the computed and experimental $J_{\text{HaN}}$ values is far from perfect. Except for sampling issues, it appears most likely that this is due to errors in the force field’s description of the intrinsic conformational preferences of individual residue types since there are systematic differences between the average $J_{\text{HaN}}$ values measured in the simulations and those measured experimentally (Figure 2.8C and 2.8D) that appear to depend on the type of the amino acid. Discrepancies between the computed and experimental $J_{\text{HaN}}$ values for capped single amino acids using the same force field have recently been reported (Jiang et al., 2014). Interestingly, the authors of that work have shown that a reparameterization of the Amber ff99SB force field to match conformational distributions observed in a coil library gave $J_{\text{HaN}}$ values for capped single amino acids in much better agreement with experimental values (Zhou et al., 2015). In the next chapter, therefore, I have examined whether that new force field, RSFF2, leads to improved agreement with experiment for $J_{\text{HaN}}$ values of two-residue peptides (Li and Elcock 2015).

An understanding of the effects that neighboring residues exert on the conformational behavior of other residues is essential for accurate predictions of the conformational ensembles of proteins (Schweitzer-Stenner 2012); such effects have been the subject of a very recent comprehensive review paper (Toal et al., 2015). NREs can be manifested in a number of ways: the PPII propensities (Shi et
al., 2005) and chemical shifts (Braun et al., 1994) of amino acids in short peptides, for example, can depend significantly upon the sequence in which they are embedded, and $^{3}J_{HNH_{\alpha}}$ values in unfolded proteins are typically higher for residues that are preceded by bulky, branched or aromatic residues (Penkett et al., 1997). NREs can also be inferred by conformational analysis of protein structure databases (Jha et al., 2005; Ting et al., 2010). Here, the focus has been on NREs between amino acids that are adjacent to each other in sequence, but it should be remembered that statistical analysis of coil conformations indicates that interactions between residues at i-1 and i+1, for example, can alter the conformational preferences of an intervening residue i (Griffiths-Jones et al., 1998), and that steric interactions of residues that are not immediate neighbors can also affect the conformational possibilities open to oligopeptides (Pappu et al., 2000).

To my knowledge, this is the first simulation effort aimed at explicitly reproducing experimental NREs in a comprehensive set of peptide systems. Implicit solvent simulations performed by the Freed and Sosnick groups on a number of two- and three-residue peptides had previously shown that the conformational properties of Ala are sensitive to the identity of neighboring residues and showed that these NREs were sensitive to the simulation force field (Zaman et al., 2003). Monte Carlo simulations performed by the Pappu group using purely steric potential functions explored the influences of neighboring Gly, Ala, Phe, Val and Pro residues on the secondary structure preferences of amino acids in pentameric host-guest peptides (Tran et al., 2005). More recently, the Liu group has used Hamiltonian-REMD simulations together with the GROMOS 53A6 force field (Oostenbrink et al., 2004) to explore the conformational free energies of a variety of peptide systems, in particular exploring the NREs exerted by Ala, Phe and Val residues on the conformational
behavior of a neighboring Ala (Xu et al., 2008). That study showed that the backbone conformation adopted by the neighboring residue could in some cases play a more significant role in influencing the conformational preferences of a residue of interest than the neighboring residue’s identity (Xu et al., 2008); an important implication of their study is that errors in the intrinsic backbone preferences of amino acids – which are apparent for the force field used here – could lead to errors in their computed NREs. Finally, the Cho group used equilibrium MD simulations of 30 ns duration each to model the conformational behavior of the two-residue peptides that they studied experimentally (Oh et al., 2012); the computed free energy differences between the PPII and β-strand conformations obtained from their MD simulations, however, yielded no correlation with the corresponding experimentally derived values (Oh et al., 2012). It is not clear to what extent the poor correlation that they obtained might have been due to the comparatively short simulation time, the force field used (Amber ff03 (Duan et al., 2003)), or to the decision to average the two $3J_{\text{HNN}}$ values measured for each two-residue peptide in order to obtain average PPII and β-strand populations.

While the correspondence between the simulation and experimental results reported here is by no means quantitative, one clear success of the simulations is in reproducing the experimental results showing that $\delta_{\text{H}\alpha}$ chemical shifts and $3J_{\text{HNH}}$ values in two-residue peptides are altered significantly by neighboring aromatic residues. Previous work had suggested that bulky, branched or aromatic residues (analyzed as one group) tended to increase the β-strand population of following residues (Penkett et al., 1997), and this effect had been interpreted in terms of electrostatic solvation free energy differences between bulky and non-bulky residues in a β-strand model of an (Ala)$_9$ peptide (Avbelj and Baldwin 2004). It is apparently in light of these earlier studies that
the Cho group has interpreted the increased $^3J_{HNH_{\alpha}}$ values caused by neighboring aromatic residues as indicating an increase in the fractional $\beta$-strand population (Jung et al., 2014). Interestingly, however, while the simulations reported here nicely reproduce the experimental changes in $^3J_{HNH_{\alpha}}$ values due to neighboring aromatic residues (Figure 2.8C), the origins of these changes are quite different: here, they are instead caused primarily by an increase in the $\alpha'$ population. Importantly, both interpretations represent plausible explanations of the experimental results: the $\phi$ angle ranges of $\alpha'$ and $\alpha$ conformations are quite similar and their $^3J_{HNH_{\alpha}}$ values are also, therefore, expected to be similar. If the increases in $\alpha'$ population predicted by the simulations were to be confirmed by direct experimentation it would provide nice support for the idea of using simulations as a means of rationalizing and interpreting experimental results.

A final important outcome of the study reported here is that the protonation state of the histidine sidechain can significantly affect both its backbone conformational preferences and those of adjacent residues. Although not explicitly commented upon by the authors, the finding that the protonation state of a residue can affect its intrinsic backbone conformational preferences is apparent also in simulation work by the Vondrášek group for histidine and for a number of other amino acids (Vymetal and Vondrášek 2013). Importantly, the finding that backbone preferences can be affected by protonation state is also supported by experimental work (Avbelj et al., 2006) showing pH-dependent changes of 0.5 Hz in the $^3J_{HNH_{\alpha}}$ values for Asp and Glu; these have been interpreted in terms of changes in protonation state causing substantial changes in the populations of $\beta$-strand conformations (Grdadolnik et al., 2011). Other experimental work has indicated that the PPII propensity of Asp depends significantly upon the protonation state of its sidechain (Brown and Zondlo 2012).
While the effects of sidechain protonation state on the intrinsic preferences of a residue have been noted elsewhere, the results reported in this chapter provide the first indication that significant effects can also be exerted on the conformational preferences of neighboring residues. The largest such effect is seen on the average α-helical population of residues C-terminal to histidine: the average equilibrium constant for the α : PPII equilibrium changes from \((0.32/0.25 = 1.3)\) with His at the N-terminus to \((0.04/0.45 = 0.10)\) with Hip at the N-terminus: in free energy terms this amounts to a relative change in the ΔG for the α:PPII interconversion of ~1.5 kcal/mol at 298 K. If validated experimentally, such a change would have obvious potential implications for attempts to understand how changes in protonation state contribute to the pH dependence of protein stability.
Figure 2.1. Representation of an example two-residue peptide (Asp-Asp). An acetyl group (Ace) is capped at the N-terminus and an N-methyl group (Nme) is capped at the C-terminus.
Figure 2.2. Representation of the phi and psi dihedral angles and a Ramachandran map with the most preferred conformations. Energies are colored in descending order from blue to red: blue regions represent highest free energies and are correspondingly the least populated conformations while red regions represent the lowest free energies and the most populated conformations.
Figure 2.3. Convergence of average Ramachandran maps in MD simulations. The calculated standard deviations of the backbone $\phi$ and $\psi$ distributions for 441 two-residue peptides. The residue at the N-terminus of each peptide is plotted along the x-axis; the residue at the C-terminus is along the y-axis. Note that Hp corresponds to the protonated form of histidine.
Figure 2.4. Ramachandran maps obtained from 3 replicate MD simulations of selected peptides. A: N-terminal and C-terminal Ramachandran maps expressed in free energy form for simulations of Ile-Pro. B. Same as A but showing results for Ile-Tyr. C. Same as A but showing results for Arg-Trp. D. Same as A but showing results for Asp-Val. Energies are colored in descending order from blue to red.
Figure 2.5. Convergence of Ramachandran maps in MD simulations for termini. The calculated standard deviations of the backbone $\phi$ and $\psi$ distributions for 441 two-residue peptides at N-terminus (A), C-terminus (B). The residue at the N-terminus of each peptide is plotted along the x-axis; the residue at the C-terminus is along the y-axis. Note that Hp corresponds to the protonated form of histidine.
Figure 2.6. Comparison of computed and experimental $3J_{HN\alpha}$ coupling constants for two-residue peptides. Plot comparing simulation and experimental $3J_{HN\alpha}$ coupling constants for all non-Gly residues in peptides that do not contain Pro, Asp or Glu; experimental data taken from literature (Jung et al., 2014); black line shows linear regression.
Figure 2.7. Comparison of simulation and experimental $^3J_{HN\alpha}$ coupling constants for residues in histidine-containing peptides. A. simulations modeled histidine using the neutral, His residue type. B. Same as A but plotting data obtained from simulations that modeled histidine using the charged, Hip residue type. Each dot represents a histidine containing peptide.
Figure 2.8. Comparison of computed and experimental $^3J_{HNH\alpha}$ coupling constants averaged by type of amino acid. A. Plot showing average $^3J_{HNH\alpha}$ coupling constant of each type of amino acid when present at the N-terminal position; all averages obtained from data on 17 peptides (all peptides that do not involve Pro, Asp or Glu); error bars indicate standard deviations. B. Same as A but showing results for each type of amino acid when present at the C-terminal position. C. Plot showing the effect of each type of amino acid, when present at the N-terminal position, on the average $^3J_{HNH\alpha}$ coupling constant of amino acids at the C-terminal position; error bars indicate standard deviations. D. Same as C but showing effect of each type of amino acid, when present at the C-terminal position, on the average $^3J_{HNH\alpha}$ coupling constant of amino acids at the N-terminal position.
Figure 2.9. Factors influencing agreement between computed and experimental $^3J_{HH\alpha}$ coupling constants. A. Plot showing correlation coefficient (blue) and mean unsigned difference (red) between simulation and experimental $^3J_{HH\alpha}$ coupling constants as a function of the standard deviations of the Ramachandran maps sampled during MD simulations. Peptides are grouped according to their rank in an ordered list of Ramachandran map standard deviations; the datapoint at far-left shows the correlation coefficient obtained for those peptides with standard deviations in the lowest 10%; the datapoint at far-right shows the same for those peptides with standard deviations in the highest 10%. B. Plot comparing simulation $^3J_{HH\alpha}$ coupling constants obtained from simulations of identical peptides using NHE (simple amide) and NME (N-methylamide) capping groups.
Figure 2.10. Comparison of computed and experimental $^3J_{HNHa}$ coupling constants for peptides grouped according to their constituent amino acids. Data points marked aliphatic (blue), for example, are for peptides that contain only Ile, Leu, or Val residues.
Figure 2.11. Residue-specific Karplus parameters improve agreement with experiment. Same as Figure 2.6 in the main text, but showing simulation values calculated using Karplus equations independently adjusted for each type of amino acid in order to maximize agreement with the experimental data. This has been achieved by implementing a simple Monte Carlo optimization scheme that allows each of the three coefficients used in the Karplus equation to vary independently for each type of amino acid; trial changes to a residue-type’s Karplus coefficients were accepted on the basis of improvements to the combined absolute error of all coupling constants of all peptides containing that type of amino acid. We note that attempts to use a single Karplus equation that was optimized for converting the MD distributions into experimental values did not produce results significantly better than those obtained using the experimentally-derived Karplus parameters of Hu and Bax (Hu and Bax 1997).
Figure 2.12. Comparison of computed and experimental $\delta_{H\alpha}$ chemical shifts.
Plot shows all non-Gly residues in peptides that do not contain Pro, Asp or Glu.
Figure 2.13. Comparison of computed and experimental $\delta_{\text{H}_\alpha}$ chemical shifts with different backbone conformations. Plot shows results computed when each peptide is restricted to one of three different backbone conformations.
Figure 2.14. Comparison of computed and experimental $\delta_{H\alpha}$ chemical shifts with neighboring residue effects. A. Plot showing the effect of each type of amino acid, when present at the N-terminal position, on the average $\delta_{H\alpha}$ value of amino acids at the C-terminal position; error bars indicate standard deviations. B. Same as C but showing effect of each type of amino acid, when present at the C-terminal position, on the average $\delta_{H\alpha}$ value of amino acids at the N-terminal position.
Figure 2.15. Ramachandran maps for the N-terminal residues sampled in MD simulations of 441 two-residue peptides. Rows identify the N-terminal residue in the two-residue peptide and columns identify the C-terminal residue. Protonated histidine is indicated by $H_p$. Energies are colored in descending order from blue to red: blue regions represent highest free energies and are correspondingly the least populated conformations while red regions represent the lowest free energies and the most populated conformations.
Figure 2.16. Ramachandran maps for the C-terminal residues sampled in MD simulations of 441 two-residue peptides. Rows identify the N-terminal residue in the two-residue peptide and columns identify the C-terminal residue. Protonated histidine is indicated by $H_P$. Energies are colored in descending order from blue to red: blue regions represent highest free energies and are correspondingly the least populated conformations while red regions represent the lowest free energies and the most populated conformations.
Figure 2.17. Average Ramachandran maps at the N-terminus. Plots showing simulation Ramachandran maps in free energy form for all 21 types of amino acids averaged over all possible C-terminal residues excluding Pro and Gly. Map at top-left, for example, shows the average Ramachandran map of the N-terminal Ala in all peptides of the form Ala-Ala, Ala-Cys, Ala-Gln, Ala-Glu, etc. Free energies are colored in descending order from blue to red.
Figure 2.18. **Average Ramachandran maps at the C-terminus.** Plots showing simulation Ramachandran maps in free energy form for all 21 types of amino acids averaged over all possible N-terminal residues excluding Pro and Gly. Map at top-left, for example, shows the average Ramachandran map of the C-terminal Ala in all peptides of the form Ala-Ala, Cys-Ala, Gln-Ala, Glu-Ala, etc. Free energies are colored in descending order from blue to red.
Figure 2.19. Fractional populations of four major backbone conformations. A. Simulation fractional populations of $\alpha$, $\alpha'$, $\beta$ and PPII conformations for all 21 types of amino acids when present at the N-terminal position; results are averaged over all possible C-terminal residues excluding Pro and Gly. B. Same as A but showing results for all 21 types of amino acids when present at the C-terminal position; results are averaged over all possible N-terminal residues excluding Pro and Gly.
Figure 2.20. Comparison of average secondary structure populations observed in this study for two-residue peptides with those obtained for single-residue peptides (Zhou et al., 2015). Average fractional populations for this work are the average over both the N-terminal and C-terminal positions.
Figure 2.21. Neighboring residue effects on the average fractional populations of four major backbone conformations. A. Plot showing effect of each type of amino acid, when present at the N-terminal position, on the average fractional populations of α, α’, β and PPII conformations of amino acids at the C-terminal position. B. Same as A but showing effect of each type of amino acid, when present at the C-terminal position, on the average fractional populations at the N-terminal position.
Figure 2.22. Comparison of simulated effect fractional populations of four major backbone conformations with that obtained from a PDB-derived coil library. **A.** Plot comparing the simulated effect of each type of amino acid, when present at the N-terminal position, on the average fractional PPII population at the C-terminal position with that obtained from analysis of a coil library. **B.** Same as A but comparing the simulated effect of each type of amino acid, when present at the C-terminal position, on the average fractional PPII population at the N-terminal position.
Figure 2.23. Neighboring residue effects on α and β fractional populations. A. Plot comparing the simulated effect of each type of amino acid, when present at the N-terminal position, on the average fractional α population at the C-terminal position with that obtained from analysis of a coil library. B. Same as A but comparing the simulated effect of each type of amino acid, when present at the C-terminal position, on the average fractional α population at the N-terminal position. C. Same as A but showing results for the average fractional β population. D. Same as B but showing results for the average fractional β population.
Figure 2.24. Populations of β-turn conformations obtained from MD simulations. A. MD-sampled populations of five types of β-turn conformation plotted as a function of the identity of the amino acid at the ‘i+1’ (i.e. N-terminal) position. B. Same as A but plotted as a function of the identity of the amino acid at the ‘i+2’ (i.e. C-terminal) position.
Figure 2.25. Propensity of each type of amino acid to be found for each type of β-turn. A. Plot showing the propensity of each type of amino acid to be found at the ‘i+1’ position for each type of β-turn. B. Same as A but showing the propensity of each type of amino acid to be found at the ‘i+2’ position.
Figure 2.26. Comparison of computed β-turn conformations with that obtained from an analysis of PDB structures. A. Plot comparing the MD-computed turn potential of each type of amino acid for the ‘i+1’ position of type β-I turns with that obtained from analysis of PDB structures. B. Same as A but showing results for the ‘i+2’ position.
Table 2.1. Residue-specific Karplus parameters derived by optimizing agreement between simulation and experiment for two-residue peptides. Values are given for each of the parameters A, B, C found in the Karplus equation; θ in all cases was set to 60°. Standard deviations are from three independent repeats of the optimization procedure: each started with initial values of A, B, and C taken from (Hu and Bax 1997). Note that no parameters are reported for Asp, Glu or Gly given the uncertainties in connecting simulation with experiment for these amino acids.

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>A (Hz)</th>
<th>B (Hz)</th>
<th>C (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>8.09 ± 0.11</td>
<td>0.91 ± 0.13</td>
<td>1.97 ± 0.04</td>
</tr>
<tr>
<td>Arg</td>
<td>7.23 ± 0.02</td>
<td>-0.20 ± 0.01</td>
<td>2.13 ± 0.00</td>
</tr>
<tr>
<td>Asn</td>
<td>4.99 ± 0.12</td>
<td>-0.66 ± 0.17</td>
<td>3.77 ± 0.06</td>
</tr>
<tr>
<td>Cys</td>
<td>6.78 ± 0.02</td>
<td>-0.74 ± 0.02</td>
<td>2.69 ± 0.00</td>
</tr>
<tr>
<td>Gln</td>
<td>5.06 ± 0.01</td>
<td>-1.04 ± 0.00</td>
<td>2.89 ± 0.01</td>
</tr>
<tr>
<td>Hip</td>
<td>3.41 ± 0.10</td>
<td>-2.08 ± 0.00</td>
<td>3.71 ± 0.01</td>
</tr>
<tr>
<td>Ile</td>
<td>9.10 ± 0.01</td>
<td>0.07 ± 0.01</td>
<td>1.48 ± 0.01</td>
</tr>
<tr>
<td>Leu</td>
<td>12.14 ± 0.01</td>
<td>5.31 ± 0.42</td>
<td>2.93 ± 0.12</td>
</tr>
<tr>
<td>Lys</td>
<td>2.52 ± 0.31</td>
<td>-5.54 ± 0.02</td>
<td>1.17 ± 0.01</td>
</tr>
<tr>
<td>Met</td>
<td>7.35 ± 0.01</td>
<td>-0.48 ± 0.37</td>
<td>2.07 ± 0.14</td>
</tr>
<tr>
<td>Phe</td>
<td>7.87 ± 0.22</td>
<td>-0.65 ± 0.01</td>
<td>1.75 ± 0.00</td>
</tr>
<tr>
<td>Ser</td>
<td>4.17 ± 0.01</td>
<td>-3.21 ± 0.01</td>
<td>2.27 ± 0.01</td>
</tr>
</tbody>
</table>
Table 2.1-continued

<table>
<thead>
<tr>
<th></th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thr</td>
<td>5.44 ± 0.07</td>
<td>-2.05 ± 0.11</td>
<td>2.64 ± 0.04</td>
</tr>
<tr>
<td>Trp</td>
<td>8.43 ± 0.01</td>
<td>-0.32 ± 0.00</td>
<td>1.14 ± 0.01</td>
</tr>
<tr>
<td>Tyr</td>
<td>11.3 ± 0.03</td>
<td>3.38 ± 0.01</td>
<td>2.59 ± 0.01</td>
</tr>
<tr>
<td>Val</td>
<td>9.48 ± 0.01</td>
<td>0.33 ± 0.02</td>
<td>1.13 ± 0.02</td>
</tr>
</tbody>
</table>
CHAPTER III
VALIDATION OF THREE AMBER FF99SB DERIVED FORCE FIELDS WITH EXPERIMENTAL NMR MEASUREMENTS ON PEPTIDE AND PROTEIN SYSTEMS

The contents of this chapter were published in 2015 in the article “A Residue-Specific Force Field (RSFF2) improves the modeling of conformational behavior of peptides and proteins” by Shuxiang Li and Adrian H. Elcock in the Journal of Physical Chemical Letters 6, 2127–2133.

Introduction

Molecular dynamics simulations using Newton's equations of motion and pairwise-additive force fields have emerged as an important complement to experimental approaches for investigating the conformational dynamics of biological macromolecules. (Baker 2015; Chen et al., 2017; Dror et al., 2012; Sakuraba and Kono 2016). Although a significant advantage of using MD is that it provides a truly atomic-level view of biomolecular dynamics, the predictive utility of the method relies heavily on the accuracy of the force fields (Beauchamp et al., 2012; Piana et al., 2014). In recent years, owing in part to the realization that force fields for proteins can show biased secondary structure preferences (Best et al., 2008; Freddolino et al., 2009; Wang and Wade 2006), considerable effort has been focused on improving the backbone and side chain dihedral angle potential functions of a number of force fields (Li and Brüschweiler 2010; Lindorff-Larsen et al., 2010; Nerenberg and Head-Gordon 2011). Although continuing improvements mean that current force fields are generally adequate for modeling globular proteins in their native states, there remain questions about their ability to accurately reproduce the thermodynamics of protein folding (Raval et al., 2012) and both the conformational propensities
(Nerenberg and Head-Gordon 2011) and scalar coupling constants of peptide systems (Graf et al., 2007; Li et al., 2015).

One very useful experimental measure of the conformational behavior of protein and peptide systems is the $^3J_{HN\alpha}$ scalar coupling constant. This measurement provides a relatively direct measure of the backbone dihedral angle, $\phi$, which can be used to validate the accuracy of conformations sampled from MD simulations (Graf et al., 2007). In an important recent work, the Cho group has reported $^3J_{HN\alpha}$ coupling constants for each residue in a complete set of 361 two-residue peptides (Jung et al., 2014). The availability of their dataset provides an excellent chance to test the abilities of current simulation force fields to describe the conformational properties of model peptide systems. In Chapter II, I have reported MD simulations of the 441 possible two-residue peptides that can be formed from the twenty common amino acids when histidine is modeled in both its neutral and protonated states (Li et al., 2015). In that work, using a force field combination that had previously been shown by the Pande group to be among the best at reproducing $^3J$ coupling constants and chemical shifts in other peptide systems, I obtained a reasonable, but far from perfect agreement between the MD-predicted $^3J_{HN\alpha}$ coupling constants and the experimental values reported by the Cho group (Jung et al., 2014). Although the termini of the two-residue peptides employed in that work are not exactly the same as those used in experiments, it has been demonstrated that differences in the termini are not likely to be a significant cause of discrepancies with experiment (see results in Chapter II).

As is considered in Chapter II, the most likely factor that causes the discrepancies between the computed and experimental $^3J_{HN\alpha}$ coupling constants is the accuracy of the force field. Therefore, in this chapter, I extend the work in Chapter II and explicitly test the abilities of three recent iterations of the AMBER
protein force field to reproduce the $^3J_{\text{HNHa}}$ coupling constants of a large number of two-residue peptides. The basis of all three force fields tested here is the influential ff99SB force field developed by the Simmerling group in order to improve the description of $\phi/\psi$ torsions in peptide and protein simulations (Hornak et al., 2006). The first force field that I test, ff99SB-ILDN (Lindorff-Larsen et al., 2010), adds to the original ff99SB parameterization a set of modified dihedral functions developed by the Shaw group to improve modeling of Ile, Leu, Asp and Asn sidechains. The second, ff99SB-ILDN-NMR (Li and Brüschweiler 2010), adds to this the modified backbone dihedral functions parameterized by the Brüschweiler group to improve reproduction of chemical shifts in protein simulations. The final force field that I test, RSFF2 (Zhou et al., 2015), represents a quite different extension of ff99SB that includes additional 1-4, 1-5 and 1-6 non-bonded functions designed to reproduce the conformational distributions of the twenty common amino acids observed a protein coil library. Based on these three AMBER ff99SB derived force fields in combination with the TIP3P water model, extensive molecular dynamics simulations on 256 two-residue peptides (totally 783 simulations, 1 µs each) and 5 folded proteins were performed to evaluate their performance in reproducing NMR scalar coupling data.

**Methods**

**Systems studied**

In this chapter, 256 blocked two-residue peptides of the form Ace-X-Y-NH$_2$ were constructed in fully extended conformations, where X and Y represent any of the standard amino acids, excluding Pro, Asp, Glu and His; the acetylated N-terminus and amidated C-terminus correspond to the capping groups used in the Cho group’s recent NMR studies (Jung et al., 2014). Pro was omitted because
it was not included in the experimental studies. Asp and Glu were omitted because they are likely to be in a mixture of protonation states at the pH studied experimentally. Finally, His was omitted because its protonated form (appropriate to the experimental pH) has not yet been explicitly parameterized in RSFF2. Figure 3.1 shows an example of the structure of a two-residue peptide (Ala-Ala) simulated in this chapter. In addition to these peptide systems, five different proteins were each simulated; these are listed, together with their corresponding PDB codes and residue counts, as follows: the Villin headpiece domain (1QQV; 67 residues), Ubiquitin (1UBQ; 76 residues), Nonspecific Lipid Transfer Protein (1BV2; 91 residues), the PPIase domain of the chaperone trigger factor (1HXV; 113 residues), and Staphylococcal nuclease (SNase) (1SNC; 149 residues). All simulations were started from the native state structure. In the exceptional case of the PPIase domain, 30 amino acids at the N-terminus are missing from the structure: these residues – for which \( ^3J_{HNHa} \) coupling constants were not reported – were built in a random coil configuration so that the construct used in the simulation was identical, chemically, to that used in the experiments. The cartoon structures of those five proteins are shown in Figure 3.2.

Molecular dynamics simulation setup

All simulations were performed using the molecular dynamics software package GROMACS version 4.6.5 (Hess et al., 2008; Van der Spoel et al., 2005). Simulations were performed with each of the following force fields: Amber ff99SB-ILDN (Lindorff-Larsen et al., 2010), Amber ff99SB-ILDN-NMR (Li and Brüschweiler 2012), and RSFF2 (Zhou et al., 2015), all in combination with the TIP3P water model (Jorgensen et al., 1983). The first two of these force fields are already implemented in GROMACS; RSFF2 was implemented using Python scripts kindly provided to us by the developers of that force field. For the 256
two-residue peptide simulations, each peptide was immersed in a 30 Å octahedral box containing explicit solvent. For proteins, each protein was immersed in a cubic box large enough so that every atom of the protein was at least 1.0 nm from the box edge. All systems were first energy minimized using steepest descent minimization for 1000 steps, gradually heated to the target temperature over the course of 420 ps, and then equilibrated for a period of 1 ns. The target temperatures were chosen to match experimental conditions: two-residue peptides were simulated at 298 K, Villin at 293 K (Vardar et al., 1999), Ubiquitin at 303 K (Wang and Bax 1996), NLTP at 298 K (Poznanski et al., 1999), PPIase at 298 K (Parac et al., 2001), and SNase at 308 K (Vuister and Bax 1993). All production simulations were carried out in the NPT ensemble for 1µs with the coordinates of the solutes saved every 0.6 ps. Following the protocol used in the original development of RSFF2, a cutoff of 9 Å was applied to short-range nonbonded interactions, the PME method (Essmann et al., 1995) was used to calculate all long-range electrostatic interactions, covalent bonds were constrained to their equilibrium lengths using the LINCS algorithm (Hess et al., 1997) and a 3 fs time step was employed throughout.

Calculation of $^3J_{HNH\alpha}$ coupling constants

To compute $^3J_{HNH\alpha}$ coupling constants, a similar method from Chapter II with some modifications was used here. The $\phi$ backbone dihedral angle of each residue of interest was computed for every snapshot from the MD simulations and converted into a $^3J_{HNH\alpha}$ coupling constant using an empirical Karplus relation (Graf et al., 2007):

$$J(\phi) = A \cos^2(\phi + \theta) + B \cos(\phi + \theta) + C$$  \hspace{1cm} (3.1)

where $A, B, C$ represent parameters, and $\theta$ denotes a phase shift value. In this
Chapter, I employed two Karplus parameter sets for computing $^{3}J_{\text{HNHa}}$ coupling constants. One set, which is used in the work of Chapter II, has the parameters due to Hu and Bax (Hu and Bax 1997) (expressed in Hz):

$$A = 7.09, B = -1.42, C = 1.55, \theta = -60^\circ.$$ 

The second set, due to Vögeli et al. (Vögeli et al., 2007), has the parameters:

$$A = 8.40, B = -1.36, C = 0.33, \theta = -60^\circ.$$ 

Previous studies have shown that the extent of correlation between simulation and experiment is effectively independent of the exact set of parameters used in the Karplus equation (Jiang et al., 2014). With the exception of the Gly residues of two-residue peptides, all calculated $^{3}J_{\text{HNHa}}$ coupling constants were compared with corresponding experimental data; Gly residues of two-residue peptides were omitted, as in our previous work, owing to uncertainties in assignment of the two H$_{\alpha}$ atoms (Jung et al., 2014).

Calculation of neighboring residue effects

Neighboring residue effects (NREs), which quantify the extent to which a residue type causes a neighboring residue type’s $^{3}J_{\text{HNHa}}$ coupling constant to deviate from its mean value, were computed as described in Chapter II. For example, the NRE exerted by a C-terminal Phe on the $^{3}J_{\text{HNHa}}$ coupling constant of an N-terminal Ala, is calculated as the difference between: (a) the $^{3}J_{\text{HNHa}}$ coupling constant of the N-terminal Ala in Ala–Phe, and (b) the mean $^{3}J_{\text{HNHa}}$ coupling constant of the N-terminal Ala in all possible peptides Ala–Ala, Ala–Cys, Ala–Gln, ..., Ala–Val. To obtain the average NRE caused by a C-terminal Phe, similar calculations for the Cys in Cys–Phe, the Gln in Gln–Phe, and so forth were carried out and the results were averaged. In order to allow an unambiguous
comparison with the experimental data reported by the Cho group, experimental NREs were recomputed here using the original data for only those peptides that were simulated.

Simulation differences for ff99SB-ILDN-NMR force field

The MD simulations reported in this chapter for ff99SB-ILDN-NMR differ from those I reported in chapter II using the same force field in the following ways: (1) the simulation times in chapter III are considerably longer (1 µs) than those in chapter II (300 ns), (2) the water model used in chapter III is TIP3P versus TIP4P-Ew in chapter II, and (3) the peptides are simulated in chapter III using the same capping groups used in the experimental studies. Despite these differences, the results I obtained here are very similar to those I reported previously (Li et al., 2015) ($r^2 = 0.92$; Figure 3.3); the principal differences involve large C-terminal sidechains for which differences in the extent of sampling issues and in the identity of the C-terminal capping group may be significant.

Results

Comparison of experimental and simulated $^3J_{HN\alpha}$ coupling constants for 256 two-residue peptides

Figure 3.4 compares the MD-computed $^3J_{HN\alpha}$ coupling constants with the corresponding experimental measurements (Jung et al., 2014) for all non-Gly residues for each of the force fields tested. All three force fields produce a reasonable correspondence with the experimental data, and all three reproduce the finding that the $^3J_{HN\alpha}$ coupling constants of C-terminal residues (red triangles) are generally significantly higher than those of N-terminal residues (blue circles) in two-residue peptides. ff99SB-ILDN-NMR ($r^2 = 0.59$) performs slightly better than ff99SB-ILDN ($r^2 = 0.56$), but the two sets of predicted values are highly correlated with each other ($r^2 = 0.96$; Figure 3.5); the close
correspondence between these independent predictions suggests that the simulations are sufficiently converged to give statistically reliable results. Compared to the two ILDN-based force fields, RSFF2 performs much better and produces an overall level of agreement with experiment that is very high (r² = 0.82). When linear regressions of the 3JHN coupling constants are performed separately for the N-terminal and C-terminal residues it is apparent that RSFF2 improves results significantly at both positions, with the r² for the N-terminal residues, in particular, improving dramatically from 0.17 and 0.22 with ff99SB-ILDN and ff99SB-ILDN-NMR, respectively, to 0.69 with RSFF2; the corresponding r² values for the C-terminal residues are 0.47, 0.44 and 0.83 respectively. Effectively identical sets of r² values are obtained when the 3JHN coupling constants are instead calculated using the earlier Karplus parametrization due to Hu and Bax (Hu and Bax 1997) (Figure 3.6). Therefore, it is not surprising to find that the 3JHN coupling constants calculated with the two sets of Karplus parameters are highly correlated with each other (Figure 3.7). Interestingly, the root-mean-square (RMS) errors in the computed 3JHN coupling constants are generally lower with the more recent Vögeli et al. parameter set (Vögeli et al., 2007) (Figure 3.8).

Comparison of experimental and simulated neighboring residue effects

for 256 two-residue peptides

As described in Chapter II, it is also possible to compare the neighboring residue effect (NRE) that each residue exerts on the 3JHN coupling constant of an adjacent residue with corresponding experimental data. Figure 3.9 compares the MD-computed NREs on 3JHN coupling constants with the corresponding experimental measurements (Jung et al., 2014) for all three force fields. Here, the level of agreement with experiment is consistently lower, but with all three force
fields the NREs exerted by N-terminal residues on neighboring C-terminal residues are significantly better reproduced than those exerted by C-terminal residues on N-terminal neighbors (see r^2 values for linear regressions listed in Figure 3.9). Given that the $^3J_{\text{HNHa}}$ coupling constants of the C-terminal residues are consistently better predicted than those of the N-terminal residues (Figure 3.4), it is possible that this indicates that a prerequisite for successful reproduction of the NRE exerted at a given residue is that the $^3J_{\text{HNHa}}$ coupling constant of that residue, and by implication its backbone conformational preferences, be accurately reproduced.

In terms of the relative abilities of the force fields to reproduce the NREs it is interesting to note that though there is an apparent increase in the overall r^2 in passing from ff99SB-ILDN to ff99SB-ILDN-NMR to RSFF2 (Figure 3.9), it only appears significant for the NREs exerted by C-terminal residues. The NRE data, grouped and averaged according to residue type, are shown in bar-graph form in Figures 3.10 and 3.11. All three force fields successfully reproduce the principal experimental finding that aromatic amino acids such as Phe and Tyr tend to increase the $^3J_{\text{HNHa}}$ coupling constants of adjacent residues, while basic amino acids Arg and Lys at the C-terminal position tend to decrease the $^3J_{\text{HNHa}}$ coupling constants of N-terminal residues (Figure 3.11). Other results, however, are not reproduced by any of the force fields: for example, the negative NRE exerted by a C-terminal Gly is not predicted, positive NREs exerted by C-terminal β-branched residues are not predicted, and an erroneously large positive NRE is predicted to be caused by a C-terminal Trp by all three force fields (Figure 3.11). Interestingly, RSFF2 significantly improves predictions of the NREs caused by Asn, regardless of position, but worsens predictions of the NREs caused by Arg and Lys at the N-terminal position (Figure 3.10). A visual comparison of the r^2 values for all of the data shown in Figures 3.4 and 3.9 is provided in Figure 3.12. From this it is
clear that RSFF2 results in a dramatic improvement in the reproduction of $^3J_{HN\alpha}$ coupling constants but a much smaller, perhaps insignificant, improvement in the reproduction of NREs.

*Comparison of experimental and simulated $^3J_{HN\alpha}$ coupling constants for five protein systems*

To explore whether the above results carry over to proteins I carried out MD simulations of five model proteins, ranging in size from 67 to 149 residues (Figure 3.2), for which structures and experimental $^3J_{HN\alpha}$ coupling constants have been reported in the literature. I simulated each protein for 1 µs, using each of the three force fields, and performed linear regressions of the computed $^3J_{HN\alpha}$ coupling constants with the corresponding experimental measurements. Three independent replicate 1 µs simulations of staphylococcal nuclease (SNase), which is the largest protein studied, were performed in order to determine the extent to which the results might be sensitive to the initial velocities assigned in the simulations. A complete set of scatter plots comparing the computed $^3J_{HN\alpha}$ coupling constants with experimental values for all of these simulations are shown in Figure 3.13. A visual comparison of the $r^2$ values for all linear regressions of simulation with experiment is shown in Figure 3.14A, and a plot of the RMS errors in the predicted $^3J_{HN\alpha}$ coupling constants is shown in Figure 3.14B; corresponding plots obtained when the calculations are repeated using the Hu and Bax parameters (Hu and Bax 1997) are provided in Figure 3.11.

The regression results for the protein simulations generally confirm the trends established above by the two-residue peptide simulations: Figure 3.14A and Figure 3.15A show that the $r^2$ values obtained with RSFF2 always exceed those obtained with either of the two other Amber ff99SB force fields, whereas those obtained with ff99SB-ILDN-NMR generally exceed those obtained with
ff99SB-ILDN. Averaged over the seven simulations performed with each force field, the increase in $r^2$ due to changing from ff99SB-ILDN to RSFF2 is $0.075 \pm 0.040$; the increase in $r^2$ due to changing from ff99SB-ILDN-NMR to RSFF2 is $0.054 \pm 0.041$ (for this latter calculation we omitted villin, for which the ff99SB-ILDN-NMR results appear anomalously poor; see Figure 3.14A).

When the RMS errors in the predicted $^3J_{HNH_a}$ coupling constants are examined, however, the trends are more equivocal (Figure 3.14B): though in four out of seven cases, the errors are lowest with RSFF2 (including in all three replicates of the SNase simulation), in other cases, notably NLTP, the errors associated with RSFF2 are no better, or are somewhat worse, than those obtained with the other force fields (see Figure 3.14B and Figure 3.15B). Two further surprises are encountered when I examine the RMS deviations (RMSDs) of the proteins from their initial structures over the course of the simulations: Figure 3.16 shows the RMSDs of the backbone heavy atoms as a function of time for all simulations, whereas Figure 3.17 presents the same results in averaged form. The first surprise is that, with all three force fields, the average backbone RMSDs for NLTP are lower than those of any of the other proteins, despite the fact that I obtain the worst reproduction of $^3J_{HNH_a}$ coupling constants for this protein (Figure 3.14A and 3.14B). The second, more minor surprise is that the average backbone RMSDs of simulations performed with the RSFF2 force field are not significantly different from those obtained with the earlier force fields. Both of these results indicate that, at least on a time scale of 1 µs, there is no simple relationship between the average RMSDs sampled during the simulations and the abilities of the simulations to reproduce experimental $^3J_{HNH_a}$ coupling constants.
**Discussion**

In recent years, NMR scalar coupling data have been used repeatedly to effectively diagnose the strengths and weaknesses of force fields for proteins (Beauchamp et al., 2012; Best et al., 2008; Lindorff-Larsen et al., 2012). The study in this chapter continues that trend and demonstrates that the very recently developed RSFF2 force field yields dramatically improved predictions of $^3J_{HNHa}$ coupling constants in two-residue peptides, and a more modest but still noticeable improvement for proteins. The clear improvement with the two-residue peptides is encouraging given the fact that both of the force fields to which I compare RSFF2 have already been shown to perform among the best in a recent NMR-based benchmark study of peptide and protein force fields (Beauchamp et al., 2012). It also provides strong support for the philosophy, advocated by the Wu group, that underlies their parametrization of RSFF2 (Zhou et al., 2015) and the related OPLS-based RSFF1 force field (Jiang et al., 2014); in particular, it is to be noted that reproduction of $^3J_{HNHa}$ coupling constants was not a specific goal of the RSFF parametrization schemes, even though good results for single amino acids were reported.

On the other hand, the observation that predictions of NREs on the $^3J_{HNHa}$ coupling constants of two-residue peptides are not hugely improved by the use of RSFF2 suggests that factors not adjusted in the RSFF2 parametrization scheme are likely to be responsible for the remaining discrepancies with experiment. Foremost among these factors for the two-residue peptides may be the description of nonbonded interactions, either those operating between side chain atoms and the capping groups of the peptides or those operating between side chain atoms on adjacent residues. Issues with describing interactions with the acetyl capping groups, for example, might explain why predictions of the $^3J_{HNHa}$ coupling constants for N-terminal residues are worse than those for C-
terminal residues with all three force fields (Figure 3.4). Possible deficiencies in the description of nonbonded interactions of adjacent side chains, on the other hand, might be responsible for difficulties in reproducing the NREs: recent work by the Zagrovic group, for example, has suggested that excessively favorable nonbonded residue–residue interactions might be a general feature of current pairwise protein force fields (Petrov and Zagrovic 2014). Regardless of their precise origins, however, it is likely that targeting errors in the computed $^3J_{HNH\alpha}$ coupling constants of small peptide systems might be a fruitful way to carry out further refinements of the force fields studied here.

The results obtained for the proteins are more difficult to summarize neatly. Of the five protein structures used to perform simulations here, two—ubiquitin and SNase—are crystal structures, whereas the remaining three are solution NMR structures. It is interesting to note, therefore, that the simulation results are consistently good for the two crystal structure proteins (Figures 3.14A and 3.14B) and that, for SNase in particular, RSFF2 leads to clear improvements in both the $r^2$ values and the RMS errors in all three replicate simulations (see also Figures 3.15A and 3.15B). It is also worth noting that the lowest RMS errors obtained from any of the comparisons with experiment are for ubiquitin using the ubiquitin-optimized Karplus parameters derived by Hu and Bax (Hu and Bax 1997) (Figure 3.15B). Although ubiquitin’s results remain the best when the Karplus parameters of Vögeli et al. (Vögeli et al., 2007) are used instead (Figure 3.14B), this does suggest that RMS errors of $\sim$0.8 Hz might be approaching the limit of what can currently be achieved from simulations of proteins, at least when simulations are performed on a time scale of 1 µs. The comparatively poor results obtained for NLTP, on the other hand, are harder to understand, especially because its backbone RMSD is the lowest of any of the proteins simulated (Figure 3.17). Interestingly, however, for this protein (and
others) there are certain residues whose $^3J_{HNHa}$ coupling constants appear to be poorly reproduced by all three force fields. It seems likely, therefore, that a detailed analysis of such outliers in a wider selection of proteins might help to identify further possible directions for force field refinement.
Figure 3.1. Representation of an example two-residue peptide (Ala-Ala). The peptide is capped with an acetyl group (Ace) at the N-terminus and amidated group (Nhe) at the C-terminus.
Figure 3.2. Cartoon images of the five model proteins simulated. Villin (PDB ID: 1QQV; Vardar et al., 1999), Ubiquitin (PDB ID: 1UBQ; Wang and Bax 1996), NLTP (PDB ID: 1BV2; Poznanski et al., 1999), PPIase (PDB ID: 1HXV; Parac et al., 2001), SNase (PDB ID: 1SNC; Vuister and Bax 1993). The images were made with VMD (Humphrey et al., 1996).
Figure 3.3. Comparison of computed $^{3}J_{\text{HNHa}}$ coupling constants obtained with the ff99SB-ILDN-NMR force field using two different water models for 256 two-residue peptides. X-axis plots results obtained in the work from Chapter II using the TIP4P-Ew water model, shorter simulation times, and a different C-terminal capping group; y-axis plots results obtained here using the TIP3P water model. In order to be consistent with our previous work, $^{3}J_{\text{HNHa}}$ coupling constants were calculated using the Karplus parameterization due to Hu & Bax (Hu and Bax 1997).
Figure 3.4. Comparison of computed $^{3}J_{HNH\alpha}$ coupling constants with experimentally measured values for 256 two-residue peptides using the Karplus parameterization due to Vögeli et al. (Vögeli et al., 2007). Panels show the results for: ff99SB-ILDN (top), ff99SB-ILDN-NMR (middle), and RSFF2 (bottom).
Figure 3.5. Comparison of computed $^{3}J_{HNH \alpha}$ coupling constants obtained with the ff99SB-ILDN and ff99SB-ILDN-NMR force fields for 256 two-residue peptides. $^{3}J_{HNH \alpha}$ coupling constants were calculated using the Karplus parameterization due to Vögeli et al (Vögeli et al., 2007).
Figure 3.6. Comparison of computed $^3J_{HN\alpha}$ coupling constants with experimentally measured values for 256 two-residue peptides using the Karplus parameterization due to Hu & Bax (Hu and Bax 1997). Panels show the results for: ff99SB-ILDN (top), ff99SB-ILDN-NMR (middle), and RSFF2 (bottom).
Figure 3.7. Comparison of computed $^3J_{HNHa}$ coupling constants obtained using two different Karplus equation parameterizations for 256 two-residue peptides. Panels show the results for: ff99SB-ILDN (top), ff99SB-ILDN-NMR (middle), and RSFF2 (bottom). y-axis plots $^3J_{HNHa}$ coupling constants calculated using the Karplus parameterization due to Vögeli et al. (Vögeli et al., 2007); x-axis plots $^3J_{HNHa}$ coupling constants calculated using the Karplus parameterization due to Hu & Bax (Hu and Bax 1997).
Figure 3.8. Comparison of root mean square (RMS) errors in the computed $^3J_{HN\alpha}$ coupling constants obtained using two different Karplus equation parameterizations for 256 two-residue peptides. Columns represent the RMS errors of the data points shown in Figures 1 of the main text and in Figure S2. The columns marked ‘J@N’, for example, show RMS errors of the computed $^3J_{HN\alpha}$ coupling constants for residues at the N-terminal position only. Columns on the left-hand half of the figure show results obtained using the Karplus parameterization due to Hu & Bax; columns on the right-hand half show results obtained using the Karplus parameterization due to Vögeli et al. (Vögeli et al., 2007).
Figure 3.9. Comparison of computed neighboring residue effects on $^{3}J_{HH\alpha}$ coupling constants with experimentally measured values for 256 two-residue peptides using the Karplus parametrization due to Vögeli et al (Vögeli et al., 2007). Panels show the results for: ff99SB-ILDN (top), ff99SB-ILDN-NMR (middle), and RSFF2 (bottom).
Figure 3.10. Comparison of computed and experimental neighboring residue effects on $^3J_{HNN\alpha}$ coupling constants averaged by type of amino acid due using the Karplus parameterization due to Vögeli et al (Vögeli et al., 2007). Plots show the average effect of each type of amino acid, when present at the N-terminal position, on the $^3J_{HNN\alpha}$ coupling constant of amino acids at the C-terminal position; error bars indicate standard deviations. Panels show the results for: ff99SB-ILDN (top), ff99SB-ILDN-NMR (middle), and RSFF2 (bottom).
Figure 3.11. Comparison of computed and experimental neighboring residue effects (NREs) on $^{3}J_{HNH_{\alpha}}$ coupling constants averaged by type of amino acid using the Karplus parameterization due to Hu & Bax (Hu and Bax 1997). Plots show the average effect of each type of amino acid, when present at the C-terminal position, on the $^{3}J_{HNH_{\alpha}}$ coupling constant of amino acids at the N-terminal position (see text); error bars indicate standard deviations. Panels show the results for: ff99SB-ILDN (top), ff99SB-ILDN-NMR (middle), and RSFF2 (bottom).
Figure 3.12. Linear regression $r^2$ values for comparisons of computed and experimental $^3J_{NN\alpha}$ coupling constants for peptide systems. Results obtained from analysis of two-residue peptide simulations. Each column represents the results of a different linear regression involving the data points shown in Figures 3.4 and 3.9. The column marked ‘3J@N’, for example, shows $r^2$ values obtained from linear regression of computed and experimental $^3J_{NN\alpha}$ coupling constants for residues at the N-terminal position only; these data points are the blue circles in Figure 3.4.
Villin

Ubiquitin

NLTP

PPlase
Figure 3.13. Scatter plots comparing computed and experimental $^3J_{\text{HHa}}$ coupling constants for the model proteins. Each row shows results for a different protein simulation, arranged in order of increasing size; for SNase, the results of the three independent replicate simulations are each shown separately. Each column shows the results obtained with a different force field: ff99SB-ILDN (left), ff99SB-ILDN-NMR (middle), and RSFF2 (right). $^3J_{\text{HHa}}$ coupling constants were calculated using the Karplus parameterization due to Vögeli et al. (Vögeli et al., 2007).
Figure 3.14. Linear regression $r^2$ values for comparisons of computed and experimental $^3J_{HNHa}$ coupling constants for protein systems using the Karplus parametrization due to Vögeli et al (Vögeli et al., 2007). A. Summary of linear regression $r^2$ values obtained from analysis of simulations of model proteins; error bars represent the standard deviations of the $r^2$ values obtained when each 1 µs trajectory is analyzed in 3 contiguous 333 ns segments. NLTP is the nonspecific lipid transfer protein from rice, PPIase is the peptidylprolylisomerase domain of trigger factor from M. genitalium, SNase is staphylococcal nuclease. Note that for SNase, results are shown for three independent replicate 1 µs simulations. B. Same as A but showing the RMS errors of the computed $^3J_{HNHa}$ coupling constants.
Figure 3.15. Linear regression $r^2$ values for comparisons of computed and experimental $^3J_{\text{HNH}a}$ coupling constants for protein systems using the Karplus parametrization due to Hu & Bax (Hu and Bax 1997). A. Summary of linear regression $r^2$ values obtained from analysis of simulations of model proteins; error bars represent the standard deviations of the $r^2$ values obtained when each 1 µs trajectory is analyzed in 3 contiguous 333 ns segments. NLTP is the nonspecific lipid transfer protein from rice, PPIase is the peptidylprolyl isomerization domain of trigger factor from M. genitalium, SNase is staphylococcal nuclease. Note that for SNase, results are shown for three independent replicate 1 µs simulations. B. Same as A but showing the RMS errors of the computed $^3J_{\text{HNH}a}$ coupling constants.
Figure 3.16. RMS deviations of the simulated protein structures from their initial structures as a function of simulation time. RMSDs were calculated using only backbone atoms after first superimposing each simulation snapshot on to the energy-minimized structure. In the special case of PPIase, for which the first 30 amino acids were built by homology modeling, these residues were omitted from the analysis.
Figure 3.17. Average RMS deviations of the simulated protein structures from their initial structures. Plot showing the average of the RMSDs shown in Figure 3.12; error bars indicate the standard deviations of the sampled RMSD values.
CHAPTER IV
DEVELOPMENT OF A MOLECULAR MODEL OF THE *E. COLI* CHROMOSOME

The contents of this chapter were published in the article “Features of genomic organization in a nucleotide-resolution molecular model of the Escherichia coli chromosome” in 2017 by William C. Hacker*, Shuxiang Li* and Adrian H. Elcock in the Nucleic Acids Research. 45, 7541-7554. (*these authors contributed equally to the paper)

The work in this chapter was carried out in collaboration with Will Hacker in Dr. Elcock’s lab. The material presented below focuses on my contribution to the project; Will’s work is mentioned as necessary for context.

Introduction

One of the long-term goals within the field of molecular biology is to understand the structures, functions and dynamics of complex biological macromolecules in cellular conditions. Developing 3D models in which each of the cell’s macromolecules is represented in structural detail can provide insight into the workings of biomolecular systems. While the structures of many of the protein constituents of cells have been solved at atomic resolution, the models of the chromosome that are currently available have a far lower level of resolution (Figure 4.1) (Fritsche et al., 2012; Hong et al., 2013; Jun and Mulder 2006; Junier et al., 2014; Le et al., 2013). For example, all of these chromosome models have been built using single beads corresponding to hundreds or thousands of base-pairs. Direct examination of a number of physical and genomic properties of the chromosome is not possible at this resolution.
To reduce this disparity, my goal in this chapter was to develop a high-resolution structural model of the complete bacterial chromosome, focusing initially on that of the Gram-negative bacterium *Escherichia coli*. In efforts to model the chromosome, a coarse-grained (CG) representation of DNA is essential because it provides a computationally inexpensive approach for constructing structures for macromolecules and enabling efficient simulations of very large systems (Ingolfsson et al., 2014). In recent years, a significant number of CG simulation models for DNA have been developed (Hinckley et al., 2014; Ilie et al., 2016; Ouldridge et al., 2011; Peter et al., 2015; Savelyev and Papoian 2010; Trovato and Tozzini 2008; Weik et al., 2016), ranging from simple-spring representations to models incorporating all-atom force fields. It should be noted that each of these DNA CG models was developed with a particular application in view. For example, Savelyev and Papoian developed CG models of DNA to investigate the DNA persistence length over a wide range of concentrations of NaCl salt buffer (Savelyev and Papoian 2010). Their results are in quantitative agreement with several experimental measurements. As another example, a CG model of DNA derived by de Pablo and coworkers was intended to reproduce the thermodynamics of melting, hybridization, bubble formation, and salt dependence of the persistence length (Knotts et al., 2007; Sambriski et al., 2009).

While many of these models use a mesoscale-level description of DNA in which two or three CG beads are used to represent each nucleotide, such models are still too computationally expensive to be applied to entire bacterial chromosomes. In this study, we chose a representation of DNA with one bead corresponding to one nucleotide because it is the coarsest model that can accurately represent physical properties of DNA such as major and minor groove widths. In addition, such a CG scheme of DNA has been used successfully in simulations exploring supercoiling in small plasmids (Trovato and Tozzini 2008).
Based on this simplified model, we have been able to construct entire chromosomes; I describe here the development of models of the chromosome of the *E.coli* in which the positions of all 4.6 million nucleotides of each DNA strand are resolved.

**Methods**

*Coarse-grained DNA models*

To build a complete *E. coli* chromosome model, different resolutions of CG DNA models were employed in this chapter. These include a 500 base-pair-per-bead (BPB) model in which one bead corresponds to 500 basepairs, a 5 BPB model, a 1 BPB model and a 1 nucleotide-per-bead (NTB) model. A simple illustration of all of these CG DNA models is shown in Figure 4.2; for comparison, an all-atom DNA model is also shown.

*Parameterization of the coarse-grained simulation DNA models*

In order to obtain physical flexibility of the CG DNA that is consistent with the stiffness of real double-stranded DNA, I parameterized the 5 BPB model and focused on attempting to reproduce the experimentally-measured persistence length (~500 Å for B-DNA at physiological ionic strength (Hagerman 1988)). The persistence length, \( P \), is a commonly used mechanical property that quantifies the stiffness of a linear polymer (Figure 4.3). It is defined as the distance after which the correlations along the tangent to the polymer axis cease (Filippov 1973). Importantly, the persistence length can be calculated from simulations by measuring the average of the squared end-to-end distance of the DNA polymer assuming that the DNA behaves as a Worm-Like chain (WLC) using the following equation (Cantor and Schimmel 1984):
\[ <R^2> = 2PL - 2P^2 (1 - e^{-L/P}) \] (4-1)

where \( R \) is the end-to-end distance, \(<...> \) indicates mean value, \( P \) is the persistence length, and \( L \) is the contour length of the polymer (i.e. the sum of the distances between adjacent beads or basepairs). With \(<R^2>\) and \( L \) measured from simulations, the corresponding value of \( P \) can be obtained using the “Goal Seek” tool in Excel.

To parameterize the persistence length of the 5 BPB model, a 10 \( \mu \)s Brownian Dynamics (BD) simulation was performed using the Elcock lab’s uiowa\_bd simulation code, using an initially linear DNA containing 100 beads (i.e. corresponding to 500 basepairs). In these simulations, all solvent molecules and salt ions are treated implicitly, with the effects of the salt being modeled very simply using the Debye-Hückel approximation:

\[ E_{DH} = \frac{3 \times 2.08 \times q_i q_j e^{-\kappa r_{ij}}}{\varepsilon r_{ij}} \] (4-2)

where \( q_i \) and \( q_j \) are the charges, \( r_{ij} \) is the distance between the two charges in Ångströms, \( \kappa \) is the Debye-Hückel screening parameter (related to the square root of the ionic strength), \( \varepsilon \) is the dielectric constant of the solvent (approximated here as that of water – i.e. 78.4), and 332.08 is the conversion factor that ensures the energy is returned in units of kcal/mol. This model can be used to mimic the effects of a simple salt such as NaCl or KCl – which, while often behaving differently, have been shown to have equivalent effects on the thermal stability of double-stranded DNA (Owczarzy et al., 2008). In most CG models of DNA, implicit salt representations are also accompanied by use of decreased net charges on the phosphate groups to mimic the effects of ‘bound’ or
condensed counterions. In the literature (Cao et al., 2011; Savelyev and Papoian 2009), net charge values assigned to the phosphate groups range from -0.25e to -1.0e, and there are models in which the phosphate charge is completely neutralized (Ouldridge et al., 2011; Trovato and Tozzini 2008). In the simulations of this project, I have chosen to use a net charge value of -0.5e per phosphate as this lies near the middle of the range of values used in the literature. Using this phosphate charge, together with some initial guesses of the model’s bonded parameters (bond force constant, \( k_b = 20 \text{ kcal/mol/Å}^2 \); no dihedral functions), six simulations with different angle force \( k_\theta \) (2.5 to 20 kcal/mol/rad\(^2\)) were performed to determine the influence of the force constant applied to angles \( (k_\theta) \) on the persistence length of the 5 BPB model. A further four simulations with varied bead charges of -10, -7.5, -2.5, and 0e were also performed to explore the charge dependence of the persistence length. The salt-dependence of the persistence length of the 5 BPB model was also investigated by performing 12 simulations at salt concentrations ranging from 1 mM to 2 M.

**Morphing simulations for generating the 5 BPB chromosome model**

In order to generate a finalized chromosome structure of the *E. coli* chromosome at a resolution of 5 BPB, an initial 5 BPB model, which is largely confined to a plane with branched plectoneme-abundant regions (PAR) and plectoneme-free regions (PFR), needed to be progressively compressed into the physical dimensions of the nucleoid. This was carried out by performing a series of “morphing” simulations of the 5 BPB structure to match the 500 BPB structure produced from Will Hacker’s simulations. Will employed a chromosome model at the 500 BPB resolution in a series of BD simulations that progressively compact the initial model into a capsule with the same dimensions of a typical *E. coli* nucleoid. The initial structure is a “bottlebrush” model of the chromosome
inspired by a corresponding model of the *C. crescentus* chromosome reported by the Laub group (Le et al., 2013). The plectoneme lengths were assigned to reproduce the exponential distribution determined experimentally for *E. coli* (Postow et al., 2004), and the branches were incorporated in a way that matches the frequencies and lengths observed in supercoiled plasmids (Boles et al., 1990). Additional simulations were performed to adjust the positioning of the beads of the 500 BPB model so that they reproduce the spatial distributions of DNA and of RNAP molecules recently visualized in live *E. coli* cells by the Kapanidis lab (Stracy et al., 2015). Thus, the chromosome models should implicitly account for the effects of macromolecular crowding (de Vries 2010) and nucleoid-associated proteins (NAPs) (Luijsterburg et al., 2006) on chromosome compaction.

Using the simulation trajectories of the 500 BPB model from Will’s work, I extracted 1330 “snapshots” as guides for the “morphing” simulations of the 5 BPB model. The aim of this process is to allow the 5 BPB model to match the behavior of the 500 BPB model during its compaction simulation. An in-house code written by Dr. Elcock was used to carry out this process. In detail, harmonic springs (force constant of $2.5 \times 10^{-5}$ kcal/mol/Å$^2$) were applied to the 5 BPB beads so that they can follow the trajectory of the 500 BPB beads as closely as possible. Additional springs (force constant of $8.4 \times 10^{-6}$ kcal/mol/Å$^2$) were applied between all pairs of 5 BPB beads that are separated by < 100 Å in the initial structure so that the models retain their branched-plectonemic “shapes” during the sometimes drastic changes that accompany compaction within the nucleoid. In total, 1330 stages of the morphing simulations were performed with 400 steps of steepest-descent energy-minimization being carried out at each stage. In the first 300 steps of energy-minimization, the 500 BPB beads moved from their positions in the snapshot $i$ to their positions in the snapshot $i+1$ while the 5 BPB beads strived to catch up. In the next 50 steps, the 5 BPB beads carried out
continued relaxation to adjust their positions determined by the 500 BPB beads. In the last 50 steps, further structural relaxation of the 5 BPB beads was allowed but without the harmonic springs between beads from the 500 BPB model and beads from the 5BPB model.

The next stage was to add increasingly realistic features into the 5 BPB model. To obtain this, a 1000 step steepest-descent energy-minimization was first performed to correct the lengths of the bonds connecting neighboring 5 BPB beads to reach the values expected of B-DNA (i.e. 17 Å). Then, 25 stages of simulation were carried out with each stage involving 100 steepest-descent energy-minimizations. The purpose of this step is to gradually introduce steric interactions between all pairs of 5 BPB beads that were not involved in bonding interactions; this is the first time that steric effects were introduced in the complete protocol for building the chromosome structures. In this procedure, the diameter of the 5 BPB beads (σ) was gradually increased from 0 to 25 Å in intervals of 1 Å. The final σ value used here guaranteed that there were no steric clashes in subsequent stage of simulation. Finally, a 250 ns BD simulation was performed for the 5 BPB model to allow modest conformational fluctuations to occur and a final 1000 step steepest-descent energy-minimization was carried out to obtain a final 5 BPB model. It should be noted that both electrostatic and steric interaction terms were used in this step.

Converting the 5 BPB model into a 1 BPB model

In preparation for making a final 1 NTB model, we first needed to make a model with a resolution of 1 BBP as an intermediate model. To do this, for each pair of bonded beads in the final 5 BPB model, four intermediate beads were inserted at even intervals. Then a simple 1000 step steepest-descent energy-minimization was performed in order to smooth out this 1 BPB model. The
resulting structure characterizes the double-helical axis of the DNA and serves as the basis for constructing a 1 NTB model.

Converting the 1 BPB model into a 1 NTB model

The 1 NTB model, which is the most detailed model described in this work, was obtained by 4,641,652 steps of structural superpositions of DNA fragments onto the beads of the final 1 BPB structure. In detail, 3 helical axis beads of the first 3 bp from an initial, idealized 1 NTB structure were superimposed on the corresponding 3 beads of the 1 BPB structure by assuming the 3 helical axis beads are linear. The rotational orientation of the first 3 bp fragment around the helical axis is effectively arbitrary. Then all additional phosphate beads are positioned in a way that continues the initially defined pitch of the helical axis. For example, to place the phosphate beads of a “new” basepair, i+1, I simultaneously superposed the helical axis beads of basepairs i, i+1, and the two phosphate beads of basepair i (i.e. 4 beads in total), and then applied the same coordinate transformation to the phosphate beads of basepair i+1. In order to ensure that each new basepair added to the growing 1 NTB model stayed close to the helical axis that was so laboriously defined by all the earlier simulations, the helical axis beads were assigned a weight three times greater than the weight attached to the phosphate beads of basepair i in the superposition. Having placed the phosphate beads of basepair i+1, the superposition process could be repeated using helical axis beads from basepairs i+1 and i+2 in order to place the phosphate beads of basepair i+2. In this way, all 4641652 pairs of phosphate beads could be placed in approximately 40 hours.

Energy minimization of the 1 NTB model

A 10,000 step energy minimization was performed with the purpose of:
a) removing any minor discontinuity that might be present where the 4641652nd basepair met the 1st basepair, b) removing any minor steric clashes, and c) ensuring that the major and minor grooves were properly formed. A so-called Gō or “native-centric” energy function which is similar to that broadly used to simulate protein folding events (Chavez et al., 2004; Clementi et al., 2000; Koga and Takada 2001), was introduced in this step in order to encourage the structure to take on the appearance of B-DNA, while remaining subject to the local environmental constraints acting on the DNA. The energy function used here is similar to previous simulations of protein folding (Elcock 2006; Frembgen-Kesner and Elcock 2009) in the Elcock lab with bonded interactions described using:

$$E_{bonded} = \sum_{bonded} K_r (r - r_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} K_\phi [1 - \cos(n(\phi - \phi_n))]$$

(4-3)

r and θ represent inter-bead bond lengths and angles, respectively. r₀ and θ₀ represent the equilibrium values of the bond lengths and angles in the initial (idealized) chromosomal build. The following parameters were used: r₀ = 6.495 Å, K_r = 20 kcal/mol/Å², θ₀ = 149.4°, K_θ = 10 kcal/mol/rad. The final summation over the dihedral angles (i.e. involving four consecutive phosphate beads on each DNA strand) expresses the energy as a sum of two cosine terms, with n = 1 and n = 3, respectively; in both cases, \(\phi^n\) represents the phase angle defining the energy maximum of the cosine term (set such that the energy minimum occurs at a dihedral angle of 19.2°), and K_φ represents the height of the energy barrier for rotating around the dihedral. For these calculations, K_φ^1 = 0.5 kcal/mol/rad, and K_φ^3 = 0.25 kcal/mol/rad.

The formation of realistic major and minor groove dimensions was rewarded through the addition of favorable Lennard-Jones potential functions
acting across the DNA strands, for which the equilibrium distances were set to those found in B-DNA. Specifically, favorable interactions were included between phosphate i on strand A, and phosphates i-2 to i+3 on strand B; for all such pairs of beads, the energy well-depth, ε, was set to 0.5 kcal/mol. For all other pairs of beads, very weakly favorable Lennard-Jones interactions were included to prevent the formation of steric clashes; for such interactions we used σ = 10 Å and ε = 0.1 kcal/mol. Finally, nonbonded electrostatic interactions were again described using equation 4-2, but with each phosphate bead assigned a charge of -0.5 e in order to again crudely mimic the effects of counterion condensation. After this final energy-minimization, the absence of any steric clashes in each of the final 1 NTB chromosome structures was confirmed using in-house code that explicitly measures the shortest distances between all beads.

Construction of 20 oriC@pole and 20 oriC@midcell models

Using the multi-scale modeling techniques mentioned above, two types of E. coli chromosome models that differ in the positioning of the origin of replication (oriC) were constructed. In this dissertation, oriC@pole is used to represent models with oriC positioned at the pole region, while oriC@midcell represents models with oriC positioned at the midcell region. For both types of model, 20 structures were constructed that use the same locations of PARs and PFRs but differ in their plectoneme positions, branches, and lengths.

Computing major and minor groove widths

The major and minor groove widths in the final 1 NTB resolution structures were calculated from the distances between phosphate beads (Figure 4.4). The major groove width is defined as the distance between phosphate i on the first strand of the DNA and phosphate i+4 on the second strand (El Hassan
and Calladine 1998). The minor groove width is defined as the distance between phosphate \(i\) on the first strand and phosphate \(i-3\) on the second strand (El Hassan and Calladine 1998). Groove widths were measured at all 4.6 million base pairs and histograms of the distributions were plotted.

Computing bending angle distributions

The bending angle of the DNA axis was defined by helical axis beads placed at basepair positions \(i-15\), \(i\), \(i+15\) and those three beads were used to calculate their angle (Figure 4.5). I then subtract the measured angle from 180° to follow the convention that an unbent DNA has a bending angle of zero and made a histogram of the bending angles measured for all 4.6 million possible values of \(i\). The histogram can then be compared with the probability distribution expected for bending angles assuming a Worm-like chain model for the DNA (Mazur 2007):

\[
ds \sim \exp\left(-\frac{P\theta^2}{2L}\right) \sin \theta \, d\theta
\]

Here \(\theta\) is the bending angle, \(L\) is the contour length of the fragment over which bending is measured (30 basepairs \(\times\) 3.4 Å = 102 Å), and \(P\) is the persistence length of the chain. The probability distribution of bending angles predicted by this equation was compared with that obtained from direct measurement of the structures and I used the Goal Seek function in Microsoft Excel to adjust \(P\) so as to minimize the absolute error between the two distributions.

Calculation of twist, writhe and linking number

Twist (Tw) is the number of complete revolutions that one DNA strand makes around the duplex axis; writhe (Wr) is the number of turns that the
duplex axis makes around the superhelical axis; linking number (Lk) is the total number of times that one DNA strand winds around the other in a closed structure (Voet 2006). For simple structures, writhe can be estimated by counting the number of times that the dsDNA crosses itself. Mathematical topology states that there is an exact relation between these three parameters:

\[ Lk = Wr + Tw \]  

(4-5)

The twist and writhe parameters for 1NTB resolution structures were computed using either the WrLINE software released by the Noy group (Sutthibutpong et al., 2015) (University of Leeds, UK) or using in-house code that is heavily based upon it. The WrLINE Python software was downloaded from http://ccpforge.cse.rl.ac.uk/gf/project/wrline. A minor modification was made to the original code to make it read the CG DNA structures and write out twist, which is computed as an intermediate step of the WrLINE procedure: in that code, twist is calculated using equation 6 of Sutthibutpong et al. (Sutthibutpong et al., 2015). The writhe calculation is much more expensive (which uses equation 5 of Clauvelin et al. (Clauvelin et al., 2012)) and the code did not complete even after three days of calculation time when applied directly to a 1NTB chromosome structure. I therefore ported the Python code into C++ and used OpenMP constructs to parallelize the most expensive parts of the calculation. To verify that our C++ code was correct, I first performed validation tests on a number of relaxed and supercoiled plasmids containing 500 or 2100 basepairs (Fig 4.5) and compared the results of my code with those of the original WrLINE code. The linking number (Lk) was calculated using OpenMP-parallelized C++ code written entirely in-house, and using the procedure outlined by equations 14, 15 and 16 in (Clauvelin et al., 2012). Correctness of the linking number code was
verified, again using the supercoiled plasmid structures, by comparison with the sum of the twist and writhe values obtained with WrLINE, i.e. using equation 4-5. The supercoil density ($\sigma$) was obtained using the relation:

$$\sigma = \frac{L_k - L_{k0}}{L_{k0}}$$

where $L_{k0}$ is the linking number of relaxed (i.e. unsupercoiled DNA) and is given by the total number of basepairs divided by 10.5 basepairs-per-turn. For the whole chromosome structure in this study, $L_{k0} = \frac{4641652}{10.5} = 442062.1$.

**Results**

*Reproducing experimental persistence length for parameterization of CG models*

It is important for the CG models to correctly mimic the flexibility of real double-stranded DNA. To achieve this, I focused on parameterization to reproduce the experimentally measured persistence length for both the 5 BPB and 1 NTB models. The reason for selecting these two CG models is that most of the simulations in this work were performed at 5 BPB resolution and the final detailed structure of the chromosome model is constructed at 1 NTB resolution.

Using some initial guesses of bonded parameters, the calculated persistence length of 5 BPB model at 150 mM salt is $\sim 420$ Å, which is in quite reasonable agreement with the corresponding experimental value ($\sim 500$ Å) (Hagerman 1988). To determine the influence of the force constant applied to angles ($k_\theta$) on the persistence length of the 5 BPB model, extra simulations with varied $k_\theta$ (3.5 to 20 kcal/mol/rad$^2$) were performed with the bead net charge fixed at -5e. The results show that a larger $k_\theta$ value decreases the flexibility of DNA and thus increases the persistence length (Figure 4.6A). A value of 3.5
kcal/mol/rad$^2$ for $k_0$ yields a persistence length (516 Å) in very good agreement with the experimentally determined value. Using this force constant, a further four simulations with varied bead charges of -10, -7.5, -2.5, and 0e were also performed to explore the charge dependence of the persistence length. It was found that while a more negative charge slightly increases the stiffness of the DNA the influence is relatively limited (Figure 4.6B). Even so, all tested charge values are able to obtain quite reasonable agreement with the experimental persistence length.

To determine how well the 5 BPB model can reproduce the experimental salt dependence of the persistence length, varied ionic strengths were used for performing simulations. The results show that the 5 BPB model does a surprisingly good job of reproducing the significant increase in the persistence length that occurs as the salt concentration is dropped below ~50 mM (Figure 4.7). The agreement with experimental measurements indicates that the 5 BPB model used here can probably describe the flexibility of DNA adequately for constructing a complete chromosome model.

**Construction of 20 oriC@pole and 20 oriC@midcell chromosome models**

Using a combination of coarse-grained simulation and high-resolution structural modeling methods together with Will Hacker, two sets of chromosome models (oriC@pole and oriC@midcell) were constructed at 1NTB resolution. The overview of the protocol for building these models is shown in Figure 4.8 and the detailed procedure is described in the Methods section. Figure 4.9 shows the final 5 BPB model (green beads) which is mapped onto the 500 BPB model (red beads). The close correspondence of the 5 BPB structure and the 500 BPB model indicates that morphing simulations have successfully followed the trajectory established by the 500 BPB model during compaction process.
An example of the final chromosome structures with 4.6 Mb is shown in Figure 4.10.A. Dimensions corresponding to the cell and nucleoid are indicated as blue and red lines, respectively. The chromosome is highly compact in the cell with PARs and PFRs intertwined through the whole structure. Figure 4.10.B shows the chromosome at increasing levels of magnification. Individual plectonemes are discernible in the left panel. Distinct major and minor grooves are observable in the right panel.

In this work, 20 oriC@pole and 20 oriC@midcell chromosome models were built that each use the same defined locations of PARs and PFRs but that each differ in their plectoneme lengths, branches and positions. The complete set of oriC@pole and oriC@midcell structures are shown in Figure 4.11 and 4.12, respectively. Specifically, the four chromosomal regions corresponding to the macrodomains (Ori, Ter, Left and Right) separate from each other. The properties of those macrodomains dwelling in well-defined locations within the nucleoid are consistent with experimental observations (Espeli et al., 2008).

Major and minor groove width distributions

To verify that the final 1NTB structures are geometrically reasonable, the major and minor groove widths involved with each base pair were calculated. The calculated average values are ~16.9 ± 0.9 Å and ~14.2 ± 0.8 Å respectively for major and minor groove widths (Figure 4.13). These results are in reasonable agreement with the average values determined for (unsupercoiled) NMR structures (Perez et al., 2008); the corresponding values of NMR structures are ~18 ± ~2 Å and ~13 ± ~2 Å for major and minor groove widths, respectively. Moreover, the results from the 1NTB models are also in good agreement with the average values obtained from MD simulations of oligonucleotides based on the CHARMM27 force field (Foloppe and MacKerell 2000; MacKerell and Banavali...
2000); the corresponding values of MD simulations are ~17 ± 2 Å and ~13-14 ± 2 Å for major and minor groove widths (Perez et al., 2008), respectively. In sum, the chromosome models at 1 NTB resolution appear to have realistic geometries.

**Bending angle distribution of the 1 NTB model**

The second physical property I determined for the final 1 NTB models is the bending angle distributions. The histograms of 4.6 million bending angles for oriC@pole and oriC@midcell models are shown in Figure 4.14A and 4.14B, respectively. Since those two types of models were constructed with identical force field parameters, there is no noticeable dissimilarity between the bending angle distributions for the oriC@pole and oriC@midcell models. When the above histograms were compared with the probability distribution derived from a Worm-like chain model of DNA, striking similarities were found between them with a predicted persistence length of about 500 Å (Figure 4.14). These results reinforce the suggestion that the chromosome structures are free of major distortions.

**Topological parameters of the 1 NTB model**

DNA supercoiling is essential for correct function of cellular processes such as transcription and replication (Drolet et al., 1993; Kouzine et al., 2013). One of the major potential advantages of constructing such a 1 NTB model of the bacterial chromosome is that it allows the effects of DNA supercoiling to be modeled and visualized in a way that cannot be captured by more coarse-grained chromosome models. To explore and describe the topological properties of the *E. coli* chromosome models in the way that is most generally used in the literature, three parameters (Tw, Wr and Lk) that are commonly used to describe and quantify DNA supercoiling were calculated.

As mentioned in the Methods, small plasmid structures with a variety of
states of supercoiling (Figure 4.15) were used to validate the accuracy of the codes that were implemented here for Tw, Wr and Lk calculation. Three of these models contained 500 base pairs and the fourth contained 2100 base pairs; note that Model B, which has a supercoil density, $\sigma$, of -0.06 has a level of supercoiling that most resembles that of the *E. coli* chromosome (Champion and Higgins 2007). Table 4.1 shows the results of calculated Tw, Wr and Lk for the four plasmid structures. It was found that the sum of the calculated Tw and Wr values was essentially equal to the calculated linking number, which validated the implemented codes for calculating topological parameters of the DNA models at 1 NTB resolution.

When the same calculation methods were applied to the 4,641,652 basepairs chromosome structure, it was found that each twist calculation took less than 10 minutes. However, the writhe and linking number calculations required ~7 hours and ~8 days, respectively, running on a 64-core AMD Opteron 6272 server. The topological parameters for the chromosome structure with a real calculation of linking number are shown Table 4.2. For other chromosome structures, I made use of the relation $Lk = Tw + Wr$ to avoid the very expensive direct calculation of linking number.

The calculated average Tw, Wr and Lk for 20 oriC@pole are 434316.81, -14118.21 and 420198.6, respectively. Based on those values, the supercoil density ($\sigma$) is obtained with a value of -0.049 ($(420198.6-442062.10) / 442062.10$, see the Methods section), which is within the range of the reported experimental values for *E. coli* (Bliska and Cozzarelli 1987; Champion and Higgins 2007; Sinden et al., 1980). The complete sets of calculated Tw, Wr and Lk values are shown in Tables 4.3 and 4.4 for oriC@pole and oriC@midcell models, respectively. The average supercoil densities are measured to be -0.049 ± 0.0004 for both oriC@pole and oriC@midcell models. Moreover, the average calculated contributions from Tw
and Wr to the total supercoil density are found to be in the ratios 35:65 and 39:61 for the oriC@pole and oriC@midcell models, respectively. These values also agree well with the 28:72 ratio based on analysis of electron micrographs of supercoiled 7kb plasmids (Boles et al., 1990). Besides, previous atomistic simulations on plasmids showed that although the Tw-Wr partition is ~50:50 for smaller plasmids (~500 basepairs), the partition is roughly on the order of 33:67 for larger plasmids (Liverpool et al., 2008; Sayar et al., 2010). These results suggest that the final chromosome models have a realistic partitioning of supercoiling into Tw and Wr.

**Discussion**

There have been extensive improvements in the scale and resolution with which vital biological processes are modeled in recent years. A complete model of a bacterial cell requires the inclusion of an accurate structural model of the chromosome. In this chapter, I have described the use of multi-scale simulation techniques and CG DNA models to construct chromosome models of the Gram-negative bacterium *E. coli* at a final resolution of one bead corresponding to one nucleotide. This work was carried out in collaboration with Will Hacker.

The challenge for building a complete chromosome model for *E. coli* is to compress its 4.6 million base pairs (Mbp), with a contour length of 1.6 mm, into a cell less than 3 µm in length in a predictable orientation (Bates and Kleckner 2005; Wang et al., 2011). To achieve this, Will employed a 500 BPB chromosome model and progressively compressed it within the nucleoid using a series of BD simulations. In my work, I used a 5 BPB model and subjected it to a series of morphing simulations so that it matched the trajectories of the 500 BPB model. Additional energy-minimizations and BD simulations were performed to regularize the 5 BPB model. Finally, I performed a series of structural
superpositions to convert the 5 BPB model into 1 NTB model.

For each CG model used here, in particular the 5 BPB and 1NTB models I was working on, the potential function parameters were first adjusted to reproduce the persistence length of double-stranded B-DNA (Hagerman 1988) by BD simulations. It should be noted that the electrostatic interaction between beads was then modeled using the Debye- Hückel approximation (see the Methods section). Although I did not explore further the salt dependence of persistence length, the result showed that the inclusion of a Debye-Hückel approximation allows the 5 BPB model to reproduce quite well the salt dependence of the persistence length of double-stranded DNA (Figure 4.7).

The high resolution model of the *E.coli* chromosome constructed here allows numerous characteristics to be analyzed. My work was focused on analyzing physical characteristics such as groove width and bending angle distributions and supercoiling; Will’s work focused on analyzing genetic features. The measured groove widths from our structures are in reasonable agreement with the average values identified for NMR structures (Perez et al., 2008), and in surprisingly good agreement with the corresponding values obtained using the CHARMM27 force field (Foloppe and MacKerell 2000; MacKerell and Banavali 2000). For the bending angle distributions, a comparison of the direct measurement of our structures with the corresponding distributions predicted by a Worm-like chain model (Mazur 2007) resulted in a persistence length \( \sim 495 \) Å, which again is consistent with the experimentally measured values (Hagerman 1988).

It is widely known that DNA in the bacterial cell rarely exists in a relaxed state, but instead exists in a supercoiled state due to a combination of DNA topoisomerase activity (Chen et al., 2013), compaction (Macvanin and Adhya 2012) and protein binding (Balandina et al., 2002). To examine the supercoiling of
our chromosome models, the topological parameters of the 1 NTB structures were calculated, albeit at considerable computational expense. Based on the calculated twist, writhe and linking number, it was found that the average supercoil density lies in the middle of the wide range of values reported for the *E. coli* chromosome (Bliska and Cozzarelli 1987; Champion and Higgins 2007; Sinden et al., 1980). All these analyses demonstrate that our final 1NTB models are reasonable at a microscopic level.
<table>
<thead>
<tr>
<th>Organism</th>
<th>Simulation method</th>
<th>Model diagram</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E.coli</em></td>
<td>Monte Carlo (MC)</td>
<td></td>
<td><em>Jun et al. Proc Natl Acad Sci USA</em> 2006</td>
</tr>
<tr>
<td><em>E.coli</em></td>
<td>Monte Carlo (MC)</td>
<td></td>
<td><em>Junier et al. Nucleic Acids Res</em> 2014</td>
</tr>
<tr>
<td><em>E.coli</em></td>
<td>Monte Carlo (MC)</td>
<td></td>
<td><em>Dorier et al. Nucleic Acids Res</em> 2013</td>
</tr>
<tr>
<td><em>E.coli</em></td>
<td>Monte Carlo (MC)</td>
<td></td>
<td><em>Fritsche et al. Nucleic Acids Res</em> 2012</td>
</tr>
<tr>
<td><em>C.crescentus</em></td>
<td>Brownian Dynamics(BD)</td>
<td></td>
<td><em>Le et al. Science</em> 2013</td>
</tr>
<tr>
<td><em>C.crescentus</em></td>
<td>Brownian Dynamics(BD)</td>
<td></td>
<td><em>Hong et al. Pro Natl Acad Sci USA</em> 2013</td>
</tr>
</tbody>
</table>

Figure 4.1. Several available bacterial chromosome models with limited resolutions or incomplete coverage of the genome.
Figure 4.2. Comparison of all-atom model with different resolutions of coarse-grained DNA models. The 5 BPB model is used in ‘morphing’ simulations, while the 1 NTB model is the final model constructed.
Figure 4.3. General properties of persistence length (P) and contour length (L).

The persistence length can be calculated by \( \langle R^2 \rangle = 2PL - 2P^2 \left( 1 - e^{-L/P} \right) \).
Figure 4.4. An illustration of the major and minor grooves. The major groove width is defined as the distance between phosphate $i$ on the first strand of the DNA and phosphate $i+4$ on the second strand. The minor groove width is defined as the distance between phosphate $i$ on the first strand and phosphate $i-3$ on the second strand (El Hassan and Calladine 1998).
Figure 4.5. An illustration of the DNA bending angle. The angle defined by helical axis beads placed at basepair positions i-15, i, i+15 is calculated and then is subtracted from 180°.
Figure 4.6. Parameterization of the 5 BPB model to match persistence length of DNA at 150 mM salt concentration. A. Angle force constant dependence of the persistence length for the 5BPB model. B. Phosphate charge dependence of the persistence length for the 5BPB model. The experimental values (Baumann et al., 1997) are shown in red.
Figure 4.7. Ionic strength dependence of the persistence length for the 5BPB model. The experimental values (Baumann et al., 1997) are shown in red.
Figure 4.8. Overview of the protocol for constructing an *E.coli* chromosome model at 1 NTB resolution. A 500 BPB model is progressively compacted into the nucleoid volume. The 5 BPB model follows the trajectories of the 500 BPB model through a series of ‘morphing’ simulations. Additional energy-minimizations and BD simulations were performed to regularize the 5 BPB model and finically convert it to a 1 BPB model and 1 NTB model.
Figure 4.9. Comparison of the 500 BPB model (red spheres) and the 5 BPB model (green spheres). The close correspondence of the 5 BPB structure and the 500 BPB model indicates that morphing simulations have successfully followed the trajectory established by the 500 BPB model during compaction.
Figure 4.10. Overview of one final *E. coli* chromosome model at 1 NTB resolution. A. A model of the 4.6 Mb *E. coli* chromosome at one-nucleotide resolution (green) pictured within a volume corresponding to the cell (black outline). Dimensions appearing above and at right are those of the nucleoid (blue) and the cell containing it (red). B. Images of the chromosome at increasing magnification. In the left panel, individual plectonemes appear; in the right panel, distinct major and minor grooves are visible.
Figure 4.11. The 20 oriC@pole structures. Twenty 1 NTB-resolution structures with the oriC@pole global orientation are depicted. The coloring scheme indicates genomic regions as described by Valens et al. (Valens et al., 2004) (green for Ori macrodomain, red for Right macrodomain, cyan for Left macrodomain, blue for the Ter macrodomain and grey for the Non-Structured regions). All the models are aligned with the most representative of the 20 structures.
Figure 4.12. The 20 oriC@midcell structures. Twenty 1 NTB-resolution structures with the oriC@midcell global orientation are depicted. The coloring scheme is same as in Figure 4.11. All the models are aligned with the most representative of the 20 structures.
Figure 4.13. Groove width distributions of the 1 NTB model. A. Distribution of minor groove (red) and major groove widths (blue) in one representative 1NTB structure of the oriC@pole models. Red dashed lines indicate the range of minor groove widths measured in NMR structures and reported by Perez et al. (Perez et al., 2008); blue dashed lines represent the same, but for major groove widths. B. Same as A but for one representative 1NTB structure of the oriC@midcell models.
Figure 4.14. **Bending angle distributions of the 1NTB model.** Distribution of bending angles obtained from one representative 1NTB structure of the oriC@pole models (red) compared with corresponding distribution obtained from assuming a Worm-like chain (WLC) model with a best-fit persistence length 496 Å (blue). **B.** Same as A but showing results for one representative of the oriC@midcell models and comparing with a WLC model distribution obtained with a best-fit persistence length of 495 Å.
Figure 4.15. Plasmid models used to validate the calculations of writhe, twist and linking number. Model A: 500 bps, no supercoiling; Model B: 500 bps, negatively supercoiled; Model C: 500 bps, positively supercoiled; Model D: 2100 bps, negatively supercoiled. Note that the calculated Writhe values for Models B-D (see Table 4.1) can be estimated by counting the number of times the dsDNA crosses itself in the above images.
Table 4.1. The calculated writhe, twist and linking number for the four test models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Size</th>
<th>$\sigma$</th>
<th>WrLINE Twist(Tw)</th>
<th>WrLINE Writhe(Wr)</th>
<th>My code C++ Writhe(Wr)</th>
<th>My code C++ Tw+Wr</th>
<th>My code C++ Linking(Lk)</th>
<th>Diff (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>500</td>
<td>0.00</td>
<td>48.07</td>
<td>0.00</td>
<td>0.00</td>
<td>48.07</td>
<td>48.00</td>
<td>0.15</td>
</tr>
<tr>
<td>B</td>
<td>500</td>
<td>-0.06</td>
<td>-0.06</td>
<td>46.38</td>
<td>-2.43</td>
<td>-2.43</td>
<td>43.95</td>
<td>44.0</td>
</tr>
<tr>
<td>C</td>
<td>500</td>
<td>+0.07</td>
<td>+0.07</td>
<td>49.52</td>
<td>2.45</td>
<td>2.45</td>
<td>51.97</td>
<td>52.0</td>
</tr>
<tr>
<td>D</td>
<td>2100</td>
<td>-0.06</td>
<td>-0.06</td>
<td>195.71</td>
<td>-7.72</td>
<td>-7.72</td>
<td>187.99</td>
<td>188.0</td>
</tr>
</tbody>
</table>
Table 4.2. Tw, Wr and Lk calculations for a whole *E.coli* model.

<table>
<thead>
<tr>
<th>Size</th>
<th>Tw</th>
<th>Wr</th>
<th>Tw + Wr</th>
<th>Lk</th>
<th>Diff (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4641652</td>
<td>-14156.10</td>
<td>434258.32</td>
<td>420102.22</td>
<td>421072.00</td>
<td>-0.23</td>
</tr>
<tr>
<td>WrLINE</td>
<td>10 min</td>
<td>16 days</td>
<td></td>
<td>Failed</td>
<td></td>
</tr>
<tr>
<td>C++ code</td>
<td>N/A</td>
<td>7 hours</td>
<td></td>
<td>8 days</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.3. Topological properties of the finial 20 oriC@pole models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Tw</th>
<th>Wr</th>
<th>Lk</th>
<th>Supercoil density</th>
<th>Wr/ΔL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>434346.11</td>
<td>-13880.20</td>
<td>420465.9</td>
<td>-0.04885</td>
<td>0.643</td>
</tr>
<tr>
<td>2</td>
<td>434387.91</td>
<td>-14173.20</td>
<td>420214.7</td>
<td>-0.04942</td>
<td>0.649</td>
</tr>
<tr>
<td>3</td>
<td>434329.52</td>
<td>-14276.30</td>
<td>420053.2</td>
<td>-0.04979</td>
<td>0.649</td>
</tr>
<tr>
<td>4</td>
<td>434312.65</td>
<td>-14131.30</td>
<td>420181.4</td>
<td>-0.04950</td>
<td>0.646</td>
</tr>
<tr>
<td>5</td>
<td>434280.68</td>
<td>-14006.20</td>
<td>420274.5</td>
<td>-0.04929</td>
<td>0.643</td>
</tr>
<tr>
<td>6</td>
<td>434323.51</td>
<td>-14195.40</td>
<td>420128.1</td>
<td>-0.04962</td>
<td>0.647</td>
</tr>
<tr>
<td>7</td>
<td>434263.89</td>
<td>-13706.70</td>
<td>420557.2</td>
<td>-0.04865</td>
<td>0.637</td>
</tr>
<tr>
<td>8</td>
<td>434326.61</td>
<td>-14189.70</td>
<td>420136.9</td>
<td>-0.04960</td>
<td>0.647</td>
</tr>
<tr>
<td>9</td>
<td>434356.49</td>
<td>-14148.80</td>
<td>420207.7</td>
<td>-0.04944</td>
<td>0.647</td>
</tr>
<tr>
<td>10</td>
<td>434359.25</td>
<td>-14112.80</td>
<td>420246.4</td>
<td>-0.04935</td>
<td>0.647</td>
</tr>
<tr>
<td>11</td>
<td>434287.21</td>
<td>-14185.00</td>
<td>420102.2</td>
<td>-0.04968</td>
<td>0.646</td>
</tr>
<tr>
<td>12</td>
<td>434304.06</td>
<td>-14349.00</td>
<td>419955.1</td>
<td>-0.05001</td>
<td>0.649</td>
</tr>
<tr>
<td>13</td>
<td>434328.52</td>
<td>-14266.40</td>
<td>420062.1</td>
<td>-0.04977</td>
<td>0.648</td>
</tr>
<tr>
<td>14</td>
<td>434278.48</td>
<td>-13946.50</td>
<td>420332.0</td>
<td>-0.04916</td>
<td>0.642</td>
</tr>
<tr>
<td>15</td>
<td>434312.48</td>
<td>-14071.60</td>
<td>420240.9</td>
<td>-0.04936</td>
<td>0.645</td>
</tr>
<tr>
<td>16</td>
<td>434334.319</td>
<td>420158.0</td>
<td>-0.04955</td>
<td>0.647</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>434304.11</td>
<td>-13847.90</td>
<td>420456.2</td>
<td>-0.04888</td>
<td>0.641</td>
</tr>
<tr>
<td>18</td>
<td>434334.48</td>
<td>-14457.50</td>
<td>419877.0</td>
<td>-0.05019</td>
<td>0.652</td>
</tr>
<tr>
<td>19</td>
<td>434307.55</td>
<td>-14087.20</td>
<td>420220.3</td>
<td>-0.04941</td>
<td>0.645</td>
</tr>
<tr>
<td>20</td>
<td>434258.32</td>
<td>-14156.10</td>
<td>420102.2</td>
<td>-0.04968</td>
<td>0.645</td>
</tr>
<tr>
<td>average</td>
<td>434316.81</td>
<td>-14118.21</td>
<td>420198.6</td>
<td>-0.04946</td>
<td>0.646</td>
</tr>
<tr>
<td>std</td>
<td>32.14</td>
<td>170.89</td>
<td>162.3</td>
<td>0.00037</td>
<td>0.003</td>
</tr>
</tbody>
</table>
Table 4.4. Topological properties of the finial 20 oriC@midcell models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Tw</th>
<th>Wr</th>
<th>Lk</th>
<th>Supercoil density</th>
<th>Wr/ΔL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>433774.31</td>
<td>-13424.20</td>
<td>420350.1</td>
<td>-0.04912</td>
<td>0.618</td>
</tr>
<tr>
<td>2</td>
<td>433725.05</td>
<td>-13071.40</td>
<td>420653.6</td>
<td>-0.04843</td>
<td>0.611</td>
</tr>
<tr>
<td>3</td>
<td>433684.69</td>
<td>-13026.50</td>
<td>420658.2</td>
<td>-0.04842</td>
<td>0.609</td>
</tr>
<tr>
<td>4</td>
<td>433712.60</td>
<td>-13177.30</td>
<td>420535.3</td>
<td>-0.04870</td>
<td>0.612</td>
</tr>
<tr>
<td>5</td>
<td>433682.11</td>
<td>-13069.80</td>
<td>420612.3</td>
<td>-0.04852</td>
<td>0.609</td>
</tr>
<tr>
<td>6</td>
<td>433659.29</td>
<td>-13003.20</td>
<td>420656.1</td>
<td>-0.04842</td>
<td>0.607</td>
</tr>
<tr>
<td>7</td>
<td>433696.36</td>
<td>-13161.30</td>
<td>420535.1</td>
<td>-0.04870</td>
<td>0.611</td>
</tr>
<tr>
<td>8</td>
<td>433695.72</td>
<td>-13290.00</td>
<td>420405.7</td>
<td>-0.04899</td>
<td>0.614</td>
</tr>
<tr>
<td>9</td>
<td>433718.85</td>
<td>-13228.20</td>
<td>420490.6</td>
<td>-0.04880</td>
<td>0.613</td>
</tr>
<tr>
<td>10</td>
<td>433676.85</td>
<td>-13039.90</td>
<td>420637.0</td>
<td>-0.04847</td>
<td>0.609</td>
</tr>
<tr>
<td>11</td>
<td>433707.57</td>
<td>-13236.70</td>
<td>420470.9</td>
<td>-0.04884</td>
<td>0.613</td>
</tr>
<tr>
<td>12</td>
<td>433683.47</td>
<td>-12847.30</td>
<td>420836.2</td>
<td>-0.04802</td>
<td>0.605</td>
</tr>
<tr>
<td>13</td>
<td>433657.98</td>
<td>-12688.90</td>
<td>420969.1</td>
<td>-0.04772</td>
<td>0.602</td>
</tr>
<tr>
<td>14</td>
<td>433705.50</td>
<td>-12823.50</td>
<td>420882.0</td>
<td>-0.04791</td>
<td>0.605</td>
</tr>
<tr>
<td>15</td>
<td>433685.36</td>
<td>-13296.80</td>
<td>420388.6</td>
<td>-0.04903</td>
<td>0.614</td>
</tr>
<tr>
<td>16</td>
<td>433665.56</td>
<td>-12865.90</td>
<td>420799.7</td>
<td>-0.04810</td>
<td>0.605</td>
</tr>
<tr>
<td>17</td>
<td>433652.55</td>
<td>-13091.80</td>
<td>420560.7</td>
<td>-0.04864</td>
<td>0.609</td>
</tr>
<tr>
<td>18</td>
<td>433730.67</td>
<td>-13228.60</td>
<td>420502.1</td>
<td>-0.04877</td>
<td>0.614</td>
</tr>
<tr>
<td>19</td>
<td>433723.91</td>
<td>-13312.80</td>
<td>420411.1</td>
<td>-0.04898</td>
<td>0.615</td>
</tr>
<tr>
<td>20</td>
<td>433663.84</td>
<td>-13114.70</td>
<td>420549.1</td>
<td>-0.04867</td>
<td>0.610</td>
</tr>
<tr>
<td>average</td>
<td>433695.11</td>
<td>-13099.94</td>
<td>420595.2</td>
<td>-0.04856</td>
<td>0.610</td>
</tr>
<tr>
<td>std</td>
<td>29.63</td>
<td>183.18</td>
<td>166.3</td>
<td>0.0008</td>
<td>0.004</td>
</tr>
</tbody>
</table>
CHAPTER V
DISCUSSION AND FUTURE STUDIES

In this dissertation, three applications of computer simulations have been presented: they explore key aspects of important biomolecules with sizes ranging from two-residue peptides, to proteins, to whole chromosome structures. In this chapter, I will summarize these projects and discuss possible future directions.

The MD simulations described in Chapter II explored the use of an all-atom force field and an explicit water model to investigate the intrinsic conformational preferences of amino acids and the extent to which they are modulated by neighboring residues. 441 two-residue peptides containing all twenty standard amino acids, were simulated. Analysis of the convergence properties indicates that a 300 ns simulation time allows reasonable Ramachandran maps to be obtained for most of the two-residue peptides in this study. Then the $J_{HNH_{\alpha}}$ coupling constants and chemical shifts $\delta_{H_{\alpha}}$ were calculated from the MD simulations and compared to experimental values (Jung et al., 2014). Generally, the combination of the Amber ff99SB-ildn-NMR force field and the TIP4P-Ew water model reproduced quite well with experimental $J_{HNH_{\alpha}}$ coupling constants and chemical shifts $\delta_{H_{\alpha}}$ for two-residue peptides. Neighboring residue effects (NREs) on the average $J_{HNH_{\alpha}}$ and $\delta_{H_{\alpha}}$ values of adjacent residues were also reasonably well-reproduced. More importantly, the large NREs exerted experimentally by aromatic residues were correctly captured. These results suggest that the analyses of intrinsic conformational preferences of amino acids and their neighboring residue effects based on these MD simulations are quite reasonable. The intrinsic conformational preference of each amino acid was determined by constructing its average Ramachandran map, each representing 5.7 $\mu$s (300 ns x 19) of simulation data. In addition, 882 individual Ramachandran
maps were also constructed for each residue at N-terminal and C-terminal positions, respectively. The fractional populations of four major backbone conformations ($\alpha$, $\alpha'$, $\beta$, and PPII) were calculated for each residue type at the both termini. The neighboring residue effects were also calculated and plotted using fractional populations of four major backbone conformations.

Owing to the large number (441) of systems simulated in Chapter II, the computational demands of that study were significant. It was principally for reasons of expense that we chose to sample conformational behavior using equilibrium MD simulations: the combined simulation time of the simulations reported in Chapter II amounts to 130 $\mu$s. In the future, it should be possible to exploit increases in computer power to make use of more effective sampling techniques that have thus far typically only been used on much smaller numbers of systems. Examples include: (a) the metadynamics method (Barducci et al., 2008) that has been used to study single amino acids with a variety of force fields (Vymetal and Vondrášek 2013), and which showed that there was little consensus among the tested force fields regarding backbone conformational preferences, (b) the Hamiltonian replica exchange molecular simulation methods that have been used to study neighbor effects on the conformational behavior of a number of small Ala, Val, or Phe-containing peptides (Xu et al., 2008), and (c) the more conventional replica exchange methods (Sugita and Okamoto 1999) used in the iterative development of the recent RSFF1 (Jiang et al., 2014) and RSFF2 (Zhou et al., 2015) force fields. Improved sampling would have the added benefit of enabling comparisons between simulation and experiment to be performed on an individual peptide-by-peptide basis with confidence. This could be important given that the Cho group has noted (Jung et al., 2014) that the NREs exerted by a given type of amino acid can vary significantly depending on the identity of the other amino acid being examined.
Although a combination of the force field (Amber ff99SB-ildn-NMR force field) and water model (TIP4P-Ew) used in Chapter II resulted in a reasonable correlation between MD-predicted and experimental measured $^3J_{HN\alpha}$ coupling constants, it was far from perfect. The inadequacy of the force field is the most likely factor that is responsible for the disagreement. Therefore, in Chapter III, I described tests of the abilities of two extensively used AMBER force fields (ff99SB-ILDN and ff99SB-ILDN-NMR), together with the recently derived residue-specific force field (RSFF2), to reproduce the $^3J_{HN\alpha}$ coupling constants of a large number of two-residue peptides. Much longer simulations (1µs) were performed on 256 two-residue peptides whose capping groups are exactly the same as those used in the experimental studies and five proteins whose experimental $^3J_{HN\alpha}$ coupling constant are available. Two sets of Karplus parameters (Hu and Bax 1997) were used to calculate the $^3J_{HN\alpha}$ coupling constants.

All three force fields produced reasonable correlations with the experimental data for the two-residue peptide systems. It was found that the $^3J_{HN\alpha}$ coupling constants of C-terminal residues were consistently better reproduced than those of N-terminal residues. Importantly, RSFF2 produced much better agreement with experiment than the two ILDN-based force fields (Figure 3.6). In terms of neighboring residue effects, however, all three force fields showed consistently low agreement with experiment. Nevertheless, RSFF2 still showed better agreement in the overall $r^2$ compared to ff99SB-ILDN and ff99SB-ILDN-NMR. Overall, for the 256 two-residue systems, RSFF2 showed a significant improvement in the reproduction of $^3J_{HN\alpha}$ coupling constants but a much smaller improvement in the reproduction of the neighboring residue effects.
When the three force fields were applied to five model proteins, it was found that RSFF2 again performed better in reproducing the experimental $^{3}\!\!J_{HN\alpha}$ coupling constants than either of the two other ff99SB-ILDN force fields (Figure 3.13, 3.14, 3.15). However, for these five proteins, there were certain residues whose $^{3}\!\!J_{HN\alpha}$ coupling constants appeared to be poorly reproduced by all three force fields. More proteins may be required to validate further the abilities of the three force fields in reproducing experimental $^{3}\!\!J_{HN\alpha}$ coupling constants and help force field refinement. Very recently, work from Mark Miller in the Elcock lab (Miller et al., 2017) has shown that modifications of nonbonded interactions of several commonly used force fields are required to improve the agreement with experimental osmotic coefficients. Thus, better reproduction of NREs might require better modeling of nonbonded sidechain-sidechain interactions.

In Chapter IV, coarse-grained (CG) models and multi-scale modeling methods were used to develop structural models of entire *E. coli* chromosomes confined within the experimentally-determined volume of the nucleoid. Using an initial chromosome structure at a resolution of 500 base-pair-per-bead (BPB) (Will Hacker’s work in the lab), two intermediate models (5 BPB and 1 BPB) were utilized to construct chromosome models with final resolutions of one-nucleotide-per-bead (1 NTB). The resulting model represents a significant increase in resolution relative to previously published coarse-grained (CG) chromosome models in which one bead usually represents hundreds or even thousands of base pairs (Fritsche et al., 2012; Hong et al., 2013; Junier et al., 2014; Le et al., 2013). The CG models used in this work were parameterized to reproduce the persistence length of double-stranded B-DNA. For both sets of chromosome models (oriC@pole and oriC@midcell), we built 20 structures for each set. Based on the final 1NTB structures, physical properties such as major
and minor groove widths, distributions of local DNA bending angles, topological parameters (Linking Number (Lk), Twist (Tw) and Writhe (Wr)) were accurately computed and compared with experimental measurements or predictions from a worm-like chain (WLC) model. These calculations indicate that the nucleotide-resolution chromosome structures are free of significant conformational strain and geometrically reasonable.

One important application of the chromosome structures we built is to explore the available interior space that is accessible to large macromolecular complexes such as ribosomes. To obtain this, we calculated the distribution of void diameters accessible within the interior of the nucleoid, averaged over the 20 structures for both the oriC@pole and oriC@midcell models. The results are shown in Figure 5.1. Broad distributions of void diameters were obtained for both types of model. The mode of void diameters is \( \sim 150 \) Å but with substantial populations of much larger void diameters (as high as \( \sim 800 \) Å). This suggests that, at least as far as excluded-volume effects are concerned, some regions of the nucleoid interior are likely to be accessible even to large macromolecular complexes. This finding was consistent with the results of recent single-molecule imaging studies of ribosomal subunits in \textit{E. coli} (Sanamrad et al., 2014).

The chromosome models we built in Chapter IV have incorporated a number of experimental restraints, but future models might include additional restraints. For example, our current models should account implicitly for the combined effects of macromolecular crowding and nucleoid-associated proteins (NAPs) because, as a population, they reproduced the experimental spatial distributions of DNA and transcribing RNAP molecules (Stracy et al., 2015). Future chromosome models might explicitly incorporate NAPs to examine their effects on DNA geometry. In addition, our current models were constructed with homogeneous DNA, so sequence-specific effects (Rohs et al., 2009) were not
included. It is also expected, therefore, that we will attempt to incorporate such sequence effects (Zhou et al., 2013) in future generations of chromosome models. Lastly, higher-resolution contact information from techniques such as HiC (Ay and Noble 2015) could also be included as it emerges, adding to the experimental data the models incorporate.

The high-resolution structural models of *E. coli* chromosome we built in Chapter IV provide an important step toward constructing a whole-cell model of an *E. coli* cell. The development of a detailed structural model for an *E. coli* cell will allow the general principles that govern biochemical processes to be revealed. In addition, we believe that the simulation methods presented here are generalizable; we hope that they can also be applied to other bacterial chromosomes.
Figure 5.1. Distributions of void sizes in the chromosome models. oriC@pole (blue) and oriC@midcell (red) (Hacker et al., 2017)
APPENDIX A: A PERL CODE FOR CALCULATING $^3J_{\text{HNHa}}$ COUPLING CONSTANTS

The following code is used to calculate coupling constants from MD trajectories computed using the GROMACS MD package.

Required input parameters are: (a) md.xtc, (b) topol.tpr, (c) number of residues in the structure

Required program: Gromacs 4.x.x version

To run this perl code, put the following source code into a script.pl file; then, in a Linux command terminal, type (for example):

perl script.pl md.xtc topol.tpr 10

#!/usr/bin/perl
use strict;
use warnings;
use constant PI => 3.14159/180;
$ARGV[0] or die "STOP! Give the md.xtc file!\n";
$ARGV[1] or die "STOP! Give the topol.tpr file!\n";
$ARGV[2] or die "STOP! Give the number of residues in the structure file !\n";
my $totalAA =$ARGV[2] - 2;
my $outfile = "calculated_Jcoupling_3HnHa.txt";
system "g_rama -f $ARGV[0] -s $ARGV[1] -o phipsi_protein.xvg ";
unless(-e "phipsi_protein.xvg")
{   die "phipsi_protein.xvg does not exist!\n";
}
open (my $INPUT," phipsi_protein.xvg ") || die "Cannot open phipsi_protein.xvg \n";
my $i=0;
my $k=0;
my $phi=0;
my $aminoacid;
my @residue;
my $J=0;
my @J_add;
my @J_aver;
my $A=7.09;
my $B=1.42;
my $C=1.55;
my $a=0;
my $b=0;
my $c=0;
while(<$INPUT>)
{
    if ($_ !~ m/^#/ and_$_ !~ m/^@/))
    {
        my @line = split (/\s+/,$_);
        if($i < $totalAA) {residue[$i] = $line[2];}
        $phi = $line[0];
        $aminoacid = substr($line[2],0,3);
        if($aminoacid =~ m/ALA/)
        {
            $a=$A; $b=$B; $c=$C;
            $phi = ($phi - 60) * PI;
            $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;
        }
        elsif($aminoacid =~ m/ARG/)
        {
            $a=$A; $b=$B; $c=$C;
            $phi = ($phi - 60) * PI;
            $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;
        }
    }
}
$J = a * \cos(\phi) * \cos(\phi) - b * \cos(\phi) + c;

}

elsif(\$aminoacid =~ m/ASN/) {
  $a=A; $b=B; $c=C;
  \$phi = ($phi - 60) * PI;
  $J = a * \cos(\phi) * \cos(\phi) - b * \cos(\phi) + c;
}

elsif(\$aminoacid =~ m/ASP/) {
  $a=A; $b=B; $c=C;
  \$phi = ($phi - 60) * PI;
  $J = a * \cos(\phi) * \cos(\phi) - b * \cos(\phi) + c;
}

elsif(\$aminoacid =~ m/CYS/) {
  $a=A; $b=B; $c=C;
  \$phi = ($phi - 60) * PI;
  $J = a * \cos(\phi) * \cos(\phi) - b * \cos(\phi) + c;
}

elsif(\$aminoacid =~ m/GLN/) {
  $a=A; $b=B; $c=C;
  \$phi = ($phi - 60) * PI;
  $J = a * \cos(\phi) * \cos(\phi) - b * \cos(\phi) + c;
}

elsif(\$aminoacid =~ m/GLU/) {
  $a=A; $b=B; $c=C;
  \$phi = ($phi - 60) * PI;
  $J = a * \cos(\phi) * \cos(\phi) - b * \cos(\phi) + c;
}
elsif($aminoacid =~ m/GLY/)  
{  $a=$A; $b=$B; $c=$C;  
    $phi = ($phi - 60) * PI;  
    $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;  
}
elsif($aminoacid =~ m/HIP/)  
{  $a=$A; $b=$B; $c=$C;  
    $phi = ($phi - 60) * PI;  
    $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;  
}
elsif($aminoacid =~ m/HIS/)  
{  $a=$A; $b=$B; $c=$C;  
    $phi = ($phi - 60) * PI;  
    $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;  
}
elsif($aminoacid =~ m/ILE/)  
{  $a=$A; $b=$B; $c=$C;  
    $phi = ($phi - 60) * PI;  
    $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;  
}
elsif($aminoacid =~ m/LEU/)  
{  $a=$A; $b=$B; $c=$C;  
    $phi = ($phi - 60) * PI;  
    $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;  
}
elsif($aminoacid =~ m/LYS/)  
{  $a=$A; $b=$B; $c=$C;
$\phi = (\phi - 60) \times \pi; 
$J = a \times \cos(\phi) \times \cos(\phi) - b \times \cos(\phi) + c; 
}
}
elsif($aminoacid =~ m/MET/)
  
  $a=A; $b=B; $c=C;
  $\phi = (\phi - 60) \times \pi; 
  $J = a \times \cos(\phi) \times \cos(\phi) - b \times \cos(\phi) + c; 

eelsif($aminoacid =~ m/PHE/)
  
  $a=A; $b=B; $c=C;
  $\phi = (\phi - 60) \times \pi; 
  $J = a \times \cos(\phi) \times \cos(\phi) - b \times \cos(\phi) + c; 

eelsif($aminoacid =~ m/PRO/)
  
  $a=A; $b=B; $c=C;
  $\phi = (\phi - 60) \times \pi; 
  $J = a \times \cos(\phi) \times \cos(\phi) - b \times \cos(\phi) + c; 

eelsif($aminoacid =~ m/SER/)
  
  $a=A; $b=B; $c=C;
  $\phi = (\phi - 60) \times \pi; 
  $J = a \times \cos(\phi) \times \cos(\phi) - b \times \cos(\phi) + c; 

eelsif($aminoacid =~ m/THR/)
  
  $a=A; $b=B; $c=C;
  $\phi = (\phi - 60) \times \pi; 
  $J = a \times \cos(\phi) \times \cos(\phi) - b \times \cos(\phi) + c; 

elsif($aminoacid =~ m/TRP/)
{ $a=$A; $b=$B; $c=$C;
    $phi = ($phi - 60) * PI;
    $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;
}
elsif($aminoacid =~ m/TYR/)
{ $a=$A; $b=$B; $c=$C;
    $phi = ($phi - 60) * PI;
    $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;
}
elsif($aminoacid =~ m/VAL/)
{ $a=$A; $b=$B; $c=$C;
    $phi = ($phi - 60) * PI;
    $J = $a * cos($phi) * cos($phi) - $b * cos($phi) + $c;
}
$k = ($i+1) % ($totalAA);
if($k == 0)
    { $k = $totalAA;
    }
$J_add[$k] += $J;
$i++;
if($i % ($totalAA * 100) == 0)
    { my $finish = $i / $totalAA;
        print "$finish frames finished\n";
    }
}
my $total_phi = $i;
my $each_phi = $total_phi/$totalAA;
print "There are totally $total_phi phi angles in this protein\n";
print "Each residue contains $each_phi phi angles\n";
if ($total_phi % $totalAA != 0)
{   die "wrong total J coupling number!\n";
}

open (my $OUT, ">$outfile") || die "Cannot create destination file\n";
for(my $j=1; $j <= $k; $j++)
{  $J_aver[$j] = sprintf ("%.3f", $J_add[$j] / $each_phi);
    printf $OUT "%8s%8s\n", $residue[$j-1], $J_aver[$j];
}
close $OUT;
print "See the results in $outfile.\n\n";
APPENDIX B: A C++ CODE FOR CALCULATING WRITHE NUMBER

The following code is used to calculate the writhe number of a chromosome model stored in "bdb" format; .bdb is an in-house structure format that is identical to the pdb format but uses 12.3f for the coordinates instead of 8.3f.

Required input parameters are: (a) model.bdb, (b) result.writhe.txt (i.e. the name of the output file)

To run this C++ code, put the following source code into a writhe.cpp and compile it with a C++ compiler with OpenMP:

g++ writhe.cpp -o writhe.exe -fopenmp

Then, in a Linux command terminal, type

./writhe.exe model.bdb result.writhe.txt

#include <iostream>
#include <fstream>
#include <vector>
#include <string>
#include <cstdlib>
#include <iomanip>
#include <cmath>
#include <omp.h>
#define CHUNK 1
#define THREAD_NUM 64

class Coordinate
{
    public:
        float x;
float y;
float z;

Coordinate operator+(const Coordinate& coor){
    Coordinate newcoor;
    newcoor.x = this->x + coor.x;
    newcoor.y = this->y + coor.y;
    newcoor.z = this->z + coor.z;
    return newcoor;
}

Coordinate operator-(const Coordinate& coor){
    Coordinate newcoor;
    newcoor.x = this->x - coor.x;
    newcoor.y = this->y - coor.y;
    newcoor.z = this->z - coor.z;
    return newcoor;
}

double operator*(const Coordinate& coor){
    return (this->x * coor.x + this->y * coor.y + this->z * coor.z);
}

};

Coordinate Cross (const Coordinate& coor1, const Coordinate& coor2);
double Size(const Coordinate& coor1);
void PrintUsage();

int main (int argc, char *argv[])
{
    char *inputFile = argv[1];
    char *outputFile = argv[2];
    if(argc != 3){

PrintUsage();
return EXIT_FAILURE;
}

std::ifstream Fin(inputFile);
if(!Fin) std::cerr<<"Cannot open input file " << inputFile << std::endl;
std::vector<Coordinate> atoms;
std::string buf;
Coordinate xyz;
int number = 1;
int interval = 1;
while (getline(Fin, buf)){
if(number % interval == 0){
  xyz.x = atof(buf.substr(31, 11).c_str());
  xyz.y = atof(buf.substr(43, 11).c_str());
  xyz.z = atof(buf.substr(55, 11).c_str());
  atoms.push_back(xyz);
}
number++;
}
atoms.push_back(atoms[0]);
Fin.close();
std::cout << "Total atom number: " << atoms.size()-1 << std::endl;
double W1 = 0;
double size = 0;
double W2 = 0;
double total = 0;
double ptotal = 0;
int i, j;
Coordinate s1, s2, V, C;

#pragma omp parallel private(j, s1, s2, V, C, W1, size, W2, ptotal)
reduction(+:total)
{
    #pragma omp for schedule(dynamic, CHUNK)
    for(i=0; i<atoms.size()-1; i++) {
        for(j=i+1; j<atoms.size()-1; j++) {
            s1 = atoms[i+1] - atoms[i];
            s2 = atoms[j+1] - atoms[j];
            V = atoms[i] - atoms[j];
            C = Cross(s1, s2);
            W1 = C * V;
            size = Size(V);
            W2 = W1 / (size * size * size);
            ptotal += W2;
        }
        if(i % 10000 == 0) {
            std::cout << inputFile << " finish:" << i << std::endl;
        }
    }
    total += ptotal;
}

total = total / (2 * M_PI);
std::ofstream Fout(outputFile);
if(!Fout)std::cerr<<"Cannot open output file " << outputFile << std::endl;
std::cout << inputFile << " Calculated writhe number: " << total << "\n";
Fout << inputFile << " Calculated writhe number: " << total << "\n";
    Fout.close();
    return 0;
}

void PrintUsage()
{
    std::cout << "\nUsage: XXX.exe <input_file> <output_file> \n
$number_of_basepairs\n\n"
;
}

Coordinate Cross (const Coordinate& coor1, const Coordinate& coor2)
{
    Coordinate newcoor;
    newcoor.x = coor1.y * coor2.z - coor1.z * coor2.y;
    newcoor.y = coor1.z * coor2.x - coor1.x * coor2.z;
    newcoor.z = coor1.x * coor2.y - coor1.y * coor2.x;
    return newcoor;
}

double Size(const Coordinate& coor1)
{
    return sqrt(coor1.x*coor1.x + coor1.y*coor1.y + coor1.z*coor1.z);
}
APPENDIX C: A C++ CODE FOR CALCULATING LINKING NUMBER

The following code is used to calculate the linking number of a chromosome model stored in "bdb" format; .bdb is an in-house structure format that is identical to the pdb format but uses 12.3f for the coordinates instead of 8.3f.

Required input parameters are: (a) model.bdb, (b) result.linking.txt (i.e. name of the output file)

To run this c++ code, put the following source code into a writhe.cpp and compile it with a C++ compiler with OpenMP:

```
g++ linking.cpp -o linking.exe -fopenmp
```

Then, in a Linux command terminal, type

```
./linking.exe    model.bdb    result.linking.txt
```

```c++
#include <iostream>
#include <fstream>
#include <vector>
#include <string>
#include <cstdlib>
#include <iomanip>
#include <cmath>
#include <omp.h>
#define CHUNK 1
#define THREAD_NUM 64
#define ZERO 1.0e-5

class Coordinate {
public:
```
float x;
float y;
float z;

Coordinate operator+(const Coordinate& coor){
    Coordinate newcoor;
    newcoor.x = this->x + coor.x;
    newcoor.y = this->y + coor.y;
    newcoor.z = this->z + coor.z;
    return newcoor;
}

Coordinate operator-(const Coordinate& coor){
    Coordinate newcoor;
    newcoor.x = this->x - coor.x;
    newcoor.y = this->y - coor.y;
    newcoor.z = this->z - coor.z;
    return newcoor;
}

Coordinate operator/(const double &a){
    Coordinate newcoor;
    newcoor.x = this->x / a;
    newcoor.y = this->y / a;
    newcoor.z = this->z / a;
    return newcoor;
}

Coordinate operator*(const double &a){
    Coordinate newcoor;
    newcoor.x = this->x * a;
    newcoor.y = this->y * a;
    newcoor.z = this->z * a;
    return newcoor;
}
newcoor.y = this->y * a;
newcoor.z = this->z * a;
return newcoor;
}
double operator*(const Coordinate& coor){
    return (this->x * coor.x + this->y * coor.y + this->z * coor.z);
}
};
const Coordinate Cross (const Coordinate& coor1, const Coordinate& coor2);
const double Link (Coordinate& t1, Coordinate& t2, Coordinate& rjoin);
const double Size(const Coordinate& coor1);
void PrintUsage();
int main (int argc, char *argv[])
{
    char *inputFile = argv[1];
    char *outputFile = argv[2];
    if(argc != 3){
        PrintUsage();
        return EXIT_FAILURE;
    }
    std::ifstream Fin(inputFile);
    if(!Fin) std::cerr<<"Cannot open input file " << inputFile << std::endl;
    std::vector<Coordinate> PDB1;
    std::vector<Coordinate> PDB2;
    Coordinate xyz;
    std::string buf;
    int number = 1;
    while (getline(Fin, buf)){
xyz.x = atof(buf.substr(31, 11).c_str());
xyz.y = atof(buf.substr(43, 11).c_str());
xyz.z = atof(buf.substr(55, 11).c_str());
if(number%2 == 1) {
    PDB2.push_back(xyz);
}
else {
    xyz.x = 0.5 * (xyz.x + PDB2.back().x);
    xyz.y = 0.5 * (xyz.y + PDB2.back().y);
    xyz.z = 0.5 * (xyz.z + PDB2.back().z);
    PDB1.push_back(xyz);
}
    number++;
}
PDB1.push_back(PDB1[0]);
PDB2.push_back(PDB2[0]);
Fin.close();
std::cout << "Total bps: " << PDB1.size()-1 << " should be equal to " << PDB2.size()-1 "std::endl;
    double total = 0;
    double ptotal = 0;
    int i, j;
    double S1, S2;
    Coordinate segment1, segment2, r, rtemp, k1, k2;
#pragma omp parallel private(j, ptotal, S1, S2, segment1, segment2, r, rtemp, k1, k2)
reduction(+:total)
{
    #pragma omp for schedule(dynamic, CHUNK)
for(i=0; i<PDB1.size()-1; i++) {
    for(j=0; j<PDB2.size()-1; j++) {
        segment1 = PDB1[i+1] - PDB1[i];
        segment2 = PDB2[j+1] - PDB2[j];
        S1 = Size(segment1);
        S2 = Size(segment2);
        k1 = segment1 / S1;
        k2 = segment2 / S2;
        rtemp = PDB1[i] - PDB2[j];
        r = rtemp + segment1;
        ptotal += Link(k1, k2, r);
        r = rtemp - segment2;
        ptotal += Link(k1, k2, r);
        r = rtemp + segment1 - segment2;
        ptotal -= Link(k1, k2, r);
        r = rtemp;
        ptotal -= Link(k1, k2, r);
    }
    if(i % 10000 == 0) {
        std::cout << inputFile << " finish:" << i << std::endl;
    }
}

total += ptotal;
}

total = total / (4 * M_PI);
std::ofstream Fout(outputFile);
if(!Fout)std::cerr<<"Cannot open output file " << outputFile << std::endl;
std::cout << std::setprecision(3) << std::fixed << inputFile << " Calculated linking number: " << total << "\n";
Fout << std::setprecision(3) << std::fixed << inputFile << " Calculated linking number: " << total << "\n";
Fout.close();
return 0;
}
void PrintUsage()
{
std::cout << "Usage: XXX.exe <input_file> <output_file>\n\n";
}
const Coordinate Cross (const Coordinate& coor1, const Coordinate& coor2)
{
    Coordinate newcoor;
    newcoor.x = coor1.y * coor2.z - coor1.z * coor2.y;
    newcoor.y = coor1.z * coor2.x - coor1.x * coor2.z;
    newcoor.z = coor1.x * coor2.y - coor1.y * coor2.x;
    return newcoor;
}
const double Size(const Coordinate& coor1)
{
    return sqrt(coor1.x*coor1.x + coor1.y*coor1.y + coor1.z*coor1.z);
}
const double Link (Coordinate& k1, Coordinate& k2, Coordinate& r)
{
    double norm_1, norm_2, norm_temp, i, j, k;
    Coordinate norm_a, norm_b;
    norm_a = Cross(r, k1);
    norm_b = Cross(r, k2);
    norm_1 = Size(norm_a);
    norm_2 = Size(norm_b);
if(norm_1 <= ZERO || norm_2 <= ZERO)
{
    return 0;
}

else
{
    norm_temp = norm_1 * norm_1 * norm_2 * norm_2;
    i = norm_b * norm_a / norm_temp;
    j = -1 * Size(r) * (r * Cross(k1, k2)) / norm_temp;
    k = atan2(j, -i);
    return k;
}

}
REFERENCES


