Statistical summary of protein structures

Yuanyuan Huang
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Bioinformatics Commons

Recommended Citation
Statistical summary of protein structures

by

Yuanyuan Huang

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

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Zhijun Wu, Co-major Professor
Robert Jernigan, Co-major Professor
Karin Dorman
Karen Abbott
Willson Stephen

Iowa State University
Ames, Iowa
2013

Copyright © Yuanyuan Huang, 2013. All rights reserved.
DEDICATION

To my parents
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER 1. General Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Dissertation Organization</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Literature Review</td>
<td>3</td>
</tr>
<tr>
<td>CHAPTER 2. Statistical Measures on Residue-Level Protein Structural Properties</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>7</td>
</tr>
<tr>
<td>2.1.1 Atomic-level Structural Properties</td>
<td>7</td>
</tr>
<tr>
<td>2.1.2 Residual-level Structural Properties</td>
<td>8</td>
</tr>
<tr>
<td>2.2 Computation Methods</td>
<td>11</td>
</tr>
<tr>
<td>2.2.1 Selection of Known Protein Structures</td>
<td>11</td>
</tr>
<tr>
<td>2.2.2 Calculation of Residue Distances and Angles</td>
<td>11</td>
</tr>
<tr>
<td>2.2.3 Calculation of Distance and Angle Distributions</td>
<td>12</td>
</tr>
<tr>
<td>2.2.4 Calculation of Angle-Angle Correlations</td>
<td>12</td>
</tr>
<tr>
<td>2.2.5 Calculation of Angle-Distance Correlations</td>
<td>13</td>
</tr>
<tr>
<td>2.2.6 Definition of Residue Distance and Angle Potentials</td>
<td>14</td>
</tr>
<tr>
<td>2.2.7 Display of Residue Angle Correlations of a Structure</td>
<td>15</td>
</tr>
<tr>
<td>2.3 Results and Discussion</td>
<td>15</td>
</tr>
<tr>
<td>2.3.1 Distribution of Residue Distances</td>
<td>15</td>
</tr>
</tbody>
</table>
2.3.2 Distributions of Residue Angles ................................. 25
2.3.3 Distributions of Residue Torsion Angles ......................... 26
2.3.4 Angles-angle Correlations ...................................... 28
2.3.5 Applications to Structural Analysis ............................... 31
2.4 Conclusion ............................................................ 38

CHAPTER 3. P.R.E.S.S. – An R-package for Exploring Residual-level Protein Structural Statistics ............................................. 41
3.1 Introduction ............................................................ 42
3.2 Graphics Interface and System Organization ......................... 43
3.3 Functional Modules .................................................... 46
3.4 Conclusion ............................................................ 54

CHAPTER 4. PRESS- PLOT: An Online Server for Protein Structural Analysis and Evaluation using Residue-level Virtual Angle Correlation Plots .......................... 58
4.1 Introduction ............................................................ 59
4.2 Implementation ........................................................ 61
4.2.1 Structural Data ..................................................... 61
4.2.2 Plot of Density Maps .............................................. 62
4.2.3 Web Interface ....................................................... 62
4.3 Results ................................................................. 65
4.3.1 Display functions ................................................... 65
4.3.2 Testings .............................................................. 68
4.4 Discussion ............................................................. 69

CHAPTER 5. General Conclusion ........................................... 73
5.1 General Discussion ..................................................... 73
5.2 Future Research ....................................................... 75

CHAPTER 6. Game Dynamic Model for Yeast Development ............. 79
6.1 Introduction ............................................................ 79
6.2 Yeast Development as a Snowdrift Game ............................. 83
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Sample virtual bond lengths</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td>Density distribution of $\alpha$-$\tau$ angle pairs</td>
<td>29</td>
</tr>
<tr>
<td>2.3</td>
<td>Density distribution of $\tau$-$\beta$ angle pairs</td>
<td>29</td>
</tr>
<tr>
<td>3.1</td>
<td>Structure analysis for previously obsoleted structures and current ones</td>
<td>54</td>
</tr>
<tr>
<td>3.2</td>
<td>Structure analysis for previously obsoleted structures and current ones</td>
<td>55</td>
</tr>
<tr>
<td>4.1</td>
<td>Assessment on PDB structures with $\alpha$-$\tau$ correlation plots</td>
<td>70</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

2.1 Ramachandran Plot ............................................. 8
2.2 Residue distances and angles .................................. 9
2.3 Distribution of virtual bond length .............................. 16
2.4 Distributions of virtual bond length in α-helices and β-sheets .............................................................. 17
2.5 Distribution of virtual bond length in short distance range ................................................................. 18
2.6 Distributions of virtual bond length in α-helices and β-sheets in short distance range ................................................................. 19
2.7 Distributions of ω-angles for residue pairs ....................... 19
2.8 Distribution of virtual bond angle .................................. 20
2.9 Distributions of virtual bond angle in α-helices and β-sheets .............................................................. 21
2.10 Distributions of virtual bond angle in different secondary structures ............................................................ 22
2.11 Correlation of residue-level 1-3-distance and virtual bond angle ............................................................ 23
2.12 Sample virtual bond lengths ........................................ 24
2.13 Distributions of virtual torsion angle in α-helices and β-sheets ............................................................ 31
2.14 Distributions of virtual torsion angle in different secondary structures ............................................................ 32
2.15 Correlation of residue-level 1-4-distance and virtual torsion angle ............................................................ 33
2.16 Contour of density distribution of α-τ angles pairs .................. 34
2.17 Contour of density distribution of τ-β angles pairs .................. 35
2.18 Contour of density distribution of α-β angle pairs .................. 36
2.19 Contour of density distribution of τ₁-τ₂ angle pairs .................. 37
3.1 P.R.E.S.S. graphics interface ...................................... 45
3.2 System organization and interface .................................. 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Distribution of virtual bond lengths between ASN and MET</td>
</tr>
<tr>
<td>3.4</td>
<td>Distribution of virtual bond angles formed by Any, ASN, and Any</td>
</tr>
<tr>
<td>3.5</td>
<td>Scattered plot of the virtual bond angles against their residue 1-3 distances</td>
</tr>
<tr>
<td>3.6</td>
<td>A matrix of scattered plots</td>
</tr>
<tr>
<td>3.7</td>
<td>3D scattered plot for the virtual torsion angle and its two neighboring virtual bond angles</td>
</tr>
<tr>
<td>3.8</td>
<td>Distributions of virtual bond energies for 1PHY (2.4Å) and 2PHY (1.4Å)</td>
</tr>
<tr>
<td>3.9</td>
<td>Distribution of virtual bond angle energies for 1PHY (2.4Å) and 2PHY (1.4Å)</td>
</tr>
<tr>
<td>3.10</td>
<td>The $\alpha$-$\tau$ correlation plots for a protein at two different resolutions</td>
</tr>
<tr>
<td>3.11</td>
<td>The $\tau$-$\beta$ correlation plots for a protein at different resolutions</td>
</tr>
<tr>
<td>3.12</td>
<td>The boxplots for $\alpha$-$\tau$ correlation</td>
</tr>
<tr>
<td>3.13</td>
<td>The boxplots for $\tau$-$\beta$ correlation</td>
</tr>
<tr>
<td>4.1</td>
<td>PRESS-PLOT organization</td>
</tr>
<tr>
<td>4.2</td>
<td>The front page of PRESS-PLOT</td>
</tr>
<tr>
<td>4.3</td>
<td>A virtual angle correlation plot for $\alpha$-helices</td>
</tr>
<tr>
<td>4.4</td>
<td>A virtual angle correlation plot for $\beta$-sheets</td>
</tr>
<tr>
<td>5.1</td>
<td>The 3D delaunay triangulation of a protein</td>
</tr>
<tr>
<td>5.2</td>
<td>The distribution of distances between two contacted pairs derived from protein triangulation</td>
</tr>
<tr>
<td>5.3</td>
<td>The distribution of distances between two specific types of contacted pairs</td>
</tr>
<tr>
<td>6.1</td>
<td>A mixed culture of two yeast strains</td>
</tr>
<tr>
<td>6.2</td>
<td>The co-existence of two yeast strains</td>
</tr>
<tr>
<td>6.3</td>
<td>The nonlinear game model for yeast development</td>
</tr>
<tr>
<td>6.4</td>
<td>Cooperator fractions at equilibrium</td>
</tr>
<tr>
<td>6.5</td>
<td>Color-coded cost values</td>
</tr>
<tr>
<td>6.6</td>
<td>Dynamic changes of cooperator fractions</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>6.7</td>
<td>Dynamic changes of cooperator fractions</td>
</tr>
<tr>
<td>6.8</td>
<td>Equilibrium fractions</td>
</tr>
<tr>
<td>6.9</td>
<td>Stability of equilibrium state</td>
</tr>
</tbody>
</table>
ABSTRACT

Every biological system has proteins, and almost all biological activities require the participation or support of a specific set of proteins. Therefore, understanding the functions of the proteins is essential to research in all biological and medical fields. To fully understand their functions, however, it is critical to know their structures and related dynamic behaviours.

There is no unique way of modelling protein structure and dynamics. Experimental techniques have been employed to collect some indirect structural data from which the structures can be deduced. These techniques are costly and time consuming and limited to certain types or sizes of proteins. Yet, they are presently the major sources available for structure determination. Theoretical approaches have been developed for modeling protein structure and dynamics, including potential energy minimization, molecular dynamics simulation, and comparative modeling. In practice, these methods are often combined. Yet, all these methods have limitations, and their modelling capabilities, even when they are combined, are not yet sufficient to meet the high quality and high quantity modelling demands from applications. New approaches and breakthroughs are actively sought.

In this work, we investigate a novel statistical approach to protein modelling. Instead of relying on physical experiments, we analyse a whole spectrum of residue-level protein structural properties statistically, for better understanding their physical and structural properties revealed in the known structural data. The data-driven and knowledge-based exploration and analysis of structural properties could take advantage of the knowledge extracted from the rich available data, and also the power of statistical methods. We first develop the statistical measures on protein residue-level structural properties. We further introduce a statistical framework for protein structural assessment, and the formulation of a novel set of residue-level statistical potentials for protein modelling and dynamics. Secondly, to allow researchers to access and manipulate a large set of statistical data on protein residue-level structural properties
and evaluation of statistical potentials, an open source package is developed and released in R, with a user-friendly GUI, accessible and executable by a public user in any R environment. Lastly, we integrate web pages and server-side programs in a one-step query workbench, making it easy for a user to submit queries and acquire results. The implementation is carried out in PHP - a popular and widely supported scripting language.
CHAPTER 1. General Introduction

1.1 Introduction

The number of protein structures determined computationally and experimentally is constantly increasing over the last decades. There are presently more than 84,000 protein structures that have been determined and deposited in the protein data bank (PDB). Advanced experimental and theoretical techniques will lead to the growth in worldwide collaborations for structure prediction and determination over the next ten years.

Experimental techniques have been employed to collect some indirect structural data from which the structures can be deduced, such as X-ray diffraction data from X-ray crystallography or the inter-atomic distance data from nuclear magnetic resonance spectroscopy (NMR). These techniques are costly and time consuming and limited to certain types or sizes of proteins. Yet, they are presently the major methods of available for structure determination (1; 2).

The knowledge of structural properties has been crucial for both theoretical and experimental approaches to protein modelling. The atomic-level structural properties of proteins, such as bond lengths, bond angles, and torsion angles, have been well studied and understood based on either small molecule knowledge or statistical analysis. The correlations among these atomic-level properties have been an important source of information as well. The information about this correlation has been employed in both experimental determination and theoretical prediction of protein structures. Potential functions have been defined for bond lengths, bond angles, and torsion angles using their known or preferred values. Energetically favourable structures can be obtained when these potentials along with other non-bond potentials are combined and minimized. In either NMR or X-ray crystallography, these properties have been used to refine an initial experiment based model, which may otherwise have little atomic detail.
In particular, in NMR, the experimental NOE (Nuclear Overhauser Effect) data is mainly for hydrogen-hydrogen distances, which is insufficient for fully determining a structure without utilizing other data on bond lengths and bond angles.

Similar properties of the residue-level, such as the distances between two residues and the angles formed by short sequences of residues, can be equally important for structural analysis and modelling, but these have not been examined and documented on a similar scale. The reason is that they are not easy to measure directly. The physics for the interactions between residues are not fully clear; and they are not as rigid as the bond lengths and bond angles, i.e., their values are likely to vary over a wide range. While these properties are difficult to measure experimentally, they can be statistically estimated in meaningful ways based on their distributions in known proteins structures.

Structural properties similar to those atomic-level properties described above can also be found at the residue-level, such as the distances between two neighbouring residues; the angles formed by three residues in sequence; and the torsion angles of four residues in sequence. Proteins are often modelled in a reduced form, with residues considered as basic units. The residue distances and angles then become crucial for the description of the model, and they can be as important as those at the atomic-level for structural determination, prediction, and evaluation. The knowledge of these distances and angles can also be used to define residue-level potential functions so that potential energy minimization and dynamics simulations can be performed more effectively and efficiently at the residue instead of atomic-level, because the number of variables is substantially reduced and the time step may be increased.

We investigate a novel statistical approach to modelling protein structure and dynamics for better understanding their physical and structural properties as revealed in the known structural data. The data-driven and knowledge-based exploration and analysis of structural properties can take advantage of the knowledge extracted from the rich available data, and also the power of statistical methods.
1.2 Dissertation Organization

This dissertation consists of three main chapters, preceded by the present general introduction and followed by a general conclusion. Each of these main chapters corresponds to a journal article. Chapter 2 introduces the statistical measures of protein residue-level structural properties. We further develop a statistical framework for residue-level protein structural assessment, and the formulation of a novel set of residue-level statistical potentials for protein modelling and dynamics in chapter 3. To allow researchers access and manipulate a large set of statistical data on protein residue-level structural properties and evaluate statistical potentials, an open source package is developed and released in R, with a user-friendly GUI, accessible and executable in any R environment. Lastly, chapter 4 formalize a web server by integrating web pages and server-side programs in a single-destination workbench, making it easy for users to visualize the quality of a protein secondary structure.

Also, we have an additional chapter 6 on evolutionary modeling. In this chapter we analyze the mathematical properties of the nonlinear game model proposed by Gore et al 2009 and determined the values of the cost factor $c$ in the payoff functions with varying experimental parameters, the glucose and the histidine concentrations in culture, using the experimentally observed yeast fractions at equilibrium. The computed yeast fractions at equilibrium based on the game theory principles and the dynamic trajectories obtained by solving a system of replicator equations indicate that both computational results and the experimental ones reported in Gore et al 2009 agreed surprisingly well. We prove that the co-existing equilibrium state of the yeast strains is asymptotically and evolutionarily stable, which is not so obvious when the game is modeled in a nonlinear form.

1.3 Literature Review

The available structures have been popularly used to discover important functional as well as structural properties of proteins. There have been remarkable accomplishments, especially for genome comparison, segment comparison, fold recognition, 3D modelling, structural classification, or ab initio prediction in structural biology. At the same time, the opportunities
for new directions of mining the protein data bank and discovering knowledge from the rich structural data have posed challenges for scientists.

Theoretical approaches have been developed for modelling protein structure and dynamics, including potential energy minimization (3; 4), molecular dynamics simulations (5; 6), and comparative modelling (7; 8). These methods are based either on physical principles or comparisons with known protein structures, to make predictions or, in other words, approximations to the structures to be determined. They are appealing to computational scientists as well as to experimentalists and they are often employed to assist in the determination and analysis of the experimental structural models as well (9; 10; 11; 12; 13).

In comparative modelling, from sequence or structural comparisons, similar proteins or protein fragments of known structures are found for given proteins or protein fragments. Their structures are then used as templates for the proteins or protein fragments to be modelled. This approach depends on the success of finding the right structural templates, which may be achieved only if there are large numbers of known structures available. However, great progress has been made with such approaches in recent years, and it has benefited from the dramatic increases in the number of reported structures (7; 8).

Potential energy minimization tries to locate the native structure of a given protein by minimizing its potential energy, with atomic potentials that describe how its parts interact. This approach has been used for structural refinement as well as prediction (4; 11). However, the global energy minimum is extremely difficult to compute, as the number of degrees of freedoms for determining a structure is impossibly large. This approach is therefore more applicable to structural refinement, where a good initial structure is provided and a local energy minimum is all that is required (14).

A molecular dynamics simulation (MD) uses the laws of physics to obtain the dynamic behavior of a protein by solving a large system of equations of motion for the atoms in the protein. In principle, this method can be applied to track many types of protein motions from local atomic vibrations to entire protein folding, but the current simulation scheme is built upon atomic motions and cannot easily reach long time scale motions within reasonable computing time. Nonetheless, MD simulation has been used widely for analyzing various protein motions
of practical interests (5; 6; 19).

While MD simulation is costly, some dynamic properties such as structural fluctuations may be evaluated in a simpler way by using normal mode analysis (NMA). NMA is based on a quadratic approximation to the energy function of the protein at its equilibrium state. The system of equations of motion can then be solved analytically, and the atomic motions can be analyzed in terms of many different modes (20; 21). Recently, an even simpler scheme called the Gaussian Network Model (GNM) has been adopted for such calculations, and applied to large proteins and protein complexes (23; 24; 25).

There is no unique way of modelling protein structure and dynamics. The methods described above are complementary and sometimes depend on each other. The advance of one method helps and benefits from advances in the others. In practice, these methods are often combined. Yet, all these methods have limitations, and their modelling capabilities, even when they are combined, are not yet sufficient to meet the high quality and high quantity of modelling demands from applications. New approaches and breakthroughs are actively sought (26; 27; 28; 29; 30).
CHAPTER 2. Statistical Measures on Residue-Level Protein Structural Properties

Yuanyuan Huang\textsuperscript{1,4}, Steve Bonett\textsuperscript{2}, Andrzej Kloczkowski\textsuperscript{3}, Robert Jernigan\textsuperscript{1,3}, and Zhijun Wu\textsuperscript{1,4}

1. Program on Bioinformatics and Computational Biology
2. Summer REU Program on Computational Systems Biology
3. Department of Biochemistry, Biophysics, and Molecular Biology
4. Department of Mathematics

Iowa State University, Ames, IA 50014, U.S.A.

This work was published in the Journal of Structural and Functional Genomics

Abstract

The atomic-level structural properties of proteins, such as bond lengths, bond angles, and torsion angles, have been well studied and understood based on either chemistry knowledge or statistical analysis. Similar properties on the residue-level, such as the distances between two residues and the angles formed by short sequences of residues, can be equally important for structural analysis and modeling, but these have not been examined and documented on a similar scale. While these properties are difficult to measure experimentally, they can be statistically estimated in meaningful ways based on their distributions in known proteins structures. Residue-level structural properties including various types of residue distances and angles are estimated statistically. A software package is built to provide direct access to the statistical data for the properties including some important correlations not previously investigated. The distributions of residue distances and angles may vary with varying sequences, but in most
cases, are concentrated in some high probability ranges, corresponding to their frequent occurrences in either α-helices or β-sheets. Strong correlations among neighboring residue angles, similar to those between neighboring torsion angles at the atomic-level, are revealed based on their statistical measures. Residue-level statistical potentials can be defined using the statistical distributions and correlations of the residue distances and angles. Ramachandran-like plots for strongly correlated residue angles are plotted and analyzed. Their applications to structural evaluation and refinement are demonstrated. With the increase in both number and quality of known protein structures, many structural properties can be derived from sets of protein structures by statistical analyses and data mining, and these can even be used as a supplement to the experimental data for structure determinations. Indeed, the statistical measures on various types of residue distances and angles provide more systematic and quantitative assessments on these properties, which can otherwise be estimated only individually and qualitatively. Their distributions and correlations in known protein structures show their importance for providing insights into how proteins may fold naturally to various residue-level structures.

**Key Words:** Protein structural analysis ; Structural bioinformatics ; Residue-level structural properties; Statistical methods.

### 2.1 Introduction

#### 2.1.1 Atomic-level Structural Properties

As we introduced in chapter 1 the detailed atomic-level structural properties of proteins, such as bond lengths, bond angles, and torsion angles, have been well studied and understood based on either chemistry knowledge or a statistical analysis of sets of structures. For example, we have learned that the bond lengths and bond angles are relatively fixed for most types of bonds, and the torsion angles rather than being continuously populated have distinct preferences. The knowledge of these properties has been crucial for both theoretical and experimental approaches to protein modelling. The correlations among these properties have been an important source of information as well. For example, a statistical analysis showed that the $\phi$-$\psi$ torsion angles around the two bonds connecting to the $C_\alpha$ atom in the backbones of the residues
Figure 2.1: *Ramachandran Plot*. The density distribution of $\phi$-$\psi$ torsion angle pairs are plotted in a 2D plane. The correlation between the angle pairs is revealed in different density regions of the plot, with high correlations corresponding to high-density regions. The torsion angles $\phi$, $\psi$, $\omega$ for a residue are indicated in (a). A sample Ramachandran Plot is displayed in (b).

have a special correlation: When the $\phi$ angle is chosen for some value, the $\psi$ angle has only a restricted range of choice, and vice versa. The information about this correlation has been employed in both experimental determination and theoretical prediction of protein structures (22; 15). The statistical distribution of the $\phi$-$\psi$ angles in known proteins has been depicted in a two-dimensional graph called the Ramachandran Plot (2.1) named after the biophysicist G. N. Ramachandran who first did such a statistical survey of structures (15). The Ramachandran Plot has been widely used for structure evaluation. By evaluating the $\phi$-$\psi$ angles for all residues in a given protein structure and putting them onto the Ramachandran Plot, one can tell whether or not the structure is well formed based on how many of the $\phi$-$\psi$ angle pairs are in the usually dense regions of the plot.

2.1.2 Residual-level Structural Properties

Structural properties similar to those atomic-level properties described above can also be found at the residue-level such as the distances between two neighboring residues; the angles formed by three residues in sequence; and the torsion angles of four residues in sequence. Proteins are often modelled in a reduced form, with residues considered as basic units. The
residue distances and angles then become crucial for the description of the model, and they can be as important as those at the atomic-level for structural determination, prediction, and evaluation. The knowledge on these distances and angles can also be used to define residue-level potential functions so that potential energy minimization and dynamics simulation can be performed more effectively and efficiently at the residue instead of atomic-level, because the number of variables is substantially reduced and the time step may be increased (17; 19). However, the residue distances and angles have not been examined and documented in similar detail as those at the atomic-level. The reason is that they are not easy to measure directly; the physics for the interactions between residues is not as clear; and they are not as rigid as the bond lengths and bond angles, i.e., their values are likely to vary over a wide range. While residue distances and angles are difficult to measure experimentally, they can nonetheless be estimated statistically, based on their distributions in sets of known protein structures. Such approaches have been used for extracting residue contact statistics starting in early 1980s (31); for developing residue-level distance-based mean-force potentials (32); for refining X-ray crystallography determined structures (33; 34); and for deriving distance and angle constraints and potentials for NMR structure refinement (35; 36; 38; 37). Several online databases have also been built for direct access to the statistical data on various types of distances or angles (39; 40). In the present work, we download a large number of high-resolution X-ray structures from PDB Data Bank (41), and collect and analyze several important residue-level structural properties including the distances between two neighboring residues; the angles formed by
three residues in sequence; and the torsion angles of four residues in sequence. We call these, respectively, the residue-level virtual bond lengths, virtual bond angles, and virtual torsion angles. We examine the statistical distributions of these virtual bonds and virtual angles in known protein structures. In a four-residue sequence, there are two virtual bond angles and one virtual torsion angle. We name them, according to their order in the sequence, the \( \alpha \)-angle, \( \tau \)-angle, and \( \beta \)-angle, where \( \tau \) is the torsion angle 2.2. In a five-residue sequence, there are three virtual bond angles and two virtual torsion angles. We name them, according to their order in the sequence, the \( \alpha \)-angle, \( \tau_1 \)-angle, \( \beta \)-angle, \( \tau_2 \)-angle, \( \gamma \)-angle, where \( \tau_1 \) and \( \tau_2 \) are the torsion angles 2.2. For these sequences, we investigate the correlations among some of associated angles and in particular, the \( \alpha \)-\( \tau \)-\( \beta \) correlations for four-residue sequences and \( \tau_1 \)-\( \beta \)-\( \tau_2 \) correlations for five-residue sequences. We show that the distributions of the residue distances and angles may vary with varying residue sequences, but in most cases, are concentrated in some high probability ranges, corresponding to their frequent occurrences in either \( \alpha \)-helices or \( \beta \)-sheets in proteins. We show that between \( \alpha \) and \( \tau \) angles and \( \tau \) and \( \beta \) angles, there exist strong correlations, which suggests that proteins follow certain rules to form their residue-level angles as well, just like those for their atomic-level \( \phi \)-\( \psi \) angles. To the authors knowledge, these properties have not been reported and documented in their details previously. In this chapter, we define some statistical measures for the residual-level properties and present statistical data collected and analyzed on the residue distances and angles. The distributions and correlations of the any types of residue distances or angles can all be retrieved and displayed using an R package (chapter 3). We show how these outcomes can be used to define various types of residue-level statistical potentials and plot residue-level Ramachandran-like plots. We demonstrate how they can be applied for example in structural evaluation and refinement. With the increase in both the number and the quality of new protein structures, many structural properties can be derived from known protein structures by statistical analyses and data mining, and be used as a supplement to the experimental data for studies on structures. Indeed, the statistical measures on various types of residue distances and angles provide systematic and quantitative assessments of these properties, which can otherwise be estimated only individually and qualitatively. Their distributions and correlations in known
protein structures reflect important aspects of protein structures, and offer insights into how proteins fold naturally into various residue-level structures.

2.2 Computation Methods

2.2.1 Selection of Known Protein Structures

Total 1052 X-ray crystallography structures are downloaded from the PDB (41), with resolution \( \leq 1.5\AA \), sequence similarity \( \leq 30\% \), and only single chains. NMR structures are not sampled since they usually are represented as ensembles of structures, but we need single structures. We could select a representative structure from each ensemble, say the average and energy minimized one, but we decide to only use X-ray crystallography structures for this work. In future, we would like to repeat the work using NMR structures. It can be interesting to compare the outcomes with those from X-ray crystallography structures. With resolution \( \leq 1.5\AA \), the structures are guaranteed to be accurate yet abundant enough for the statistical analysis to be conducted. With sequence similarity \( \leq 30\% \), the proteins are in the so-called homological twilight zone(44), and the structures will be less likely to be repeated and hence over sampled. Multi-chain structures are often multi-copies of single chains, and are excluded as well to avoid the duplication of the data.

2.2.2 Calculation of Residue Distances and Angles

Let \((R_i, R_{i+1})\) be a sequence of two residues located at positions \(x_i\) and \(x_{i+1}\) in \(\mathbb{R}^3\). Let \(u = x_{i+1} - x_i\). Then, the residue-level 1-2-distance for this sequence is \(d_{i,i+1} = \|u\|\), where \(\|\cdot\|\) is the Euclidean norm, \(\|u\| = \sqrt{u_1^2 + u_2^2 + u_3^2}\) for any \(u = (u_1, u_2, u_3)^T\) in \(\mathbb{R}^3\). Let \((R_i, R_{i+1}, R_{i+2})\) be a sequence of three residues located at positions \(x_i, x_{i+1}, x_{i+2}\) in \(\mathbb{R}^3\). Let \(u = x_{i+1} - x_i\), \(v = x_{i+2} - x_{i+1}\). Then, the residue-level 1-3-distance for this sequence is

\[
d_{i,i+2} = \|u + v\| = \sqrt{\|u\|^2 + \|v\|^2 - 2\|u\|\|v\|\cos \alpha},
\]

where \(\alpha\) is the virtual bond angle of this sequence. Let \((R_i, R_{i+1}, R_{i+2}, R_{i+3})\) be a sequence of four residues located at positions \(x_i, x_{i+1}, x_{i+2}, x_{i+3}\) in \(\mathbb{R}^3\). Let \(u = x_{i+1} - x_i\), \(v = x_{i+2} - x_{i+1}\),
\[ w = x_{i+3} - x_{i+2}. \] Then, the residue-level 1-4-distance for this sequence is

\[
d_{i,i+3} = \| u + v + w \|
= \sqrt{\| u \| + \| v \| + \| w \| - 2 \| u \| \| v \| \cos \alpha - 2 \| v \| \| w \| \cos \beta - 2 \| u \| \| w \| \cos \theta}
\]

where \( \cos \theta = \sin \alpha \sin \beta \cos \tau - \cos \alpha \cos \beta \), and \( \alpha, \beta, \tau \) are virtual bond and torsion angles of this sequence.

### 2.2.3 Calculation of Distance and Angle Distributions

For a pair of residues, the distances between the two residues in the downloaded structures can be collected from the distance database, **D-database**, and a density distribution of the distances can be formed using the collected distances. A distance interval is divided into small bins. The density distribution of a distance in any of these bins is defined as the number of distances in that bin divided by the total number of distances in the entire distance interval.

For a three-residue sequence, the virtual bond angles formed by the same type of sequences in the downloaded structures can be collected from the angle database, **A-database**, and a density distribution of the angles can be formed using the collected angles. For a four-residue sequence, the virtual torsion angles formed by the same type of sequences in the downloaded structures can be collected from the angle database, **T-database**, and a density distribution of the angles can be formed using the collected angles. An 180° angle interval is divided into small bins for the virtual bond angles, and a 360° angle interval is divided into small bins for the virtual torsion angles. The density distribution of an angle in any of these bins is defined as the number of angles in that bin divided by the total number of angles in the entire angle interval.

### 2.2.4 Calculation of Angle-Angle Correlations

For a given sequence of four residues, the corresponding \( \alpha-\tau-\beta \) angle sequences in the downloaded structures can be collected from the **ATA-database**. The density distribution of the angle sequences on the same residue sequence can be formed using the collected angle sequences. An 180° angle interval for \( \alpha \) is divided into small bins. A 360° angle interval for \( \tau \) is divided
into small bins. An $180^\circ$ angle interval for $\beta$ is divided into small bins. Multiply these intervals to form a subspace $[0^\circ, 180^\circ] \times [0^\circ, 360^\circ] \times [0^\circ, 180^\circ]$ in $\mathbb{R}^3$, which is divided into small boxes. The density distribution of a sequence of $\alpha$-$\tau$-$\beta$ angles in any of these boxes is defined as the number of the angle sequences in that box divided by the total number of the angle sequences in the entire angle subspace.

For a sequence of five-residues, the corresponding $\alpha$-$\tau$-$\beta$-$\gamma$ angle sequences in all the downloaded structures can be collected from the TAT-database. The density distribution of $\tau_1$-$\beta$-$\tau_2$ angles can be formed using the collected angle sequences. A $360^\circ$ angle interval for $\tau_1$ is divided into small bins. An $180^\circ$ angle interval for $\beta$ is divided into small bins. A $360^\circ$ angle interval for $\tau_2$ is divided into small bins. Multiply these intervals to form a subspace $[0^\circ, 360^\circ] \times [0^\circ, 180^\circ] \times [0^\circ, 360^\circ]$ in $\mathbb{R}^3$, which is divided into small boxes. The density distribution of a sequence of $\tau_1$-$\beta$-$\tau_2$ angles in any of these boxes is defined as the number of the angle sequences in that box divided by the total number of the angle sequences in the entire angle subspace.

Two more special density distributions can be calculated: the density distributions of the $\alpha$-$\tau$ angles and the $\tau$-$\beta$ angles in the $\alpha$-$\tau$-$\beta$ angle sequences. They are simply the projections of the $\alpha$-$\tau$-$\beta$ density distribution on the $\alpha$-$\tau$ and $\tau$-$\beta$ subspaces, but they demonstrate more clear correlations of the angles. The density distribution of $\alpha$-$\beta$ angle pairs in $\alpha$-$\tau$-$\beta$ angle sequences and $\tau_1$-$\tau_2$ angle pairs in $\tau_1$-$\beta$-$\tau_2$ angle sequences have also been calculated. All these density distributions can be plotted in a 2D plane for the corresponding two angles.

### 2.2.5 Calculation of Angle-Distance Correlations

For a specific sequence of three residues, their virtual bond angles and corresponding 1-3-distances in the downloaded structures can be collected in the angle database, A-database. The correlations between the virtual bond angle and the corresponding 1-3-distance of this specific residue sequence can be demonstrated using the density distribution of the virtual bond angles vs. their corresponding 1-3-distances in an angle-distance space. An $180^\circ$ angle interval is divided into small bins for the virtual bond angles. A 20Å distance interval is divided into small bins for the corresponding 1-3-distances. Multiply the angle interval with the distance
interval to obtain a subspace $[0^\circ, 180^\circ] \times [0\text{Å}, 20\text{Å}]$ in $R^2$. The subspace is divided into small squares. The density of the angle-distance pairs in each of these squares is defined as the number of the angle-distance pairs in that square divided by the total number of angle-distance pairs in the entire angle-distance subspace.

For a specific sequence of four residues, their virtual torsion angles and corresponding 1-4-distances in all the downloaded structures can be collected from in the angle database, T-database. The correlation between the virtual torsion angle and the corresponding 1-4-distance of this specific residue sequence can be demonstrated using the density distribution of the virtual torsion angles vs. their corresponding 1-4-distances in an angle-distance space. A $360^\circ$ angle interval is divided into small bins for the virtual torsion angles. A $20\text{Å}$ distance interval is divided into small bins for the corresponding 1-4-distances. Multiply the angle interval with the distance interval to obtain a subspace $[0^\circ, 360^\circ] \times [0\text{Å}, 20\text{Å}]$ in $R^2$. The subspace is divided into small squares. The density of the angle-distance pairs in each of these squares is defined as the number of the angle-distance pairs in that square divided by the total number of angle-distance pairs in the entire angle-distance subspace.

### 2.2.6 Definition of Residue Distance and Angle Potentials

Let $P[R_i, R_{i+1}](D_{i,i+1})$ be the density distribution function for the distance $D_{i,i+1}$ between residues $R_i$ and $R_{i+1}$. A potential energy function for this distance can be defined as

$$E[R_i, R_{i+1}](D_{i,i+1}) = -kT \ln P[R_i, R_{i+1}](D_{i,i+1}).$$

Let $P[R_i, R_{i+1}, R_{i+2}](A_{i,i+1,i+2})$ be the density distribution function for the angle $A_{i,i+1,i+2}$ formed by residues $R_i$, $R_{i+1}$, $R_{i+2}$. A potential energy function for this angle can be defined as

$$E[R_i, R_{i+1}, R_{i+2}](A_{i,i+1,i+2}) = -kT \ln P[R_i, R_{i+1}, R_{i+2}](A_{i,i+1,i+2}).$$

Let $P[R_i, R_{i+1}, R_{i+2}, R_{i+3}](T_{i,i+1,i+2,i+3})$ be the density distribution function for the torsion angle $T_{i,i+1,i+2,i+3}$ formed by residues $R_i$, $R_{i+1}$, $R_{i+2}$, $R_{i+3}$. A potential energy function for
this angle can be defined as

$$E[R_{i}, R_{i+1}, R_{i+2}, R_{i+3}](T_{i,i+1,i+2,i+3}) = -kT \ln P[R_{i}, R_{i+1}, R_{i+2}, R_{i+3}](T_{i,i+1,i+2,i+3}).$$

With these potential energy functions, the distributions of the virtual bond length potentials, virtual bond angle potentials, and virtual torsion angle potentials can be calculated and displayed for a given structure over its residue sequence.

2.2.7 Display of Residue Angle Correlations of a Structure

The contours of density distributions of $\alpha$-$\tau$ and $\tau$-$\beta$ angle pairs can be plotted in 2D $\alpha$-$\tau$ and $\tau$-$\beta$ angle planes. Regions of different densities are outlined with colours in different gradients. They are defined as Most Favoured, Favoured, and Allowed, corresponding to regions of high 50%, 75%, and 90%, respectively.

The $\alpha$-$\tau$ or $\tau$-$\beta$ angle pairs for every sequence of four residues of a given structure can be computed and plotted in the $\alpha$-$\tau$ or $\tau$-$\beta$ plane, on top of the contour of the general $\alpha$-$\tau$ or $\tau$-$\beta$ density distribution function. The structure is considered to be well formed in terms of its virtual bond angels and virtual torsion angles if most of the plotted dots are in the high-density regions of the $\alpha$-$\tau$ or $\tau$-$\beta$ density distribution contour.

2.3 Results and Discussion

2.3.1 Distribution of Residue Distances

The residue-level virtual bond lengths or in other words, the residue-level, so-called 1-2-distances, for all the neighboring residue pairs in the downloaded protein structures are computed and accumulated. This data comprises the distribution of the virtual bond lengths over a range of distances (Figure 2.3). The distributions of virtual bond lengths for specific residue pairs are also calculated and can be accessed through our R-package PRESS (chapter 3). The collected virtual bond lengths range from 2.73 Å to 4.26 Å, after removing a few large outliers (Table 2.1), but the average length is 3.80 Å with a standard deviation equal to 0.05
Figure 2.3: *Distribution of virtual bond length.* The residue-level 1-2-distances are grouped into small bins. The number of distances in each distance bin is plotted. The mean value of these distances is 3.80Å with standard deviation equal to 0.05Å.
Figure 2.4: Distributions of virtual bond length in α-helices and β-sheets. The number of distances between residue pairs in α-helices or β-sheets in each distance bin is plotted. The distributions are similar to the general one in 2.3. Note that a residue pair is counted as in α-helices or β-sheets if both residues are in α-helices or β–sheets.

<table>
<thead>
<tr>
<th>R1</th>
<th># R1</th>
<th>R2</th>
<th># R2</th>
<th>Length</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLU</td>
<td>345</td>
<td>GLU</td>
<td>346</td>
<td>2.7268Å</td>
<td>2ZWS</td>
</tr>
<tr>
<td>LYS</td>
<td>116</td>
<td>PRO</td>
<td>117</td>
<td>2.7564Å</td>
<td>2SNS</td>
</tr>
<tr>
<td>GLY</td>
<td>47</td>
<td>VAL</td>
<td>48</td>
<td>2.7647Å</td>
<td>2VOV</td>
</tr>
<tr>
<td>THR</td>
<td>4</td>
<td>GLU</td>
<td>5</td>
<td>2.7738Å</td>
<td>3ELN</td>
</tr>
<tr>
<td>GLY</td>
<td>8</td>
<td>GLY</td>
<td>9</td>
<td>2.7879Å</td>
<td>3F04</td>
</tr>
<tr>
<td>GLY</td>
<td>1001</td>
<td>GLU</td>
<td>1002</td>
<td>2.7920Å</td>
<td>2O9U</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>TRP</td>
<td>1092</td>
<td>ALA</td>
<td>1093</td>
<td>4.1051Å</td>
<td>1LU4</td>
</tr>
<tr>
<td>ALA</td>
<td>649</td>
<td>ASN</td>
<td>650</td>
<td>4.1497Å</td>
<td>1N7O</td>
</tr>
<tr>
<td>THR</td>
<td>276</td>
<td>GLU</td>
<td>277</td>
<td>4.2562Å</td>
<td>1GV9</td>
</tr>
<tr>
<td>LYS</td>
<td>292</td>
<td>LYS</td>
<td>293</td>
<td>4.9025Å</td>
<td>2VK2</td>
</tr>
<tr>
<td>LYS</td>
<td>195</td>
<td>TYR</td>
<td>196</td>
<td>5.5915Å</td>
<td>2BW4</td>
</tr>
<tr>
<td>ASN</td>
<td>263</td>
<td>PRO</td>
<td>264</td>
<td>9.3897Å</td>
<td>2JLI</td>
</tr>
</tbody>
</table>

Table 2.1: Sample virtual bond lengths. The collected virtual bond lengths have been sorted. Shown here are the first and last six lengths listed (in Length), with their corresponding residue names (in R1 and R2), residue numbers (in # R1 and # R2), and protein ID (in Protein).
Figure 2.5: *Distribution of virtual bond length in short distance range.* The number of residue-level 1-2-distances in each distance bin is plotted over a short distance range. There is a small peak around 2.95Å.
Figure 2.6: *Distributions of virtual bond length in α-helices and β-sheets in short distance range.* The number of residue-level 1-2-distances in α-helices and β-sheets in each distance bin is plotted in a short distance range.

Figure 2.7: *Distributions of ω-angles for residue pairs.* The angle interval [-180°, 180°] is divided into small bins. The number of ω-angles for residue pairs with distance < 3.1Å (or > 3.1Å) in each angle bin is plotted in the whole angle range.
Figure 2.8: Distribution of virtual bond angle. The residue-level virtual bond angles are grouped into small bins. The number of angles in each angle bin is plotted. Two frequency peaks around 90° and 120° respectively can be identified.
Figure 2.9: *Distributions of virtual bond angle in α-helices and β-sheets.* The number of angles for the residue triplets in α-helices or β-sheets in each angle bin is plotted. There are only single large peaks in each case. Note that a residue triplet is counted as in α-helices or β-sheets if the first and third residues are in α-helices or β-sheets.

Å. This average length remains about the same for the residue pairs in α-helices or β-sheets, as shown in Figure 2.4, where the number of distances between residue pairs in α-helices or β-sheets in each distance bin is plotted, and the two distributions are nearly the same as the general one in Figure 2.3. There is also a large number of short distances in the region from 2.7 Å to 3.6 Å, which are invisible in Figures 2.3 and Figure 2.4 because these are still relatively few compared to those greater than 3.6 Å. However, when we look on a finer scale, we can see that they are concentrated around 2.9 Å, forming another second peak in the distribution graphs (Figures 2.5 and Figure 2.6). The reason for the two peaks is that the four atoms $C_\alpha$, C, N, $C_\alpha$ are always nearly planar but can exist in two forms: 1) the dominant trans structures with the distance between the two $C_\alpha$ atoms equal to 3.8 Å approximately and 2) a few cis structures with the $C_\alpha$-$C_\alpha$ distance around 2.9 Å. To verify this, we have calculated the $\omega$-angles of the residue pairs with distance $<3.1$ Å and $>3.1$ Å separately. The distributions of these angles (in Figure 2.7) show that the angles corresponding to residue distances $<3.1$ Å are concentrated around 0° degree, while those corresponding to residue distances $>3.1$ Å are around 180°, which justifies our conjecture. However, it seems that the number of residue pairs with short residue distances ($<3.1$ Å) is quite small in α-helices or β-sheets (Figure 2.6).
Figure 2.10: Distributions of virtual bond angle in different secondary structures. The densities of the virtual bond angles for the residue triplets in α-helices, β-sheets, and other random coils are plotted in one graph, with different colours. The density in α-helices is the highest, followed by that in β-sheets.
Figure 2.11: Correlation of residue-level 1-3-distance and virtual bond angle. The residue-level 1-3-distances and the corresponding virtual bond angles are plotted as dots in a distance-angle plane. The red dots represent the distance-angle pairs with the corresponding residue triplets having a cis structure around one of their virtual bonds.
Figure 2.12: *Sample virtual bond lengths*. The residue-level virtual torsion angles are grouped into small bins. The number of angles in each angle bin is plotted. Two peaks around 55.9° and 195.2° can be identified clearly in this plot.
fact, based on our calculated $\omega$-angle values, there seem to be only a small number of residue pairs in $\alpha$-helices or $\beta$-sheets having cis peptide bond structures, with only 5 in $\alpha$-helices and 23 in $\beta$-sheets. Note that the distributions of the virtual bond lengths for specific pairs of residues can also be calculated and analyzed in the same way as above. They can be accessed through our R-package (chapter 3). For example, among all types of residue pairs, the GLY-PRO pair has the smallest average distance, which is around 3.72 Å with standard deviation equal to 0.26 Å (because of the greater number of occurrences of cis peptide bonds). The distances for residue pairs with PRO as the second residue have a mean value smaller than 3.80 Å with standard deviations ranging from 0.15 Å to 0.30 Å, which is about 3-6 times the standard deviation of the distances between general residue pairs. Interestingly, the distances formed by the residue pairs with PRO as the first residue do not have so much variation, providing an example where the distribution of the distances for a specific residue pair is not symmetric. This points up the much greater likelihood of the cis peptide bond immediately preceding PRO residues—a well-known fact.

2.3.2 Distributions of Residue Angles

The residue-level virtual bond angles for all the three-residue sequences from the downloaded protein structures are computed and accumulated. The distribution of the virtual bond angles is shown in Figure 2.8. The distributions of virtual bond angles for specific residue triplets have also been calculated and can be accessed through our R-package PRESS (chapter 3). The collected virtual bond angles range from $51.3^\circ$ to $177.0^\circ$ as shown in Figure 2.8, and have a distinct distribution when plotted separately for $\alpha$-helices or $\beta$-sheets (Figure 2.9). The sequential triplets in $\alpha$-helices have an average virtual bond angle around $92.7^\circ$ with a standard deviation of $5.37^\circ$, while those in $\beta$-sheets are somewhat more variable (standard deviation of $12.63^\circ$) with an average angle around $123.1^\circ$. Shown in Figure 2.10 are the density plots for the virtual bond angles in $\alpha$-helices, $\beta$-sheets, random coils in one graph. We see that the plots for the angles in $\alpha$-helices and $\beta$-sheets form two large peaks. It further shows that the two large peaks in the general distribution plot in Figure 2.8 are the sum of the angles in $\alpha$-helices and $\beta$-sheets, respectively, which are due to the high frequency occurrences of the
residue sequences in these two types of secondary structures in proteins. It is interesting to note however that the angles in other cases (random coils) seem to be distributed mostly in the same two large frequency peaks as well. The residue-level 1-3-distance and the corresponding virtual bond angle for a three-residue sequence should agree with the cosine law: Let \((R_i, R_{i+1}, R_{i+2})\) be a sequence of three residues located at positions \(x_i, x_{i+1}, x_{i+2}\) in \(\mathbb{R}^3\). Let \(u = x_{i+1}x_i, v = x_{i+2}x_{i+1}\). Then, the residue-level 1-3-distance for this sequence is

\[
d_{i,i+2} = \|u + v\| = \sqrt{\|u\|^2 + \|v\|^2 - 2\|u\|\|v\| \cos \alpha}
\]

, where \(\alpha\) is the virtual bond angle of the sequence. However, if we treat the virtual bond lengths \(\|u\|\) and \(\|v\|\) as random variables, the correlation between the residue-level 1-3-distance and the corresponding virtual bond angle may not be identifiable clearly. Indeed, the correlations form two clearly separated curves as shown in Figure 2.11, i.e., with a fixed distance, there may be two values of angles, and vice versa. The reason for this is that one of the virtual bonds in the residue triplets may form cis or trans structures with distinct bond lengths, resulting in different correlations between the residue-level 1-3-distance and the virtual bond angle. Indeed, when we examine all the distance-angle pairs in Fig. 11 and put a red point for each occurrence of a cis structure, we have found these points in the upper-left strip and the rest for trans structures in the right-most strip. Interestingly enough, for the trans cases there is a greater variability generally the same angle value can be found for a wider range of distances or the same distance can be observed for a wider range of angles.

### 2.3.3 Distributions of Residue Torsion Angles

The residue-level virtual torsion angles for all four-residue sequences in the downloaded protein structures are computed and accumulated. The distribution of the virtual torsion angle over a large angle range is shown (Figure 2.12). The distributions of virtual torsion angles for specific quadruplet of residues are also calculated and can be accessed through our R-package PRESS (chapter 3). The collected virtual torsion angles have a range from 0° to 360°. The mean value is 137.6° with a standard deviation of 90.26°. Two high frequency peaks are observed: one around 55.9° corresponding to the residues in \(\alpha\)-helices, and another around
195.2° corresponding to those in β-sheets. The peak for the virtual torsion angles formed by residues in α-helices exhibit less variability. If the distributions of virtual torsion angles are plotted for their occurrences in α-helices and β-sheets separately, there is then a single peak in each graph (Figure 2.13). Each of these peaks corresponds to one of the two peaks seen in Figure 2.12, suggesting that the two peaks in Figure 2.12 are basically high frequency virtual torsion angles for the α-helices and β-sheets, respectively. The first peak occurs around 55.9°, which implies that the virtual torsion angles in α-helices are on average around 55.9°. The second peak is near 195.2°, meaning that the virtual torsion angles in β-sheets are around 195.2°. Figure 2.14 shows the density plots for the virtual torsion angles in α-helices, β-sheets, or other random coils in one graph. We see that the plots for the angles in α-helices and β-sheets form two large peaks. It further shows that the two large peaks in the general distribution plot in Figure 2.12 should mainly be formed by the angles in α-helices and β-sheets, respectively, which correspond to the high frequency occurrences of the residue sequences in these two types of secondary structures in proteins. It is interesting to note though that the angles in other random coils seem to be distributed mostly in the two large frequency peaks as well, but unlike the virtual bond angles there are significant numbers of cases at intermediate values and even outside the α-helix and β-sheet peaks. Let \((R_i, R_{i+1}, R_{i+2}, R_{i+3})\) be a sequence of four residues located at positions \(x_i, x_{i+1}, x_{i+2}, x_{i+3}\) in \(\mathbb{R}^3\). Let \(u = x_{i+1} - x_i, v = x_{i+2} - x_{i+1}, w = x_{i+3} - x_{i+2}\). Then, the residue-level 1-4-distance for this sequence is

\[d_{i,i+3} = \|u + v + w\| = \sqrt{\|u\| + \|v\| + \|w\| - 2\|u\|\|v\|\cos \alpha - 2\|v\|\|w\|\cos \beta - 2\|u\|\|w\|\cos \theta}\]

where \(\cos \theta = \sin \alpha \sin \beta \cos \tau - \cos \alpha \cos \beta\), and \(\alpha, \beta, \tau\) are virtual bond and torsion angles of this sequence. However, since the virtual bond lengths \(\|u\|, \|v\|, \|w\|\) and bond angles \(\alpha, \beta\) are all random variables now, the correlation between the residue-level 1-4-distance and the corresponding virtual torsion angle may not be so clearly identifiable. Indeed, as shown in Figure 2.15, one virtual torsion angle may correspond to multiple residue-level 1-4-distances, and vice versa. However, from 0° to 180°, the angle-distance pairs tend to concentrate from lower left to upper right, roughly forming a positive correlation between the residue 1-4-distances and their
corresponding virtual torsion angles. From 180° to 360°, the pairs concentrate from upper left
to lower right, forming a decreasing strip of dots. In either case, the residue 1-4-distances seems
to be roughly correlated with their virtual torsion angles as shown by the above formula.

2.3.4 Angles-angle Correlations

The residue-level angle-angle correlations or, in other words, the correlations among the
residue-level virtual bond angles and torsion angles in the downloaded protein structures are
computed and documented. We exhibit, for sequences of angles, \( \alpha-\tau-\beta \), the correlations between
\( \alpha-\tau, \tau-\beta, \) and \( \alpha-\beta \) angle pairs and for sequences, \( \alpha-\tau_1-\beta-\tau_2-\gamma \), the correlations between \( \tau_1-\tau_2 \)
angle pairs in Figure 2.16-2.19. We form two large data sets, one containing all \( \alpha-\tau-\beta \) angle
sequences and another containing all \( \alpha-\tau_1-\beta-\tau_2-\gamma \) angle sequences. Each data set has 229,812
angle sequences. These are used to generate the density distributions of the angle-pairs (\( \alpha-\tau, \tau-\beta, \alpha-\beta \) and \( \tau_1-\tau_2 \)). The background plots in Figure 2.16-2.19 are the contours of the density
distributions of the angle pairs. The scattered dots correspond to the angle pairs found in
10 sampled protein structures. The background contours are plotted in different gradients,
with a darker gradient representing higher density regions. From high to low density, there are
50%, 75%, and 90% density regions, named Most Favoured, Favoured, and Allowed regions,
respectively. The plots show that there are distinct density distribution regions for \( \alpha-\tau \) and \( \tau-\beta \)
angle pairs. That means that these angle pairs are highly dependent of each other in well-formed
protein structures or in other words, if one angle in \( \alpha-\tau \) or \( \tau-\beta \) angle pair is fixed, the choice
for the other is restricted. These correlations are certainly important structural properties of
proteins, but have not been investigated thoroughly. However, for \( \alpha-\beta \) and \( \tau_1-\tau_2 \) angle pairs,
the correlations are not so strong: One angle does not impose as strong a restriction on the
possible choices for the other. The scattered dots for the angle pairs from 10 sampled structures
further confirm the correlations of the angle pairs reflected in the general estimations, i.e., the
distributions of the angle pairs in these structures (represented by the dots) agree with those
in all the downloaded structures (represented by the background contours). We have also used
two differently coloured dots for the angle pairs in \( \alpha \)-helices and \( \beta \)-sheets, respectively. Then,
the differently coloured dots are distributed in different high-density regions. The following are
<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Allowed region</th>
<th>Favoured region</th>
<th>Most Favoured region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1TJY</td>
<td>91.37%</td>
<td>75.72%</td>
<td>55.59%</td>
</tr>
<tr>
<td>1UJP</td>
<td>93.51%</td>
<td>80.09%</td>
<td>63.20%</td>
</tr>
<tr>
<td>1WCK</td>
<td>83.46%</td>
<td>65.41%</td>
<td>29.32%</td>
</tr>
<tr>
<td>2BOG</td>
<td>86.50%</td>
<td>72.99%</td>
<td>52.19%</td>
</tr>
<tr>
<td>2DSX</td>
<td>95.92%</td>
<td>73.47%</td>
<td>38.78%</td>
</tr>
<tr>
<td>2E3H</td>
<td>80.00%</td>
<td>62.67%</td>
<td>22.67%</td>
</tr>
<tr>
<td>2FG1</td>
<td>82.22%</td>
<td>68.15%</td>
<td>48.89%</td>
</tr>
<tr>
<td>2O8L</td>
<td>84.51%</td>
<td>68.15%</td>
<td>33.33%</td>
</tr>
<tr>
<td>2P4F</td>
<td>94.59%</td>
<td>87.57%</td>
<td>60.00%</td>
</tr>
<tr>
<td>3IIS</td>
<td>96.62%</td>
<td>89.86%</td>
<td>77.70%</td>
</tr>
</tbody>
</table>

Table 2.2: *Density distribution of α-τ angle pairs.* The table gives the percentiles of α-τ angle pairs in the allowed, favoured, and most favoured regions of the general α-τ density distribution contour for 10 sampled protein structures.

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Allowed region</th>
<th>Favoured region</th>
<th>Most Favoured region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1TJY</td>
<td>91.37%</td>
<td>77.96%</td>
<td>56.55%</td>
</tr>
<tr>
<td>1UJP</td>
<td>93.51%</td>
<td>85.71%</td>
<td>66.23%</td>
</tr>
<tr>
<td>1WCK</td>
<td>83.46%</td>
<td>70.68%</td>
<td>36.09%</td>
</tr>
<tr>
<td>2BOG</td>
<td>87.23%</td>
<td>73.72%</td>
<td>52.55%</td>
</tr>
<tr>
<td>2DSX</td>
<td>97.96%</td>
<td>81.63%</td>
<td>40.82%</td>
</tr>
<tr>
<td>2E3H</td>
<td>77.33%</td>
<td>56.00%</td>
<td>30.67%</td>
</tr>
<tr>
<td>2FG1</td>
<td>90.37%</td>
<td>73.33%</td>
<td>51.85%</td>
</tr>
<tr>
<td>2O8L</td>
<td>84.98%</td>
<td>64.32%</td>
<td>36.62%</td>
</tr>
<tr>
<td>2P4F</td>
<td>94.59%</td>
<td>84.32%</td>
<td>61.62%</td>
</tr>
<tr>
<td>3IIS</td>
<td>96.62%</td>
<td>85.14%</td>
<td>77.03%</td>
</tr>
</tbody>
</table>

Table 2.3: *Density distribution of τ-β angle pairs.* The table gives the percentiles of τ-β angle pairs in the allowed, favoured, and most favoured regions of the general τ-β density distribution contour for 10 sampled protein structures.

more specific explanations of the plots.

The contour of the density distribution of α-τ angle pairs is plotted in Figure 2.16. The contour is displayed in three types of regions named as Most Favoured, Favoured, and Allowed, each containing high 50%, 75%, 90% of all collected α-τ angle pairs. In addition, the α-τ angle pairs sampled from 10 arbitrarily selected proteins are overlaid as points over the contour of the general α-τ density distribution. The red triangles represent the α-τ angle-pairs in α-helices, the blue squares in β-sheets, and the black dots in other type of secondary structures. In total 1756 scattered points for the sampled structures (PDB ID: 1TJY, 1UJP, 1WCK, 2BOG,
2DSX, 2E3H, 2FG1, 2O8L, 2P4F, 3IIS) are shown, of which 1566 points (∼ 89.2%) are in Allowed regions, 1319 points (∼ 75.1%) in Favoured regions, and 901 points (∼ 51.31%) in Most Favoured regions. Table 2.2 shows the percentile of α-τ angle pairs in Allowed, Favoured, and Most Favoured regions for each of the 10 sampled protein structures. Most of these proteins, especially the proteins with more helical structures such as 3IIS, have high percentages of points in these regions, but 2E3H, with no helical structures, is an exception. The density distribution contour of τ-β angle-pairs is plotted in Figure 2.17. The contour is displayed in three types of regions named as Most Favoured, Favoured, and Allowed regions, each containing high 50%, 75%, and 90% of all collected τ-β angle pairs. In addition, the τ-β angle pairs sampled from the same 10 arbitrarily selected proteins are plotted as points on top of the contour of the general τ-β density distribution. The red triangles represent the τ-β angle pairs in α-helices, the blue squares in β-sheets, and the black dots in other type of secondary structures. In total 1756 scattered points from the same sampled structures are shown, of which 1567 points (∼ 89.8%) are in Allowed regions, 1338 points (∼ 76.2%) in Favoured regions, and 941 points (∼ 53.59%) in Most Favoured regions. Table 2.3 gives the percentile of τ-β angle pairs in Allowed, Favoured, and Most Favoured regions for each of 10 sampled protein structures. Most of these proteins, especially the proteins with more helical structures, have high percentages of points in these regions, while proteins 2E3H and 1WCK with more sheets have lower percentiles of points in these regions than average. The density distribution contour of α-β angle-pairs is plotted in Figure 2.18, with regions also named as Most favoured, Favoured, Allowed, each corresponding to high 50%, 75%, 90% of all collected α-β angle pairs. In addition, the α-β angle pairs sampled from the 10 arbitrarily selected proteins are plotted as points on top of the contour of the general α-β density distribution. The red triangles represent the α-β angle-pairs in α-helices, the green squares in β-sheets, and the black dots in other type of structures. In total 1756 scattered points for the same sampled structures are shown, of which 1681 points (∼ 95.7%) are in Allowed regions, 1608 points (∼ 91.6%) in Favoured regions, and 645 points (∼ 36.73%) in Most Favoured regions. The density distribution contour of τ₁-τ₂ angle pairs is plotted in Figure 2.19, with regions again named as Most favoured, Favoured, Allowed, each corresponding to high 50%, 75%, 90% of all collected τ₁-τ₂ angle pairs. In addition, the τ₁-τ₂
Figure 2.13: Distributions of virtual torsion angle in α-helices and β-sheets. The number of angles for the residue quadruplets in α-helices or β-sheets in each angle bin is plotted. There are only single large peaks in both cases. Note that a residue quadruplet is counted as in α-helices or β-sheets if the first and third residues are in α-helices or β-sheets.

angle pairs sampled from the same 10 selected proteins are plotted as points on top of the contour of the general τ₁-τ₂ density distribution. The red triangles represent the τ₁-τ₂ angle pairs in α-helices, the blue squares in β-sheets, and the black dots in other type of secondary structures. In total there are 1737 scattered points for the sampled structures, of which 1551 points (≈ 89.3%) are in Allowed regions, 1309 points (≈ 75.4%) in Favoured regions, and 902 points (≈ 51.93%) in Most Favoured.

2.3.5 Applications to Structural Analysis

The distributions of residue-level virtual bonds, bond angles, and torsion angles can be used to define residue-level statistical potentials. The energy distributions over sequences of virtual bond lengths, virtual bond angles, or virtual torsion angles for a given protein structure can then be evaluated with corresponding potential energy functions. If the potential energy is high for a virtual bond length (or a virtual bond angle or a virtual torsion angle), it implies that the virtual bond length (or the virtual bond angle or the virtual torsion angle) is not energetically favourable. The density distributions of α-τ and τ-β angle pairs show strong correlations. Therefore, a given structure can also be evaluated by comparing its α-τ and τ-β
Figure 2.14: Distributions of virtual torsion angle in different secondary structures. The densities of the virtual torsion angles for the residue quadruplets in α-helices, β-sheets, and other random coils are plotted in one graph, with different colours. The density in α-helices is the highest, followed by that for β-sheets.
Figure 2.15: Correlation of residue-level 1-4-distance and virtual torsion angle. The residue-level 1-4-distances and the corresponding virtual torsion angles are plotted as dots in a distance-angle plane. The blue dots represent the distance-angle pairs with the corresponding residue quadruplets in $\beta$-sheets while the red dots in $\alpha$-helices.
Figure 2.16: Contour of density distribution of $\alpha$-$\tau$ angles pairs. The contour of the density distribution of the $\alpha$-$\tau$ angle-pairs is plotted. The plot is divided into three regions named as most favoured, favoured, and allowed, each containing high 50%, 75%, and 90% of all $\alpha$-$\tau$ angle pairs. In addition, the $\alpha$-$\tau$ angle pairs sampled from 10 arbitrarily selected proteins are plotted as points. The red triangles represent the $\alpha$-$\tau$ angle-pairs in $\alpha$-helices, the blue squares in $\beta$-sheets, and the black dots in other type of secondary structures.
Figure 2.17: Contour of density distribution of $\tau$-$\beta$ angles pairs. The contour of the density distribution of the $\tau$-$\beta$ angle-pairs is plotted. The plot is divided into three regions named as most favoured, favoured, and allowed, each containing 50%, 75%, and 90% of all $\tau$-$\beta$ angle-pairs. In addition, the $\tau$-$\beta$ angle pairs sampled from 10 arbitrarily selected proteins are also plotted as points on top of the contour of the general $\tau$-$\beta$ density distribution. The red triangles represent the $\tau$-$\beta$ angle pairs in $\alpha$-helices, the blue squares in $\beta$-sheets, and the black dots in other type of secondary structures.
Figure 2.18: Contour of density distribution of $\alpha$-$\beta$ angle pairs. The contour of the density distribution of the $\alpha$-$\beta$ angle pairs is plotted. The plot is divided into three regions named as most favoured, favoured, and allowed, each containing high 50%, 75%, and 90% of all $\alpha$-$\beta$ angle pairs. In addition, the $\alpha$-$\beta$ angle pairs sampled from 10 arbitrarily selected proteins are plotted as dots over the contour of the general $\alpha$-$\beta$ density distribution. The red triangles represent the $\alpha$-$\beta$ angle pairs in $\alpha$-helices, the blue squares in $\beta$-sheets, and the black dots in other type of secondary structures.
Figure 2.19: *Contour of density distribution of $\tau_1\text{-}\tau_2$ angle pairs.* The contour of the density distribution of the $\tau_1\text{-}\tau_2$ angle pairs is plotted. The plot is divided into three regions named as most favoured, favoured, and allowed, each containing high 50%, 75%, 90% of all the $\tau_1\text{-}\tau_2$ angle pairs. In addition, the $\tau_1\text{-}\tau_2$ angle pairs sampled from 10 arbitrarily selected proteins are plotted as dots overlaid on the contour of the general $\tau_1\text{-}\tau_2$ density distributions. The red triangles represent the $\tau_1\text{-}\tau_2$ angle-pairs in $\alpha$-helices, the blue squares in $\beta$-sheets, and the black dots in other type of secondary structures.
angle pairs with their general distributions. A 2D plot can be obtained by positioning these angle pairs, as points, in the corresponding 2D contours of their general density distribution functions. The 2D contours have three density regions, called Most Favoured, Favoured, and Allowed, corresponding to high 50%, 75%, and 90% of all the angle pairs in the surveyed structures. A structure is considered to be well-formed in terms of its $\alpha$-$\tau$ (or $\tau$-$\beta$) angle pairs if the percentiles of the $\alpha$-$\tau$ (or $\tau$-$\beta$) angle pairs falling in the corresponding density regions are close to their general distributions. Chapter 3 and chapter 4 will show how the statistical potentials and the angle correlation plots can be used for structural analysis, for example, how they can be used effectively for distinguishing a well-resolved structure from a poorly determined one. In these figures, we see that the energy distributions along the residue sequences for two structures of different resolutions are clearly different. The potential energies for the better resolved structure are lower in average. The distributions of the $\alpha$-$\tau$ (or $\tau$-$\beta$) angle pairs of the structures show even greater contrast. The better resolved structure has many more angle pairs distributed in the high density regions of the corresponding angle distribution contour, while the poorly resolved structures have many angle pairs falling outside of these high density regions.

2.4 Conclusion

The statistical distributions of residue-level distances and angles in known protein structures provide a valuable source of information for estimating these residue-level structural properties of proteins, which are not otherwise accessible experimentally. However, these statistical measures rely upon the quality as well as quantity of the sampled known structures. We have downloaded around one thousand high-quality structures from the Protein Data Bank, which should be sufficient to obtain reliable statistical estimates of the distributions of virtual bond lengths, virtual bond angles, virtual torsion angles, and some of their correlations, but of course there is the possibility that for some cases of specific residue sequences, the values might deviate from the overall characteristic distributions. In chapter 3, we provided information about the size of the data set for each estimate. One of the most important results from this study shows that residue-level angles, especially the neighboring virtual bond angles and virtual torsion an-
gles, exhibit strong correlations. For example, if a virtual bond angle is fixed, then the choice for the virtual torsion angle adjacent to it will be highly restricted. Such a correlation is similar to the correlation shown in Ramachandran Plots between the backbone atomic $\phi$-$\psi$ torsion angles, and can be equally important for understanding the structural properties of proteins at their residue-levels and even for evaluating the quality of individual structures. However, as we have shown, the correlations between two virtual bond angles, when separated by one virtual torsion angle, are not so strong, and so are the correlations between two virtual torsion angles, when separated by one virtual bond angle. The reason may be due to the fact that at residue-level, these angles are relatively further apart and therefore behave more independently of one another. In addition, where the angles are closely correlated, they tend to have smaller deviations when they are in $\alpha$-helices than in $\beta$-sheets. This implies that even at the residue-level, $\alpha$-helices are more rigid (or stable) than $\beta$-sheets. We have demonstrated that the statistical distributions of the residue-level distances or angles can be used to define various statistical potentials, but further refinements are required to make them computationally and physically meaningful. For example, the potentials for virtual bond lengths of different pairs of residues, or for virtual bond angles of different triplets of residues, or for virtual torsion angles of different sequence quadruplets of residues, need to be scaled appropriately before they can be combined. These are only potentials for short-range interactions. In order to define a relatively complete energy function for a protein, potentials for long-range interactions also must be included. The useful tool from this study is a residue-level Ramachandran-type of plot for correlations between pairs of neighboring virtual bond angles and virtual torsion angles. Several examples have been given in the present paper, but these differ from the atomic-level Ramachandran Plot in an important way, because the density distribution contours of these residue-level angles show relatively larger deviations. Thus their use requires specifying more precisely what density regions should be permitted for high-quality structures. Further evaluations are needed to decide generally what these evaluation criteria should be. The atomic-level structural properties of proteins, such as bond lengths, bond angles, and torsion angles, have been thoroughly studied and understood based on either chemistry knowledge or statistical analysis. Similar properties at the residue-level, such as the distances between two residues
and the angles formed by short sequences of residues, can be equally important for structural analysis and modelling, but these have not previously been examined and tabulated as carefully and thoroughly. While these properties are difficult to measure experimentally, they can be estimated statistically based on observed distributions in known proteins structures, as demonstrated in this paper. In this paper, we have conducted a statistical analysis of protein residue-level local structural properties. A software package was built, and this provides direct access to the statistical data for these properties including correlations among them. The distributions of residue distances and angles may vary residue sequence, but in most cases, these are concentrated in some high probability ranges, which correspond to their frequent occurrences in either $\alpha$-helices or $\beta$-sheets in proteins. Strong correlations among neighbouring residue angles, similar to those between neighbouring torsion angles at the atomic-level, are revealed based on statistical measures. Residue-level statistical potentials can be defined using the statistical distributions and correlations of the residue distances and angles. Ramachandran-like plots for strongly correlated residue angles are plotted and analyzed. Their applications for structural evaluation and refinement are demonstrated. With the increase in both the number and quality of determined protein structures, many structural properties can be derived from known protein structures with statistical analysis and data mining, and then be used as a supplement to experimental data for refining or evaluating structures. Indeed, the statistical measures of various types of residue distances and angles afford systematic and quantitative assessments on these properties. Their distributions and correlations in known protein structures inform us about the important limitations of conformations of proteins and can even offer some insights into how proteins fold or change their conformations.
CHAPTER 3.  P.R.E.S.S. – An R-package for Exploring Residual-level Protein Structural Statistics

Yuanyuan Huang\textsuperscript{1,4}, Steve Bonett\textsuperscript{2}, Andrzej Kloczkowski\textsuperscript{3},
Robert Jernigan\textsuperscript{1,3}, and Zhijun Wu\textsuperscript{1,4}

1. Program on Bioinformatics and Computational Biology
2. Summer REU Program on Computational Systems Biology
3. Department of Biochemistry, Biophysics, and Molecular Biology
4. Department of Mathematics

Iowa State University, Ames, IA 50014, U.S.A.

This work was published in the \textit{Journal of Bioinformatics and Computational Biology}

Abstract

P.R.E.S.S. is an R package developed to allow researchers to get access to and manipulate on a large set of statistical data on protein residue-level structural properties such as residue-level virtual bond lengths, virtual bond angles, and virtual torsion angles. A large set of high-resolution protein structures are downloaded and surveyed. Their residue-level structural properties are calculated and documented. The statistical distributions and correlations of these properties can be queried and displayed. Tools are also provided for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, new tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement. P.R.E.S.S. has been released in R as an open source software package, with a user-friendly GUI, accessible and executable by a public user in any R environment. P.R.E.S.S. can also be downloaded directly

Key Words: Protein structure analysis; protein residual-level structural properties; structural bioinformatics; statistical potentials; structural correlation plots; R-package.

3.1 Introduction

In this chapter, we describe a related piece of work on developing a software package called P.R.E.S.S. for direct access to the statistical data on the residue-level structural properties we have collected and analyzed. The software is developed in R (42; 43) and has been released as an open source package, with a user-friendly GUI, accessible and executable by a public user in any R environment. With this software, the distributions and correlations of given types of residue distances or angles can all be retrieved and displayed. Tools are also provided in P.R.E.S.S. for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement. We describe the organization of the software, the data source, the computational methods, and all the functional modules. We provide examples to demonstrate the use of the software.

Note that in chapter 2, we focused on the statistics of the residue-level structural properties. The PRESS software was mentioned with a list of functions, but it was not fully tested and demonstrated. In this paper, we devote solely to the description of PRESS as a publically available software tool. In Section 3.2, we describe the system organization, and provide a full account on the databases in the system, the computational units, and the graphics interface with their functional modules and uses. We also present some new test results on more than 1,000 obsolete low resolution structures in PDB, showing how our residue angle correlation plots can be applied to distinguishing them from their currently replaced high resolution ones. We summarize this chapter and make some concluding remarks in Section 3.4.
3.2 Graphics Interface and System Organization

P.R.E.S.S. can be divided into two ends, front end and back end. The back end includes the parts for downloading structural data, calculating residue distances and angles, and saving the distances and angles. The front end is responsible for providing all data retrieving and analysis functions using the distance and angle data calculated and saved in the back end. In the back end, there are three major components: 1). Download the structural data from PDB, which can be done automatically or manually, but currently only semi-automatically. 2). Compute and collect residue-level distances and angles. 3). Save the distances and angles into five databases. They are database for virtual bond lengths, database for virtual bond angles, database for virtual torsion angles, database for virtual angle sequences of four-residue sequences, and database for virtual angle sequences of five-residue sequences, named B, A, T, ATA, and TAT databases, respectively. More specific information on the content of each of these databases is given below.

**B-database:** Stores the virtual bond lengths for all the neighboring pairs of residues for each downloaded structure. Each record in the database contains the following information:

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Residue 1</th>
<th>Residue 2</th>
<th>1-2- distance</th>
</tr>
</thead>
</table>

**A-database:** Stores the virtual bond angles formed by all the connected triplet of residues of each downloaded structure. All residue-level 1-3-distances are also saved. Each record in the database contains the following information:

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Residue 1</th>
<th>Residue 2</th>
<th>Residue 3</th>
<th>Bond Angle</th>
<th>1-3- distance</th>
</tr>
</thead>
</table>

**T-database:** Stores the virtual torsion angles (τ) formed by all the connected quadruplets of residues of each downloaded structure. All residue-level 1-4-distances are also saved. Each record in the database contains the following information:

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Residue 1</th>
<th>Residue 2</th>
<th>Residue 3</th>
<th>Residue 4</th>
<th>τ</th>
<th>1-4 distance</th>
</tr>
</thead>
</table>

**ATA-database:** Stores the α-τ-β angle sequences for all the four-residue sequences in each downloaded structure. Each record in the database contains the following information:
### TAT-database:
Stores the $\alpha\tau_2\beta\tau_2\gamma$ angle sequences for all the five-residue sequences in each downloaded structure. Each record in the database contains the following information.

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Residue 1</th>
<th>Residue 2</th>
<th>Residue 3</th>
<th>Residue 4</th>
<th>$\alpha$</th>
<th>$\tau$</th>
<th>$\beta$</th>
</tr>
</thead>
</table>

In the R package there are two major components: the GUI (graphics user interface) and the computational unit. The GUI takes a query from the user and passes it to the computational unit. The computational unit has a collection of routines, responsible for various computational tasks. It retrieves the data from the databases in the back end, performs certain calculations, and returns the results to the GUI. The interface then displays the results.

More specifically, the GUI shows a window of six functional panels (Figure 3.1), each accepting a specific type of queries: 1). Queries on virtual bond lengths for two residues. 2). Queries on virtual bond angles for three residues. 3). Queries on virtual torsion angles and ATA correlations ($\alpha\tau\beta$ angle sequences) for four residues. 4. Queries on TAT correlations ($\alpha\tau_2\beta\tau_2\gamma$ angle sequences) for five residues. 5. Structural analysis and evaluation. 6. Help information.

The computational unit has mainly the following routines:

- Calculate the distribution of the virtual bond length ($B$) between a given pair of residues.
- Calculate the distribution of the virtual bond angle ($A$) for a given sequence of three residues.
- Calculate the distribution of the residue-level 1-3-distance for a given sequence of three residues.
- Calculate the distribution of the virtual torsion angle ($T$) for a given sequence of four residues.
- Calculate the distribution of the residue-level 1-4-distance for a given sequence of four residues.
Figure 3.1: **P.R.E.S.S. graphics interface.** PRESS has a graphics interface with six functional panels corresponding to six functional routines, each providing a specific structural computing or analysis function.

- Calculate the correlation between the residue-level 1-3-distance and the virtual bond angle for a given sequence of three residues.

- Routine for calculating the correlation between the residue-level 1-4-distance and the virtual torsion angle for a given sequence of four residues.

- Calculate the correlations of the $\alpha$-$\tau$-$\beta$ ($\text{ATA}$) angle sequence on a given sequence of four residues.

- Calculate the correlation of the $\alpha$-$\tau_1$-$\beta$-$\tau_2$-$\gamma$ ($\text{TAT}$) angle sequence on a given sequence of five residues.

- Evaluate the statistical potentials for the virtual bond lengths or the virtual bond angles for a given protein structure.

- Evaluate the $\alpha$-$\tau$ ($\text{AT}$) and $\tau$-$\beta$ ($\text{TB}$) angle pairs for a given protein structure and display them in $\alpha$-$\tau$ ($\text{AT}$) and $\tau$-$\beta$ ($\text{TB}$) density distribution contour plots.

The overall system organization of P.R.E.S.S. is shown in Figure 3.2. The data source and computational methods used in P.R.E.S.S. are discussed in chapter 2.
Figure 3.2: System organization and interface. The system has a GUI, which takes queries on distributions or correlations of residue-level distances or angles and passes them to the computational unit. The computational unit retrieves the data from the distance or angle databases documented from the structural database, and computes the requested distributions or correlations. The results are returned to the interface and displayed.

### 3.3 Functional Modules

One of the functions of P.R.E.S.S. is to retrieve the virtual bond lengths for a given pair of residues and find the distribution of the particular bond length over a certain distance range. The found distribution can be displayed in a graph as shown in Figure 3.3. The residue pair to be searched for can be specified from a pull-down menu. Each residue can be a specific or any type. For the latter, any type is considered for that residue. The bin size of the distribution graph can be adjusted. The graph can be displayed to show either the frequency or density of the bond lengths.

One of the functions of P.R.E.S.S. is to retrieve the virtual bond angles for a given sequence of three residues and find the distribution of the particular bond angle over a certain angle range. The found distribution can be displayed in a graph as shown in Figure 3.4. The residue triplet to be searched for can be specified from a pull-down menu. Each residue can be a specific or any type. For the latter, any type is considered for that residue. The bin size of the distribution graph can be adjusted. The graph can be displayed to show either the frequency or density of the bond angles.

When the distribution of a virtual bond angle is queried, an option is available for display-
Figure 3.3: Distribution of virtual bond lengths between ASN and MET. This snapshot shows the distribution graph for the virtual bond lengths between ASN and MET. The users can not only move the slider to adjust the bin size of the histogram, but also switch between frequency and density displays.

Figure 3.4: Distribution of virtual bond angles formed by Any, ASN, and Any. This snapshot shows the distribution graph for the virtual bond angles formed by a residue sequence Any, ASN, and Any.
The virtual torsion angles for a given sequence of four residues can be retrieved. The distribution of the particular torsion angle can be displayed over a certain angle range. The residue quadruplet to be searched for can be specified from a pull-down menu. Each residue can be a specific or any type. For the latter, any type is considered for that residue. The bin size of the distribution graph can be adjusted. The graph can be displayed to show either the frequency or density of the torsion angles as shown in Figure 3.6. The correlation between the virtual torsion angle and the corresponding residue 1-4 distance for a given sequence of four residues can also be displayed.

The angle sequence $\alpha-\tau_1-\beta-\tau_2-\gamma$ for a given sequence of five residues can be retrieved. The density distributions of the virtual bond angle $\beta$ and its neighbouring two virtual torsion angles can be displayed. The density distribution of $\tau_1-\beta-\tau_2$ can also be displayed as a 3D plot (Figure 3.7) in the $\tau_1-\beta-\tau_2$ space.

One of important functions of P.R.E.S.S. is to evaluate the statistical potentials on the virtual bonds or virtual bond angles for a given structure. The potentials are defined in terms of the statistical distributions of the virtual bond lengths and virtual bond angles. The virtual bond length potential can be evaluated for every neighboring pair of residues of the given structure. Therefore, the distribution of the potential energy along the residue sequence of the structure can be obtained and displayed to show how flexible the virtual bonds are along the
Figure 3.6: A matrix of scattered plots of the distributions and correlations of the virtual torsion angle and its two neighboring virtual bond angles. The plots show the distribution and correlation graphs for the virtual torsion angle and its two neighboring virtual bond angles for residues LYS, Any, ASN, Any. The matrix of plots is 3 by 3. The graph in each is defined as follows: Square(1,1) = distribution of virtual bond angle 1; Square(2,2) = distribution of the virtual torsion angle; Square(3,3) = distribution of virtual bond angle 2; Square(1,2) = correlation between virtual bond angle 1 and the virtual torsion angle; Square(1,3) = correlation between the bond angle 1 and virtual bond angle 2; Square(2,3) = correlation between the virtual torsion angle and virtual bond angle 2.

Figure 3.7: 3D scattered plot for the virtual torsion angle and its two neighboring virtual bond angles. This snapshot shows the density distribution of the $\alpha$-$\tau$-$\beta$ angle triplets for residue sequence LYS, Any, ASN, and Any in the $\alpha$-$\tau$-$\beta$ space. A lowess approximation to the distribution is also plotted.
sequence. The higher the potential energy is for a specific bond, the lower the probability of the bond length is in the distribution of the bond length in known proteins, and hence the more deviated it must be from its average value the bond length (Figure 3.8).

The virtual bond angle potential can be evaluated for every sequence of three residues of the given structure as well. The distribution of the potential energy along the residue sequence of the structure can also be obtained and displayed to show how flexible the virtual bond angles are along the sequence. It has the same property as that for the bond length energy for structural evaluation (Figure 3.9).

The statistical potentials for residue distances and angles are short-range potentials, but they may be combined with some long-range statistical potentials such as residue contact potentials(31) to define a residue-level, coarse-grained statistical potential energy function. Such a function may prove to be valuable for coarse-grained protein modeling including potential energy minimization, molecular dynamics simulation, and ab-initio structural prediction(5; 6; 46; 17; 15). We restrict our work in this paper on structural analysis, but will pursue other possible extensions in our future efforts.

One of the most important functions of P.R.E.S.S. is that it can evaluate the correlations
Figure 3.9: *Distribution of virtual bond angle energies for 1PHY (2.4 Å) and 2PHY (1.4 Å).* The energy levels of the virtual bond angles of two structures 1PHY and 2PHY are plotted in solid lines. The minimal possible energies are shown as the dashed line. These two structures are determined with different resolutions for the same protein. The better-resolved structure (2PHY) has lower potential energies in average than the poorly determined one (1PHY).

of the virtual bond and torsion angles and display a residue-level Ramachandran-like plot for a given structure. Two of the angle-angle correlation plots are proven to be especially valuable. One is the $\alpha$-$\tau$ correlation plot or the AT-plot for short. Another one is the $\tau$-$\beta$ correlation plot or the TB-plot for short. Given a protein structure, the $\alpha$-$\tau$ or $\tau$-$\beta$ angle pairs can be computed along the residue sequence for the structure. Each angle pair can be plotted as a dot in the $\alpha$-$\tau$ or $\tau$-$\beta$ space. The distribution of the dots over the contour of the general $\alpha$-$\tau$ or $\tau$-$\beta$ density distribution can then be evaluated to show how the angle pairs in the given structure correlated against to their average correlations in known proteins. These plots can be used effectively to differentiate high-quality structures from low-quality structures at the residue level as the Ramachandran plots for structural evaluation at the atomic level, as shown in Figure 3.10 and Figure 3.11.

Note that the above residue-level angle-angle correlation plots differ from the atomic-level Ramachandran Plot(47) in an important way, because the density distribution contours of these residue-level angles show relatively larger deviations. Therefore, their evaluations on the quality of the structures may not be as strict as the atomic level Ramachandran Plot. However,
Figure 3.10: The $\alpha$-$\tau$ correlation plots for a protein at two different resolutions. The X-ray structure of the periplasmic galactose binding protein from salmonella typhimurium, 1GBP (3.0Å), shown in (a) compared with the later refined structure, 3GBP (2.4 Å), shown in (b). The background contours are generated from the general density distributions of the $\alpha$-$\tau$ angle pairs in known proteins. Regions of different densities are outlined with colours in different gradients. They are defined as Most Favoured (high 50% density), Favoured (high 75% density), and Allowed (high 90% density) regions. The scattered triangles correspond to the $\alpha$-$\tau$ angle pairs in the given protein structures. The lines in (a) indicate that there are 59.86% of the triangles of the $\alpha$-$\tau$ angles pairs in 1GBP falling in the 90% region, 46.37% of triangles falling in the 75% region, and only 31.83% of the triangles falling in the 50% region. On the other hand, In (b), there are 93.44% of the triangles of the $\alpha$-$\tau$ angles pairs in 3GBP falling in the 95.03% region, 76.23% of triangles falling in the 81.13% region, and 61.59% of the triangles falling in the 50% region.

they do show how neighboring residue angles should be correlated in general, which cannot be detected by the atomic-level Ramachandran Plot. In this respect, the residue-level angle-angle correlation plots can be used as a complementary tool to the atomic-level Ramachandran Plot for both residue as well as atomic level structural evaluation.

We have applied the $\alpha$-$\tau$ or $\tau$-$\beta$ angle correlation plots to evaluating a large set of obsoleted structures in PDB and compare the plots with those for the current superseded structures. As of Feb. 8, 2012, there are total 1,654 obsoleted protein structures superseded by their succesors according to a report from PDB (ftp://ftp.wwpdb.org/pub/pdb/data/status/obsolete.dat). For each pair of obsoleted and replaced structures, we checked the percentage of the angle pairs in the Allowed, Favoured, and Most Favoured regions of the plots, and examined the average percentages for all the structural pairs. The results are summarized in Table 3.1 and Table 3.2, for the structural pairs with RMSD values in between 0 and 1Å, 1 and 3Å, 3 and 5Å,
Figure 3.11: The $\tau$-$\beta$ correlation plots for a protein at different resolution. The X-ray structure of the periplasmic galactose binding protein from salmonella typhimurium, 1GBP (3.0Å), shown in (a) compared with the later refined structure, 3GBP (2.4Å), shown in (b). The background contours are generated from the general density distributions of the $\tau$-$\beta$ angle pairs in known proteins. Regions of different densities are outlined with colours in different gradients. They are defined as Most Favoured (high 50% density), Favoured (high 75% density), and Allowed (high 90% density) regions. The scattered triangles correspond to the $\tau$-$\beta$ angle pairs in the given protein structures. The lines in (a) indicate that there are 63.32% of the triangles of the $\tau$-$\beta$ angle pairs in 1GBP falling in the 90% region, 45.33% of the triangles falling in the 75% region, and only 27.34% of the triangles falling in the 50% region. On the other hand, in (b), there are 92.38% of the triangles of the $\tau$-$\beta$ angle pairs in 3GBP falling in the 90% region, 82.78% of the triangles falling in the 75% region, and 60.60% of the triangles falling in the 50% region.
Table 3.1: Structure analysis for previously obsoleted structures and current ones: compare the average percentages of angle pairs in different density regions of α-τ correlation plots. The size column gives the number of paired structures – previously obsoleted and current ones. The other columns show the average percentages of angle pairs that fall into the allowed region (90%), the favoured region (75%), and the most favoured region (50%) for the previously obsoleted structures and the corresponding current ones in different groups of paired structures with different RMSD values.

<table>
<thead>
<tr>
<th>RMSD</th>
<th>size</th>
<th>pre90%</th>
<th>cur90%</th>
<th>pre75%</th>
<th>cur75%</th>
<th>pre50%</th>
<th>cur50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>922</td>
<td>86.41</td>
<td>88.42</td>
<td>71.42</td>
<td>73.75</td>
<td>47.19</td>
<td>49.38</td>
</tr>
<tr>
<td>(0, 1A)</td>
<td>542</td>
<td>89.56</td>
<td>89.69</td>
<td>75.36</td>
<td>75.51</td>
<td>51.00</td>
<td>51.16</td>
</tr>
<tr>
<td>(1A, 3A)</td>
<td>37</td>
<td>83.04</td>
<td>86.39</td>
<td>68.66</td>
<td>72.11</td>
<td>44.06</td>
<td>48.52</td>
</tr>
<tr>
<td>(3A, 5A)</td>
<td>17</td>
<td>86.42</td>
<td>86.95</td>
<td>69.26</td>
<td>71.20</td>
<td>43.47</td>
<td>43.23</td>
</tr>
<tr>
<td>(5A, ∞)</td>
<td>136</td>
<td>84.09</td>
<td>86.31</td>
<td>67.69</td>
<td>69.84</td>
<td>42.62</td>
<td>45.03</td>
</tr>
</tbody>
</table>

and beyond 5Å. The structural pairs for which the RMSD values cannot be computed due to various reasons are considered as a separate group (RMSD: NA).

As we can see in Table 3.1 and Table 3.2, the angle plots for the current structures, all having higher resolutions than their previous ones, have higher percentages of angle pairs in in their high-density regions. They indicate clearly that the angle plots can be used as powerful tools to distinguish low quality structures from high quality ones in terms of their resolutions(22). We have also performed a t-test for the paired data. The p-values of the one-tailed student t-test supported our statement that the current structures have indeed more points in the high density regions of the angle correlation plots than the previosly obsoleted structures. We have also observed that for the structural pairs with RMSD values in between 1 and 3Å, the differences in the angle correlation plots between the superseded and replaced ones are the most notable (Figure 3.12 and Figure 3.13).

3.4 Conclusion

In this chapter, we have reported our recent work for the development of an R package, called P.R.E.S.S., which allows researchers to get access to and manipulate on a large set of statistical data on protein residue-level structural properties such as residue-level virtual bond lengths, virtual bond angles, and virtual torsion angles. We have downloaded and surveyed a large set of high-resolution protein structures, and calculated and documented an important set of their
Table 3.2: Structure analysis for previously obsoleted structures and current ones: compare the average percentages of angle pairs in different density regions of $\tau$-$\beta$ correlation plots. The size column gives the number of paired structures – previously obsoleted and current ones. The other columns show the average percentages of angle pairs that fall into the allowed region (90%), the favoured region (75%), and the most favoured region (50%) for the previously obsoleted structures and the corresponding current ones in different groups of paired structures with different RMSD values.

<table>
<thead>
<tr>
<th>RMSD</th>
<th>size</th>
<th>pre90%</th>
<th>cur90%</th>
<th>pre75%</th>
<th>cur75%</th>
<th>pre50%</th>
<th>cur50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>922</td>
<td>86.44</td>
<td>88.27</td>
<td>71.83</td>
<td>74.01</td>
<td>48.03</td>
<td>50.22</td>
</tr>
<tr>
<td>(0,1A)</td>
<td>542</td>
<td>89.60</td>
<td>89.76</td>
<td>75.35</td>
<td>75.57</td>
<td>51.86</td>
<td>52.00</td>
</tr>
<tr>
<td>(1A, 3A)</td>
<td>37</td>
<td>82.95</td>
<td>86.14</td>
<td>68.65</td>
<td>72.68</td>
<td>44.42</td>
<td>49.53</td>
</tr>
<tr>
<td>(3A, 5A)</td>
<td>17</td>
<td>85.54</td>
<td>86.15</td>
<td>68.81</td>
<td>71.12</td>
<td>45.12</td>
<td>46.08</td>
</tr>
<tr>
<td>(5A, ∞ )</td>
<td>136</td>
<td>83.86</td>
<td>84.94</td>
<td>67.60</td>
<td>69.70</td>
<td>43.22</td>
<td>45.72</td>
</tr>
</tbody>
</table>

Figure 3.12: The boxplots of percentages of points in $\alpha$-$\tau$ correlation regions for obsoleted structures and their successors (1Å < RMSD < 3Å). For the structural pairs with RMSD values less than 3Å but bigger than 1Å, the average differences between the structural pairs can be observed from their differences in median and mean.
Figure 3.13: The boxplots of percentages of points in $\tau$-$\beta$ correlation regions for obsoleted structures and their successors ($1 \text{Å} < \text{RMSD} < 3 \text{Å}$). For the structural pairs with RMSD values less than 3Å but bigger than 1Å, the average differences between the structural pairs can be observed from their differences in median and mean.

Residue-level structural properties in P.R.E.S.S. With P.R.E.S.S., the statistical distributions and correlations of these properties can be queried and displayed. Tools are also provided for modeling and analyzing a given structure in terms of its residue-level structural properties. In particular, new tools for computing residue-level statistical potentials and displaying residue-level Ramachandran-like plots are developed for structural analysis and refinement.

We have discussed the principle for the development of P.R.E.S.S. for statistical analysis on protein structures. We have described the system organization and interface of the software, and provided detailed information on how the structural data was collected and documented in P.R.E.S.S., and how all the statistical results were calculated. We have described the major computational and analysis functions of P.R.E.S.S. and demonstrated them in many examples. P.R.E.S.S. has been released in R as an open source software package, with a user-friendly GUI, accessible and executable by a public user in any R environment.

The statistical distributions of residue-level distances and angles in known protein structures provide a valuable source of information for estimating these residue-level structural properties of proteins, which are not otherwise accessible experimentally(35; 44; 45; 46). However, these statistical measures rely upon the quality as well as quantity of the sampled known structures.
We have downloaded around one thousand high-quality structures (resolution higher than 1.5 Å) from the PDB, which should be sufficient to obtain reliable statistical estimates of the distributions of virtual bond lengths, virtual bond angles, virtual torsion angles, and some of their correlations, but of course there is the possibility that for some cases of specific residue sequences, the values might deviate from the overall characteristic distributions. In P.R.E.S.S., we have provided information about the size of the data set for each estimate.

The useful tool from this study is a residue-level Ramamchandran-type of plot for correlations between pairs of neighboring virtual bond angles and virtual torsion angles. Several examples have been given in the present paper, but these differ from the atomic-level Ramachandran Plot in an important way, because the density distribution contours of these residue-level angles show relatively larger deviations. Thus their use requires specifying more precisely what density regions should be permitted for high-quality structures. Further evaluations are needed to decide generally what these evaluation criteria should be.

Finally, the major contributions of this work include two parts: (1) the statistical data on residue-level distances and angles; these structural properties have not been investigated and documented before, but have been made accessible through PRESS, which is certainly valuable for increasing our knowledge on protein residue level structural properties and for assisting the scientists in their structural modeling practices. (2) the statistical tools for structural evaluation including the statistical potentials on residue-level distances and angles and the residue-level angle-angle correlation plots; the usefulness of these two sets of tools are still subject to further testing, improvement, and validation.
CHAPTER 4. PRESS-PLOT: An Online Server for Protein Structural Analysis and Evaluation using Residue-level Virtual Angle Correlation Plots

Yuanyuan Huang\textsuperscript{1,3}, Kejue Jia\textsuperscript{1}, Robert Jernigan\textsuperscript{1,2}, and Zhijun Wu\textsuperscript{1,3}

1. Program on Bioinformatics and Computational Biology
2. Department of Biochemistry, Biophysics, and Molecular Biology
3. Department of Mathematics

Iowa State University, Ames, IA 50014, U.S.A.

This work was submitted to \textit{BMC Bioinformatics}

Abstract

Predicting the 3D structure of a protein given the protein’s residue sequence is one of the “Holy Grails” of structural bioinformatics. Developing efficient tools for protein structural analysis is a necessary step towards achieving this goal. A statistical structural analysis package called PRESS has been developed and released in R. In this package, the distributions and correlations of various types of protein residue distances and angles in known protein structures are calculated and documented. A set of residue-level structural analysis tools are also developed and made available, among which, a set of so-called residue-level virtual angle correlation plots are proven to be particularly valuable for analyzing and evaluating either experimentally determined or theoretically predicted structures. Based on the data and tools in PRESS, especially the residue-level virtual angle correlation plots, a web-server called PRESS-PLOT is further developed for easy access and display of the plots for structural analysis and evaluation. A structure to be analyzed and evaluated can be submitted to the server by either giving its structural ID in PDB or uploading its structural file in the PDB format. The residue-level
virtual bond angles and torsion angles of the structure are then computed. The neighboring virtual bond angle and torsion angle pairs are displayed as scattered points in a 2D graph and compared against the 2D contour map of the density distribution of such angle pairs in known protein structures, as given in the background of the 2D graph. The virtual angle pairs that can be analyzed and evaluated include \( \alpha - \tau \) and \( \tau - \beta \) angle pairs as they appear in either general structures or specific secondary structures such as \( \alpha \)-helices, \( \beta \)-sheets, or their turns. As a verification of PRESS-PLOT, more than 1000 obsoleted structures (with lower resolutions) in PDB are evaluated using PRESS-PLOT and compared with their current superseded versions (with higher resolutions). The results show that PRESS-PLOT clearly distinguishes high-quality structures (the current ones) from low-quality structures (the obsolete ones) in its angle correlation plots.

**Key Words:** Protein structure analysis; protein residual-level structural properties; structural bioinformatics; structural correlation plots; web server.

### 4.1 Introduction

With the enormous number of protein structures already determined and deposited in PDB, statistical data mining becomes feasible and potentially fruitful for structural bioinformatics: The statistical properties of many structural properties can now be derived from the database of known protein structures. The distributions or correlations of these properties in the structures can be computed and then used for subsequent structural inferences. They provide a wealth of information for recovering general structural properties beyond individual experimental outcomes. They can be used to develop computational tools for structural analysis as well as structural determination including structural assessment, refinement, and prediction (30).

The atomic-level structural properties of proteins, such as the backbone torsion angles \( \phi \), \( \psi \), and \( \omega \), which are among the main determinants of a protein fold, have been well studied and understood based on either chemistry knowledge or statistical analysis. For example, it is well known that the allowed range of \( \omega \) angle is very restrictive, while \( \phi \) and \( \psi \) angles are closely correlated to each other. The latter is a key indicator for the correct fold of a structure,
and is often demonstrated via a so-called Ramachandran plot, a 2D contour map of the density
distribution of the $\phi$-$\psi$ angle pairs in known protein structures. The Ramachandran Plot has
been widely adopted for structural analysis and evaluation, with its 2D contour map used as a
reference for the correct formation of the $\phi$-$\psi$ angle pairs in the structure (47; 22).

Structural properties similar to those at the atomic level can also be found at the residue
level such as the distances between two neighbouring residues (called virtual bonds); the angles
formed by three residues in sequence (called virtual bond angles); and the torsion angles of
four residues in sequence (called virtual torsion angles) (Figure 2.2 (a)). They can be as
important as those at the atomic level for structural analysis and evaluation, especially when
reduced models for proteins are considered where residues are the basic units (18). Due to the
difficulty of measuring the residue distances and angles, either experimentally or theoretically,
a statistical approach to the study of these properties becomes crucial and necessary. Much
work has been done along this line in the past(31; 32; 35; 37; 39). In particular, in chapter
2 we conducted a detailed survey of residue-level protein structural properties using a large
set of known protein structures in PDB. An R package called PRESS (Protein REsidual-level
Structural Statistics) provides access to the structural properties and to the structural analysis
tools developed in chapter chapter 3. Among the analysis tools developed is a set of so-called
residue-level virtual angle correlation plots, similar to Ramachadran Plots for atomic-level angle
correlations. These residue-level angle correlation plots contain 2D contour maps of density
distributions of certain virtual bond angle and torsion angle pairs in the surveyed structures.
They can be used to analyze and evaluate candidate protein structures, either experimentally
determined or theoretically predicted, by comparing the bond angles and torsion angles of the
test structure with the 2D contour maps computed from the known structures. These angle
correlation plots provide a unique and valuable set of tools for residue-level structural analysis
and assessment, and are expected to have a useful impact in current protein modeling practices.

This chapter develops a web-server called PRESS-PLOT for easy access and display of the
virtual angle correlation plots in PRESS, especially for online access via WWW (World Wide
Web). A structure to be analyzed and evaluated can be submitted to the server by either giving
its structural ID in PDB or uploading its structural file in the PDB format. The residue-level
virtual bond angles and torsion angles of the structure are then computed. The neighboring virtual bond angle and torsion angle pairs are displayed as scattered points in a 2D graph and compared against the 2D contour map of the density distribution of such angle pairs in known protein structures. The virtual angle pairs that can be analyzed and evaluated include $\alpha$-$\tau$ and $\tau$-$\beta$ angle pairs as they appear in either general structures or specific secondary structures such as $\alpha$-helices, $\beta$-sheets, or their turns. As a verification of PRESS-PLOT, more than 1000 obsolete structures (with lower resolutions) in PDB are evaluated and compared with their current superseded versions (with higher resolutions). The results show that PRESS-PLOT clearly distinguishes high-quality structures (the current ones) from low-quality structures (the obsoleted ones) in its angle correlation plots.

4.2 Implementation

PRESS-PLOT is derived from PRESS structural data and functions for structural analysis and evaluation using residue-level virtual angle correlation plots. Different from PRESS, PRESS-PLOT is focused on structural assessment. It has a web interface for online access. It also evaluates the virtual angle correlations for specific as well as general secondary structures. The development of PRESS-PLOT is motivated by the successful application of residue-level virtual angle correlation plots to structural assessment and justified by extensive testing on a large set of current vs. obsolete structures in PDB.

4.2.1 Structural Data

As in PRESS, a total of 1052 X-ray crystallography structures are downloaded from PDB, with resolution $\leq$ 1.5Å, sequence similarity $\leq$ 30%, and only single chains. The angle sequences $\alpha$-$\tau$-$\beta$ for all four residue sequences in the structures are calculated and stored in a database named **ATA-database**. Each record in the database contains the following information:

| ID | $R_1$, $S_1$ | $R_2$, $S_2$ | $R_3$, $S_3$ | $R_4$, $S_4$ | $\alpha$ | $\tau$ | $\beta$ | SS |

where ID is the structural ID in PDB, $R_j$ is the type of the residue in the sequence, $S_j$ is the secondary structure type of $R_j$, $\alpha$, $\tau$, $\beta$ are the corresponding virtual bond and torsion angles,
and SS is the type of the secondary structure of the whole residue sequence. The last item is
determined by the following rules: A four residue sequence R$_1$-R$_2$-R$_3$-R$_4$ is considered to be in

- α-helix: if $R_1$, $R_2$, $R_3$, $R_4$ are in α-helix
- head of α-helix: if $R_2$, $R_3$, $R_4$ are in α-helix
- tail of α-helix: if $R_1$, $R_2$, $R_3$ are in α-helix
- β-sheet: if $R_1$, $R_2$, $R_3$, $R_4$ are in β-sheet
- head of β-sheet: if $R_2$, $R_3$, $R_4$ are in β-sheet
- tail of β-sheet: if $R_1$, $R_2$, $R_3$ are in β-sheet

where the secondary structure type of each residue is identified by using the program DSSP (16).

With the identification of the secondary structure type, PRESS-PLOT is capable of evaluating
the virtual angle correlations when they are in specific types of secondary structures, while
PRESS evaluates the correlations without specifying the secondary structure types of the angle
pairs.

### 4.2.2 Plot of Density Maps

The 2D countour maps of the density distributions of α-τ and τ-β angle pairs are ploted
in 2D α-τ and τ-β planes, respectively. The maps are displayed in a special graphical form
similar to that for the Ramachadran Plots. Each map has three different density regions, with
high 50%, 75%, and 90% of density, called most favored, favored, and allowed regions,
and plotted in dark, less dark, and light colors, respectively. The region with lower 10%
density is called disallowed region and colored in white (Figure 3.10 3.11 ). The maps for the
distributions in certain secondary structure conditions are plotted similarly, with the density
percentages adjusted slightly for those different density regions.

### 4.2.3 Web Interface

PRESS-PLOT is a web-based integrated online service dedicated to protein structural as-
assessment. It helps the user visualize the quality of a given structure in terms of its residue-level
Figure 4.1: PRESS-PLOT organization. PRESS-PLOT can be broken down to two parts: The front end and the back end. The front end takes user’s input structure and passes it to the query handling unit of the back end. The latter carries out preprocessing and directs the structure to the computing unit of the back end for required calculation and plot generation. The query handling unit takes the final results from the computing unit and renders them to the front end for display.
PRESS-PLOT can be broken down into two major components: front-end dynamic web pages and back-end computing components (Figure 4.1). The front-end web pages are designed in MVC (Model, View, and Control) pattern, which provides a high refactoring ability and is also simple for maintenance. The result generated by PRESS-PLOT is graphical data. It is important that any result is presented to the user immediately. For a faster query response, AJAX (Asynchronous JavaScript and XML) technique is adopted on the web pages. It allows the web pages to update the result without refreshing all the elements on the pages. The back-end computing components are composed of two sub-units: the query handling unit and the computing unit. The query handling unit is responsible for pre-processing and transferring user queries to the computing unit. After the results are generated, it also renders the results and outputs plots onto the web pages. The query handling unit is implemented in PHP, one of the most popular and widely supported scripting languages. The computing unit implements the core computing functions. It accepts the query information from the query handling unit, computes the virtual angle data for the input structure, and generates the final graphical results. It is implemented in R, an open source environment for statistical computing.

The 2D countour maps of the density distributions of $\alpha$-$\tau$ and $\tau$-$\beta$ angle pairs are plotted in 2D $\alpha$-$\tau$ and $\tau$-$\beta$ planes, respectively. The maps are displayed in a special graphical form similar to that for the Ramachadran Plots. Each map has three different density regions, with high 50%, 75%, and 90% of density, called most favored, favored, and allowed regions, and plotted in dark, less dark, and light colors, respectively. The region with the lower 10% of density is called the disallowed region and colored in white (Figure 3.10 3.11). The maps
for the distributions in certain secondary structure conditions are plotted similarly, with the
density percentages adjusted slightly for those different density regions.

4.3 Results

PRESS-PLOT is developed to provide an online server for structural assessment using
the PRESS virtual angle correlation plots. In addition, it further extends the PRESS angle
correlation plots to angle pairs in specific secondary structures, which can be more accurate for
specific structural types and practical for more detailed structural analysis. PRESS-PLOT is
tested on a large set of structures in PDB, showing that higher-resolution structures in general
have better evaluations in PRESS-PLOT angle correlation plots.

4.3.1 Display functions

A structure to be evaluated can be submitted to PRESS-PLOT by either providing the
PDB ID of the structure or uploading the structural file in the PDB format. The structure is
then evaluated for their $\alpha$-$\tau$ and $\tau$-$\beta$ angle correlations. A total of seven groups of evaluation
results, in both graphics and text forms, are generated:

1. general $\alpha$-$\tau$ plot
2. general $\tau$-$\beta$ plot
3. $\alpha$-$\tau$ plot for angle pairs in $\alpha$-helices
4. $\tau$-$\beta$ plot for angle pairs in $\alpha$-helices
5. $\alpha$-$\tau$ plot for angle pairs in $\beta$-sheets
6. $\tau$-$\beta$ plot for angle pairs in $\beta$-sheets
7. $\alpha$-$\tau$ plot for angle pairs in heads of $\alpha$-helices
8. $\tau$-$\beta$ plot for angle pairs in heads of $\alpha$-helices
9. $\alpha$-$\tau$ plot for angle pairs in heads of $\beta$-sheets
10. $\tau$-$\beta$ plot for angle pairs in heads of $\beta$-sheets
11. $\alpha$-$\tau$ plot for angle pairs in tails of $\alpha$-helices
12. $\tau$-$\beta$ plot for angle pairs in tails of $\alpha$-helices
The front page is designed so that the user can either type in the structural ID in PDB or upload the structural file in PDB format. In the former case, the system will search the PDB site for the structural file. The output correlation plots are displayed as thumbnails in the bottom of the front page. Each can be expanded in the main window by moving cursor over it.

13. $\alpha$-$\tau$ plot for angle pairs in tails of $\beta$-sheets
14. $\tau$-$\beta$ plot for angle pairs in tails of $\beta$-sheets

The first group of results is displayed in the window as default. The remaining groups are listed as small icons in the bottom of the window and can be selected to be drawn in the window. Each plot shows the selected type of angle pairs in the given structure as scattered points in the corresponding density map. The percentages of the points in different density regions are summarized in the graph. Figure 4.2 shows the front page of PRESS-PLOT when a general $\alpha$-$\tau$ coorelation plot is displayed. Figure 3.10 3.11 shows the general $\alpha$-$\tau$ and $\tau$-$\beta$ correlation plots for structures 1GBP and 3GBP. Figure 4.3 and 4.4 show the $\alpha$-$\tau$ correlation plots for a structure 3PTE in two specific secondary structure types: one in $\alpha$ helices and another in $\beta$ sheets.

In the first group of plots, all $\alpha$-$\tau$ ($\tau$-$\beta$) angle pairs of the given structure are calculated and plotted as scattered points in the $\alpha$-$\tau$ ($\tau$-$\beta$) plane. The background of the $\alpha$-$\tau$ ($\tau$-$\beta$) plane
Figure 4.3: A virtual angle correlation plot for α-helices. All α-τ angle pairs in the α-helices of structure 3PTE are displayed as scattered points and compared with the background contour map of the density distribution of α-τ angle pairs in the α-helices of all surveyed structures.

is the contour map of the density distribution of the α-τ (τ-β) angle pairs in general structures that include all types of secondary structures. If the percentages of the α-τ (τ-β) angle pairs of the given structure in *most favored*, *favored*, and *allowed* regions are near or above 50%, 75%, and 90%, respectively, the structure is considered to be well formed in terms of α-τ (τ-β) angle correlations.

In the second group of plots, all α-τ (τ-β) angle pairs in α-helices of the given structure are calculated and plotted as scatter points in the α-τ (τ-β) plane. The background of the α-τ (τ-β) plane is the contour map of the density distribution of the α-τ (τ-β) angle pairs in α-helices. Likewise, in the third group of plots, all α-τ (τ-β) angle pairs in β-sheets of the given structure are calculated and plotted as scattered points in the α-τ (τ-β) plane. The background of the α-τ (τ-β) plane is the contour map of the density distribution of the α-τ (τ-β) angle pairs in β-sheets. The remaining groups of plots are generated similarly for α-τ (τ-β) angle pairs in heads or tails of α-helices or β-sheets.
Figure 4.4: A virtual angle correlation plot for $\beta$-sheets. All $\alpha$-$\tau$ angle pairs in the $\beta$-sheets of structure 3PTE are displayed as scattered points and compared with the background contour map of the density distribution of $\alpha$-$\tau$ angle pairs in the $\beta$-sheets of all surveyed structures.

4.3.2 Testings

PRESS-PLOT is applied to evaluate a large set of obsolete structures in PDB. The results are compared with those for the current superseded structures. Up to early 2012, there are a total of 705 obsolete protein structures superseded by their successors with the same sequence length according to a report from PDB (48). For each pair of obsolete and replaced structures, the percentages of the virtual angle pairs in most favored, favored, and allowed regions of the virtual angle correlation plots are examined. The average percentages for the structural pairs with RMSD values between 0 and 1Å, 1 and 3Å, 3 and 5Å, and beyond 5Å are calculated and summarized in Table 4.1. The structures are grouped according to the RMSD values of the structural pairs. For each group of structures, the average percentages of their $\alpha$-$\tau$ angle pairs and $\beta$-$\tau$ angle pairs in different density regions in angle correlation plots are summarized.

Table legends:

RMSD – RMSD range for obsoleted and superceded structural pairs; size – # of structural
pairs with given RMSD range; obs#AT – average percentage of $\alpha - \tau$ angle pairs of obsoleted structures in high #\% region; sup#AT – average percentage of $\alpha - \tau$ angle pairs of superceded structures in high #\% region; obs#BT – average percentage of $\beta - \tau$ angle pairs of obsoleted structures in high #\% region; sup#BT – average percentage of $\beta - \tau$ angle pairs of obsoleted structures in high #\% region; obs#helixAT – average percentage of helical $\alpha - \tau$ angle pairs of obsoleted structures in high #\% region; sup#helixAT – average percentage of helical $\alpha - \tau$ angle pairs of superceded structures in high #\% region; obs#helixBT – average percentage of helical $\beta - \tau$ angle pairs of obsoleted structures in high #\% region; sup#helixBT – average percentage of helical $\beta - \tau$ angle pairs of superceded structures in high #\% region; obs#sheetAT – average percentage of $\alpha - \tau$ angle pairs in $\beta$ sheets of obsoleted structures in high #\% region; sup#sheetAT – average percentage of $\alpha - \tau$ angle pairs in $\beta$ sheets of superceded structures in high #\% region; obs#sheetBT – average percentage of $\beta - \tau$ angle pairs in $\beta$ sheets of obsoleted structures in high #\% region; sup#sheetBT – average percentage of $\beta - \tau$ angle pairs in $\beta$ sheets of superceded structures in high #\% region.

As can be seen in Table 4.1, the current structures with higher resolutions than their previous ones all have higher percentages of virtual angle pairs in high-density regions of the virtual angle correlation plots, showing that PRESS-PLOT can distinguish low quality structures from high quality ones very well. In particular, for the structure pairs with RMSD values in between 1 and 3Å, the differences in the virtual angle correlation plots between the superseded and replaced ones are the most notable. A simple explanation for this is that if two structures are very similar (with RMSD <1Å), their virtual angle correlations are certainly expected to be about the same, and therefore, their PRESS-PLOT evaluations would be similar. On the other hand, if two structures are very different (with RMSD >3Å), they may differ in their tertiary structures but still have similar secondary structures and hence similar local structures. The latter would keep the virtual angle correlations of the two structures similar.

4.4 Discussion

Atomic-level structural analysis tools such as the Ramachandran plot assessment refer to pdb explicitly have been used successfully for protein structural analysis and evaluation.
<table>
<thead>
<tr>
<th>RMSD</th>
<th>[0,1)</th>
<th>[1,3)</th>
<th>[3,5)</th>
<th>[5,Inf)</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>527</td>
<td>32</td>
<td>16</td>
<td>130</td>
<td>705</td>
</tr>
<tr>
<td>obs90AT</td>
<td>89.57</td>
<td>83.04</td>
<td>86.42</td>
<td>84.10</td>
<td>88.19</td>
</tr>
<tr>
<td>sup90AT</td>
<td>89.70</td>
<td>86.39</td>
<td>86.96</td>
<td>86.32</td>
<td>88.86</td>
</tr>
<tr>
<td>obs75AT</td>
<td>75.36</td>
<td>68.67</td>
<td>69.27</td>
<td>67.70</td>
<td>73.51</td>
</tr>
<tr>
<td>sup75AT</td>
<td>75.51</td>
<td>72.12</td>
<td>71.20</td>
<td>69.85</td>
<td>74.22</td>
</tr>
<tr>
<td>obs50AT</td>
<td>51.00</td>
<td>44.07</td>
<td>43.48</td>
<td>42.63</td>
<td>48.97</td>
</tr>
<tr>
<td>sup50AT</td>
<td>51.17</td>
<td>48.52</td>
<td>43.24</td>
<td>45.04</td>
<td>49.74</td>
</tr>
<tr>
<td>obs90BT</td>
<td>89.60</td>
<td>82.96</td>
<td>85.55</td>
<td>83.87</td>
<td>88.15</td>
</tr>
<tr>
<td>sup90BT</td>
<td>89.77</td>
<td>86.15</td>
<td>86.16</td>
<td>84.95</td>
<td>88.63</td>
</tr>
<tr>
<td>obs75BT</td>
<td>75.36</td>
<td>68.65</td>
<td>68.81</td>
<td>67.61</td>
<td>73.48</td>
</tr>
<tr>
<td>sup75BT</td>
<td>75.58</td>
<td>72.68</td>
<td>71.13</td>
<td>69.70</td>
<td>74.26</td>
</tr>
<tr>
<td>obs50BT</td>
<td>51.86</td>
<td>44.42</td>
<td>45.12</td>
<td>43.23</td>
<td>49.78</td>
</tr>
<tr>
<td>sup50BT</td>
<td>52.01</td>
<td>49.53</td>
<td>46.09</td>
<td>45.72</td>
<td>50.60</td>
</tr>
<tr>
<td>obs99helixAT</td>
<td>88.85</td>
<td>78.73</td>
<td>78.47</td>
<td>77.91</td>
<td>86.14</td>
</tr>
<tr>
<td>sup99helixAT</td>
<td>88.94</td>
<td>81.55</td>
<td>78.82</td>
<td>80.76</td>
<td>86.87</td>
</tr>
<tr>
<td>obs95helixAT</td>
<td>82.38</td>
<td>67.35</td>
<td>74.44</td>
<td>68.61</td>
<td>78.98</td>
</tr>
<tr>
<td>sup95helixAT</td>
<td>82.60</td>
<td>74.98</td>
<td>76.04</td>
<td>73.45</td>
<td>80.42</td>
</tr>
<tr>
<td>obs80helixAT</td>
<td>62.05</td>
<td>36.74</td>
<td>46.00</td>
<td>44.33</td>
<td>57.27</td>
</tr>
<tr>
<td>sup80helixAT</td>
<td>62.98</td>
<td>50.06</td>
<td>46.44</td>
<td>49.60</td>
<td>59.55</td>
</tr>
<tr>
<td>obs99helixBT</td>
<td>88.46</td>
<td>77.09</td>
<td>79.51</td>
<td>76.07</td>
<td>85.46</td>
</tr>
<tr>
<td>sup99helixBT</td>
<td>88.69</td>
<td>80.47</td>
<td>78.12</td>
<td>79.28</td>
<td>86.34</td>
</tr>
<tr>
<td>obs95helixBT</td>
<td>82.02</td>
<td>63.51</td>
<td>68.26</td>
<td>67.19</td>
<td>78.14</td>
</tr>
<tr>
<td>sup95helixBT</td>
<td>82.15</td>
<td>73.49</td>
<td>70.16</td>
<td>70.24</td>
<td>79.29</td>
</tr>
<tr>
<td>obs80helixBT</td>
<td>61.57</td>
<td>33.58</td>
<td>42.96</td>
<td>43.93</td>
<td>56.62</td>
</tr>
<tr>
<td>sup80helixBT</td>
<td>62.12</td>
<td>49.52</td>
<td>44.32</td>
<td>48.30</td>
<td>58.59</td>
</tr>
<tr>
<td>obs95sheetAT</td>
<td>80.63</td>
<td>49.98</td>
<td>56.08</td>
<td>65.48</td>
<td>75.89</td>
</tr>
<tr>
<td>sup95sheetAT</td>
<td>80.78</td>
<td>47.87</td>
<td>72.27</td>
<td>65.42</td>
<td>76.26</td>
</tr>
<tr>
<td>obs85sheetAT</td>
<td>73.15</td>
<td>41.81</td>
<td>50.23</td>
<td>57.34</td>
<td>68.29</td>
</tr>
<tr>
<td>sup85sheetAT</td>
<td>73.32</td>
<td>41.63</td>
<td>66.77</td>
<td>57.30</td>
<td>68.78</td>
</tr>
<tr>
<td>obs65sheetAT</td>
<td>56.11</td>
<td>30.59</td>
<td>40.39</td>
<td>43.05</td>
<td>52.18</td>
</tr>
<tr>
<td>sup65sheetAT</td>
<td>56.27</td>
<td>30.50</td>
<td>48.79</td>
<td>42.32</td>
<td>52.36</td>
</tr>
<tr>
<td>obs95sheetBT</td>
<td>80.39</td>
<td>49.03</td>
<td>55.27</td>
<td>64.41</td>
<td>75.45</td>
</tr>
<tr>
<td>sup95sheetBT</td>
<td>80.68</td>
<td>49.37</td>
<td>71.52</td>
<td>65.75</td>
<td>76.30</td>
</tr>
<tr>
<td>obs85sheetBT</td>
<td>72.41</td>
<td>42.35</td>
<td>47.15</td>
<td>58.49</td>
<td>67.90</td>
</tr>
<tr>
<td>sup85sheetBT</td>
<td>72.62</td>
<td>41.11</td>
<td>57.66</td>
<td>57.76</td>
<td>68.11</td>
</tr>
<tr>
<td>obs63sheetBT</td>
<td>55.38</td>
<td>27.47</td>
<td>36.83</td>
<td>42.64</td>
<td>51.34</td>
</tr>
<tr>
<td>sup65sheetBT</td>
<td>55.49</td>
<td>29.15</td>
<td>42.98</td>
<td>43.18</td>
<td>51.74</td>
</tr>
</tbody>
</table>

Table 4.1: Assessment on PDB structures with α-τ correlation plots. Total 705 pairs in PDB, each having a previously obsoleted and currently superceded structure, are evaluated not only with the general α-τ correlation and β-τ correlation plots in PRESS-PLOT but also evaluated with the correlations for α-helices and β-sheets.
Residue-level structural properties are as important as those at atomic level for protein modeling but are more difficult to measure. PRESS-PLOT provides a presumptive-sounding valuable set of tools to analyze and evaluate protein structures based on their residue-level virtual angle correlations. The effectiveness of the tools are clearly demonstrated in their ability to distinguish the low resolution obsoleted structures from their superseded high-resolution counter parts.

PRESS-PLOT is derived from the PRESS angle-based structural assessment function, but it provides more details about the angle correlations: It examines the virtual angle pairs in specific secondary substructures as well as the whole structure, extending the correlation plots from the original two to fourteen. PRESS-PLOT utilizes various advanced web technologies and makes it possible for the users to get access to the PRESS-PLOT structural evaluation tools easily from anywhere on the internet, with zero software installation or command line to learn. The users can submit their structures and obtain the PRESS-PLOT evaluations immediately in both graphic and text forms.

PRESS-PLOT is effective for secondary structure assessment, because the virtual angle correlations are short-range restrictions (within four connected residues) and relate directly to the correct fold of the secondary structures. If there are two structures with the same secondary structural components, but different tertiary orders, their PRESS-PLOT evaluations would be about the same, because their local virtual angle correlations would remain the same. Tools for tertiary structural assessment may be developed by combining certain long range constraints such as residue contact potentials (31).

The current implementation of PRESS-PLOT is based on a survey of a large set of X-ray structures in PDB, and therefore, applies to general structures, with X-ray structures as references. The implementation based on a special type of structures, such as the structures of a special protein family or the structures determined by NMR, could be interesting and particularly effective for the structures of that type.

The residue-level virtual angle correlations are not as restrictive as those at the atomic level such as the φ-ψ angle correlations in Ramachandran plot. For both atomic and residue level accuracies, one may use the Ramachandran plot as well as PRESS-PLOT as a pair of
complementary assessment tools. After all, the PRESS-PLOT assessment is statistically based. The results need to be examined with caution: There could be exceptions: some angle pairs in the most favored regions may not be really favored in a particular structure; some in disallowed regions may be just due to a special arrangement in that structure.
CHAPTER 5. General Conclusion

5.1 General Discussion

Instead of relying on physical experiments, we investigate a whole spectrum of residue-level protein structural properties statistically, in order to understand the interactions or relationships among amino-acid residues, with which we can construct or predict protein structures at the residue-level. By statistically, we mean to examine a large number of known protein structures, and obtain the statistical distributions and correlations of key residue-level protein structural properties such as the residue distances or angles in given residue sequences and the contact frequencies of residue pairs, etc. With these statistical results, we may analyze and assess given protein structures, develop statistical potentials, and build structural and dynamic models, all at the protein residue-level.

This work is fundamental in the field of protein modeling, yet challenging, for it requires a huge amount of data analysis and computation. In chapter 2 we described the statistical measurements and presented the statistical analysis results on residue-level protein structural properties. We downloaded a large set of high-quality structures from PDB and organized them into proper forms for further computation and analysis. We developed R codes for extracting the data from the structural files, for computing the virtual bond lengths, virtual bond angles, and virtual torsion angles for a given sequences of residues, and for collecting these various residue distances and angles into several databases. We conducted a thorough analysis on various important residue-level structural properties, including the distributions for the residue-level virtual bond lengths, virtual bond angles, and virtual torsion angles for all possible residue sequences and correlations among them. Many important residue-level structural properties thus are revealed, which have never been systematically analyzed and documented before. Of
particular interest are the strong correlations between certain neighboring virtual bond and virtual torsion angles. They have never been described before, but proved to be extremely valuable for the understanding of the residue-level folding mechanism of proteins, and can be further utilized to make certain residue-level virtual angle correlation contour maps for protein residue-level structural analysis and assessment. The results from this analysis have been well received in the structural biology community.

In chapter 3, we described the development of our statistical package in R for residue-level protein structural assessment, and the formulation of a novel set of residue-level statistical potentials for protein modeling. Based on the distributions of various types of residue distances and angles, we derived a set of protein residue-level statistical potentials with a method for defining the mean-force potentials in statistical physics. They include the potentials for virtual bond lengths, virtual bond angles, and virtual torsion angles for all possible residue sequences. We organized all of our R code into a software package called PRESS, for Protein Residual Level Structural Statistics. This package has a user-friendly graphics interface for computing and displaying the distributions and correlations of those various residue distances and angles for given residue sequences and can be installed conveniently in any R environment. With the modules for statistical potentials in the package, the virtual bond energy, virtual bond angle energy, and virtual torsion angle energy for a given protein can immediately be evaluated and displayed along the protein residue sequence, and be used for protein residue-level structural analysis and refinement.

In chapter 4, we described the development of an online server called PRESS-PLOT for protein residue-level structural assessment using angle-angle correlation plots. The strong correlations between neighboring virtual bond and torsion angles are demonstrated as certain contour maps in 2D angle planes, and can be employed to analyze and assess a given structure. They are very similar to the Ramachandran Plots that are widely used for atomic level protein structural assessment. They have regions described as most favored, favored, allowed, and disallowed related angle pairs, so they can be used to see how the angle pairs of a given protein structure are distributed. These plots have proven to be valuable for protein residue-level structural evaluation. This web server called PRESS-PLOT can generate general angle
correlation plots as well as refined ones, and it is easily accessible for scientists. The users can upload their structural files and obtain any of these plots for their own structures.

5.2 Future Research

The residue-level statistical properties that we presented in this study are general and can be extended in several important directions for future research. These include the following:

- The sequentially long-range residue-residue correlation (e.g., the contact interactions) could be investigated to develop contact potentials.

- The work here could be extended to a complete statistical potential energy function with both short-range (neighboring) and long-range residue interactions.

- A well-developed protein residue-level statistical potential, which can be used not only for structural assessment and prediction but also for physical modeling such as potential energy minimization and molecular dynamics simulation.

More specifically, the virtual bond lengths, virtual bond angles, and virtual torsion angles reflect only short-range interaction among residues. In order to have a complete description of protein residue interactions, we have to include some long-range interactions. For this purpose, we can use for example the residue contact potentials developed based on the contact frequencies. The contacted pairs of residues can be obtained with a method of triangulation for 3D points. As showed in Figure 5.1, we can get a partition of the convex hull of the amino acids into tetrahedrons for a given structure using the triangulation techniques. If two residues represented by two vertices within one tetrahedron are not in the sequential neighbourhood (index of residues $i, j$, such that $j - i > 4$), we say they are contacted. The overall distribution of distances between two contacted residues has a peak around $6.5\AA$ (see Figure 5.2). The distances between two contacted residues in type $R_i$ and type $R_j$ ($j - i > 4$) can also be computed and collected for a reference of contact frequencies $p_{C[R_i,R_j]}(D_{ij})$, where $D_{ij}$ is the contact distance (see Figure 5.3). So the statistical potentials of contacted residues of any
The 3D delaunay triangulation of a protein. We have applied the technique of delaunay triangulation on a protein structure (ID: 1JM1) to get a partition of the convex hull of the amino acids into tetrahedrons whose vertices are the \( C_\alpha \) points from the given sequence. The triangulation forms a partition of 3D space. The tetrahedral cells (four-face bodies) are such that two cells do not intersect.
Figure 5.2: The distribution of distances between two contacted pairs derived from protein triangulation. We have applied the technique of delaunay triangulation on ~ 1000 X-ray determined protein structures with high resolution (< 1.5 Å) and obtained a collection of distances between two contacted pairs. The Gaussian-like distributions of all the contact distances and the individual distributions of specific contact distances between different types of residues can be used for defining specific types of contact potentials (see Figure 5.3).
Figure 5.3: The distribution of distances between two specific types of contacted pairs. (a) Distance distribution of two contacted hydrophobic pairs (b) Distance distribution of contacted hydrophobic-hydroxyl pairs

types can be defined as $P_{C}[R_i, R_j](D_{ij}) = -kT \ln(p_{C}[R_i, R_j](D_{ij})/Z)$, where $Z$ is the partition function, $k$ is Boltzmann constant, and $T$ is temperature.
CHAPTER 6. Game Dynamic Model for Yeast Development

Yuanyuan Huang, and Zhijun Wu
Program on Bioinformatics and Computational Biology
Department of Mathematics
Iowa State University, Ames, IA 50014, U.S.A.
This work was published in the Bulletin of Mathematical Biology

Abstract

Game theoretic models, along with replicator equations, have been applied successfully to the study of evolution of populations of competing species, including the growth of a population, the reaching of the population to an equilibrium state, and the evolutionary stability of the state. In this paper, we analyze a game model proposed by Gore et al. (65) in their recent study on the co-development of two mixed yeast strains. We examine the mathematical properties of this model with varying experimental parameters. We simulate the growths of the yeast strains and compare them with the experimental results. We also compute and analyze the equilibrium state of the system and prove that it is asymptotically and evolutionarily stable.

Key Words: Evolutionary game theory; evolution of cooperation; replicator equations; Nash-equilibrium; evolutionarily stable states; co-existence of mixed yeast strains

6.1 Introduction

Game theory, originally developed in economics, has been applied to many other scientific fields including ecology, psychology, sociology, and political sciences. In 1940s, John von Neumann and Oskar Morgenstein (49) investigated the zero-sum games and proved the existence
of the optimal strategies. In 1950s, John Nash (50) considered more general games and showed the existence of the equilibrium solution, called the Nash equilibrium. Later in 1970s, John Maynard Smith and George R. Price (51) introduced game theory to biology and proposed the concept of evolutionarily stable state, which laid the foundation for evolutionary game theory. The latter, along with replicator equations, has been applied successfully to the study of evolution of populations of competing species, including the growth of a population, the reaching of the population to an equilibrium state, and the evolutionary stability of the state (52). The evolution of cooperation has been an area of intensive research in the past several decades (53). General rules have been studied in terms of kin selection, direct and indirect reciprocity, network selection, and group selection (54). Special strategies have been discovered such as the tit for tat strategies in the repeated prisoner’s dilemma games (55) and the restrictions of dispersal, movement, and interaction for the co-existence of competing species (56; 57; 58; 59). The cooperative behaviors have been explained in game theoretic models and verified in experiments with model bacteria and yeast systems (60; 61; 62; 63; 64; 65). In particular, Gore et al. (65) investigated the evolution of cooperation using a deliberately designed system of two competing yeast strains. What special about this work is the following: First, the designed system depends on several experimental parameters. Therefore, by varying the parameters, the system’s behavior can be changed from a mutual benefit game to a snowdrift game and from a snowdrift game to a prisoner’s dilemma game. Then, cooperation can be investigated in different game settings. Second, the authors showed that if a simple matrix game model was used, there might be only two types of games, the mutual benefit game or prisoner’s dilemma game. The snowdrift game is not implied in this model, while in the experiments the system behaved most time like a snowdrift game. Therefore, the authors derived a nonlinear payoff function for the game, i.e., a nonlinear game model, with a more accurate description of the system’s behaviors.

While Gore et al. (65) focused on experimental results, this paper is a further theoretical stretch of their work. In this paper, we examine the mathematical properties of the proposed game model with varying experimental parameters. We simulate the growths of the yeast strains and compare them with the experimental results. We also compute the equilibrium
states of the system and analyze their stabilities. We show that the equilibrium state of the snowdrift game is evolutionarily stable, which is not so obvious when the model is given in a nonlinear form.

In what follows in this section, we introduce some basic concepts in game theory. In Section 6.2, we describe the yeast experiments in Gore et al. (65). We describe their model and show some of their results. In Section 6.3, we describe the methods for computing the growths and equilibrium states of the yeast system. We present our computational results and compare them with the experimental ones. In Section 6.4, we analyze the model and prove the asymptotic and evolutionary stabilities of the equilibrium states. We conclude the paper in Section 6.5 and make some additional remarks.

For introduction, let’s consider a simple two-player two-strategy game. Then, the outcomes of the game can be measured by a pair of payoff functions that depend on the strategies of the players. Let \( \pi_1 \) and \( \pi_2 \) be the payoff functions for player 1 and 2, respectively. Let \( x = (x_1, x_2)^T \) be the player 1’s choice of strategies, playing strategy 1 with probability \( x_1 \) and strategy 2 with probability \( x_2 \), and \( x_1 + x_2 = 1 \). Let \( y = (y_1, y_2)^T \) be the player 2’s choice of strategies, playing strategy 1 with probability \( y_1 \) and strategy 2 with probability \( y_2 \), and \( y_1 + y_2 = 1 \). Then, the payoffs for player 1 and 2 will be \( \pi_1(x, y) \) and \( \pi_2(y, x) \), respectively. If \( \pi_1 \) is equal to \( \pi_2 \), the game is called symmetric, i.e., the payoff will be the same for a pair of competing strategies no matter which player chooses which strategy. For a biological game, a player could be a certain species; a strategy could be a certain genotypic or phenotypic feature; and the payoff function could be the fitness function defined in terms of resource gain or reproduction success.

In any case, each player competes to maximize his payoff, but can only achieve it when the other player can no longer reduce his gain with a counter strategy. If both players reach such a point, the game is said to reach equilibrium or more specifically Nash equilibrium, and the strategies of the players are called the optimal strategies. Consider a symmetric game. Define \( S_x = x \geq 0 : \sum_i x_i = 1 \). Then, a game reaches Nash equilibrium if there are optimal strategies \( x^* \) in \( S_x \) for player 1 and \( y^* \) in \( S_y \) for player 2 such that

\[
(1) \quad \pi(x^*, y^*) \geq \pi(x, y^*), \quad \forall x \in S_x \text{ and } \pi(y^*, x^*) \geq \pi(y, x^*), \quad \forall y \in S_y.
\]

Usually, the payoffs for all the pairs of competing strategies can be given in a matrix \( U \).
with $U_{ij}$ being the payoff for strategy $i$ against strategy $j$. The payoff function $\pi$ can then be defined as the average payoff of the player when he chooses strategy $x$ against the other player’s strategy $y$, i.e., $\pi(x, y) = x^TUy = \sum_i(x_i\sum_j U_{ij}y_j)$. In this case, the game is called a matrix game. Otherwise, the payoff function can be in a general nonlinear form.

A game is called a prisoner’s dilemma game (PD) if each player chooses to be a cooperator or defector to counter the other: A cooperator gains substantially (with a payoff, say 4) if the opponent also cooperates, but loses the most (with a payoff, say -1) if the opponent defects. A defector gains the most (with a payoff, say 5) if the opponent cooperates, and does not lose that much as the cooperator (with a payoff, say 1) if the opponent defects. So a safe guard is to always play a defector, for nobody wants to take a risk to cooperate but become a loser when the opponent defects.

A game is called a mutual benefit game (MB) if each player chooses to be a cooperator or defector, but the payoff outcomes are very different from PD: A cooperator gains (with a payoff, say 5) no matter what the opponent chooses to be. A defector always loses (with a payoff, say 3) no matter what the opponent is. As a result, both players will choose to play cooperators.

A snowdrift game is played when two players act as two drivers whose cars are blocked by the snow on the road. Each of them needs to decide if he should go out to remove the snow or stay in the car to wait until the other driver clears the road. In this game, a defector (who stays in the car) takes advantage (with a payoff, say 5) if the opponent cooperates (removing the snow), but the worst scenario is when the opponent also defects and then both get stuck in the car (with a payoff, say 0). A cooperator benefits (with a payoff, say 3) if the opponent also cooperates, but would benefit more if he were a defector. Besides, he may get worse (with a payoff, say 1) if the opponent defects. In such cases, the cooperators are attempted to be defectors, and the defectors have to cooperate sometimes if there is no cooperator. As a result, a good scenario would be for the player to sometimes cooperate and sometimes defect, to obtain an optimal amount of benefit.
6.2 Yeast Development as a Snowdrift Game

Gore et al. designed an experiment with two mixed yeast strains in sucrose culture. The wild type cells, considered as cooperators, hydrolyze sucrose into glucose and fructose with protein invertase. They consume only a small portion (about 1%, called the capture rate) of the products. The rest of the glucose and fructose diffuses into the culture. The mutant cells, considered as defectors, do not have the gene $SUC2$ that produces protein invertase and therefore, cannot make glucose or fructose. The wild type cells are also made histidine auxotroph (with defected $HIS3$ genes) so they have an additional fitness cost due to histidine anabolic deficiency. The cost is high when the histidine concentration in culture is low (see Figure 6.1). In general, the growth of the mutant strain depends on the wild type strain, but the mutant strain will outgrow the wild type if the culture is maintained with high concentration glucose and low concentration histidine, because the wild type cells incur a fitness cost due to invertase production as well as histidine anabolic deficiency. In this case, the interaction between the wild type cell and the mutant cell is more like a prisoner’s dilemma game (PD),
Figure 6.2: The co-existence of two yeast strains. (a) If started with a high population, the cooperator fraction decreases. If started with a low portion, the cooperator fraction increases. Somewhere in between, there must be an equilibrium state of nonzero cooperator fraction. (b) If the cost is high (in low histidine concentration), the equilibrium fraction tends to be low (RED). If the cost is low (in high histidine concentration), the equilibrium fraction tends to be high (BLUE) (65).

and the mutant cells, the defectors, eventually take over the whole population, assuming the high glucose concentration of the culture is maintained. On the other hand, if the glucose concentration in culture is low and the histidine concentration is high, the wild type cells will eventually win, because they have the advantage of consuming their own glucose or fructose directly while the “cost of living” due to histidine anabolic deficiency is not so high. This makes the interaction between the two types of cells a mutual benefit game (MB). In Gore et al., the authors were able to change several experimental parameters including the glucose and histidine concentrations in culture. The authors found that for a large range of parameter values, the interaction between the two types of yeast cells maintained a snowdrift game (SG) instead, and the two strains co-existed at equilibrium. More specifically, with a reasonable concentration of glucose and histidine in culture, the authors observed the increase of the wild type strain when it started from a small fraction and the decrease from a large fraction, and it reached an equilibrium fraction no matter what fraction it started (see Figure 6.2).

Note that by the equilibrium of the yeast system in experiment we mean a relative stable state of the system after a sufficiently long time. It is an approximately estimated equilibrium
Figure 6.3: *The nonlinear game model for yeast development.* With low efficiency $\epsilon$ and high cost $c$, cheaters win (PD). With high efficiency $\epsilon$ and low cost $c$, cooperators win (MB). In between, the two strains co-exist, reaching equilibrium (SG)(65).
state, and in the yeast experiment, was roughly the state after 6 days.

To build a game model for a biological system, the population is usually assumed well mixed so that each individual can interact with any others. An individual is designated as one of the players, say player 1, while the population as another, say player 2. The types of species in the population can be considered as the types of strategies, and each individual can choose to be a certain type of species (i.e., a certain strategy) or to be a mixed type of species (i.e., a mixed strategy). Such a game is called a population game or an evolutionary game or a game against the field (66; 67). For the system of two yeast strains, assume that all yeast cells are well mixed. Then, the system can be considered to have two strategies: $f$ fraction of wild types (cooperators) and $(1 - f)$ fraction of mutants (defectors). Each individual cell can be considered to play a mixed strategy, with probability $f^o$ to be a wild type and $(1 - f^o)$ to be a mutant. Clearly, when the system reaches equilibrium, $f = f^o$. Let $\epsilon$ be the capture rate and $f$ the fraction of the cooperators (the wild types). Let the payoff (growth rate) of the cooperators be $P_C$ and of the defectors (the mutants) be $P_D$. Then, a simple game model for the yeast system of interest can be defined such that

$$P_C = \epsilon + f(1 - \epsilon)c, \quad P_D = f(1 - \epsilon).$$

where $c$ is the cost for cooperation. Consider $P_C$ and $P_D$ to be the functions of $f$ and $h$, with $h = 1 - f$. Then, $P_C(f, h) = f(1 - c) + h(\epsilon - c)$ and $P_D(f, h) = f(1 - \epsilon)$. Let the strategy of an individual cell be $(f^o, h^o)^T$, where $h^o = 1 - f^o$. Then, the average payoff for the cell against a population of strategy $(f, h)^T$ can be written as

$$\pi = f^o P_C(f, h) + h^o P_D(f, h).$$

Let $x = (f^o, h^o)^T$ and $y = (f, h)^T$. Define a payoff matrix,

$$U = \begin{bmatrix} 1 - c & \epsilon - c \\ 1 - \epsilon & 0 \end{bmatrix}.$$

Then, the average payoff in (3) can be written in the following form,

$$\pi(x, y) = x^T U y.$$

Based on the general properties of two by two matrix games, it is not hard to see that this payoff function defines a PD game if the cost is high ($c > \epsilon$) and a MB game if the cost is
However, in Gore et al., it was observed that the yeast strains behaved more of a snowdrift game. Therefore, the matrix game does not seem to be a suitable model for the yeast system of interest. Indeed, Gore et al. proposed a nonlinear game model that matched the observed behaviors of the yeast strains. In this model, the payoff functions $P_C$ and $P_D$ are defined in the following form:

$$P_C(f, h) = [g + (1 - g)(\epsilon + f(1 - \epsilon))]^\alpha - c, \quad P_D(f, h) = [g + (1 - g)f(1 - \epsilon)]^\alpha,$$

where an additional parameter $g$ is also introduced to account for the dependency of the payoffs on the glucose concentration $g$ of the culture, and the payoff functions are made nonlinear by using a fractional power $\alpha = 0.15$. The glucose concentration $g$ of the culture is measured in milligrams per 100 milliliters and is maintained via routine artificial adjustment in the experiment. Consider the following two equations:

$$\begin{align*}
(6) \quad [g + (1 - g)(1 - \epsilon)]^\alpha &= 1 - c, \quad \alpha = [g + (1 - g)\epsilon]^{\alpha} - c.
\end{align*}$$

The first one holds when the whole population becomes cooperators ($f = 1$) at equilibrium ($P_C = P_D$). The second one holds when the whole population becomes defectors ($f = 0$) at equilibrium ($P_C = P_D$). The two equations define two curves in the $c - \epsilon$ plane. The payoff functions define a MB game when the $(c, \epsilon)$ pair is above the first curve and a PD game when it is below the second curve. When the $(c, \epsilon)$ pair is in between the two curves, a snowdrift game (SG) is played as it is observed in the experiment (see Figure 6.3).

### 6.3 Computer Simulation

Suppose that the growth rates of cooperators and defectors are given as the formulas in (5). Then, the average payoff function for a cell of strategy $x = (f^o, h^o)^T$ against a population of strategy $y = (f, h)^T$ can be written as

$$\begin{align*}
(7) \quad \pi(x, y) &= f^o P_C(f, h) + h^o P_D(f, h) \\
&= f^o[g + (1 - g)(\epsilon + f(1 - \epsilon))]^\alpha - f^o c + h^o[g + (1 - g)f(1 - \epsilon)]^\alpha.
\end{align*}$$

By the definition of Nash equilibrium and the fact that $x^* = y^*$ when the system reaches equilibrium, we then have $y^*$ as an equilibrium state of the system if and only if

$$\begin{align*}
(8) \quad \pi(y^*, y^*) &\geq \pi(y, y^*), \forall y \in S_y.
\end{align*}$$
Figure 6.4: Cooperator fractions at equilibrium. The experimentally observed cooperator fractions at equilibrium with varying glucose and histidine concentrations are shown in a color-coded graph. From this graph, we see that in low histidine and high glucose, the cooperator fractions are almost zero and the defectors win, implying that the game is PD. On the other hand, in high histidine and low glucose, the cooperator fractions are almost one and the cooperators win, implying that the game is MB. In between, the two strains co-exist, and the game is SG (65).
Figure 6.5: *Color-coded cost values.* The values for cost $c$ are calculated using Formula (10). From this graph, we see that the $c$ values are high for low histidine and high glucose (red regions) and are low for high histidine and low glucose (blue regions). With high cost, the system certainly tends to be like a PD game, while with low cost, it is more of a MB game. In between, the system behaves as a SG game with (the wild type) cooperators co-existing with (the mutant) defectors.

Figure 6.6: *Dynamic changes of cooperator fractions.* Shown in (b) are the simulated trajectories of cooperator fractions with varying values of glucose concentration, where the value for $\epsilon = 0.01$ and histidine concentration is fixed to 0.05x. The simulations are all started with an initial cooperator fraction $f_0 = 10^{-1}$. Then, if glucose concentration is relatively low (0.001%), the defector fraction grows and the cooperator fraction decreases to an equilibrium fraction. When the glucose concentration is increased, the cooperator fraction at equilibrium decreases (not as demanded yet costly for co-existence). When it is increased to 0.03%, the cooperator fraction at equilibrium is almost zero (below $10^{-4}$ after 5 days). All these results agree well with the experimental observations as displayed in (a) (65).
where $S_y = y \geq 0 : \sum_i y_i = 1$. Define $S^*_y = y \in S_y : \sum_{i \in \text{supp}(y^*)} y_i = 1$, where $\text{supp}(y^*) = i : y^*_i > 0$. Then, we also have an important property for $y^*$ as stated in the following theorem.

**Theorem 3.1** If a population game defined by a payoff function $\pi$ reaches (Nash) equilibrium with an optimal strategy $y^*$, then $\pi(y^*, y^*) = \pi(y, y^*), \forall y \in S^*_y$. This theorem is called the equal payoff theorem in evolutionary game theory. A proof can be found in standard textbooks such as (67). The theorem implies that if $y^*$ is an optimal strategy, then $\pi(y, y^*) = \pi(y', y^*), \forall y, y' \in S^*_y$. For a two-strategy game, if $y^* = (f^*, h^*)^T$ is an optimal strategy and $f^*, h^* > 0$, then $S^*_y = S_y$. Let $y_C = (1, 0)^T$ and $y_D = (0, 1)^T$. Then, $y_C, y_D$ are in $S^*_y$ and $\pi(y_C, y^*) = \pi(y_D, y^*)$. It follows that for the payoff function $\pi$ defined in (7),

$$ (9) \quad [g + (1 - g)(\epsilon + f^*(1 - \epsilon))]^\alpha - c = [g + (1 - g)f^*(1 - \epsilon)]^\alpha. $$

This is equivalent to the condition when $P_C(f^*, h^*) = P_D(f^*, h^*)$.

For computer simulation, the equation in (9) gives an important condition for the computation of the equilibrium fraction $f^*$ of the yeast game if the parameters $g$, $\epsilon$, and $c$ are given. Indeed, in (65), $\epsilon$ was estimated to be around 0.01. Then, if glucose and histidine concentrations were set to certain values, the corresponding equilibrium fraction could be observed after the system settled down for several days (Rigorously speaking, the observed equilibrium fraction in finite time is only an approximated equilibrium fraction). Figure 6.4 shows the experimentally observed cooperator fractions $f^*$ at equilibrium with varying glucose and histidine concentrations in a color-coded graph. From this graph, we see that in low histidine and high glucose, the cooperator fractions $f^*$ are almost zero and the defectors win, implying that the game is PD. On the other hand, in high histidine and low glucose, the cooperator fractions $f^*$ are almost one and the cooperators win, implying that the game is MB. In between, the two strains co-exist, and the game is SG.

Conversely, using condition (9), if $g$ and $\epsilon$ are given and $f^*$ is known, the cost $c$, dependent of the histidine concentration, may be determined. We call this process an inverse game: determine the parameters that define the payoff function, given the optimal strategies at equilibrium. It is called the inverse game because it corresponds to an inverse problem in mathematical terms (68). Such a problem can be as important as a direct problem, for often in biological games, we can observe or measure the equilibrium states at least approximately, but we do not know
anything about the payoff rules, which in fact are what we want to learn and discover. From condition (9), the cost $c$ can be determined given $\epsilon$, $g$, and equilibrium fraction $f^*$ using the following formula:

$$
(10) \quad c = [g + (1 - g)(\epsilon + f^*(1 - \epsilon))]^\alpha - [g + (1 - g)f^*(1 - \epsilon)]^\alpha.
$$

Figure 6.5 shows the values of $c$ in a color-coded graph calculated using formula (10) with the values of $f^*$ shown in Figure 6.4. Note again that these equilibrium fractions are observed ones in experiment and are approximately estimated values. From this figure, we see that the values of $c$ are high for low histidine and high glucose (red regions) and are low for high histidine and low glucose (blue regions). With high cost, the system certainly tends to be like a PD game, while with low cost, it is more of a MB game. In between, the system behaves as a SG game with (the wild type) cooperators co-existing with (the mutant) defectors.

Once all the parameters in the payoff functions are determined, the equilibrium fraction $f^*$ and hence $h^*$ can be computed based on Theorem 3.1 or more specifically, on condition (9) for the yeast problem. However, in order to learn how the yeast strains have developed from some initial state to an equilibrium state, we have to solve a system of replicator equations that describes the dynamic properties of the population changes of the species.

A general principle for constructing a system of replicator equations for a given population of species is that the growth rate for each species relative to its current population is proportional to the payoff advantage of this species over the average payoff of the whole population. Let $y_i$ be the population fraction of species $i, i = 1, \ldots, n$, where $n$ is the total number of species. Then, the replicator equations can be written in the following form

$$
(11) \quad \frac{y_i'}{y_i} = \pi(e_i, y) - \pi(y, y), \quad i = 1, \ldots, n, \quad \sum_i y_i = 1.
$$

For the yeast problem, $n = 2$, $y = (f, h)^T$, $\pi(e_1, y) = P_C(f, h)$, $\pi(e_2, y) = P_D(f, h)$, and $\pi(y, y) = fP_C(f, h) + hP_D(f, h)$. We then have

$$
(12) \quad f' / f = P_C(f, h) - fP_C(f, h) - hP_D(f, h),
$$

$$
(13) \quad h' / h = P_D(f, h) - fP_C(f, h) - hP_D(f, h).
$$

Given the fact that $f + h = 1$, the above equations can be reduced to one:

$$
(14) \quad f' = f(1 - f)[P_C(f, 1 - f) - P_D(f, 1 - f)].
$$

Let $P(f) = P_C(f, 1 - f) - P_D(f, 1 - f)$. Then, equation (14) can also be written as
Figure 6.7: Dynamic changes of cooperator fractions. Shown in (b) are the simulated trajectories of cooperator fractions with varying values of histidine concentration, where the value for \( \epsilon = 0.01 \) and glucose concentration is fixed to 0.003%. The simulations are all started with an initial cooperator fraction \( f_0 = 10^{-1} \). Then, if histidine concentration is relatively high (> 0.2x), the cooperators maintain their population to equilibrium. When the histidine concentration is decreased, the cooperator fraction decreases from its initial fraction to equilibrium, and in general, the lower the histidine concentration, the smaller the cooperator fraction at equilibrium (costly to maintain). When it is decreased to 0.005x, the cooperator fraction at equilibrium is close to \( 10^{-5} \) in 5 days, while the experimentally observed one is close to \( 10^{-4} \). Overall, these results agree well with the experimental observations as displayed in (a) (65).

\[(15) \ f' = f(1 - f)P(f), \]

where

\[(16) \ P(f) = [g + (1 - g)(\epsilon + f(1 - \epsilon))]^{\alpha} - c - [g + (1 - g)f(1 - \epsilon)]^{\alpha} \]

depends on the parameters \( \epsilon, g, \) and \( c \). By fixing the values for these parameters, we can solve a corresponding equation (15) for any initial fraction \( f_0 \), to obtain a trajectory for \( f \) changing from \( f_0 \) in certain time period. In particular, we can also find possible fixed points (or limiting points) for \( f \) from equation (15) when the growth rate of \( f \) is zero, i.e., either \( f = 0 \), or \( f = 1 \), or \( f \) is such that \( P(f) = 0 \). It follows from the following theorem that a Nash equilibrium point must be a fixed point, but a fixed point may not necessarily be a Nash equilibrium point.

**Theorem 3.2** A strategy \( y^* \) is a fixed point of the system of replicator equations (11) if it is a Nash-equilibrium strategy for the game defined by the payoff function \( \pi \).

The proof for this theorem is straightforward, since if \( y^* \) is an equilibrium strategy, by Theorem 3.1, \( \pi(y^*, y^*) = \pi(y, y^*) \), \( \forall y \in S_y^* \) including \( \forall e_i \in S_y^* \), and then, the equations in (11) are all equal to zero when \( y = y^* \). More detailed and related discussions on the fixed points of the replicator equations can be found in standard textbooks such as Webb (67).
Figure 6.8: *Equilibrium fractions.* Shown in (b) are the computed cooperator fractions at equilibrium with varying values for glucose and histidine concentrations, while the value for $\epsilon = 0.01$. The curves for histidine concentrations 1.0x and 0.2x match with experimentally observed equilibrium fractions well, decreasing relatively slowly as the glucose concentration increases. When the histidine concentration further decreases, the equilibrium fraction is supposed to approach to zero quickly as glucose concentration increases. Indeed, for example, when the histidine concentration is decreased to below 0.02x, the equilibrium fractions become zero for all glucose concentrations higher than 0.005%. Otherwise, a true fractional cooperator equilibrium strain is found. All these results agree well with the experimentally observed equilibrium fractions as displayed in (a) (65).

Figure 6.9: *Stability of equilibrium state.* There are three fixed points: $f = 0, f = 1, f = f^*$ such that $P(f) = 0$ for equation (15). Assume that $0 < f^* < 1$. Then, $f'$ is positive and $f$ increases to $f^*$ for $0 < f < f^*$ and $f'$ is negative and $f$ decreases to $f^*$ for $f^* < f < 1$. The point $f^*$ must be a stable point.
The following are the computational results we have obtained from our simulation. These results agree well with the experiment results reported in (65) and hence validate the proposed game model. In addition, the simulation allows us to not only predict all possible equilibrium states of the system but also track the dynamic changes of the yeast system in a continuous time period. Figure 6.6 (b) shows the solution curves for the equation in (15) with varying values of glucose concentration of the culture, while the value for $\epsilon = 0.01$ and histidine concentration is fixed to 0.05x, where x means times the normal histidine concentration, which is $20\mu g/mL$. The simulations are all started with an initial cooperator fraction $f_0 = 10^{-1}$. Then, if glucose concentration is relatively low (0.001%), the defector fraction grows and the cooperator fraction $f$ decreases to an equilibrium fraction. When the glucose concentration is increased, the cooperator fraction at equilibrium decreases (not as demanded yet costly for co-existence). When it is increased to 0.03%, the cooperator fraction at equilibrium is almost zero (below $10^{-4}$ after 5 days). All these results agree well with the experimental observations as displayed in Figure 6.6 (a).

Figure 6.7 (b) shows the solution curves for the equation in (15) with varying values of histidine concentration instead, while the value for $\epsilon = 0.01$ and glucose concentration is fixed to 0.003%. The simulations are all started with an initial cooperator fraction $f_0 = 10^{-1}$. Then, if histidine concentration is relatively high (> 0.2x), the cooperators maintain their population to equilibrium. When the histidine concentration is decreased, the cooperator fraction $f$ decreases from its initial fraction to equilibrium, and in general, the lower the histidine concentration, the smaller the cooperator fraction at equilibrium (costly to maintain). When it is decreased to 0.005x, the cooperator fraction at equilibrium is close to $10^{-5}$ in 5 days, while the experimentally observed one is close to $10^{-4}$. Overall, these results agree well with the experimental observations as displayed in Figure 6.7 (a).

Note that in Figure 6.6 and Figure 6.7, the computed trajectories are shown only in 6 days, in order to compare with the experimental ones. They also end slightly higher than the experimental curves, because they are supposed to converge to the observed equilibrium fractions as the time goes to infinity.

Figure 6.8 (b) shows the equilibrium fractions $f^*$ obtained using condition (9) with varying
values for glucose and histidine concentrations, while the value for $\epsilon = 0.01$. The condition (9) holds when $0 < f^* < 1$ and two yeast strains co-exist. Therefore, the curves for histidine concentrations 1.0x and 0.2x match with experimentally observed equilibrium fractions well, decreasing relatively slowly as the glucose concentration increases. When the histidine concentration further decreases, the equilibrium fraction is supposed to approach to zero quickly as glucose concentration increases. Indeed, for example, when the histidine concentration is decreased to below 0.02x, the equilibrium fractions become zero for all glucose concentrations higher than 0.005%. Otherwise, a true fractional cooperator equilibrium strain is found. All these results agree well with the experimentally observed equilibrium fractions as displayed in Figure 6.8 (a).

6.4 The Stability of Equilibrium

We focus on the stability of the equilibrium states of the snowdrift game for the yeast problem. To be self-contained, we first introduce some stability terms and theorems without proofs and refer the readers to (66; 67) for more details.

Definition 4.1 A solution $y^*$ for a system of replicator equations (11) is called a fixed point if $y^*_i = y^*_i \left[ \pi(e_i, y^*) - \pi(y^*, y^*) \right] = 0, \forall i = 1, \ldots, n$.

Definition 4.2 A fixed point $y^*$ for a system of replicator equations (11) is said to be asymptotically stable if for any $\epsilon > 0$, there is $\delta > 0$ such that $||y(t)y^*|| < \epsilon$ for all $t > 0$ whenever $||y(0)y^*|| < \delta$ and if $y(t) \rightarrow y^*$ as $t \rightarrow \infty$.

Note that by this definition, asymptotically stable for a fixed point $y^*$ simply means that a small deviation or perturbation $y(t)$ from $y^*$ will remain arbitrarily close to $y^*$ for all $t$ and converge to $y^*$ as $t$ goes to infinity. However, it is hard to verify the asymptotic stability of a fixed point directly from its definition. The following theorem, called Lyapunov Theorem, is often used in stability analysis.

Theorem 4.1 A fixed point $y^*$ for a system of replicator equations (11) is said to be asymptotically stable if there is a function $V(y) \geq 0$ with $V(y) = 0$ if and only if $y = y^*$ such that $V'(y) \leq 0$ with $V'(y) = 0$ if and only if $y = y^*$. 
Definition 4.3 A strategy $y^*$ for a population game defined by a payoff function $\pi$ is called a Nash equilibrium point if $\pi(y^*, y^*) \pi(y, y^*)$, $\forall y \in S_y = y \geq 0 : \sum_i y_i = 1$.

Definition 4.4 A Nash equilibrium point $y^*$ for a population game defined by a payoff function $\pi$ is said to be evolutionarily stable if for any $y$ in $S_y$, $y \neq y^*$, $\pi(y, \epsilon y + (1 - \epsilon)y^*) < \pi(y^*, \epsilon y + (1 - \epsilon)y^*)$, for any $\epsilon > 0$ sufficiently small.

Note that evolutionary stability is a stronger stability term than asymptotic stability. It was first proposed by (51) for measuring the stability of equilibrium states of a biological system and later justified as a general stability measure. However, the definition does not provide a simple way for verifying evolutionary stability. The following theorem is one among many others that can be used more directly for formal proofs.

Theorem 4.2 A Nash equilibrium point $y^*$ for a population game defined by a payoff function $\pi$ is said to be evolutionarily stable if and only if $\pi(y^*, y) > \pi(y, y)$ for any $y$ in $S_y$, $y \neq y^*$, and $y$ in a small neighborhood of $y^*$.

Note again that as a system of replicator equations corresponds to a population game, the fixed points for a system of replicator equations and their asymptotic stabilities have close relationships with the Nash equilibrium points for the corresponding population game and their evolutionary stabilities. In fact, as stated in the following theorems, an evolutionarily stable Nash equilibrium point must be an asymptotically stable fixed point; an asymptotically stable fixed point must be a Nash equilibrium point; and a Nash equilibrium point must be a fixed point. However, the reverse of any of these statements may not always valid.

Theorem 4.3 An asymptotically stable fixed point for a system of replicator equations is a Nash equilibrium point for the corresponding population game.

Theorem 4.4 An evolutionarily stable Nash equilibrium point for a population game is an asymptotically stable fixed point for the corresponding system of replicator equations.

Theorem 4.5 Let $F$ be the set of fixed points, $N$ the set of Nash equilibrium points, $A$ the set of asymptotically stable fixed points, and $E$ the set of evolutionarily stable Nash equilibrium points for a given population game and its corresponding system of replicator equations. Then, $F \supset N \supset A \supset E$.

We now consider the game model for the yeast problem. Consider the replicator equation
(15) for the model. By the definition of fixed points, this equation has three fixed points: 
\( \mathbf{f} = 0, \mathbf{f} = 1, \mathbf{f} = \mathbf{f}^* \) such that \( P(\mathbf{f}) = 0 \). It is easy to verify that \( P'(\mathbf{f}) < 0 \) and therefore, \( P(\mathbf{f}) \) is a monotonically decreasing function. Assume that \( 0 < \mathbf{f}^* < 1 \). Then, \( f' \) is positive and \( f \) increases to \( \mathbf{f}^* \) for \( 0 < f < \mathbf{f}^* \) and \( f' \) is negative and \( f \) decreases to \( \mathbf{f}^* \) for \( \mathbf{f}^* < f < 1 \), as can be depicted in Figure 6.9. The point \( \mathbf{f}^* \) must be a stable point. We state this fact more formally in the following theorem.

**Theorem 4.6** The fractional fixed point \( \mathbf{f}^* \) for the replicator equation in (15) of the yeast game is asymptotically stable.

**Proof** Let \( V(\mathbf{f}) = P^2(\mathbf{f}) \). Then, \( V(\mathbf{f}) > 0 \) for all \( \mathbf{f} \neq \mathbf{f}^* \) but \( V(\mathbf{f}^*) = 0 \). Also, since

\[
P'(\mathbf{f}) < 0, \quad \mathbf{f}' = \mathbf{f}(1 - \mathbf{f})P(\mathbf{f}),
\]

we have

\[
V'_t(\mathbf{f}) = V'_t f = 2\mathbf{f}(1 - \mathbf{f})P^2(\mathbf{f})P(\mathbf{f}) < 0
\]

for all \( \mathbf{f} \neq \mathbf{f}^* \) but \( V'_t(\mathbf{f}^*) = 0 \). It follows from Theorem 4.1 that \( \mathbf{f}^* \) is an asymptotically stable fixed point.

Let \( \mathbf{y}^* = (\mathbf{f}^*, \mathbf{h}^*)^T \), where \( \mathbf{h}^* = 1 - \mathbf{f}^* \). Then, by Theorem 3.1, \( \mathbf{y}^* \) is an optimal mixed strategy for the yeast game, since \( \pi(\mathbf{y}^*, \mathbf{y}^*) = \pi(\mathbf{e}_i, \mathbf{y}^*) = \pi(\mathbf{y}, \mathbf{y}^*), \forall \mathbf{y} \in S_y \). In other words, \( \mathbf{y}^* \) is a fractional Nash equilibrium point for the yeast game, corresponding to the co-existing state of the yeast strains. In the following theorem, we prove that this state is also evolutionarily stable.

**Theorem 4.7** The optimal mixed strategy \( \mathbf{y}^* \) of the yeast game defined by the nonlinear payoff function in (7) is an evolutionarily stable strategy.

**Proof** Let \( \mathbf{f} \) be any cooperator fraction and \( \mathbf{f} \neq \mathbf{f}^* \). Since \( P(\mathbf{f}) \) is a monotonically decreasing function, we have \( (\mathbf{f}^* - \mathbf{f})P(\mathbf{f}) > 0 \). The latter is equivalent to

\[
(\mathbf{f}^* - \mathbf{f})[P_C(f, h)P_D(f, h)] > 0, \quad (\mathbf{f}^* - \mathbf{f})P_C(f, h)(\mathbf{f}^* - \mathbf{f})P_D(f, h) > 0,
\]

which further reduces to

\[
\mathbf{f}^*P_C(f, h) + (1 - \mathbf{f}^*)P_D(f, h) > \mathbf{f}P_C(f, h) + (1 - \mathbf{f})P_D(f, h).
\]
The left-hand side of the above inequality is exactly the payoff function \( \pi(y^*, y) \), where \( y^* = (f^*, h^*)^T \) and \( y = (f, h)^T \), while the right-hand side is \( \pi(y, y) \). We then have \( \pi(y^*, y) > \pi(y, y) \), for all \( y \neq y^* \). By Theorem 4.2, \( y^* \) is an evolutionarily stable strategy.

### 6.5 Concluding Remarks

In this paper, we have examined several mathematical properties of the nonlinear game model proposed by Gore et al. (65). While Gore et al. (65) focused on experimental results, this paper is a further theoretical stretch of their work. The contributions of this paper can be summarized as follows.

First, we have analyzed the mathematical properties of the nonlinear game model proposed by Gore et al. (65) and determined the values of the cost factor \( c \) in the payoff functions with varying experimental parameters, the glucose and the histidine concentrations in culture, using the experimentally observed yeast fractions at equilibrium. The calculations were not complicated, but the process of determining one or more parameters that define the game model using the dynamic properties such as the equilibrium states were emphasized. Termed as an inverse game, the process can be important for general game theoretic modeling of biological systems. Considered as an inverse problem, issues such as the existence of the inverse solutions, the uniqueness of the solutions, and the stability of the problem can be important for general models, either linear or nonlinear (68).

Second, with the determined cost factor \( c \), the payoff function is well defined. Then, we have been able to compute the yeast fractions at equilibrium based on the game theory principles and obtain their dynamic trajectories by solving a system of replicator equations. We have obtained these results with a fixed glucose concentration but varying histidine concentrations and also with a fixed histidine concentration but varying glucose concentrations. The results were not just simple recoveries of the experimental outcomes. In fact, they can be generated with arbitrary parameter values and in continuous time intervals.

Third, we have compared our computational results with the experimental ones reported in Gore et al. (65). Both sources of results agreed well, confirming the theoretical model as well as the simulation results. In Gore et al. (65), a nonlinear game model is proposed because
it admits a snowdrift game for a large range of parameter values. It is therefore important to verify computationally if this model can indeed lead to similar equilibrium states and dynamic trajectories of the yeast strains as observed in experiments.

Fourth, with the model validated computationally, we have been able to prove that the co-existing equilibrium state of the yeast strains is asymptotically and evolutionarily stable, which is not so obvious when the game is modeled in a nonlinear form. The stability, especially the evolutionary stability, is a critical issue for evolutionary game dynamics. It is a rigorous measure for how strong a given biological system is against any mutant invasion, but the proof for such a property often requires some mathematical techniques. On the other hand, the proof is only based on the theoretical model, and the true stability of the system is yet to be verified by experiment, which may not be easy to do as well.

Finally, the paper is written in a more expository form, with both biological and mathematical accounts, while certain level of mathematical rigorousness is maintained. The paper invites interests from biological and mathematical communities in evolutionary game theory and its applications in biological modeling. Further extension of this work can be interesting as well such as applying the studied game model to the yeast strains in a 2D plane. The outcomes can be different from what we have so far observed either experimentally or computationally, because of the effect of the scale of the local movement and interaction (56; 57; 58; 59).


I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis.

First and foremost, Dr. Wu and Dr. Jernigan for providing me the flexibility to select problems that I was interested in; I thank them for spending time to discuss my research, and to teach me a proper way of writing; I thank them for the support of my work and the belief in the quality of my work.

I would also like to thank my committee members Dr. Dorman, Dr. Abbott and Dr. Willson for discussing and contributing research ideas. I thank my committee members for their efforts and contributions to this work.

I would like to thank Stephen Bonett and Kejue Jia for developing the package and web server with their outstanding computer skills. Without Stephen and Kejue’s help, we can not accomplish this work.

I would like to thank Prof. Jeff Gore for kindly giving us the Matlab code for generating some of the graphical results in their paper and for providing helpful suggestions and advices for how to proceed in our computation.

I would like to thank my group members. I thank them for spending time to discuss research and share their views, to stay up late and work together at office. I thank Yiping Hao, Min Wang, and Rick Zhou for collaborating on evolutionary games.

I would like to thank Trish Stauble for helping me throughout PhD, and being a good friend.

I would like to thank BCB program for funding my first year of PhD. I would like to thank Baker Center for funding my research. I would like to thank BCB program, Baker Center, and Math department for partially supporting my travel to various conferences.

Finally, I would like to thank my parents Yilin and Zhen for always being there supporting
me, and sharing the joy and bitterness.