Combining Empirical and Experimental Data in Protein Structure Determination

Yi Zhang

University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/theses
COMBINING EMPIRICAL AND EXPERIMENTAL DATA IN PROTEIN STRUCTURE DETERMINATION

A Thesis Presented
by
YI ZHANG

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL AND COMPUTER ENGINEERING

September 2009

Electrical and Computer Engineering
COMBINING EMPIRICAL AND EXPERIMENTAL DATA IN PROTEIN STRUCTURE DETERMINATION

A Thesis Presented
by
YI ZHANG

Approved as to style and content by:

Ramgopal R. Mettu, Chair

Aura Ganz, Member

Weibo Gong, Member

C. V. Hollot, Department Head
Electrical and Computer Engineering
ABSTRACT

COMBINING EMPIRICAL AND EXPERIMENTAL DATA IN PROTEIN STRUCTURE DETERMINATION

SEPTEMBER 2009

YI ZHANG
B.Sc., SMITH COLLEGE
M.S.E.C.E., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ramgopal R. Mettu

In this thesis, we develop an algorithm for the problem of finding sidechain conformations for protein structures. The goal of our algorithm is to incorporate the experimental data into a linear programming approach for sidechain prediction in order to improve structure accuracy. In order to do so, we modified an existing prediction method with a linear programming framework to incorporate the experimental data from X-ray crystallography. We tested and compared our algorithm with existing prediction software and experimentally solved structures. This thesis shows that while our method is a feasible approach to obtain lower energy levels, the assumptions made for the energy function need to be further developed in order to more accurately correlate predictions with experimental data.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
</tbody>
</table>

## CHAPTER

1. **INTRODUCTION** ................................................................. 1
   1.1 Protein Structures ............................................................ 1
   1.2 Rotamer and Rotamer Library ................................................. 4
   1.3 Problem Definition .................................................................. 6
   1.4 Related Work ......................................................................... 6
   1.5 Thesis Contributions ............................................................ 7

2. **OVERVIEW OF X-RAY CRYSTALLOGRAPHY** ............................ 8
   2.1 X-ray Crystallography Methodology ............................................ 9
   2.2 Principle of X-ray crystallography and Electron Density Map (EDM) 10
      2.2.1 Electron Density Map (EDM) and its Interpretation ............ 10
   2.3 Difficulties with X-ray Crystallography ................................... 12

3. **STRUCTURE PREDICTION WITHOUT EXPERIMENTAL DATA** ........ 15
   3.1 Energy Function ..................................................................... 15
      3.1.1 SCWRL Energy Function .................................................... 16
   3.2 Linear Programming (LP) and its Use in Protein Structure
      Determination ........................................................................... 18
3.2.1 Integer Linear Programming .............................................. 19
3.2.2 LP used in Protein Structure Determination ....................... 19

4. STRUCTURE PREDICTION WITH ELECTRON DENSITY DATA ........................................ 22
4.1 Modified Energy Function ....................................................... 22
4.2 Modified Linear Programming Model ................................. 24
4.3 Implementation .............................................................. 25
  4.3.1 Implementation Procedures ............................................. 25
4.4 Implementation Result ....................................................... 29
  4.4.1 Accuracy .................................................................. 30
  4.4.2 χ Angle Errors ........................................................... 35
  4.4.3 Energy .................................................................. 37

5. CONCLUSION ................................................................. 42

BIBLIOGRAPHY ................................................................. 44
LIST OF TABLES

Table | Page |
------|------|
4.1  | Proteins used for our experiments. Information is listed with the four letter protein ID name, the resolution of its electron density map, and the total number of rotamers from the rotamer library for the entire protein. | 27 |
4.2  | Absolute accuracy comparison residue by residue. Listed by the residue names, the total number of a certain type of residue in the 14 proteins that we chosen and the absolute accuracy of a certain type of residue. Absolute accuracies of our modified energy function (weight {0, 1, 1} and weight {0.4, 1, 0.6}), the original energy function without EDM data (weight {1, 1, 0}), and the SCWRL software. Our modified energy functions overall have a little improvement than the original SCWRL function, but not as good as the SCWRL software. | 33 |
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Amino acid [21]. An amino acid contains an amino group, a hydrogen, a α- carbon, a carboxylic group and a sidechain group.</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Linear chain of amino acids. Amino acids bond together by peptide bonds to form a linear chain.</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Six dihedral angles for protein 3D structure. φ and ψ are backbone dihedral angles, and χ₁ to χ₄ are sidechain dihedral angles [11].</td>
<td>3</td>
</tr>
<tr>
<td>1.4 Rotamers. The blue helix represents the backbone of a protein. Numbers (1-5) are the Cα’s positions in the protein. Therefore, each number represents a residue. On the third Cα, there are several rotamer choices as shown on the figure. Each of those rotamer choices can be a possible sidechain representation [5].</td>
<td>4</td>
</tr>
<tr>
<td>1.5 χ angles distribution. Distribution of χ₁ (horizontal) and χ₂ (vertical) angles’ combination for more than 1719 HIS residues. Densely populated regions are shown with contours [22].</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Procedures of X-ray crystallography. Protein crystals will be grown and examined by X-rays to form the diffraction pattern for electron density map calculation.</td>
<td>9</td>
</tr>
<tr>
<td>2.2 A beam of X-rays go through protein crystal and generate a diffraction pattern on the detector [10].</td>
<td>10</td>
</tr>
<tr>
<td>2.3 Bragg’s Law of diffraction [23].</td>
<td>11</td>
</tr>
<tr>
<td>2.4 2-dimensional electron density map example. The numbers are the electron density ρ calculated from Equation 2.1, and the location of each number on the plane corresponds to the position vector (x, y). Contours are drawn around the areas with electron density greater than certain values. Contours are added by crystallographers in order to locate atom positions [16]. The real 3D electron density map has iso-surfaces instead of contours.</td>
<td>12</td>
</tr>
</tbody>
</table>
2.5 Interpretation of EDM. Based on contour lines which are drawn according to electron densities, crystallographers manually identify each atom’s position in the EDM [8].

2.6 EDM with different resolution. (a) is the EDM with resolution 5.0 Å, (b) is the EDM with resolution 3.0 Å, and (c) is the EDM with resolution 1.5 Å. EDM has more details with higher resolution (lower number). With resolution 1.5 Å or higher, individual atom can be identified in the EDM [1].

3.1 An example for explaining LP SCP model. The protein has two residues Residue 1 and Residue 2. Residue 1 has two rotamer choices and Residue 2 has three rotamer choices. The red rotamer choices are the optimal rotamer choices for their residues.

4.1 Volume matching score calculation example. $V_{map}$ is the EDM mesh surface. A rotamer $r$ is inserted into map for a volume matching score calculation. There are three atoms overlap the map (atom $A$, atom $B$, and atom $F$.) The volume matching score is calculated in the Equation 4.2.

4.2 Implementation of the algorithm. Each one of the 14 proteins we used in this thesis was tested according to this diagram. First, the EDM was given by EDS, and the mesh surface was created by Matlab. Then using the BALL library and Matlab, we calculated the self and pairwise energy score. Finally, after the result returned by the LP solver, we can reconstruct the protein 3D structures.

4.3 Mesh with different isovalues (generated by PYMOL with protein 1a6m)

4.4 Absolute accuracy comparison protein by protein. Absolute accuracy of our modified energy function (weight {0, 1, 1} and weight {0.4, 1, 0.6}), the original energy function without EDM data (weight {1, 1, 0}), and the SCWRL software. Our modified energy functions overall have a little improvement by using EDM data, but our accuracy is not as good as the SCWRL software.
4.5 Absolute accuracy comparison residue by residue. Absolute accuracy of our modified energy function (weight set \{0, 1, 1\} and weight set \{0.4, 1, 0.6\}), the original energy function without EDM data (weight set \{1, 1, 0\}), and the SCWRL software. Our modified energy functions overall have a little improvement than the original energy function, but not as good as the SCWRL software. .......................................................... 32

4.6 SCWRL accuracy comparison protein by protein. SCWRL accuracies of our modified energy function (weight \{0, 1, 1\} and weight \{0.4, 1, 0.6\}), the original energy function without EDM data (weight \{1, 1, 0\}), and the SCWRL software. Our modified energy functions overall have a little improvement than the original SCWRL function, but are not as good as the SCWRL software. .................. 34

4.7 $\chi_1^{\text{error}}$ Comparison protein by protein. Average $\chi_1$ angle errors of our modified energy functions (weight \{0, 1, 1\} and weight \{0.4, 1, 0.6\}), the original energy function (weight \{1, 1, 0\}), and the SCWRL software. Our modified energy functions overall have a slight improvement than the original SCWRL energy function, but still are not as good as the SCWRL software. .................. 36

4.8 $\chi_2^{\text{error}}$ Comparison protein by protein. Average $\chi_2$ angle errors of our modified energy functions (weight \{0, 1, 1\} and weight \{0.4, 1, 0.6\}), original energy function without EDM data (weight \{1, 1, 0\}), and the SCWRL software. All energy functions have similar average $\chi_2$ angles, SCWRL software does a little better prediction than others......................................................... 38

4.9 Energy comparison protein by protein. Energies of sidechain conformations returned by our modified energy function (weight \{0, 1, 1\} and weight \{0.4, 1, 0.6\}), original energy function without EDM data (weight \{1, 1, 0\}), SCWRL software, and the experimental structures. ................................................. 41
CHAPTER 1
INTRODUCTION

Proteins accomplish most of the functions of the living cell. They build up and maintain the tissues in a living cell. Proteins take about 20% of human weight, and all of our body parts are made from proteins except for bile and urine. Muscles, organs, and the immune system are made up mostly of protein. Therefore, obtaining and analyzing the 3D molecular structures of proteins is very important and helpful for understanding basic biochemical processes. Various algorithms have been developed to predict protein 3D structures; however, protein structures are still determined experimentally. In this thesis, we propose a prediction algorithm that combines pure modeling data and experimental data form X-ray crystallography in a linear programming (LP) framework.

1.1 Protein Structures

Proteins are made of amino acids (Figure 1.1) arranged in a linear chain and joined together by peptide bonds (Figure 1.2). As shown in Figure 1.1, an amino acid is made of an amino group, a carboxylic acid group, a hydrogen, an α-carbon, and a sidechain group (R) that is attached to the α-carbon. Each amino acid is called a residue in proteins. There are 20 different amino acids called standard amino acids or proteinogenic amino acids that naturally exist in nature (as opposed to being manufactured in a laboratory). Those amino acids play a major role in forming proteins in human bodies. This linear chain formed with amino acids will fold itself into a 3D structure. There are six dihedral angles describe this 3D structure as
shown in Figure 1.3. Dihedral angles $\phi$ and $\psi$ correspond to the backbone, and dihedral angles $\chi_1$ to $\chi_4$ correspond to the sidechain. Currently, determining the 3D structures of proteins still relies on experimental methods. In Chapter 2, details about experimental methods will be discussed.

![Figure 1.1. Amino acid](image1.png)

**Figure 1.1.** Amino acid [21]. An amino acid contains an amino group, a hydrogen, a $\alpha$-carbon, a carboxylic group and a sidechain group.

![Figure 1.2. Linear chain of amino acids](image2.png)

**Figure 1.2.** Linear chain of amino acids. Amino acids bond together by peptide bonds to form a linear chain.

Each protein has two parts: the backbone (mainchain) – the same make up of atoms in all residues (all the make up of atoms except for the sidechain (R) as shown in Figure 1.1 and Figure 1.2), and the sidechain atoms (R).

Ideally, we would want to be able to identify the 3D structures by given the amino acid sequence. Unfortunately, existing algorithms cannot determine these structures
Figure 1.3. Six dihedral angles for protein 3D structure. $\phi$ and $\psi$ are backbone dihedral angles, and $\chi_1$ to $\chi_4$ are sidechain dihedral angles [11].

accurately and efficiently from just amino acid sequences, and scientists routinely rely upon laboratory methods in order to determine protein structures.

Most protein structures are determined experimentally by using X-ray crystallography method, which will be explained in further detail in Chapter 2. Proteins first need to be crystalized, and then be examined by X-rays to retrieve the diffraction data. The electron density map can be generated according to the diffraction data and is interpreted by crystallographers for determining 3D structures of proteins. Without any statistical data and computational models, this process requires extensive manual work and is very time consuming.

Most sidechains have one or more conformations. Therefore, determining the sidechain positioning in proteins becomes very important for protein structure determination. We call this problem of finding the conformation of each residue the sidechain positioning (SCP) problem. SCP problem generally assume there is a rigid
backbone, known amino acid sequence, and some sidechain conformation library. Sidechain conformation will be discussed with further detail in Section 1.2.

1.2 Rotamer and Rotamer Library

As mentioned before, a sidechain can have several conformations. Depending on the C-C axis angle of rotation (also called the dihedral angle or the angle of torsion), every distinct arrangement around the \( \alpha \)-carbon \((C_\alpha)\) is called a rotational isomer (rotamer) \([6]\). Each rotamer is a single conformation represented by a set of values, and there is one rotamer for each dihedral angle degree of freedom as shown in Figure 1.4.

![Figure 1.4. Rotamers. The blue helix represents the backbone of a protein. Numbers (1-5) are the \( C_\alpha \)'s positions in the protein. Therefore, each number represents a residue. On the third \( C_\alpha \), there are several rotamer choices as shown on the figure. Each of those rotamer choices can be a possible sidechain representation \([5]\).](image)

A rotamer library considers a discrete set of rotamers for each residue type. In nature, a conformation can be in any position in a 3D structure. Given the increased sample size of numbers of solved protein structures, the most common sidechain conformations of each residue type can be identified using statistical analysis. Many
research over the years have found that sidechain conformations are not evenly distributed [7], in fact, most dihedral angles (χ angles) occur in tight clusters around certain values as shown in Figure 1.5. Therefore, rotamer libraries cluster observed 

![Figure 1.5. χ angles distribution. Distribution of χ₁ (horizontal) and χ₂ (vertical) angles' combination for more than 1719 HIS residues. Densely populated regions are shown with contours [22].](image)

conformations or divide dihedral angle spaces into bins based on statistical analysis of sidechain conformations in solved protein structures. A rotamer in a rotamer library is generally an average conformation over a certain region of dihedral angle space or the local minimum on a potential energy map. A rotamer library generally contains information about rotamers’ positions, χ angles, and the frequency of a certain conformation.

A rotamer library can be backbone-independent, secondary-structure-dependent, or backbone-dependent. The difference between those three types of rotamer libraries is whether the χ angles and the frequencies of rotamers depend on the local backbone conformation or not. Since the backbone information is assumed to be given in this thesis, we used the Dunbrack and Cohen backbone-dependent rotamer library [2].
1.3 Problem Definition

The objective is to solve the sidechain positioning (SCP) problem of a protein with known amino sequence, fixed backbone and experimental information.

In order to solve the SCP problem, one rotamer per residues need to be picked from the rotamer library (Dunbrack and Cohen backbone-dependent rotamer library) which minimizes the total energy of the protein. Therefore, the problem can be described as finding the conformation $R^* = (r_1^*, r_2^*, ..., r_N^*)$ in all possible conformations such that,

$$E(R^*) = \min_k E(R_k)$$  \hspace{1cm} (1.1)

where $E(R_k)$ is the total energy of the protein’s $k$th sidechain conformation and $r_i$ is the chosen rotamer for $i$th residue among $N$ residues. In this thesis, the energy $E$ is calculated by using the SCWRL energy function which will be discussed in Chapter 3 and Chapter 4.

1.4 Related Work

In general, there are two approaches to determine the 3D structures of proteins: the experimental method and the prediction method. Each approach has its advantages and disadvantages. The experimental method is the traditional method and remains to be the main method for determining 3D structures of proteins. In Chapter 2, a commonly used experimental method (X-ray crystallography) will be introduced and discussed. The prediction method is a new approach for protein structure determination which this thesis expands upon. In Chapter 3, the general prediction method approach with an energy function will be introduced and discussed. We adopted the SCWRL energy function [3] which is a simple and commonly used energy function in many prediction methods. The SCWRL energy function will be explained in Chapter 3. However, the mathematical model of the prediction method has been proven to be not only NP-complete [18] but also inapproximable [4], which means an optimal
solution or a good solution is not guaranteed to be found within a polynomial-time. We adopted the linear programming (LP) model which was especially developed for the SCP problem [12]. However, the optimal LP SCP model is also NP-hard, a condition that will explained in Section 3.2.1. The estimation LP method of the optimal solution is guaranteed to be found in a polynomial time which will be introduced in Section 4.2. Both the LP method and the LP SCP model will also be introduced in Chapter 3. All the aforementioned methods and information are incorporated in this thesis to form a new prediction method of 3D protein structure determination.

1.5 Thesis Contributions

The purpose of this thesis is to generate an algorithm that combines the experimental and prediction methods in order to have better predictions for solutions to the sidechain positioning (SCP) problem given backbones and a set of rotamer choices for each residue. Recall Equation 1.1, we need to find the set of rotamers $R^*$ for all residues which has the minimum energy for the given protein. It was mentioned in the previous section that the mathematical model for finding the optimal solution is NP-hard. Therefore, in Chapter 4, an algorithm based on linear programming relaxation will be introduced for computing sidechain conformations within a polynomial-time.

In Chapter 4, we will also discuss the modification of the SCWRL energy function that we adopted to incorporated the experimental data into our prediction method. The implementation of our algorithm and experimental results will be discussed in Section 4.3 and Section 4.4.
CHAPTER 2

OVERVIEW OF X-RAY CRYSTALLOGRAPHY

X-ray crystallography [1, 6, 9] is one approach among several experimental methods for determining proteins 3D structures. It is still the main method for protein structure determination. Over 80% of current solved proteins’ structures are solved by X-ray crystallography.

X-ray crystallography is a technology which uses X-ray to determine atoms’ positions inside the crystal. To be more specific, X-ray crystallography analyzes the patterns of electron diffraction caused by X-rays, and generates the electron density map. Using the electron density map, we obtain information about atom positions; consequently, information about the examined structure inside the crystal can be reconstructed. The molecular is crystallized prior to performing X-ray crystallography. Crystals can be seen as arrays of repeating identical molecules. Using this unique structure characteristic of crystals, most diffraction caused by X-ray when it crosses through crystals cancel each other out, and only diffraction in certain directions will add together to produce diffracted beams on a photographic plate or detector.

The procedure of using X-ray crystallography to reconstruct the protein 3D structure is discussed in Section 2.1. The principle of X-ray crystallography will be explained in Section 2.2. Section 2.2.1 contains details about the electron density map and how to rebuild the crystal from it. Finally, Section 2.3 will discuss the obstacles and disadvantages of X-ray crystallography.
2.1 X-ray Crystallography Methodology

X-ray crystallography used to reconstruct protein 3D structures and its procedures are shown in Figure 2.1. Protein crystals of pure quality must be grown for the X-ray radiation to examine it. In order to produce protein crystals of quality pure enough for X-ray examination, factors, such as the pH value, temperature, protein concentration, are controlled for a large sample of proteins. This procedure normally requires a large set of testing before obtaining the satisfied crystal samples.

![Diagram of X-ray crystallography](image)

**Figure 2.1.** Procedures of X-ray crystallography. Protein crystals will be grown and examined by X-rays to form the diffraction pattern for electron density map calculation.

Once pure protein crystals are ready to be examined, a beam of X-rays with short wave lengths is fired through the crystals and generate a diffraction pattern as shown in Figure 2.2. Protein crystals are rotated during this procedure in order to get the diffraction pattern from all directions. The diffraction pattern is measured and by using inverse Fourier transform and phase estimation methods, the electron density maps is generated. Then the electron density map will be interpreted by a crystallographer to reconstruct the proteins’ 3D structures. Details about the electron density map and its interpretation are discussed in the Section 2.2.1.
2.2 Principle of X-ray crystallography and Electron Density Map (EDM)

According to Bragg’s Law: diffraction occurs only when the difference in distance equals the X-ray’s wavelength, as illustrated in Figure 2.3. In addition, the relation between the reflection angle $\theta$, the distances between the planes $d$, and the wavelength $\lambda$ is also analyzed by Bragg and expressed as: $2d \cdot \sin \theta = \lambda$. Therefore, by using Bragg’s law, the unit cell of the crystal can be determined. Furthermore, with the measurements of distance between diffraction spots and the position where X-ray hits the detector, the electron density can be calculated as discussed in Section 2.2.1.

2.2.1 Electron Density Map (EDM) and its Interpretation

An electron density map contains 3D lattice of points covering a unit cell or repeating units in the protein crystals. Each density $\rho(r)$ in the map is calculated
Figure 2.3. Bragg’s Law of diffraction [23].

giving the average electron density in the crystallized protein structure as shown in Figure 2.4.

\[ \rho(r) = \frac{1}{V} \sum_{H} |F_H| \exp(i\phi_H - 2\pi i H \cdot r) \] (2.1)

\(|F_H|\) is the magnitude of the structure factor \(F_H\), \(H\) is the frequency vector, and \(r\) is the position vector \((x, y, z)\). If \(H\) is known, \(|F_H|\) can be calculated directly from each diffraction point’s magnitude. However, the phase \(\phi_H\) cannot be experimentally determined, and this issue is also referred to as the phase problem [1, 9]. Scientists have to use various methods to estimate the phase [9]. We assume the EDM is given with solved \(\phi_H\).

Finally, the electron density map will be interpreted by crystallographers to determine the 3D structures of proteins. First, contour lines are defined as shown in Figure 2.4; and then, crystallographers fit atoms one by one into the EDM as shown in Figure 2.5. A crystallographer needs to figure out the backbone trace though the entire EDM before rotating the rotamer for each residue in order to get the best fit in the EDM. This process is a trial-and-error process, therefore the interpretation procedure may takes several weeks with a poor electron density map. Without any statistical data, a crystallographer must be well trained and experienced in order to
Figure 2.4. 2-dimensional electron density map example. The numbers are the electron density $\rho$ calculated from Equation 2.1, and the location of each number on the plane corresponds to the position vector $(x, y)$. Contours are drawn around the areas with electron density greater than certain values. Contours are added by crystallographers in order to locate atom positions [16]. The real 3D electron density map has iso-surfaces instead of contours.

efficiently interpret the EDM. In general, the higher resolution, the better the EDM quality as shown in Figure 2.6. However, if the resolution is too high, the EDM contains too much information (noisy) for interpretation. In this research, we use thirteen protein EDMs with a resolution ranging from 0.85 Å to 1.15 Å and one protein EDM with a lower resolution at 1.8 Å.

2.3 Difficulties with X-ray Crystallography

X-ray crystallography is very time-intensive and labor-intensive. The unique well-ordered crystal structures, eg. structures with repeating units, accounts for the reason to use. However, well-ordered crystals are really difficult to grow. Protein molecules have irregular globular shapes and require spacing between them in forming the crystal through some solvent such as water. Therefore, it is very difficult if not impossible to pack all molecules into a crystal without forming large spaces between them. Sometimes, the solvent takes more than half of the volume of a crystal, and as a result,
Figure 2.5. Interpretation of EDM. Based on contour lines which are drawn according to electron densities, crystallographers manually identify each atom’s position in the EDM [8].

Figure 2.6. EDM with different resolution. (a) is the EDM with resolution 5.0 Å, (b) is the EDM with resolution 3.0 Å, and (c) is the EDM with resolution 1.5 Å. EDM has more details with higher resolution (lower number). With resolution 1.5 Å or higher, individual atom can be identified in the EDM [1].
the protein molecules are only in contact with each other in small regions. Growing a crystal may also take months, and ideally a 97% threshold of structure purity is required for pure crystals to be examined. In order to get satisfactory protein crystals, scientists have to prepare a huge amount protein samples and spend months to grow and test those crystals.

Several software, such as the SwissModel [20] and MODELLER [19], were developed to assist crystallographers in interpreting the EDM computationally. However, the majority of the work still needs to be done manually. Sometimes an EDM may takes a well experienced crystallographer weeks to interpret. The goal of this thesis is to address this issue by incorporating additional modeling information.
CHAPTER 3
STRUCTURE PREDICTION WITHOUT EXPERIMENTAL DATA

In the last decade, X-ray crystallography has enabled us to obtain information about protein sequence data experimentally; however, the experimental method alone is limited in its capacity to meet our objective of predicting 3D protein structures due to the obstacles mentioned in Section 2.3. Therefore, computational methods were developed to predict protein structures, and the advancement of the computational methods has become more important today than ever.

The two main challenges of protein structure prediction are: the large number of possible protein structures and the lack of base physical knowledge of protein structures. Therefore, a primary objective of prediction methods is finding the unique protein structure efficiently and accurately. Today, prediction methods for backbone prediction are fairly accurate; however, satisfactory methods are still lacking for sidechain positioning problems (SCP). Recall the SCP definition (Equation 1.1), the goal is to find a rotamer choice from the rotamer library for each residue, such that the energy of the given protein is minimized. In order to find such a rotamer set; first, we need to define an energy function on protein conformations for evaluation.

3.1 Energy Function

As mentioned earlier, the physical knowledge of protein structures is still not well known, hence there is none accurate enough definition of protein’s energy which can provide high-resolution structures. The energy $E(R)$ is calculated using different energy functions $[12, 3]$ for different purposes. In general, the energy function is defined
with two terms: the self energy term \((S_i(r_i))\) and the pairwise energy term \((P_{ij}(r_i, r_j))\) as shown in Equation 3.1.

\[
E(R) = \sum_i S_i(r_i) + \sum_{i<j} P_{ij}(r_i, r_j) \tag{3.1}
\]

\(r_i\) and \(r_j\) are the chosen rotamer for residue \(i\) and \(j\). \(R = \{r_1, r_2, \ldots, r_N\}\) is the set of chosen rotamers for residues 1 through \(N\). In another words, each self energy is a singleton energy score for a particular rotamer choice, and each pairwise energy is an energy score for a pair of rotamer choices from two different residues.

With respect to different environments and purposes, the self and pairwise energies are also defined differently. The SCWRL energy function is a simple and commonly used energy function in prediction methods, and it will be introduced in Section 3.1.1.

Once the energy function is established, the next step is finding the optimal sidechain positions of a protein structure according to the defined energy function. Recall the SCP problem definition in Equation 1.1, now the problem can be described as finding the sidechain conformation \(R^* = (r_1^*, r_2^*, \ldots, r_N^*)\), where \(r_1^*, r_2^*, \ldots, r_N^*\) are the chosen rotamers for residue 1 to \(N\), such that the sum of \(R^*\)'s self and pairwise energies is the minimum among all possible sidechain conformations.

There are several computational methods developed for finding the optimal solution of Equation 1.1, and the linear programming method adopted for this project will be introduced in Section 3.2.

### 3.1.1 SCWRL Energy Function

Recall the energy function Equation 3.1, an energy function has two terms: self energy and pairwise energy. The self energy term \(S_i(r_i)\) in SCWRL energy function is defined as shown in Equation 3.2.

\[
S_i(r_i) = -K \log \left( \frac{Pr(r_i|\phi_i, \psi_i)}{\max_{\forall r_i \in D_i} Pr(r_i|\phi_i, \psi_i)} \right) + \sum_{i,j} \sum_{a \in r_i} \sum_{b \in B_j} E_{ab}(a, b) \tag{3.2}
\]
Where $K$ is a weight factor (in this project $K$ is set to be 6 for emphasizing the energy term). The log-probability term consists of the probability ($Pr(r_i|\phi_i, \psi_i)$) of a certain rotamer choice ($r_i$) from residue $i$. This term is given by the backbone dependent library according to the backbone dihedral angles $\phi_i$ and $\psi_i$. This log-probability argument is normalized with the highest rotamer probability of each residue $i$ (defined as $\max_{r_i \in D_i} Pr(r_i|\phi_i, \psi_i)$, where $D_i$ is the rotamer set for residue $i$). In this way, the energy of the rotamer with the highest probability is zero, and the lower probability a rotamer has, the higher the energy will be assigned to this rotamer. $E_{ab}(a, b)$ also known as the van der waals energy, is an expression that describes the interaction energy between atom $a$ and atom $b$, where $a$ is an atom from a certain rotamer choice $r_i$ of residue $i$ and $b$ is an atom from the backbone atoms of residue $j$ ($B_j$). $E_{ab}(a, b)$ is calculated as Equation 3.3.

$$E_{ab}(a, b) = \begin{cases} 0 & |a - b| > R_{ab} \\ 10 & |a - b| < 0.8254R_{ab} \\ 57.273 \left(1 - \frac{|a - b|}{R_{ab}}\right) & 0.8254R_{ab} \leq |a - b| \leq R_{ab} \end{cases} \quad (3.3)$$

where $|a - b|$ is the distance between atom $a$ and atom $b$’s centers. $R_{ab}$ is the sum of atom $a$ and atom $b$’s radiiuses. Therefore, if there is no collision between atoms $a$ and $b$, the energy $E_{ab}(a, b)$ is zero; while a higher energy is assigned to $E_{ab}(a, b)$ as the intersecting distance of $a$ and $b$’s radiiuses increases. The worse case is finding two atoms located in the same position (not a possible scenario in reality) implied when $a$ and $b$ are completely imposed on one another.

The SCWRL pairwise energy is defined as shown in Equation 3.4.

$$P_{ij}(r_i, r_j) = \sum_{a \in r_i} \sum_{b \in r_j} E_{ab}(a, b) \quad (3.4)$$
Where $r_i$ and $r_j$ are the chosen rotamer choices for residue $i$ and residue $j$, and $i \neq j$. In addition, $a$ is an atom from rotamer $r_i$ and $b$ is an atom from rotamer $r_j$. $E_{ab}(a, b)$ is defined in Equation 3.3.

### 3.2 Linear Programming (LP) and its Use in Protein Structure Determination

Linear programming is one of the most important research areas in optimization. LP is widely applied in math, operational research, micro economic and many other fields. The main purpose of LP is to find the minimum or maximum of a given mathematical model with given constraints. To be more specific, given a linear function:

$$f(x_1, x_2, ..., x_n) = c_1x_1 + c_2x_2 + ... + c_nx_n + d$$

LP can find a point $(x_1, x_2, ..., x_n)$ in the function such that the function has the minimum or maximum value subject to some constrains. Mathematically, a general LP problem can be expressed in the following canonical format [15]:

Minimize $c^T x$

Subject to $A x \leq b$

Where $x \geq 0$

$x = \{x_1, x_2, x_3, ..., x_n\}$ is the vector of variables that need to be determined, $c$ and $b$ are vectors of coefficients, and $A$ is a matrix of coefficients.

$c^T x$ is the expression that needs to be optimized which is called the objective function. $A x \leq b$ and $x \geq 0$ are the constraints. Given $c^T$, $A$, and $b$, LP returns the optimal $x$. 

18
3.2.1 Integer Linear Programming

If all variables in $x$ are integers, we call this special type LP as an integer linear programming (ILP) [12, 15]. ILP is proven to be NP-hard [15]. 0-1 integer programming or binary integer programming (BIP) is the special case of ILP which requires all variables in $x$ to be either 0 or 1. The BIP is also proven to be NP-hard [15]. However, LP can efficiently solve the problem in polynomial time when the integer solution constraint is removed [12, 15].

3.2.2 LP used in Protein Structure Determination

LP is used as a tool for finding the optimal sidechain conformation such that the protein energy has the minimum energy as mentioned in Equation 3.2. The LP sidechain positioning (SCP) problem now can be expressed in the following format [12]:

Minimize $E(R) = \sum_i \sum_k S_i(r_{ik}) \cdot x_{ik} + \sum_{i<j} \sum_k \sum_l P_{ij}(r_{ik}, r_{jl}) \cdot y_{ikjl}$

Subject to

$\sum_k x_{ik} = 1 \quad \forall i$

$\sum_k y_{ikjl} = x_{jl} \quad \forall i, j, l : i < j$

where $x_{ik}, y_{ikjl} \in \{0, 1\}$ (3.5)

$r_{ik}$ and $r_{jl}$ and the $k$-th and $l$-th rotamer choices from the rotamer sets for residue $i$ and $j$. $x_{ik}$ is corresponding to the rotamer $r_{ik}$, and is either 0 or 1. Similarly, $y_{ikjl}$ is corresponding to the rotamer pair $(r_{ik}, r_{jl})$, and is either 0 or 1. Notice that the minimum condition is written such that the self energy contribution calculated only for the chosen rotamers (ie., $x_{ik} = 1$). Also, the pairwise energy in the minimum condition is considered if and only if both rotamers in the pair of $r_{ik}$ and $r_{jl}$ are chosen (ie. $y_{ikjl} = 1$). An example of this is illustrated in the following (Figure 3.1):
A simple protein has two residues: “Residue 1” and “Residue 2”. Residue 1 has two rotamer choices \( r_{11} \) and \( r_{12} \). Residue 2 has three rotamer choices \( r_{21}, r_{22}, \) and \( r_{23} \) as shown in Figure 3.1. If \( r_{11} \) and \( r_{22} \) are the optimal rotamer choices for Residue 1 and Residue 2, then the corresponding \( x_{11} \) and \( x_{22} \) are indicated as 1, and other \( x_{ik} \) \((x_{12}, x_{21}, x_{23})\) are indicated as 0. In addition, the corresponding \( y_{1122} \) is indicated as 1, and other \( y_{ikjl} \) \((y_{1121}, y_{1221}, y_{1222})\) are indicated as 0. Therefore, the energy \( E(R) \) is:

\[
E(R) = \sum_i \sum_k S_i(r_{ik}) \cdot x_{ik} + \sum_{i<j} \sum_k \sum_l P_{ij}(r_{ik}, r_{jl}) \cdot y_{ikjl} \\
= S_1(r_{11}) + S_2(r_{22}) + P_{1,2}(r_{11}, r_{22}) \\
= E(R^*).
\]

The energy \( E(R) \) is the optimal minimum energy \( E(R^*) \) with the optimal sidechain conformation \( R^* = \{r_{11}, r_{22}\} \).

While solving this linear program is NP-hard since it is a BIP (introduced in Section 3.2.1.) If we relax the constraint in Equation 3.5 by allowing \( x_{ik}, y_{ikjl} \in [0, 1] \), then we can find the optimal solution in polynomial time. However, this solution may not correspond to a feasible sidechain conformation. [12] shows this relaxation yields integral solutions without experimental data. In Chapter 4, we show this relaxation almost always yields an integral solution even with electron density map data incorporated.
Figure 3.1. An example for explaining LP SCP model. The protein has two residues Residue 1 and Residue 2. Residue 1 has two rotamer choices and Residue 2 has three rotamer choices. The red rotamer choices are the optimal rotamer choices for their residues.
CHAPTER 4
STRUCTURE PREDICTION WITH ELECTRON DENSITY DATA

Our goal is to incorporate experimental data to the prediction method in order to have a better prediction of protein sidechain structures. In order to do so, we adopted the SCWRL energy function [3] which applied data of each rotamer from the backbone dependent library into the backbone and calculated the energy according to each sidechain atom’s position, probability, interaction with the nonlocal backbone and other sidechain atoms which are all discussed in Section 3.1.1. Subsequently, the electron density map (EDM) data is incorporated in the self energy term in the SCWRL energy function, which will be discussed in further detail with regard to the original SCWRL energy function and our modified energy function. Once our energy function is established, we used the LP SCP model introduced in Section 3.2.2 for an estimation of the optimal rotamer set. Section 3.2 will discuss the modified LP SCP model which is guaranteed to return a solution within polynomial time.

4.1 Modified Energy Function

We adopted both self energy and pairwise energy terms in the SCWRL energy function. In order to incorporate EDM data into this energy function, we defined a volume matching score for each rotamer choice of each residue and defined an EDM energy term $E_{EDM}(r_i)$ as shown in Equation 4.1 which is similar to the log-probability term in the SCWRL self energy term introduced in Equation 3.2.
\[ E_{\text{EDM}}(r_i) = -K \log \left( \frac{V(r_i)}{\max_{\forall i \in D_i} V(r_i)} \right) \] (4.1)

Where \( V(r_i) \) is the volume matching score for the chosen rotamer choice \( r_i \) of residue \( i \), and \( \max_{\forall i \in D_i} V(r_i) \) is the highest volume matching score in residue \( i \). The volume matching score is defined as the fraction of the matched rotamer volume of the EDM divided by the total volume of this rotamer. The EDM is interpreted by generating a mesh surface of it. The example in Section 2.2.1 is a 2D example of the EDM interpretation. In real 3D EDM, instead of contour lines, iso-surfaces such as mesh are made according to isovalues of electron densities. More information about the mesh iso-surface will be discussed in Section 4.3. An example of \( V(r_i) \) calculation is shown in Figure 4.1. \( V_{\text{map}} \) represents the EDM, and the rotamer has 7 atoms A to G. The matching volume score \( V(r) \) is calculated as:

\[
V(r) = \frac{V_{\text{overlap}}}{V_r} = \frac{\text{vol}(A) + \text{vol}(B) + \text{vol}(F)}{\text{vol}(A) + \text{vol}(B) + \ldots + \text{vol}(G)}
\]
The volume matching score is a number expressed in a fraction between zero and one which is similar to the probability definition. Therefore, the energy of EDM can be defined similarly to the log-probability energy term in the SCWRL self energy term.

The EDM energy is added to the self energy term, and weights $\alpha, \beta, \gamma$ are added before each term in the self energy. The modified self energy term is as shown in Equation 4.2.

$$S_i(r_i) = -\alpha K \log \left( \frac{P(r_i|\phi_i, \psi_i)}{\max_{\psi_i \in D_i} P(r_i|\phi_i, \psi_i)} \right) + \beta \sum_{i,j} \sum_{a \in r_i} \sum_{b \in B_j} E_{ab}(a, b) + \gamma E_{EDM}(r_i) \quad (4.2)$$

$\alpha$, $\beta$, and $\gamma$ are constants added for adjusting the weight of each energy term. The pairwise energy was kept as the same from the SCWRL pairwise energy term shown in Equation 3.4.

### 4.2 Modified Linear Programming Model

As mentioned in Section 3.2, LP is used as a tool for finding the optimal sidechain conformation that would result in the minimum energy of the given protein. However, the LP model that returns the optimal solution is NP-hard. Therefore, we relaxed constraints $x_{ik} \in \{0, 1\}$ to $0 \leq x_{ik} \leq 1$, and $y_{ikjl} \in \{0, 1\}$ to $0 \leq y_{ikjl} \leq 1$ in order to get the solution in polynomial time. The final LP model is as shown below:

Minimize

$$E(R) = \sum_i \sum_k S_i(r_{ik}) \cdot x_{ik} + \sum_i \sum_{k < l} \sum_k \sum_{l} P_{ij}(r_{ik}, r_{jl}) \cdot y_{ikjl}$$

Subject to

$$\sum_k x_{ik} = 1 \quad \forall i$$

$$\sum_k y_{ikjl} = x_{jl} \quad \forall i, j, l : i < j$$

where $0 \leq x_{ik} \leq 1$ \quad (4.3)

This new model will return $x = \{x_{11}, x_{12}, ..., x_{NM}\}$ in fractional forms between 0 and 1. The larger $x_{ik}$ value means it is more likely to be the optimal rotamer choice.
Sometimes more than one rotamer can be chosen per residue, i.e., it will return an ensemble of structures. In some cases, it is desirable to find an ensemble of structures. For example, if two rotamer choices return the same energy score, both of them will be picked and this is possible since protein is always in an active mode, and so sidechain conformations may be in different positions but have the same energy level. In this project, we kept the rotamers with the largest $x_{ik}$ values till the sum of those rotamers’ $x_{ik}$ values is greater or equal to 0.8. More details about the implementation of this model will be discussed in Section 4.3.

4.3 Implementation

Once we have the algorithm, the next step is testing our model. We chose 13 proteins from the SCWRL benchmark [7] which have electron density maps with resolutions ranging from 0.85 Å to 1.15 Å. We also included protein ubiquitin (1UBQ) with 1.8 Å resolution EDM since it is a small and commonly used protein. Also, performed a sensitivity analysis to see how our model would be affected by lower resolutions. Table 4.1 displays the details on the resolution and number of rotamers of all 14 proteins.

In this section, procedures beginning with taking the EDM to finding the prediction of the sidechain conformation is described step by step. Then the results from our algorithm will be analyzed and discussed in 4.4.

4.3.1 Implementation Procedures

The procedures of the implementation is illustrated in Figure 4.2. In this section, all procedures and softwares that we used will be discussed in detail following the order of the diagram in Figure 4.2. First we retrieved EDM data from Electron Density Server (EDS) website in the .cns format. The .cns (or EXPLOR) format is one of several EDM ASCII formats which can be read by a TCL script originally used.
Figure 4.2. Implementation of the algorithm. Each one of the 14 proteins we used in this thesis was tested according to this diagram. First, the EDM was given by EDS, and the mesh surface was created by Matlab. Then using the BALL library and Matlab, we calculated the self and pairwise energy score. Finally, after the result returned by the LP solver, we can reconstruct the protein 3D structures.
Table 4.1. Proteins used for our experiments. Information is listed with the four letter protein ID name, the resolution of its electron density map, and the total number of rotamers from the rotamer library for the entire protein.

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Resolution (Å)</th>
<th>Number of Rotamers</th>
</tr>
</thead>
<tbody>
<tr>
<td>3PYP</td>
<td>0.85</td>
<td>2,066</td>
</tr>
<tr>
<td>1BPI</td>
<td>0.92</td>
<td>1,124</td>
</tr>
<tr>
<td>1RB9</td>
<td>0.92</td>
<td>630</td>
</tr>
<tr>
<td>1NLS</td>
<td>0.94</td>
<td>3,013</td>
</tr>
<tr>
<td>2FDN</td>
<td>0.94</td>
<td>407</td>
</tr>
<tr>
<td>1AHO</td>
<td>0.96</td>
<td>1,032</td>
</tr>
<tr>
<td>1BYI</td>
<td>0.97</td>
<td>2,959</td>
</tr>
<tr>
<td>1A6M</td>
<td>1.00</td>
<td>3,008</td>
</tr>
<tr>
<td>2ERL</td>
<td>1.00</td>
<td>445</td>
</tr>
<tr>
<td>1CTJ</td>
<td>1.10</td>
<td>1,117</td>
</tr>
<tr>
<td>2IGD</td>
<td>1.10</td>
<td>962</td>
</tr>
<tr>
<td>1A7S</td>
<td>1.12</td>
<td>3,396</td>
</tr>
<tr>
<td>3DLC</td>
<td>1.15</td>
<td>3,904</td>
</tr>
<tr>
<td>1UBQ</td>
<td>1.80</td>
<td>1,596</td>
</tr>
<tr>
<td>AVG</td>
<td>1.06</td>
<td>1,833</td>
</tr>
</tbody>
</table>

for a software called VMD [14]. The script was modified to output a .txt file which contains all the EDM information for later use in Matlab.

Then we used Matlab to create a mesh surface from the .txt format EDM that we created before. In addition, Matlab needs an isovalue to generate the mesh. The isovalue we used is 1.0. As shown in Figure 4.3, if the isovalue is too high (Figure 4.3(c)), the mesh cannot cover the entire protein, and if the isovalue is too low (Figure 4.3(a)), the mesh has too much information than needed. The 1.0 isovlaue (Figure 4.3(b)) seems to be a reasonable number to optimally cover the area of interest.

Matlab uses 3 points to define a triangle surface in a mesh and can be stored into two matrices: faces and vertices. Matrix vertices contains information of all the coordinates (x, y, z) of each point in the mesh. Each face in the matrix, faces, the indexes of the three vertices. We stored both faces and vertices into .txt files for further uses.
Meanwhile, we use the BALL library (a C++ code library [13]) which reads all the information of a protein PDB file (PDB is a standard ASCII file format which contains all information about a protein such as the coordinates and names of each atoms, residue names, chemical bond types, etc.) This library contains various packages for biological uses such as reading/writing the PDB file, using Dunbrack and Cohen backbone-dependent and back-bone independent rotamer library, displaying protein structures, etc. We pre-stored some information about each residue (residue names, range in coordinates, number of rotamers and $\alpha$ carbon–$C_\alpha$ positions) into .txt files for further uses.

When we obtained the mesh surface and all the information from the protein PDB file, we used ANN (a C++ code library created by David Mount and Sunil Arya [17]) to define which residue each point belongs to efficiently and stored all those information into a .txt file for further uses. ANN stands for *Approximately Nearest Neighbor* library. This library efficiently searches multidimensional keys in some data structure such as the $k$-$d$ tree with given conditions such as: within a certain distance.

Once we gathered all the information from BALL, ANN, and Matlab, we compare the EDM with each rotamer choice residue by residue in Matlab. Recall the definition

**Figure 4.3.** Mesh with different isovalues (*generated by PYMOL with protein 1a6m*)
of the volume matching score $V(r_i)$ in Section 4.1. We calculated the $V(r_i)$ for each rotamer choice, $r_i$, and stored this score into a .txt file.

Then, we calculated and stored the self energy and pairwise energy of each rotamer choice by using the modified SCWRL energy function defined in Section 4.1 with the BALL library and ANN. The $\alpha$, $\beta$, and $\gamma$ values were first set to be $\{0, 1, 1\}$. This weight set is completely dependent on the electron density map data without any consideration of statistical analysis on rotamer probabilities. However, with this set sometimes one residue will be assigned more than one rotamer choice with our algorithm. In order to have exactly one rotamer per residue, we adjusted the weight set to be $\{0.4, 1, 0.6\}$. More details will be discussed in Section 4.4. Next, we submitted those energy scores into MOSEK by using the LP format described in Section 3.2.2.

Finally after MOSEK returns the result of $x$ values, we chose rotamers $r_i$ which has highest $x_{ii}$, and the summation of those $x_{ii}$ is greater or equals to 0.8. At this point, we finished our prediction of SCP problem. Next section will show the results and comparisons we did with our modified energy function, original SCWRL energy function, and the SCWRL software. The original SCWRL energy function is exactly our modified energy function with $\alpha = 1$, $\beta = 1$, $\gamma = 0$. In the next section, the original SCWRL energy function will be represented as the weight set $\{1, 1, 0\}$, and our modified energy function will be represented as the weight set $\{0, 1, 1\}$ for $\alpha = 0$, $\beta = 1$, $\gamma = 1$ and the weight set $\{0.4, 1, 0.6\}$ for $\alpha = 0.4$, $\beta = 1$, $\gamma = 0.6$.

### 4.4 Implementation Result

We tested all 14 proteins with two sets of $\alpha$, $\beta$, and $\gamma$ (weight set $\{0, 1, 1\}$ and weight set $\{0.4, 1, 0.6\}$). For each of the scenarios: accuracy, average $\chi$ angle ($\chi_1$, $\chi_2$) errors, and energy, we compared our energy functions with the original SCWRL energy function which can be seen as our energy function with weight set $\{1, 1, 0\}$. 
Each scenario is summarized in this section. The SCWRL software was developed based on the SCWRL function and with several additional terms added on the energy function [3]. We will compare our results against the SCWRL sidechain prediction algorithm, we note that this software utilizes the energy function we use but also incorporates additional optimization. All results will be analyzed and discussed in the Conclusion section in Chapter 5.

4.4.1 Accuracy

A rotamer in the rotamer library is defined as the best rotamer choice if the average of its $\chi_1$ and $\chi_2$ angles is the closest to the experimental sidechain conformation’s average of $\chi_1$ and $\chi_2$ angles. With respect to the best rotamer choices, we defined two types of accuracies: absolute accuracy and SCWRL accuracy.

**Absolute Accuracy.** If the algorithm has chosen the best rotamer choice, we say the prediction is correct. Otherwise, the prediction is incorrect. The average absolute accuracy of the prediction by using the original energy function without EDM data (weight set $\{1, 1, 0\}$) is 39.44%, the average accuracy by using our energy function with weight set $\{0, 1, 1\}$ is 40.42%, and the average accuracy by using our energy function with weight set $\{0.4, 1, 0.6\}$ is 42.14%. Therefore the average improvements of our energy function with two weight sets are 0.98% and 2.70%. Details about the absolute accuracy are analyzed protein by protein as shown in Figure 4.4.

Our modified energy functions have same or better accuracies than the original energy function with most proteins. However, the average accuracy of using SCWRL software is 54.94%, which is better than our prediction with both weight sets as shown in Figure 4.4.

We also analyzed the absolute accuracy according to residue types as shown in Figure 4.5 and Table 4.2. The average absolute accuracy of all residues by using the original energy function without EDM data (weight set $\{1, 1, 0\}$) is 37.29%, the
Figure 4.4. Absolute accuracy comparison protein by protein. Absolute accuracy of our modified energy function (weight \{0, 1, 1\} and weight \{0.4, 1, 0.6\}), the original energy function without EDM data (weight \{1, 1, 0\}), and the SCWRL software. Our modified energy functions overall have a little improvement by using EDM data, but our accuracy is not as good as the SCWRL software.
average absolute accuracy of all residues by using our energy function with weight set \{0, 1, 1\} is 38.78\%, and the average absolute accuracy by using our energy function with weight set \{0.4, 1, 0.6\} is 40.40\%. In general, the weight sets that incorporated EDM data had similar or better accuracies than the energy function without EDM data in most residues, but our energy functions’ accuracies were about 10-15\% lower than the SCWRL software which had a 52.24\% absolute accuracy.

**Figure 4.5.** Absolute accuracy comparison residue by residue. Absolute accuracy of our modified energy function (weight set \{0, 1, 1\} and weight set \{0.4, 1, 0.6\}), the original energy function without EDM data (weight set \{1, 1, 0\}), and the SCWRL software. Our modified energy functions overall have a little improvement than the original energy function, but not as good as the SCWRL software.

**SCWRL Accuracy.** If the algorithm returned a rotamer choice with both its $\chi_1$ and $\chi_2$ angles being within 40 degrees to the best rotamer choice, we say the prediction is correct. Otherwise, the prediction is incorrect. This type of accuracy was defined
### Table 4.2. Absolute accuracy comparison residue by residue.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Size</th>
<th>Weight 1,1,0%</th>
<th>Weight 0,1,1%</th>
<th>Weight 0.4,1,0.6%</th>
<th>SCWRL Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>68</td>
<td>10.29%</td>
<td>2.94%</td>
<td>13.24%</td>
<td>11.76%</td>
</tr>
<tr>
<td>ASN</td>
<td>79</td>
<td>24.05%</td>
<td>18.99%</td>
<td>22.78%</td>
<td>32.91%</td>
</tr>
<tr>
<td>ASP</td>
<td>107</td>
<td>31.78%</td>
<td>17.76%</td>
<td>24.30%</td>
<td>44.86%</td>
</tr>
<tr>
<td>CYS</td>
<td>46</td>
<td>50.00%</td>
<td>56.52%</td>
<td>56.52%</td>
<td>73.91%</td>
</tr>
<tr>
<td>GLN</td>
<td>58</td>
<td>5.17%</td>
<td>10.34%</td>
<td>13.79%</td>
<td>24.14%</td>
</tr>
<tr>
<td>GLU</td>
<td>95</td>
<td>13.68%</td>
<td>8.42%</td>
<td>7.37%</td>
<td>18.95%</td>
</tr>
<tr>
<td>HIS</td>
<td>34</td>
<td>38.24%</td>
<td>41.18%</td>
<td>35.29%</td>
<td>41.18%</td>
</tr>
<tr>
<td>ILE</td>
<td>101</td>
<td>48.51%</td>
<td>71.29%</td>
<td>67.33%</td>
<td>82.18%</td>
</tr>
<tr>
<td>LEU</td>
<td>121</td>
<td>67.77%</td>
<td>61.98%</td>
<td>65.29%</td>
<td>77.69%</td>
</tr>
<tr>
<td>LYS</td>
<td>98</td>
<td>3.06%</td>
<td>2.04%</td>
<td>3.06%</td>
<td>11.22%</td>
</tr>
<tr>
<td>MET</td>
<td>22</td>
<td>9.09%</td>
<td>4.55%</td>
<td>4.55%</td>
<td>27.27%</td>
</tr>
<tr>
<td>PHE</td>
<td>69</td>
<td>73.91%</td>
<td>68.12%</td>
<td>69.57%</td>
<td>69.57%</td>
</tr>
<tr>
<td>PRO</td>
<td>78</td>
<td>52.56%</td>
<td>51.28%</td>
<td>51.28%</td>
<td>74.36%</td>
</tr>
<tr>
<td>SER</td>
<td>99</td>
<td>48.48%</td>
<td>41.41%</td>
<td>52.53%</td>
<td>63.64%</td>
</tr>
<tr>
<td>THR</td>
<td>102</td>
<td>43.14%</td>
<td>64.71%</td>
<td>64.71%</td>
<td>88.24%</td>
</tr>
<tr>
<td>TRP</td>
<td>22</td>
<td>45.45%</td>
<td>54.55%</td>
<td>54.55%</td>
<td>54.55%</td>
</tr>
<tr>
<td>TYR</td>
<td>56</td>
<td>60.71%</td>
<td>57.14%</td>
<td>58.93%</td>
<td>66.07%</td>
</tr>
<tr>
<td>VAL</td>
<td>108</td>
<td>45.37%</td>
<td>64.81%</td>
<td>62.04%</td>
<td>77.78%</td>
</tr>
<tr>
<td>AVG</td>
<td></td>
<td>37.29%</td>
<td>38.78%</td>
<td>40.40%</td>
<td>52.24%</td>
</tr>
</tbody>
</table>

Listed by the residue names, the total number of a certain type of residue in the 14 proteins that we chose and the absolute accuracy of a certain type of residue. Absolute accuracies of our modified energy function (weight \{0, 1, 1\} and weight \{0.4, 1, 0.6\}), the original energy function without EDM data (weight \{1, 1, 0\}), and the SCWRL software. Our modified energy functions overall have a little improvement than the original SCWRL function, but not as good as the SCWRL software.
by the SCWRL[3], and commonly used in the literature for sidechain conformation comparisons.

The average SCWRL accuracy of the prediction by using original energy function without EDM data (weight set \{1, 1, 0\}) is 48.52%, the average accuracy by using our energy function with weight set \{0, 1, 1\} is 50.05%, and the average accuracy by using our energy function with weight set \{0.4, 1, 0.6\} is 52.39%. Therefore the average improvements of our energy functions are 1.53% and 3.86%. Details about the SCWRL accuracy are analyzed protein by protein as shown in Figure 4.6.

![Figure 4.6](image.png)

**Figure 4.6.** SCWRL accuracy comparison protein by protein. SCWRL accuracies of our modified energy function (weight \{0, 1, 1\} and weight \{0.4, 1, 0.6\}), the original energy function without EDM data (weight \{1, 1, 0\}), and the SCWRL software. Our modified energy functions overall have a little improvement than the original SCWRL function, but are not as good as the SCWRL software.
In general, our modified energy functions have same or better accuracies than an energy function without EDM data. However, the average SCWRL accuracy by using the SCWRL software is 64.62%, which is better than our prediction with both weight sets as shown in Figure 4.4.

4.4.2 \( \chi \) Angle Errors

We calculated the \( \chi \) angle errors according to the best rotamer choices which are defined in Section 4.4.1. The best rotamer choices have very small \( \chi \) angle errors in contrast to the actual experimental solved sidechain conformations, normally by a difference of 0-5 degrees. It is more reasonable here to use the best rotamer choices in the rotamer library for comparison than to use the actual experimental sidechain conformations, since the best prediction we can achieve are those rotamer choices in the rotamer library not the actual conformations solved by experiments. The average \( \chi \) angle errors \( \overline{\chi_1^{\text{error}}}, \overline{\chi_2^{\text{error}}} \) were calculated as the following:

\[
\overline{\chi_{1,2}^{\text{error}}} = \frac{\sum_{i=1}^{n} \chi_{1,2}^{\text{error}}(i)}{n}
\] (4.4)

Where \( n \) is the number of total rotamer choices that our algorithm has chosen. For some rotamers there is no \( \chi_2 \) angle, and we did not count it when calculating for the \( \overline{\chi_2^{\text{error}}} \).

\( \chi_1 \) Angle Error. The \( \overline{\chi_1^{\text{error}}} \) of the 14 proteins calculated by using the original energy function without EDM data (weight set \{1, 1, 0\}) is 51.68°, the average \( \overline{\chi_1^{\text{error}}} \) of the 14 protein calculated by using our modified energy function with weight set \{0, 1, 1\} is 46.02°, and the average \( \overline{\chi_1^{\text{error}}} \) of the 14 protein calculated by using our modified energy function with weight set \{0.4, 1, 0.6\} is 44.81°. Therefore, the average \( \chi_1 \) improvements by using our energy functions are 5.66° and 6.87°. Details regarding \( \overline{\chi_1^{\text{error}}} \) is analyzed protein by protein and is shown in Figure 4.7.
Figure 4.7. $\bar{\chi}_1^{\text{error}}$ Comparison protein by protein. Average $\chi_1$ angle errors of our modified energy functions (weight $\{0, 1, 1\}$ and weight $\{0.4, 1, 0.6\}$), the original energy function (weight $\{1, 1, 0\}$), and the SCWRL software. Our modified energy functions overall have a slight improvement than the original SCWRL energy function, but still are not as good as the SCWRL software.
Our modified energy function with both weight sets have similar or better average $\chi_1$ angles than the Original energy function in most proteins except with protein 2FDN. However, the average $\chi_1$ angle error of using the SCWRL software is 31.36°, which is better than our prediction with both weight sets as shown in Figure 4.7.

$\chi_2$ Angle Error. The $\bar{\chi}_2$ error of the 14 proteins calculated by using the original energy function without EDM data (weight set $\{1, 1, 0\}$) is 28.92°, the average $\bar{\chi}_2$ error angle error of the 14 protein calculated by using our modified energy function with weight set $\{0, 1, 1\}$ is 31.24°, and the average $\bar{\chi}_2$ error of 14 protein by using our modified energy function with weight error $\{0.4, 1, 0.6\}$ is 30.53°. Therefore, there is no average improvement of our prediction in the $\chi_2$ angle scenario. Details about $\bar{\chi}_2$ error are analyzed protein by protein as shown in Figure 4.8. For the average $\chi_2$ comparison, all energy functions return a similar result. We do not have an improvement in this comparison, but both of our modified energy functions had better average $\chi_2$ angles in several proteins, and similar average $\chi_2$ angles in a number of proteins. The SCWRL software did a little better than other energy functions in this comparison with a 23.77° $\bar{\chi}_2$ error.

4.4.3 Energy

We calculated the energy $E(R)$ as the set up in Equation 3.1. We changed the self energy term in Equation 3.2 to:

$$S_{i}(r_{i}) = \sum_{i,j} \sum_{a \in r_i} \sum_{b \in B_j} E_{ab}(a, b)$$ (4.5)

We discarded the log-frequency term since it is a pseudo energy score for prediction and not the actual natural energy representation. The $E_{ab}(a, b)$ is the same as defined in Equation 3.3. We kept the pairwise energy term as defined in Equation 3.4.
Figure 4.8. $\bar{\chi}^2_{\text{error}}$ Comparison protein by protein. Average $\chi^2$ angle errors of our modified energy functions (weight $\{0, 1, 1\}$ and weight $\{0.4, 1, 0.6\}$), original energy function without EDM data (weight $\{1, 1, 0\}$), and the SCWRL software. All energy functions have similar average $\chi^2$ angles, SCWRL software does a little better prediction than others.
If there are more than one rotamer choice chosen for a certain residue, we want the average energy \( \overline{E(R)} \) of all possible sidechain conformations as shown in Equation 4.6.

\[
\overline{E(R)} = \frac{\sum_k E(R_k)}{N} \quad (4.6)
\]

For \( N \) possible sidechain conformations, \( E(R_k) \) is the energy of the \( k \)th possible sidechain conformation. In order to calculate this average energy more efficiently, we actually did the calculation as shown in Equation 4.7.

\[
\overline{E(R)} = \sum_i \sum_t \frac{S_i(r_{it})}{N_i} + \sum_{i<j} \sum_t \sum_l \frac{P_{ij}(r_{it}, r_{jl})}{N_i N_j} \quad (4.7)
\]

For each residue \( i \), \( S_i(r_{it}) \) is the self energy of the chosen rotamer \( r_{it} \), and \( N_i, N_j \) are the numbers of rotamers for residue \( i \) and \( j \). \( P_{ij}(r_{it}, r_{jl}) \) is the pairwise energy for the chosen pair of rotamers \( (r_{it}, r_{jl}) \) of residue \( i \) and residue \( j \).

The steps performed to obtain the expression of Equation 4.7 derived from Equation 4.6 is shown below:

\[
\overline{E(R)} = \frac{\sum_k E(R_k)}{N} = \frac{\sum_k \left\{ \sum_i S_i(r_i) + \sum_{i<j} P_{ij}(r_i, r_j) \right\}}{N} = \frac{\sum_i \sum_k S_i(r_i)}{N} + \sum_{i<j} \sum_k \frac{P_{ij}(r_i, r_j)}{N} \quad (4.8)
\]

Let \( m_i = \frac{N}{N_i} \), thus for each rotamer \( r_i \) at residue \( i \) which has \( N_i \) possible rotamers; therefore, there are \( m_i \) possible conformations according to this rotamer choice \( r_i \). Thus:
\[ \sum_k S_i(r_i) = m_i S_i(r_{i1}) + m_i S_i(r_{i2}) + \ldots + m_i S_i(r_{iN_i}) \]

\[ = m_i \sum_t S_i(r_{it}) \]

\[ = \frac{N}{N_i} \sum_t S_i(r_{it}) \] (4.9)

Similarly: let \( m_{ij} = \frac{N}{N_iN_j} \), then for each pair of rotamers \((r_i, r_j)\) (Note: there are \(N_i \times N_j\) such pairs.) There are \(m_{ij}\) possible conformations according to the pair of rotamers \((r_i, r_j)\). Thus:

\[ \sum_k P_{ij}(r_i, r_j) = m_{ij} \sum_t \sum_l P_{ij}(r_{it}, r_{ij}) \]

\[ = \frac{N}{N_iN_j} \sum_t \sum_l P_{ij}(r_{it}, r_{jt}) \] (4.10)

Apply Equation 4.9 and Equation 4.10 into Equation 4.8:

\[ \bar{E}(R) = \sum_i \sum_t \frac{1}{N} \frac{N}{N_i} S_i(r_{it}) + \sum_{i<j} \sum_t \sum_l \frac{1}{N} \frac{N}{N_iN_j} P_{ij}(r_{ik}, r_{jl}) \]

\[ = \sum_i \sum_t \frac{S_i(r_{it})}{N_i} + \sum_{i<j} \sum_t \sum_l \frac{P_{ij}(r_{it}, r_{jt})}{N_iN_j} \] (Equation 4.7)

For all proteins, both of our modified energy functions return the sidechain conformations with lower energies than that of the original energy function’s sidechain conformations, SCWRL software’s sidechain conformations, and the experimental sidechain conformations. Details are shown in Figure 4.9.
Figure 4.9. Energy comparison protein by protein. Energies of sidechain conformations returned by our modified energy function (weight \{0, 1, 1\} and weight \{0.4, 1, 0.6\}), original energy function without EDM data (weight \{1, 1, 0\}), SCWRL software, and the experimental structures.
Our modified energy function, which incorporates EDM data, has a consistent improvement in energies compared to an original energy function without EDM data. This observation proves our idea about incorporating the electron density map data into the prediction method in order to have a better prediction of protein sidechain conformation.

Our modified energy function does have a little improvement under the accuracy and $\chi_1$ angle error scenarios compared to the original energy function, especially with the weight set $\{0.4, 1, 0.6\}$. This observation also proves the idea about combining information can help the prediction. The SCWRL software added several terms in the original energy function, such as the electrostatic interaction between the rotamer and other parts of the protein and the hydrogen bond energy, which returns sidechain conformations with better accuracy and less $\chi$ angle errors, however those terms also results in sidechain conformations with higher energies.

A postulate of statistical physics is that the native state of a protein has minimal energy. Since our energy function gives that the experimentally solved structure has a higher energy than our prediction, this calls into question about our choice of energy function. Our original energy function is very crude since it only approximates van der Waals interaction energy, and does not, for example account the electrostatic interaction energy or solvent contributions. We believe that if the energy function is more faithful to the underlying physics, then our algorithm should return sidechain conformations with higher accuracy and lower $\chi$ angle errors. In addition, in this
thesis, the mesh and rotamer comparison is a very weak comparison which does not fully describe whether a rotamer perfectly fits into the iso-surface.
BIBLIOGRAPHY


