DEVELOPMENT OF A PREDICTION METHOD FOR AMPHIPATHIC α-HELICES FROM PROTEIN PRIMARY STRUCTURE

by

Mamta Bajaj

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professors Jitender S. Deogun, Etsuko Moriyama, and Hideaki Moriyama

Lincoln, Nebraska

May, 2005

DEVELOPMENT OF A PREDICTION METHOD FOR AMPHIPATHIC α-HELICES FROM PROTEIN PRIMARY STRUCTURE

Mamta Bajaj, M.S.

University of Nebraska, 2005

Advisor: Jitender S. Deogun, Etsuko Moriyama, and Hideaki Moriyama

Three-dimensional structures of proteins are directly related to their functions. Therefore, development of prediction methods for protein structures is one of the most studied areas in computational biology. The primary structure of proteins (the amino acid sequence) is folded into secondary structures (e.g., α-helices and β-sheets). Polypeptide chains with secondary structures are further folded into higher order three-dimensional structures. Predicting secondary structures is thus usually the first step for understanding the protein structures. Many secondary structure prediction methods have been developed. However, only a few methods are available for predicting amphipathic α-helices. Amphipathic α-helices have both hydrophobic and hydrophilic sides. These helices are often found at biologically active protein regions, usually at the surface areas where one side contacts the outside of the protein (aqueous in nature) and the other side faces the hydrophobic inside of the protein. Locating these helices helps in predicting protein functions, such as DNA-binding proteins, and also predicting the tertiary structure of proteins. Experimentally determined three-dimensional coordinate information is available in the Protein Data Bank (PDB). In order to utilize such information effectively for precise quantitative analysis, in this thesis, methods were developed for systematizing the secondary structural information and for identifying amphipathic α-helices based on protein structure information contained in the PDB. Using these methods developed, 556 α-helices with ten amino acids or longer were identified from 160 PDB protein entries. Among these α-helices, 26 were found

to be amphipathic. A simple set of statistics was developed to discriminate amphipathic α-helices from non-amphipathic α-helices by examining the distributions of hydrophobic to hydrophilic amino acid ratios. The difference between these statistics estimated from the data set containing only amphipathic α-helices and one without secondary structure was significant and could be used to predict amphipathic α-helices simply from protein primary structures (amino acid sequences).

ACKNOWLEDGEMENTS

I want to thank Drs. Etsuko Moriyama and Hideaki Moriyama for their continuous guidance and support throughout this thesis. I would also like to thank Dr. Jitender S. Deogun for advising me throughout this thesis and the MS program at CSE. Thanks to Dr. Stephen Scott for serving on my committee.

I would also like to thank my friends from Moriyama Lab, Pooja K. Strope, Cory Strope, Yavuz Yavuz, Stephen Opiyo, and Chendhore Sai Veerappan for sharing ideas and creating a fun environment for research at the lab. Thanks to my brother, Prashant Bajaj, for supporting me throughout the thesis. This thesis would not have been possible without the patience, encouragement, and support from my parents.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Outline

The structure of proteins is divided into four layers. Primary structure is the linear amino acid sequence of a polypeptide chain. Secondary structures include α-helices and β-sheets. Higher order structures (tertiary and quaternary structures) allow proteins to function. It is, therefore, important to understand structural details of proteins. Protein structures can be experimentally determined by nuclear magnetic resonance (NMR) and diffraction (e.g., X-ray) methods. However, these methods are time consuming and expensive. Furthermore, solving the structures of some types of proteins has been particularly difficult (e.g., transmembrane proteins). Therefore, prediction methods have been developed as alternative ways to obtain protein structure information.

Secondary structure prediction is by far the most actively studied and successful in the field of structural bioinformatics. Although many methods have been developed for secondary structure prediction, very few methods are available particularly for predicting amphipathic α-helices. Amphipathic α-helices are those that extend hydrophilic side-chains from one side and hydrophobic side-chains from the opposite side. This secondary structure

is often found in the biologically active proteins and peptides, e.g., DNA-binding proteins, usually at surface areas where one side faces the aqueous medium and the other side the internal area of the protein. Predicting the location of amphipathic α-helices is, therefore, an important step towards predicting functions of proteins. Identifying amphipathic α-helices also helps when predicting the tertiary structure of proteins.

Existing methods for identifying amphipathic α-helices include the "helical wheel diagram" [16], the "helical-net diagram" [4], and the "helical hydrophobic moment" [5]. Both of "helical wheel" and "helical-net" diagrams are simple two-dimensional visualization methods that identify amphipathic α-helices with hydrophobic and hydrophilic arcs along the wheel (helical wheel) or along the cylinder (helical-net). The "helical hydrophobic moment" numerically expresses the helical amphipathicity of a protein segment. This method detects the periodicity in hydrophobicity values by using discrete Fourier transform. Currently the "helical hydrophobic moment" is the only quantitative method that is applicable for prediction. This method, however, is not suitable for a large scale database search as the method is not length invariant and it is less 0 in defining the angle periodicity.

The main goal of this project was to develop a new method that predicts amphipathic α-helices given a primary protein structure (amino acid sequence). In order to achieve this goal following three major problems were required to be solved:

1. Systematizing the available structural information in the public database, so that it can be used for quantitative analysis;

2. Development of a method to identify amphipathic α-helices quantitatively from structural information; and

3. Development of statistics that can be used to discriminate amphipathic α-helices from non-amphipathic α-helices.

The Protein Data Bank (PDB) [1] is the database that contains information on experi-

mentally determined three-dimensional structures of proteins and other biological macromolecules. In addition to the physical coordinate data of atoms and the primary sequence information, the database also contains annotations including secondary structure information. However, these annotations are given by the researchers and are not defined consistently nor are written in a systematic format. Such inconsistency prevents us from performing precise quantitative analysis of structural data and from using the information efficiently for developing structural prediction methods.

My first objective of this thesis was, therefore, to determine the secondary structures from each protein structural coordinate data using a consistent definition and method. The two major secondary structures (α-helix and β-strand) were determined based on the exact calculation of rotation angles, ϕ and ψ, in each dipeptide. Once the exact positions of secondary structures within proteins were obtained, the second objective was to identify amphipathic α-helices from the available protein structural data. In general, amphipathic α-helices are objectively described by researchers depending on their interests in protein functions, and PDB annotations do not always contain such information. Thus, a new method had to be developed for defining and identifying this particular type of α-helix. In this thesis, amphipathic α-helices were identified entirely based on structural information and this is a very new approach. Using atomic coordinate information, each protein structure in PDB was reconstructed using a three-dimensional grid system. Based on the "neighboring" information for each cell in a protein, a "surface area map" was generated. This information enabled us to identify amphipathic α-helices using only structural data.

My third objective was to develop a set of statistics that can identify amphipathic α-helices by using only amino acid sequence information. A small data set including six each of positive and negative samples was collected manually from PDB based on biological annotations. The positive samples contained strictly amphipathic α-helices, and the negative samples were protein segments that did not contain α-helix or β-sheet. Based on a series

of preliminary analysis using this small data set, a set of simple statistics that can be used to identify amphipathic α-helices was developed. These statistics represent a bias in the distribution of hydrophobic amino acids along the amino acid sequence. Negative data set did not show any bias in the distribution of hydrophobic and hydrophilic amino acids.

My forth and final objective was to develop a discrimination method for predicting amphipathic helices based on the statistics developed before. Larger data sets were prepared using the secondary structure and amphipathic α-helices identification methods developed earlier. Independently a set of simulated binary sequences was prepared to represent all of the hydrophobic and hydrophilic sequence space for a certain length of peptides. This data set was used to examine the behavior of statistics in completely random peptide properties.

In summary, the outcomes from this thesis are: (1) protein structural information was systematized for precise quantitative analysis; (2) an identification method was developed for amphipathic α-helices based only on structural information; and (3) a set of statistics that can be used for discriminating amphipathic α-helices was developed.

1.2 Organization of the thesis

The remainder of this thesis is organized as follows. Chapter 2 describes the background information on proteins, protein structures, amphipathic α-helices, and their commonly used visualization and quantitative methods. This chapter also explains the format and problems associated with the public domain structural database, Protein Data Bank (PDB). In Chapter 3, methods are developed for identifying secondary structures and amphipathic α-helices based on structural information found in PDB. The preliminary analysis and development of statistics for predicting amphipathic α-helices using amino acid sequence information are discussed in Chapter 4. Finally Chapter 5 concludes this thesis with overall discussion and future works.

Chapter 2

Background

2.1 Proteins

Living organisms are composed of cells. A cell is constituted of about 70 percent of water, 15 percent of proteins, 7 percent of nucleic acids, as well as carbohydrates (3%), lipid (2%), inorganic minerals (1%), and other miscellaneous organic molecules [13]. Proteins are, therefore, the most abundant macromolecules in the cells. Proteins are designed to bind simple ions as well as large complex molecules such as sugars, fat, nucleic acids, and other proteins. They provide structural rigidity to the cell, regulate the concentrations of metabolites, control flow of material through membranes, and catalyze chemical reactions. Proteins also cause motion, act as sensors and switches, and control gene functions [11].

2.2 Protein structures

Amino acids are the biochemical building blocks of proteins. Twenty different amino acids are joined together by peptide bonds to form a polypeptide (Appendix A lists the 20 amino acids). The amino acid sequence of a polypeptide chain is the primary structure of a protein.

The formation of hydrogen bonds between amino acids in the polypeptide chain produces secondary structures: e.g., α-helices and β-sheets. An α-helix is a coiled (spring like) conformation of consecutive amino acids as shown in Figure 2.1. It contains 3.6 residues per turn. α-helices are most commonly found at the surface of the protein cores [2].A β-sheet

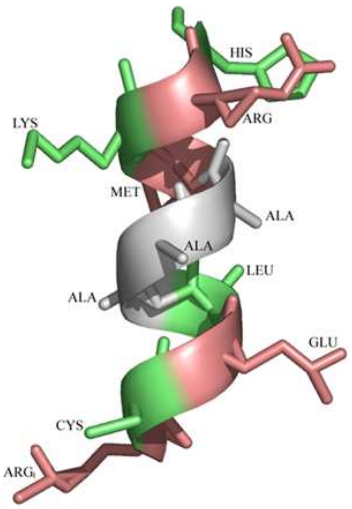


Figure 2.1: Amino acid arrangement in an α-helix of a lysozyme (PDB entry 1LYZ). The chemical structure of each amino acid is shown. PyMol [3] is used to create the 3D visualization.

is a pleated structure formed with two or more hydrogen bonded β-strands as shown in Figure 2.2. Unlike α-helices, which are comprised of residues from a continuous polypeptide segment, β-sheets are formed very often from β-strands that occur at distant portions of the polypeptide sequence. β-sheets could be present in either parallel or antiparallel forms based on the relative directions of two interacting β-strands.

More detailed explanation of α-helix and β-strand are given later in this section. In addition to these two major types, "turns" are also considered as part of secondary structure as it helps forming the major secondary structures. These secondary structures are linked by loops that lack secondary structure and then folded into a tertiary structure. The association of two or more polypeptide chains is called a quaternary structure. Most proteins found in nature have quaternary structures. The three dimensional structure of each protein allows

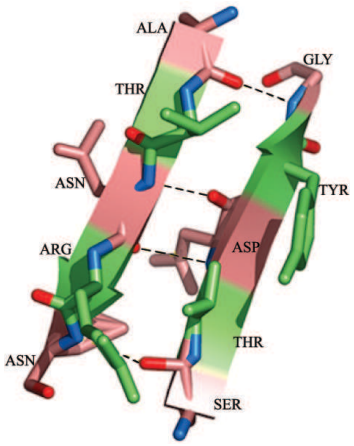


Figure 2.2: Amino acid arrangement in a β-sheet of a lysozyme (PDB entry 1LYZ). The dotted lines represent the hydrogen bonds holding two β-strands. PyMol is used to create this 3D visualization.

it to perform each unique function. The four layers of protein structures are summarized in Figure 2.3.

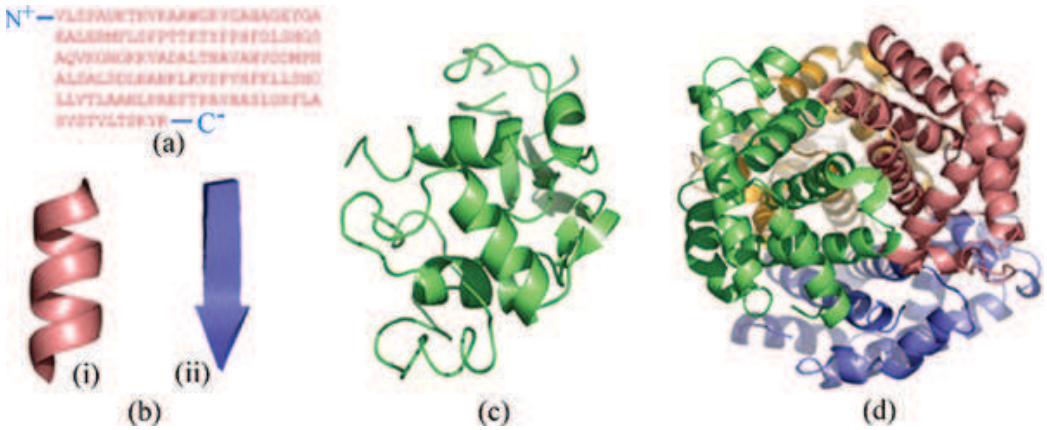


Figure 2.3: Four layers of protein structure. A primary structure (a), secondary structures ((b) (i) α-helix and (ii) β-strand), a tertiary structure (c), and a quaternary structure (d) are illustrated. PyMol is used to create these 3D visualization.

An amino acid is any molecule that contains both "amino" and "carboxylic acid" 0 groups as shown in Figure 2.4. Different amino acids are distinguished by their different side chains denoted by R. A side chain is a chemical structure in a polymer that projects from the repeating backbone. These side chains are bonded to the "α-carbon" of the back-

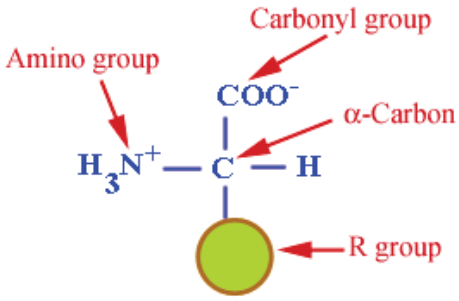


Figure 2.4: General structure of an amino acid, also called the "main chain" or the "backbone". The R group represents the "side chain" that is specific to each amino acid.

bone providing each amino acid its particular chemical identity. Appendix A shows the chemical structures of all 20 amino acids. The amino acids are linked linearly through peptide bonds, also called amide bonds (Figure 2.5). Peptide bonds are formed by a dehydration synthesis reaction between the carboxyl group of the first amino acid and the amino group of the second amino acid. Adding additional amino acids to the growing peptide chain produces a polypeptide chain.

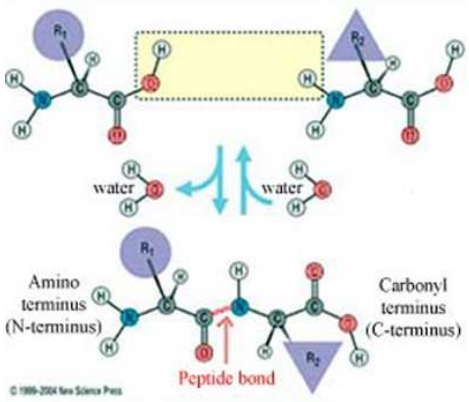


Figure 2.5: Peptide bond formation. Its formation requires loss of a water whereas hydrolysis (the opposite reaction) requires addition of a water molecule (courtesy [14]).

A peptide bond has a property that plays an important role in the rigidity and folding of a polypeptide chain. It has a partial double bond character (resonance structure) caused by the delocalization of bonding electrons rapidly moving between the oxygen and nitrogen atoms. This gives the C-N single bond (shown in the red line in Figure 2.5) a "partial

double bond" character. Because of this resonance structure, the carbonyl carbon, carbonyl oxygen, and amide nitrogen atoms are coplanar and the free rotation of the C-N bond is limited. However, the N-C_α (amide nitrogen and α-carbon) and C_α-C (α-carbon and carbonyl carbon) bonds are single bonds and free rotation around these bonds is allowed. The angle between two groups on either side of a bond is called torsion angle (also known as dihedral angle). By convention, the torsion angle for the N-C_α bond is called phi (ϕ), whereas the torsion angle of rotation around C_α-C bond is called psi (ψ). Figure 2.6 shows the relationships of torsion angles and dipeptides.

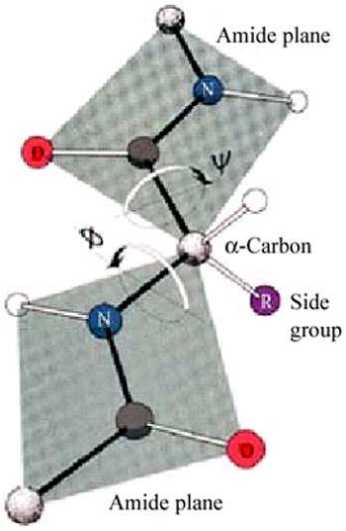


Figure 2.6: Torsion angles in a peptide unit. The rotations about the N-C_α bond is Phi (ϕ) and the C_α-C bond is Psi (ψ) (courtesy [14]).

As described before, certain repeating patterns of hydrogen-bonds between C=O and NH groups of amino acids in the polypeptide chain form either α-helices or β-strands. These different patterns can be identified on the combinations of torsion angles. It is an α-helix when the consecutive residues have the ϕ and ψ angle pair approximately -57° and -47° as defined by International Union of Pure and Applied Chemistry (IUPAC: Figure 2.7). α-helix can be either right-handed or left-handed depending on the screw direction of the chain. However, right-handed helices are more frequently observed in nature. β-strands

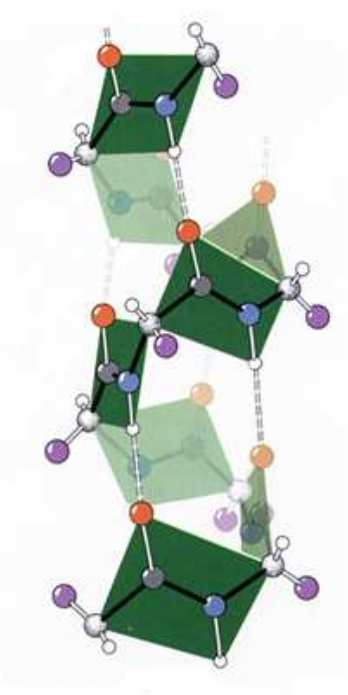


Figure 2.7: An α-helix, which has 3.6 peptide units per turn (courtesy [17]). Dashed lines represent the hydrogen bonds.

are, on the other hand, identified with torsion angles ϕ = -119° and ψ = 113° (as defined by IUPAC; Figure 2.8).The torsion angles are plotted on a conformational map called Ra-

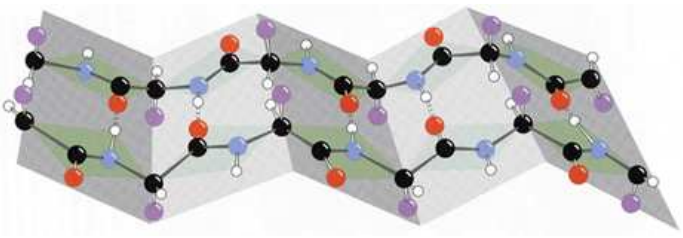


Figure 2.8: A two-stranded antiparallel pleated β-sheet. Dashed lines indicate hydrogen bonds (courtesy [17]).

machandran plot. Ramachandran plot for proteins is a useful visualization method that easily identifies secondary structures as shown in Figure 2.9.

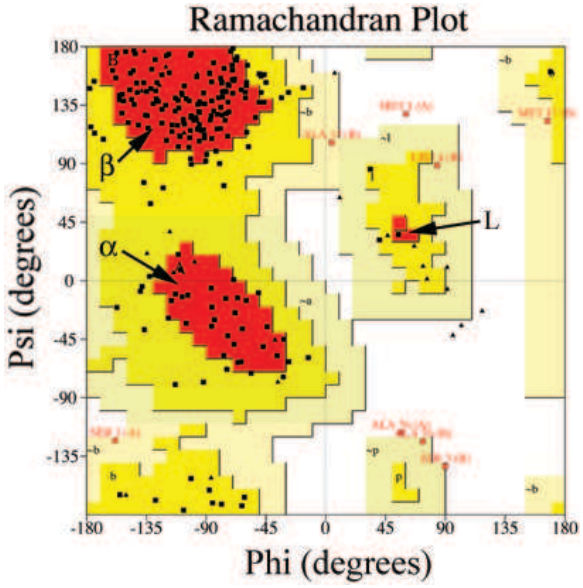


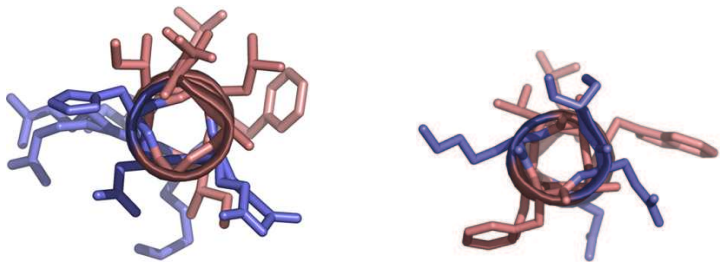
Figure 2.9: Ramachandran plot representing major secondary structures. The red areas 0 as α, β, and L correspond to conformational angles found for the right-handed α-helices, β-strands, and left-handed α-helices, respectively (courtesy [10]).

2.3 Amphipathic α-helix and its biological significance

The foldings and functions of proteins depend on the chemical characteristics of amino acid side-chains (the list of amino acids and their side chains is given in Appendix A). The chemical properties of amino acids play important roles in interactions between them and with water. The chemical properties of amino acids can be divided into two main categories: hydrophobic and hydrophilic (See Appendix A). There are eleven hydrophobic residues: alanine (A), cysteine (C), phenylalanine (F), glycine (G), isoleucine (I), leucine (L), methionine (M), proline (P), valine (V), tyrosine (Y), and tryptophan (W). These molecules are non-polar and uncharged, and they tend to avoid contacting water. The nature of these residues is the basis for the hydrophobic effect.The hydrophobic effect causes the polypeptide folded into a compact conformation. This results in minimizing the total hydrophobic surface area and allows van der Waals interactions between the hydrophobic groups. Hydrophobic residues are usually packed in the core of the protein.

On the other hand, there are nine hydrophilic amino acid residues: aspartic acid (D), glutamic acid (E), histidine (H), lysine (K), asparagine (N), glutamine (Q), arginine (R), serine (S), and threonine (T). These molecules are polar and charged in nature and they tend to interact with water through hydrogen bonds. The hydrogen bonding enables the molecule to dissolve in water. These two opposite characters of residues, hydrophobic and hydrophilic, enable a protein to assume its functional conformation.

An amphipathic molecule contains both hydrophobic and hydrophilic groups. Amphipathic α-helices extend hydrophilic side-chains from one side and hydrophobic side-chains from the opposite side. One example of such amphipathic α-helices is shown in Figure 2.10. It clearly shows that hydrophobic amino acids (shown in red) are located on one side of the α-helix and 0 amino acids (shown in blue) are located on the other side. Such distribution bias is not found in non-amphipathic α-helices.



(a) (b)

Figure 2.10: Vertical views of α-helix segments for (a) an amphipathic helix (generated from a PDB entry 1AJG: a myoglobin) and (b) a non-amphipathic helix (generated from a PDB entry 1H87: a lysozyme). PyMol is used to create these 3D visualization. Red represents the hydrophobic residues and blue represents the hydrophilic residues.

Due to the conformation of α-helix, the hydrophobic and hydrophilic residues are distributed 3-4 residues apart in the sequence, thus producing hydrophobic and hydrophilic faces. This kind of distribution can stabilize, for example, helix-helix packing found in lysozyme [14]. This arrangement of amino acids also allows the structure to create a barrier between aqueous and hydrophobic environments upon folding and therefore, amphi-

pathic α-helices frequently occur on the surface of proteins. Due to these unique chemical properties, amphipathic α-helices play various important structural and functional roles in proteins such as DNA-binding proteins, fibrous proteins, as well as receptor binding segments of polypeptide hormones, polypeptide venoms, and polypeptide antibiotics [7].

2.4 Methods to detect amphipathic α-helices

As mentioned before, currently there are only a few methods specifically developed for predicting amphipathic α-helices from amino acid sequences. The most commonly used methods, "helical wheel" [16] and "helical-net" [4] diagrams, rely on visualization techniques to detect amphipathic α-helices. "Hydrophobic moment" [5], on the other hand, is a widely used quantitative method to detect the amphipathic α-helices. All of these three methods assume that the α-helix region is already known, and particular conformational properties of amino acid sequences within an α-helix is utilized for the amphipathic α-helix detection. These three methods are described next.

2.4.1 "Helical wheel" and "helical-net" diagrams

These graphical methods project the three-dimensional structures of α-helices onto two-dimensional diagrams. One of the most commonly used such methods is "helical wheel" developed by Shiffer and Edmundson [16]. This method is simply based on the property of α-helix that there are 3.6 amino acid residues per complete turn (as shown in Figure 2.1). The angle between two residues is, therefore, 100°. "Helical wheel" visualizes an α-helix by looking down perpendicularly at the center and projecting the amino acids on a unit circle as shown in Figure 2.11. In this diagram, the residues on the α-helix will appear like a wheel. If the α-helix is amphipathic, hydrophobic residues position on one side of the wheel and hydrophilic residues on the other side. This method is very simple but has some

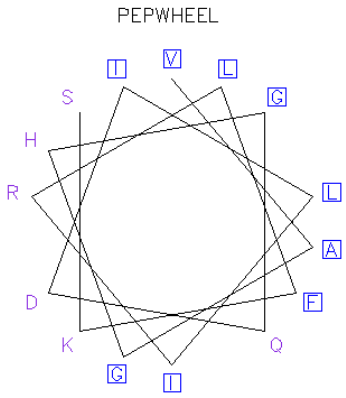
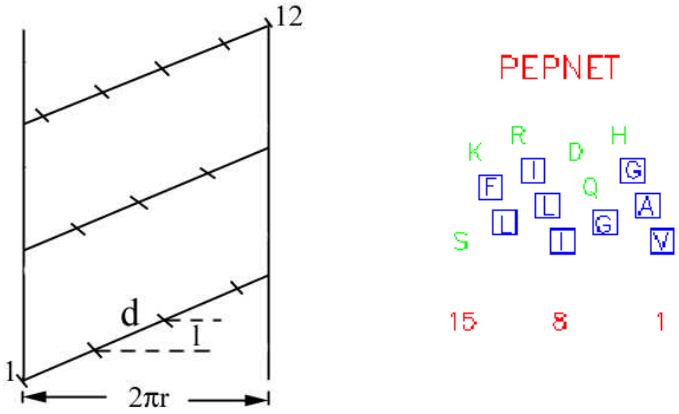


Figure 2.11: Helical wheel diagram. The pepwheel program from EMBOSS [15] is used to create the diagram of the myoglobin protein segment (1AJG). The amino acids non-polar are marked with squares.

drawbacks. It does not reflect the character of the amino acid residues in any detail (e.g., polar character, charged, or the size of the residue) or their longitudinal arrangement. To overcome these limitations, the "helical-net" diagram was developed.

The "helical-net" diagram developed by Dunhill [4] generates a longitudinal representation of the amino acids along the α-helix. An α-helix is represented as a cylinder with the residues winding around it. The radius of the cylinder (r) is the distance from the center of the α-helix to the α-carbon atom of the backbone. The amino acids are visualized as a graphical projection of the side chain positions wrapped around a cylindrical surface. The amino acids are positioned on the cylinder by calculating the distance (d) between the adjacent α-carbons and longitudinal shift (l) per residue as shown in Figure 2.12(a). The cylinder is then split open along a single line parallel to its axis and flattened into a rectangle. This arrangement gives the appearance of a net. The helical net diagram shows amphipathic α-helices with separated hydrophobic and hydrophilic arcs along the cylinder. An example is given in Figure 2.12(b).

The use of these visualization techniques is, however, limited as the hydrophilic and hydrophobic arc boundaries are not well defined. These diagrams are useful in detecting amphipathic α-helices where single short sequence is used and the amphipathic structure is



(a) (b)

Figure 2.12: Helical net diagram. The mechanism of helical net calculation is shown in (a). d represents the distance between the adjacent α-carbons, l represents the longitudinal shift per residue, and r is the radius of the cylinder. An example of the helical net diagram using the program pepnet from EMBOSS is shown in (b). The α-helix is the same myoglobin protein segment (1AJG) as used in Figure 2.11. The amino acids non-polar are marked with squares.

well defined. When considering long sequences and locating amphipathic α-helix regions, these graphical techniques are not easy to apply and often difficult to interpret the results.

2.4.2 Hydrophobic moment

One of the most common approaches to quantitatively detect amphipathic α-helices is the "hydrophobic moment" developed by Eisenberg, *et al.* [5]. "Hydrophobic moment" quantifies the property of amphipathic α-helices by combining a hydrophobicity scale with the "helical wheel." It avoids the visual interpretation problem of "helical wheel" and "helical net" diagrams by considering each amino acid as being represented by a vector whose direction points orthogonally out from the backbone and whose sign and magnitude are defined based on its hydrophobicity value. A mean of "net" vector, termed as the "hy-

drophobic moment", is then calculated as follows:

$$\langle \mu_H \rangle = |\sum_{i=1}^{N} \vec{H}_i|/N \qquad , \tag{2.1}$$

where $\vec{H}_i$ represents the vector that has the hydrophobicity value associated with the side chain of of the amino acid i within an α-helix and *N* is the length (number of amino acids) of the helix [5].

In a general form, the "hydrophobic moment" is defined using an angular frequency δ [6] and given in the following equation:

$$\mu_\delta = \left\{ \left[\sum_{i=1}^{N} H_i \sin(\delta i) \right]^2 + \left[\sum_{i=1}^{N} H_i \cos(\delta i) \right]^2 \right\}^{1/2} \qquad , \tag{2.2}$$

where H_i is the hydrophobicity of the residues and δ is the angular periodicity at which the successive side chains emerge from the backbone. It is calculated with $\delta = 2\pi/m$, where *m* is the number of residues per turn. For α-helices, *m* = 3.6 and δ = 100. Therefore, the conventional hydrophobic moment is computed as μ_{100} / *N*. In Figure 2.13, the hydrophobic moment is plotted along a 144 amino acid region (entire sequence length = 153) that includes a 15 amino acid amphipathic region (from positions 21 to 35, the same region used in Figures 2.11 and 2.12). The window size of ten amino acids (N = 10) is used for this plot. The region between the amino acid positions 21 and 35 has continuously high hydrophobic moments indicating the possible existence of an amphipathic helix.

The angular periodicity δ in the equation 2.2 is a variable that can take any value between 0 and 180 degrees. Then the hydrophobic moment is interpreted as the modulus of the discrete Fourier transform. A strong component of periodicity at δ is indicated by a large value of μ at a particular δ. Eisenberg *et al.* [6] thus used the equation 2.2 to examine the plot between μ and δ, called hydrophobic moment profile. This plot showed the large

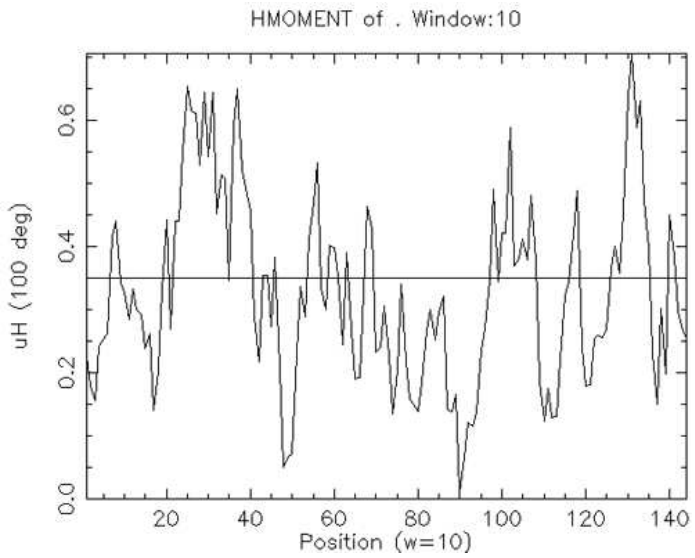


Figure 2.13: Graphical output of the hydrophobic moment plot. The angular periodicity of 100 degrees for α-helix and 10 amino acids for the window size (w) are used. The same α-helix region used in Figures 2.11 and 2.12 is included from positions 21 to 35 (the entire sequence includes 153 amino acids of a myoglobin, 1AJG). The program hmoment from EMBOSS is used to generate the plot.

maximum at δ = 100° for an amphipathic α-helix as expected.

The hydrophobic moment assumes the 100° angular periodicity for α-helices. However, due to various amino acid compositions and environmental factors affecting proteins, the angle frequently deviates from 100°, and it affects the identification of amphipathic helices by using this method.

Furthermore, Fourier transform based methods are, in general, not good for comparing sequences of different lengths. A short sequence will more likely reveal a periodic pattern by chance. Therefore, hydrophobic moment measures are not length invariant, and short sequences are more likely to have higher hydrophobic moment than longer sequences. In order to avoid this problem, usually hydrophobic moment is plotted using a window shifting procedure (as shown in Figure 2.13) with a fixed window size. But this produces another problem for deciding the optimal window size.

The definition of hydrophobic moment assumes that hydrophobic and hydrophilic residues

are assigned positive and negative values respectively. Some hydrophobicity scales have only positive values. Therefore, the choice of hydrophobicity scales could affect the results and sometimes an important periodicity could be masked from the analysis. Another problem associated with hydrophobic moment is choosing a criteria to decide the amphiphilicity of an α-helix. A cut-off boundary is usually calculated depending on the mean hydrophobicity. However, the choice of any cut-off value is arbitrary and some regions with high hydrophobic moments may not be identified depending on the cut-off value used.

2.5 Protein Data Bank (PDB)

The Protein Data Bank (PDB) was established by Brookhaven National Laboratories in 1971 as the single worldwide archive of structural data of biological macromolecules [1]. It contains the atomic information, general information required for all deposited structures and information specific to the method of structure determination.

The first section of each PDB entry is the title section (Figure 2.14). The title section of a PDB entry begins with a single line containing the identifier HEADER and continues until the end of the lines containing the identifier REMARK. The HEADER record uniquely identifies a PDB entry with the ID (1AJG in Figure 2.14). The COMPND record describes the macromolecular contents of an entry and includes the molecule name, synonyms, and other detailed specifications relevant to functions of the macromolecule. The AUTHOR record contains the names of the people responsible for the contents of the entry. After this usually there are multiple lines of the REMARK records. These lines present experimental details, annotations, comments, and information not included in the other records. REMARK 2, for example, shows the resolution in angstroms. REMARKs 4-999 are used to include free text annotation.

Figure 2.15 shows the remaining part of the same PDB entry. The first section after the

```
HEADER    OXYGEN TRANSPORT                        02-MAY-97   1AJG

COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MYOGLOBIN;
COMPND   3 CHAIN: NULL;
COMPND   4 BIOLOGICAL_UNIT: MONOMER;
COMPND   5 OTHER_DETAILS: CARBONMONOXY MYOGLOBIN, CO LIGAND BOUND TO
COMPND   6 FE OF THE HEME, HEME BOUND TO NE2 OF HIS 93
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: PHYSETER CATODON;
SOURCE   3 ORGANISM_COMMON: SPERM WHALE
KEYWDS    OXYGEN TRANSPORT, RESPIRATORY PROTEIN, HEME
    * *
AUTHOR    T.Y.TENG,V.SRAJER,K.MOFFAT
    * *
    * *
REMARK   1 REFERENCE 1
REMARK   2
REMARK   2 RESOLUTION. 1.7  ANGSTROMS.
    * *
    * *
REMARK   4
REMARK   4 1AJG COMPLIES WITH FORMAT V. 2.2, 16-DEC-1996
REMARK 200
REMARK 200 EXPERIMENTAL DETAILS
REMARK 200  EXPERIMENT TYPE                : X-RAY DIFFRACTION
REMARK 200  DATE OF DATA COLLECTION        : 30-JUN-1993
REMARK 200  TEMPERATURE          (KELVIN) : 40
REMARK 200  PH                             : 6.0
REMARK 200  NUMBER OF CRYSTALS USED        : 1
    * *
    * *
REMARK 200 METHOD USED TO DETERMINE THE STRUCTURE: DIFFERENCE FOURIER
REMARK 200 SOFTWARE USED: X-PLOR
REMARK 200 STARTING MODEL: PDB ENTRY 1MBO
REMARK 200
REMARK 200 REMARK: DATA WERE COLLECTED IN DARK USING AN OPEN FLOW
REMARK 200 NITROGEN/HELIUM CRYOSTAT FOR COOLING.
    * *
    * *
REMARK 500 THE FOLLOWING ATOMS THAT ARE RELATED BY CRYSTALLOGRAPHIC
REMARK 500 SYMMETRY ARE IN CLOSE CONTACT.  SOME OF THESE MAY BE ATOMS
REMARK 500 LOCATED ON SPECIAL POSITIONS IN THE CELL.  ATOMS WITH
REMARK 500 NON-BLANK ALTERNATE LOCATION INDICATORS ARE NOT INCLUDED
REMARK 500 IN THE CALCULATIONS.
    * *
    * *
```

Figure 2.14: A part of the title section of a PDB entry, 1AJG

REMARK records is the primary structure section. SEQRES records contain the amino acid or nucleic acid sequence of residues in each chain of the macromolecule concerned. The secondary structure section (including HELIX, SHEET, and TURN) describes helices, sheets, and turns found in protein and polypeptide structures. Finally, the coordinate section contains the collection of atomic coordinates as well as the MODEL records for the proteins. The ATOM records present the atomic coordinates for standard residues. Sometimes there are atoms missing from the coordinate information and usually (but not always) such missing atoms are mentioned in the REMARKS record. The format of the ATOM records is given in Figure 2.16. In this thesis, all of the atomic coordinate information were obtained from the ATOM records in each PDB entry. For more detailed information on PDB file format, see Appendix B.

```
DBREF  1AJG     1   153  SWS    P02185   MYG_PHYCA        1    153
SEQRES   1   153  VAL LEU SER GLU GLY GLU TRP GLN LEU VAL LEU HIS VAL
SEQRES   2   153  TRP ALA LYS VAL GLU ALA ASP VAL ALA GLY HIS GLY GLN
SEQRES   3   153  ASP ILE LEU ILE ARG LEU PHE LYS SER HIS PRO GLU THR
SEQRES   4   153  LEU GLU LYS PHE ASP ARG PHE LYS HIS LEU LYS THR GLU
SEQRES   5   153  ALA GLU MET LYS ALA SER GLU ASP LEU LYS LYS HIS GLY
SEQRES   6   153  VAL THR VAL LEU THR ALA LEU GLY ALA ILE LEU LYS LYS
SEQRES   7   153  LYS GLY HIS HIS GLU ALA GLU LEU LYS PRO LEU ALA GLN
SEQRES   8   153  SER HIS ALA THR LYS HIS LYS ILE PRO ILE LYS TYR LEU
SEQRES   9   153  GLU PHE ILE SER GLU ALA ILE ILE HIS VAL LEU HIS SER
SEQRES  10   153  ARG HIS PRO GLY ASP PHE GLY ALA ASP ALA GLN GLY ALA
SEQRES  11   153  MET ASN LYS ALA LEU GLU LEU PHE ARG LYS ASP ILE ALA
SEQRES  12   153  ALA LYS TYR LYS GLU LEU GLY TYR GLN GLY
  **
  **
HELIX    1   A SER      3  GLU     18  1                                  16
HELIX    2   B ASP     20  SER     35  2                                  16
HELIX    3   C HIS     36  LYS     42  3                                   7
  **
  **
ATOM      1  N   VAL     1      -3.710  15.394  13.509  1.00 21.43           N
ATOM      2  CA  VAL     1      -3.436  15.765  14.923  1.00 20.18           C
ATOM      3  C   VAL     1      -2.648  14.656  15.604  1.00 18.66           C
ATOM      4  O   VAL     1      -3.101  13.491  15.671  1.00 19.96           O
ATOM      5  CB  VAL     1      -4.738  16.044  15.723  1.00 20.85           C
ATOM      6  CG1 VAL     1      -4.419  16.271  17.199  1.00 20.94           C
ATOM      7  CG2 VAL     1      -5.434  17.275  15.166  1.00 21.23           C
ATOM      8  N   LEU     2      -1.470  15.007  16.112  1.00 14.27           N
ATOM      9  CA  LEU     2      -0.657  14.020  16.779  1.00 11.76           C
ATOM     10  C   LEU     2      -1.232  13.824  18.164  1.00 10.73           C
  **
```

Figure 2.15: The primary structure, secondary structure, and coordinate sections of a myoglobin PDB entry, 1AJG.

Note that the PDB records are structured, but there are also a lot of flexibility in the format as well as in the information content. For example, the keyword “AMPHIPATHIC” may be found in the KEYWDS (keywords) record or embedded somewhere within the REMARK records. Furthermore, inclusion of such biological or biochemical information totally depends on researchers who submit the structural data.

COLUMNS	DESCRIPTION
1 – 6	Record name ATOM
7 – 11	Atom serial number
13 – 16	Atom name
17	Alternate location indicator
18 – 20	Residue name
22	Chain identifier
23 – 26	Residue sequence number
27	Code for insertion of residues
31 – 38	Orthogonal coordinates for X in Angstroms
39 – 46	Orthogonal coordinates for Y in Angstroms
47 – 54	Orthogonal coordinates for Z in Angstroms
55 – 60	Occupancy
61 – 66	Temperature factor

Figure 2.16: The format of the ATOM records in PDB.

Chapter 3

Development of structural recognition methods

PDB entries contain annotations including secondary structure information. However, how such information is determined totally depends on the researchers who submit the structural data. It is desirable, therefore, to identify secondary structures from atomic coordinate data for each protein using consistent definitions and methods before we perform quantitative analysis on structural data. Similarly, amphipathic α-helices need to be identified from structural information of each protein. In this chapter, development of these two methods are described.

3.1 Method for identifying secondary structures

3.1.1 Defining the α-helix and β-strand

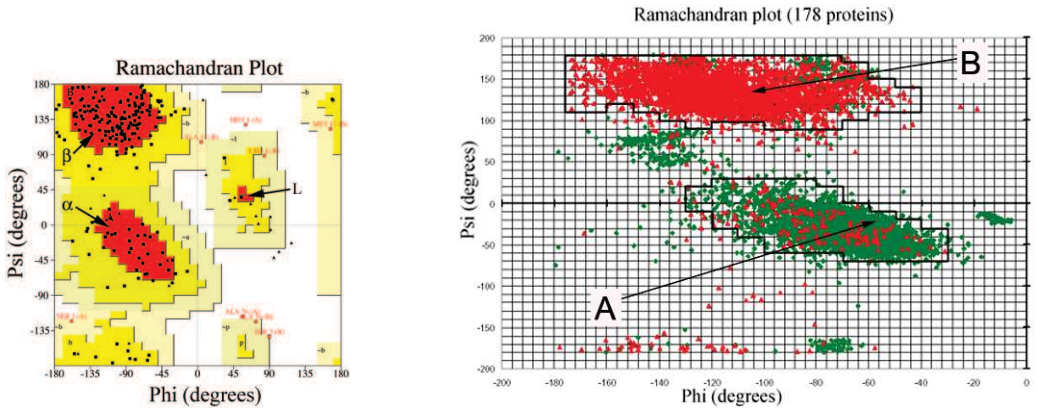
Before identifying secondary structures, their good definitions were required. The two major secondary structures: α-helices and β-strands can be determined based on the exact calculation of rotation angles, ϕ and ψ, within each two consecutive amino acids as de-

scribed in Chapter 2 (section 2.2). Certain combinations of ϕ and ψ torsion angles define α-helix and β-strand. As described earlier, for example, the definition by IUPAC identifies amino acids with (ϕ, ψ) = (-57°, -47°) as α-helices and (-119°, 113°) as β-strands. However, such single-point definition is too rigid, and in practice, torsion angles obtained from the actual protein data are continuous. It is thus necessary to define α-helices and β-strands with torsion angle combinations with some ranges allowed. In order to find such allowable ranges of torsion angles, it is useful to examine the real protein structural data and the actual distribution of torsion angles. Such torsion angle distributions were obtained from 121,870 residues from 463 known x-ray protein structures by Morris *et al.* [12]. Based on their study, "core", "allowed", and "generous" (or "disallowed") distributions of ϕ-ψ combination have been identified for secondary structures. In Figure 3.1(a), these distributions are given as a density contour map on a Ramachandran plot.

The areas shaded in red in Figure 3.1(a) are the most favorable ϕ-ψ "core" areas for α-helices and β-strands. With high resolution structures, over 90% of the residues should be located in these most favored regions. These red "core" areas shown in Figure 3.1(a) were used with some adjustments to define α-helices and β-strands for this study. The amino acids that were not classified as α-helices and β-strands were classified as "non-structure" residues. The precise ranges of torsion angles used to determine α-helices and β-strands are listed in Appendix C and these ranges are illustrated in Figure 3.1(b) with black boundaries.

3.1.2 Identifying α-helices and β-strands from PDB records

The atomic coordinate information available in PDB protein entries was used for calculating torsion angels. In this study, only monomer (single subunit) proteins were used for simplicity. As described in Chapter 2, proteins can form quaternary structures containing more than one polypeptide chains (subunits), and such proteins are called multimers (e.g.,



(a) (b)

Figure 3.1: Torsion angles and secondary structure definitions. (a) The most favorable "core" areas for α-helices and β-strands determined by Morris *et al.* [12] (figure obtained by PROCHECK [10]) are shown as the high density "core" area with red in this Ramachandran plot. The labels α, β, and L point the areas for right-handed α-helices, β-strands, and left-handed α-helices, respectively. (b) The Ramachandran plot obtained from the 178 monomer proteins used in this study. The plot shows only ϕ angles greater than zero. Green and red dots represent the residues annotated as α-helices and β-strands, respectively, in the original PDB records. The regions A and B, each surrounded by a boundary, depict the ψ/ϕ areas used to identify α-helices and β-strands, respectively, in this study.

two subunit proteins are called dimers). Such multimers were excluded from this study. 160 monomer proteins were obtained from the June 2004 release of PDB.

Using the atomic coordinate information (ATOM records) from each PDB entry, the ϕ and ψ torsion angles were calculated from nitrogen (N), α-carbon (C^{α}), and carbon (C) atoms for each consecutive two amino acids (dipeptide) as shown in Figure 3.2. In order to calculate the torsion angle ϕ of amino acid 2, C of the amino acid 1 (C_1), and N, C^{α}, and C of amino acid 2 (N_2, C_2^{α}, and C_2) are considered, whereas the calculation of the ψ angle requires N, C^{α}, and C of the amino acid 2 (N_2, C_2^{α}, and C_2) and N of the amino acid 3 (N_3) as shown in Figure 3.2. In general, the torsion angle, τ, is an angle defined by four atoms: i, j, k, and l as shown in Figure 3.3. It is the angle between two intersecting planes (τ in the figure). In order to calculate the torsion angle ϕ, let i, j, k, and l represent atoms C_1, N_2,

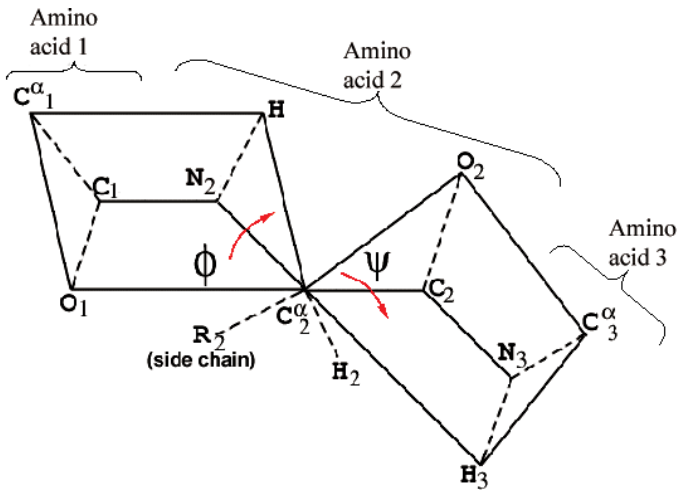


Figure 3.2: Atoms used for torsion angles. N, C, C^{α}, H, and O represent a nitrogen, a carbon, α-carbon, a hydrogen, and oxygen, respectively. ϕ and ψ are the two torsion angles. The atoms on the solid lines are used for calculating torsion angles. The numbers 1, 2, and 3 with atomic symbols represent which amino acid, amino acids 1, 2, or 3, they belong to.

C_2^{α}, and C_2, respectively . Calculation of the ψ angle involves N_2, C_2^{α}, and C_2, and N_3, but representing these four atoms with i, j, k, and l similarly, the procedure becomes the same. As shown in Figure 3.3, **a** is the vector from the atom i to the atom j, **b** is the vector from the atom j to the atom k, and **c** is the vector from the atom k to the atom l. The vector $\mathbf{n}_{ab}$ $= \mathbf{a} \times \mathbf{b}$ is perpendicular to the plane defined by the three atoms i, j, and k. This vector $\mathbf{n}_{ab}$ is called the *normal* of the plane (also known as the normal vector). $\mathbf{n}_{bc} = \mathbf{b} \times \mathbf{c}$ is the *normal* of the plane defined by atoms j, k, and l. Then the torsion angle is calculated as τ $=\cos^{-}1((\mathbf{n}_{ab} \cdot \mathbf{n}_{bc}) \,/\, (\mathbf{n}_{ab}\, \mathbf{n}_{bc}))$.

Using the calculated torsion angles and based on the table given in Appendix C, for each amino acid, the secondary structure (α-helix, β-strand, or non-structure) was identified. Table 3.1 shows an example of calculated torsion angles and their secondary structure classifications. The secondary structures identified by the new method ("New" in the table) are located approximately in the same regions as annotated by researchers in the PDB entry ("PDB" in the table). All other PDB entries used in this study consistently showed such

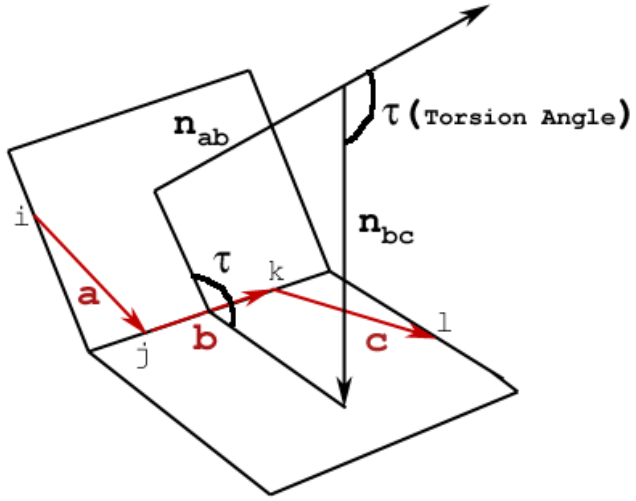


Figure 3.3: Representation of the torsion angle (τ) and four atoms i, j, k, and l. **a**, **b**, and **c** are three vectors. The vector $\mathbf{n}_{ab}$ represents the normal of the plane defined by the vectors **a** and **b**. The vector $\mathbf{n}_{bc}$ represents the normal of the plane defined by the vectors **b** and **c**.

approximate match between the two annotations. Differences seen were shifts of secondary structure regions by one or two amino acids.

The distribution of torsion angles calculated from the 178 proteins used in this study were densely clustered in two major areas A and B as shown in Figure 3.1(b). The secondary structure annotations obtained from the original PDB entries (as presented with red and green dots) show some overlapping distributions of torsion angles for α-helices and β-strands and there are a few small islands with lower densities. The definitions used in this study (illustrated with black boundaries) did not include these small islands, and the overlapped possibilities of α-helices and β-strands were ignored. If ϕ-ψ values of an amino acid were within the defined areas A and B in Figure 3.1(b), they were classified as α-helix or β-strands. The ϕ-ψ values outside of these areas were classified as "non-structure".

After identifying the secondary structure for each amino acid, in order to incorporate flexibilities, some adjustments were attempted. A single amino acid with the β-strand or non-structure identifier was frequently observed within a stretch of α-helix amino acids. The same situation was found also with a single β-strand identifier. At first, the identifier of

Table 3.1: Torsion angles and secondary structure classifications

Position	Amino acid	Torsion angles		Secondary structure annotations	
		ϕ	ψ	New	PDB
1	VAL	-	121.640	non	non
2	LEU	-78.261	151.277	non	non
3	SER	-83.083	171.941	non	alpha
4	GLU	-63.630	-38.395	alpha	alpha
5	GLY	-60.816	-41.211	alpha	alpha
6	GLU	-65.995	-42.917	alpha	alpha
7	TRP	-63.241	-36.440	alpha	alpha
8	GLN	-64.436	-37.872	alpha	alpha
9	LEU	-67.185	-39.774	alpha	alpha
10	VAL	-61.682	-50.959	alpha	alpha
11	LEU	-75.547	-23.652	alpha	alpha
12	HIS	-66.337	-46.546	alpha	alpha
13	VAL	-72.228	31.715	alpha	alpha
14	TRP	-66.836	-32.082	alpha	alpha
15	ALA	-62.137	-31.926	alpha	alpha
16	LYS	-70.247	-34.044	alpha	alpha
17	VAL	-65.414	-36.123	alpha	alpha
18	GLU	-61.946	-14.752	alpha	alpha
19	ALA	-76.035	-21.599	alpha	non

The table shows only a part of the sequence from a PDB entry 1AJG.
Secondary structure annotations from the original PDB record and the new annotations given in this study are listed. “alpha”: α-helix, “non”: non-structure.

the single non-α-helix amino acid was switched to the α-helix when it was found within a α-helix region. Similarly, if there was a single α-helix or a single non-structure amino acid located within a stretch of β-strand amino acids, then the identifier of such non-β-strand amino acids were switched to the β-strand. However, this adjustment approach generated unreasonably long α-helix and β-strand regions in a protein. Therefore, we decided not to use this adjustment technique.

On the other hand, very short α-helix or β-strand regions cannot be considered as structurally significant and such regions cannot contribute to functions of a protein. Therefore, any consecutive amino acid regions with the α-helix or β-strand identifier were considered as part of non-structure regions if such regions are three amino acids or shorter and if these

short regions were surrounded by amino acids with the “no structure” identifier.

3.2 Method for identifying amphipathic α-helices

After obtaining the exact positions of secondary structures within the proteins, the next step was to identify amphipathic α-helices. As described before, biological information included in the PDB annotations depends largely on the researcher’s interest, and even if amphipathic α-helices exist in a protein, it is not always mentioned. In this study, a new method was developed to identify this particular type of α-helices directly based on the atomic coordinate information available in PDB protein entries. Using the atomic coordinate information, each protein structure in the PDB was reconstructed using a three-dimensional grid system. A “surface area map” was generated by identifying amino acids at the protein surface based on the “neighboring” information of atoms located in each grid cell. This “surface area map” enabled us to identify amphipathic α-helices as in this type of α-helices, hydrophilic amino acids tend to appear on the protein surface and hydrophobic amino acids away from the surface.

3.2.1 Modeling the protein structure using unit cubic cells

Consider a protein molecule. The ranges of x, y, and z coordinates can be found from the PDB coordinate data: $x_1 < x < x_2$, $y_1 < y < y_2$, and $z_1 < z < z_2$, where x_1, y_1, and z_1 are the minimum values of x, y, and z-coordinates, respectively, and x_2, y_2, and z_2 are maximum values of x, y, and z-coordinates, respectively, found in the protein structure. The unit of these x, y, and z-coordinates is Å. A cubic container that is large enough for the protein molecule can be represented by the following eight coordinates: (x_1, y_1, z_1), (x_2, y_1, z_1), (x_1, y_2, z_1), (x_2, y_2, z_1), (x_1, y_1, z_2), (x_2, y_1, z_2), (x_1, y_2, z_2), and (x_2, y_2, z_2). Figure 3.4 shows the protein molecule put in the cubic container. The depth (X), width

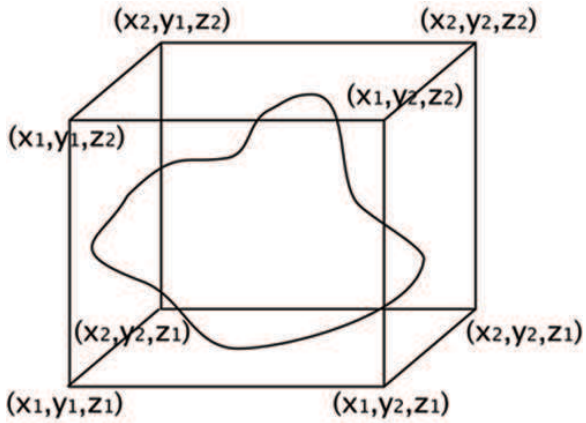


Figure 3.4: Representation of a protein molecule in the cubic container.

(Y), and height (Z) of this cubic container can be calculated as $X = x_2 - x_1$, $Y = y_2 - y_1$, and $Z = z_2 - z_1$. 3Å of a margin space was added to the all six sides of the protein. This increased the lengths of X, Y, and Z by 6Å. The new depth, width, and height of the container are called X', Y', and Z', where $X' = X + 6$, $Y' = Y + 6$, and $Z' = Z + 6$. Next, this cube was sliced with 1Å interval along the z-coordinate as shown in Figure 3.5. The

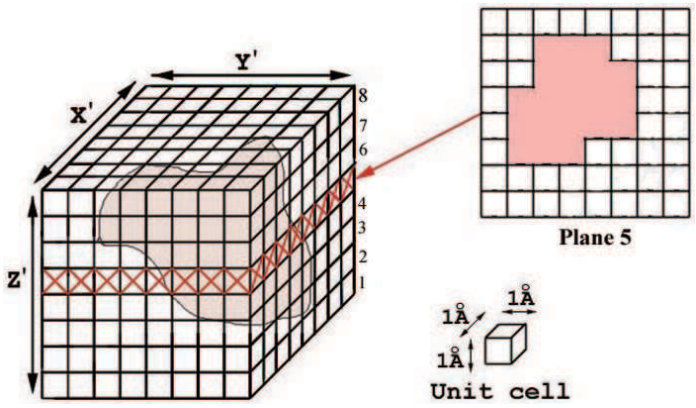


Figure 3.5: Slicing a cubic container containing a protein molecule. The matrix plane on the right is the two-dimensional representation of the slice 5 viewed from the above. The shaded area in the matrix shows the shape of the protein molecule in this plane using the grid cell as a unit.

small 1Å $\times$ 1Å $\times$ 1Å cube is called a "unit cell" or just simply a "cell". The number of cells contained in each thin slice with 1Å height is X' $\times$ Y'. The number of slices obtained

from the cubic container is Z'.

The structure of the protein molecule can be reconstructed by finding the cells that are occupied by any atoms belonging to the protein. Considering the margin space at the bottom of the z-coordinate, the lowest value of the z-coordinate is $z_1 - 3$. Starting from the xy plane at $z = z_1 - 3$, the first slice includes all of the atoms located between $z = z_1 - 3$ and $z = z_1 - 2$. All of these atoms are projected and plotted on a two-dimensional xy plane. Thus each slice can be viewed as a two-dimensional plane where atoms contained within each slice are projected on. This two-dimensional matrix view of each slice is called the matrix plane. This process is illustrated in Figure 3.6. Within each slice, the number

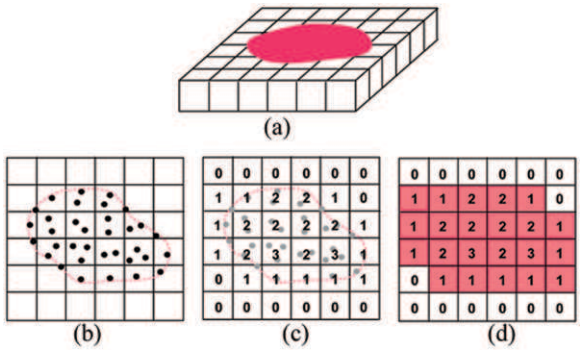


Figure 3.6: Reconstruction of a protein slice at the atomic level using a matrix plane. (a) Shaded area on the slice illustrates the shape of the protein molecule within this slice. (b) The dots represent the atoms in the slice projected on the two-dimensional matrix plane viewed from the above. (c) The number in each cell represents the number of atoms occupying the cell. (d) The shaded area represents the approximated shape of the protein in this slice.

of atoms located is counted for each cell, the number of atoms counted for each cell is illustrated in Figure 3.6 (b and c). After all of the atoms located in a slice were searched, the distribution of the non-0 number on the matrix plane shows the approximated shape of the protein in this slice (in Figure 3.6d).

In reality, each atom is not simply a point, but it occupies a certain area, which can be considered as a circle on a two-dimensional plane. Therefore, atomic radii need to be taken into consideration when atoms are plotted. The atomic radius is the distance

between the outermost region where an atom can occupy and its nucleus. These radii were used to determine the average bond length between two atoms. In general, amino acids were consisted of five basic atoms: carbon (C), nitrogen (N), oxygen (O), sulfur (S), and hydrogen (H). Since atomic information of hydrogen is not available in PDB, hydrogen was ignored from the calculations. The atomic radii used in this study are listed in Appendix D.

If the radius of an atom is 1Å and this atom is at the center of the cell, then nine cells on a plane will be covered by this "atomic area" as shown in Figure 3.7(a). The size of the

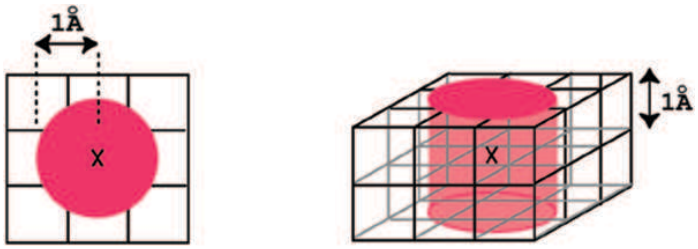


Figure 3.7: (a) An atom at the position "x" in a plane covers nine cells shown by shaded area. (b) An atom at the position "x" has a three-dimensional atomic area (shaded area) in the shape of a cylinder, which covers 18 cells across two slices.

"atomic area" depends on the atomic radius, and hence the number of cells covered by the atomic area varies among atoms. If a cell is covered even with any small part of an atomic area, it is considered to be covered. In a more realistic model, the atomic area should be considered in a three-dimensional space. As a simpler model, each three-dimensional atomic area was considered as a cylinder, instead of a sphere, as shown in Figure 3.7(b). Therefore, the same number of cells is covered on each slice the atomic cylinder covers. In the example of Figure 3.7(b), if the atom is located at the "x" (between the two slices), and if the radius of the atom is 1Å, the two slices are covered by this atomic cylinder, since the number of cells covered by this atom on each slice is consistently nine cells, and a total 18 cells are considered to be covered by this single atom.

3.2.2 Finding the surface cells from the protein molecule modeled

After reconstructing the structure of the protein using the grid system, a “surface area map” was generated by searching the cells locating at the surface of the protein molecule. Protein molecules exist always with surrounding water molecules. Closely bound water molecules around a protein are, therefore, considered as a part of the protein structure, and these water molecules indeed stabilize the protein structure. A unit cell is considered at the surface if it contacts with water molecules, in another words, if a water molecule can occupy the empty spaces around the cell. The diameter of a water molecule (H_2O) is approximately 2.8Å. Based on this, a “probe” of 3Å in diameter was chosen.

The “surface cell search” was done in the following three steps. The first step is the smoothing process. The purpose of this process was to smoothen the shape of the protein structure on each plane by using a probe of the size equivalent to a water molecule. In the next step, the cells in a matrix plane are further examined to verify if the existing empty space within a structure is connected to the surface of the protein structure. Finally, the “surface area map” is generated based on the “neighboring” cell information for each cell in the protein.

In the first step, cells that are occupied with any atoms were found. All of the six directions (four directions on the same plane as well as three “up” and three “down” cells in the neighboring six planes) of each occupied cell were searched. This was done to identify small empty spaces around the cell and to smooth out the structure of the protein. Consider that the cell marked with “x” in Figure 3.8 was found to be occupied by atoms. The numbers 1, 2, and 3 show the three neighboring cells on the right in the same plane. If the neighboring cell 3 is occupied (shaded in the figure) and the cells 1 and 2 are empty as shown in Figure 3.8(a), the empty space between the cells “x” and 3 is 2Å and shorter than the probe (3Å). Similarly, if the neighboring cells 2 and 3 are both occupied as shown in Figure 3.8(b), there is only 1Å of the empty space between the cells x and 2. In both

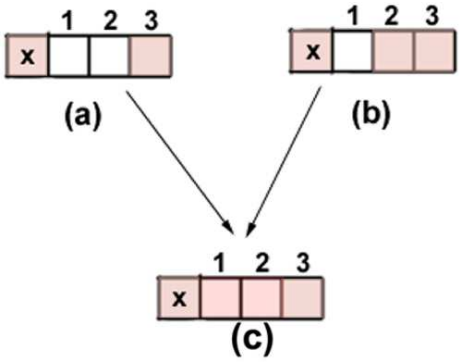


Figure 3.8: Representation of the smoothing process. Shaded boxes represent the cells occupied with any atoms. The cell marked with "x" represents the cell to be examined. The cells numbered with 1, 2, and 3 are the three neighboring cells on the right. In the above two cases shown in (a) and (b), the 3Å-diameter probe cannot enter the space next to the cell x. For both cases, the empty cells next to the cell x will be filled in as shown in (c). In this figure, the process of smoothing is described only for one direction, the right direction for the cell "x".

cases, there is not enough empty space where the probe can be entered. In these cases, the status of the "empty" cells (the cells 1 and 2 in Figure 3.8(a) and the cell 1 in Figure 3.8(b)) were switched to "occupied" as shown in Figure 3.8(c). This process of searching not big-enough spaces and switching the "empty" cell status to "occupied" was repeated for all the six directions. After going through all of the occupied cells on the plane, all of the "too small spaces" were filled in and as a result, the structure of the protein slice would be smoothed out. Figure 3.9 shows a matrix plane before and after the smoothing process. The red cells were "occupied" originally, and the blue cells were filled in after the smoothing process. This process was applied to all of the planes from the protein molecule.

In the second step, all the cells were further verified by examining especially if any empty spaces surrounded by the non-empty cells are connected to the outside. Such "internal space" areas are found on a matrix plane as shown in Figure 3.10 (areas 1, 2, and 3). The "internal space" areas are the empty areas (marked with 0 digits) surrounded by non-0 digit markers from the four directions on the same plane (not considering the diagonal directions). Even if these internal space areas look "closed" on the two-dimensional plane, if

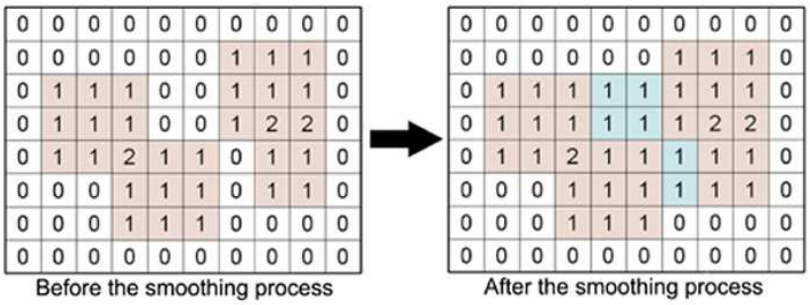


Figure 3.9: Smoothing process of the protein structure on a plane. The red cells were occupied before smoothing process running the probe. After running the probe over the plane, the blue cells were filled in and the structure of the protein in this matrix plane was smoothen.

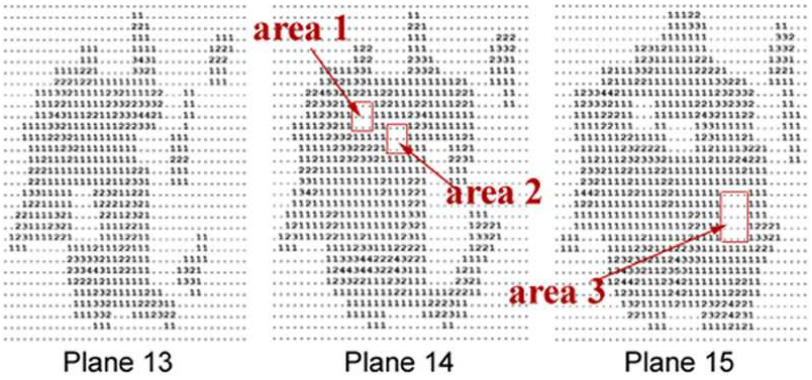


Figure 3.10: Three consecutive matrix planes generated from a PDB entry 1ADS. For simplicity, '0' cells are shown in '.'. It illustrates how internal space areas change the size and shape between planes.

we consider the three-dimensional structure, some of them may not be "closed", but "open" as shown in the example of Figure 3.11. Since we have already used the "probe" to examine all of the open spaces previously if they are big enough to contain a water molecule, if we find any "opening" to the outside for these internal space areas, such internal spaces can be filled with water molecules. Such internal spaces can be considered as "open" with "surface" areas. On the other hand, if some of the internal space areas are completely "closed", such closed internal spaces cannot have contacts with water, and we do not have to consider any internal "surface" areas for such closed internal spaces. For example, Figure 3.10 shows the three consecutive matrix planes with some internal spaces. Scanned

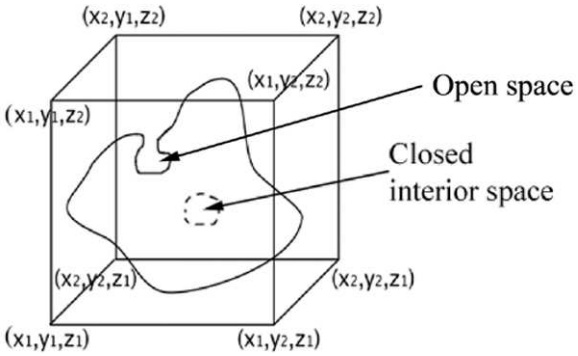


Figure 3.11: Representation of a closed interior space and an open space (viewed from the side of the protein). The dashed line shows the closed interior opening.

vertically across the planes, each of the internal spaces is examined if it has any opening to the outside. The areas 2 and 3 in Figure 3.10 are connected to the outside on the planes 13 and 14, respectively, for example. On the other hand, the area 1 is not considered to be open in the three dimensional structure if in any connected neighboring planes higher than the plane 15, it is closed with non-0 digits.

Finally, cells were examined by gathering the empty cell information from the neighboring cells. First, cells that are occupied with any atoms were found. All of the six directions (one cell in each of four directions on the same plane as well as one “up” and one “down” cells in the neighboring two planes) of each occupied cell were searched. In Figure 3.12 the cell marked with “x” is occupied by atoms. If any one of the six neighboring cells is empty, the cell x is considered to be at the surface. This process was repeated

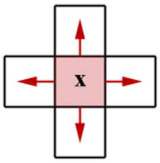


Figure 3.12: Representation of a surface decision process. The shaded box “x” represents the occupied cell with any atom(s).

for each occupied cell in all of the planes from the protein molecule. After all of the cells were verified for their locations (surface or not), a surface area map can be generated from

each plane as shown in Figure 3.13.

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	1	1	0
0	1	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	2	2	0
0	1	1	2	1	1	1	1	1	0
0	0	0	1	1	1	1	1	1	0
0	0	0	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Figure 3.13: A surface area map. Shaded cells are at the surface.

3.2.3 Identifying amino acids at the protein surface

Now each cell is marked with the number of atoms located in the cell as well as if it is at the surface or not. As described before, one atom occupies more than one cells (e.g., nine cells for the atom with 1Å diameter). In order to identify if each atom is at the surface or not, we assumed that if any one cell belonging to an atom was located at the "surface," the entire atom was considered to be on the surface of the protein.

Once the atoms existing on the protein surface were known, the next problem was to determine which amino acids should be considered to be on the surface. The charged atoms are usually present on the side chain of an amino acid and tend to attract water molecules. These atoms are thus more likely to be found on the surface of the proteins, and they play an important role in positioning an amino acid at the surface of the protein. Hence, the charged atoms like nitrogen (N) and oxygen (O) present in the side chain was considered in deciding an amino acid to be at the surface. At least one of the charged atoms should be at the surface in order to decide an amino acid to be at the surface.

After identifying amino acids at the surface, the next question was how to identify α-helix as amphipathic. After various attempts, an α-helix was identified to be amphipathic if

more than 20% of its amino acids exist on the surface. In an ideal amphipathic α-helix, 50% of the amino acids in the helix would be hydrophobic (facing the inside of the protein) and the other 50% would be hydrophilic (facing the outside of the protein). But, in practice, not always the entire helix is amphipathic, but only a part of the α-helix could have the amphipathic property. Considering, for example, a possibility of having a half of one side of the helix exposed to the surface, 25% could be more realistic. Finally, we decided to use 20% as the threshold. An α-helix was identified as amphipathic if more than 20% of its amino acids were at the surface.

3.2.4 Results

The above method was applied to 556 α-helices (found from 160 proteins by the secondary structure identification method developed earlier), and 26 of them were identified as amphipathic. All of these 26 identified amphipathic α-helices were visually inspected by using PyMol. Three of these amphipathic α-helices are shown in Figure 3.14 and the rest of the 23 α-helices are included in Appendix E. As explained before, the 20$\%$ threshold was rather arbitrarily chosen. However, due to this low threshold, a few exceptional amphipathic α-helices were also identified. These α-helices contain only a part of the helix as amphipathic as shown in Figure 3.14(b). None of the identified amphipathic α-helices were annotated as amphipathic in the PDB records.

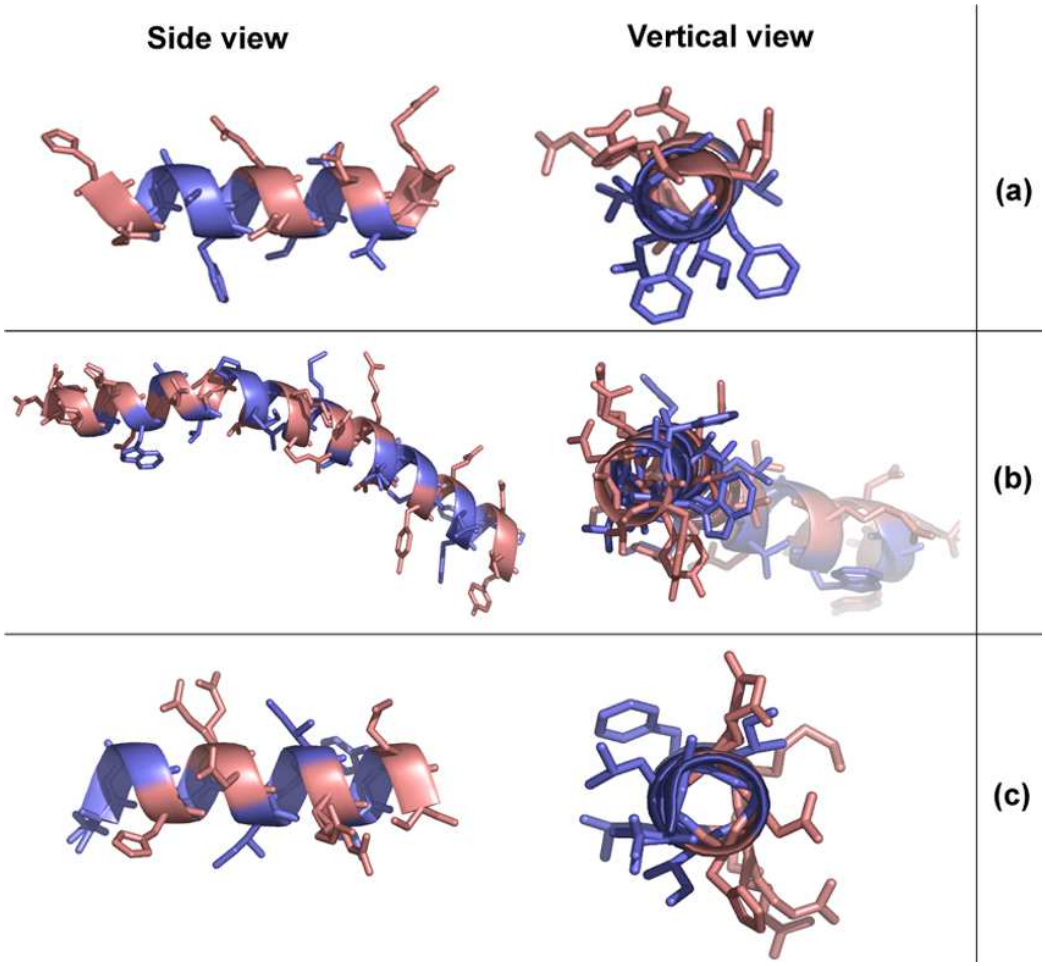


Figure 3.14: Examples of amphipathic α-helices identified by the new method. (a) 1A8L (oxidoreductase), (b) 1AH7 (hydrolase), and (c) 1AJG (myoglobin). The hydrophobic residues are represented in red and hydrophilic in blue. PyMol is used to create 3D visualizations. See Appendix E for the remaining 23 amphipathic α-helices identified by the new method.

Chapter 4

Development of a prediction method for amphipathic α-helices

4.1 Development of statistics

As described in Chapter 2, amphipathic α-helices contain both hydrophobic and hydrophilic residues. The hydrophilic side-chains extend from one side of the α-helix and hydrophobic side-chains from the opposite side, dividing the α-helix into two sides: hydrophobic and hydrophilic. Based on this property, a set of simple statistics was developed to quantify a bias in hydrophobic and hydrophilic amino acid distributions between two sides.

In order to identify such bias, each sequence was divided into two sides, called A and B sides. Because we do not know how the amphipathic characteristic (having hydrophobic and hydrophilic sides) should be observed from each peptide, each sequence was divided into four ways as shown in Figure 4.1. Using each of the first four amino acids as a starting position, the sequence can be divided into two sides in four possible ways. This is based on the possible location of each amino acid on an assumed α-helix.

In Figure 4.1, the helix is viewed from above. As described in Chapter 2, an α-helix

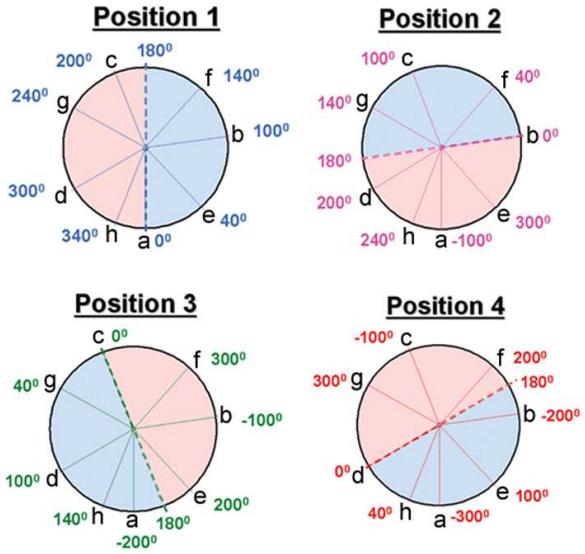


Figure 4.1: Four possibilities of dividing a sequence into the sides A and B. The angles at each position are at the interval of 100°, 0° being the initial angle with the respective first, second, third, and fourth amino acid starting points. The amino acids are shown in lower case alphabets (a, b, c, d, e, f, g, and h in the order of the sequences). The dashed line in each position divides the α-helix sequence into two sides A (blue side) and B (red side). The solid lines show the angle at which each residue is placed.

contains 3.6 residues per turn. One complete turn is 360 degrees. Thus each amino acid occurs at approximately every 100°. For example, consider a sequence of seven amino acids (shown as “abcdefgh” in Figure 4.1). If “a” is the starting amino acid, the initial angle is 0°. The next residue b is placed at 100°, the third residue c at 200°, and so forth. If the helix is divided into A and B sides using the 0-180 degree line (the dashed line), residues located at the angle between 0° and 180° (a, e, b, and f) are considered to be in the side A (shown in blue), and residues located between the angle 180° and 360° (c, g, d, and h) are in the side B (shown in red). This is illustrated in the “Position 1” of Figure 4.1. When the second amino acid b is considered to be at the angle 0°, the residues are divided into the side A (b, f, c, and g) and the side B (d, h, a, and e) shown in the “Position 2” of Figure 4.1. The negative angles indicate that these amino acids locate before the one at 0°. Similarly, the third and fourth amino acids are used as the starting points.

The following statistics were calculated from each position (1, 2, 3, or 4) from each sequence:

1. **The total number of hydrophobic amino acids: $\mathbf{N}_1$**

   The total number of hydrophobic amino acids is counted from each side, A or B. The following eleven amino acids are considered to be hydrophobic: alanine (A), cysteine (C), phenylalanine (F), glycine (G), isoleucine (I), leucine (L), methionine (M), proline (P), valine (V), tyrosine (Y), and tryptophan (W).

2. **The total number of hydrophilic amino acids: $\mathbf{N}_2$**

   The total number of hydrophilic amino acids is counted from each side, A or B. The following nine amino acids are considered to be hydrophilic: aspartic acid (D), glutamic acid (E), histidine (H), lysine (K), asparagine (N), glutamine (Q), arginine (R), serine (S), and threonine (T).

3. **The number of consecutive hydrophobic or hydrophilic amino acid region: $\mathbf{C}_1$ or $\mathbf{C}_2$**

   A consecutive hydrophobic amino acid region is identified if there are three or more hydrophobic amino acids consecutively in one side. The number of such regions is counted from each side. The number of consecutive hydrophilic regions is counted similarly.

4. **The length of the longest consecutive hydrophobic or hydrophilic amino acid region: $\mathbf{L}_1$ or $\mathbf{L}_2$**

   The longest consecutive regions with either hydrophobic or hydrophilic amino acids are identified from each side and the length (the number of amino acids) of this region is recorded.

5. **The percentage of hydrophobic or hydrophilic amino acids: $\mathbf{P}_1$ or $\mathbf{P}_2$**

   The percentage of hydrophobic or hydrophilic amino acids for each side is calculated as follows: $N_1 \times 100 \ / \ N_A$ or $N_2 \times 100 \ / \ N_A$ for the side A, $N_1 \times 100 \ / \ N_B$ or N_2 x $100 \ / \ N_B$ for the side B, where N_A and N_B are the numbers of amino acids in the sides A and B, respectively.

6. **The ratio of hydrophobic to hydrophilic amino acids in each side: $\mathbf{R}_A$ or $\mathbf{R}_B$**

   The ratio (R_A for the side A or R_B for the side B) is calculated as $N_1 \ / \ N_2$ (or $P_1 \ / \ P_2$) for each side. When N_2 is 0, the ratio cannot be computed, and in such a case, INF is shown. In the practical computation, this 0 violation was avoided by introducing a large constant. If N_2 is 0, the ratio was 0 given as '1000'.

7. **The overall ratio of hydrophobic amino acids: R**

   The ratio, $R_A \ / \ R_B$, was represented using the natural logarithm is instead calculated as $\mathbf{R = log(R}_A \ / \ \mathbf{R}_B\mathbf{) = log(R}_A\mathbf{) - log(R}_B\mathbf{)}$. $\mathbf{log(R}_A\mathbf{)}$ and $\mathbf{log(R}_B\mathbf{)}$ can be also calculated as $\mathbf{log(N}_1\mathbf{) - log(N}_2\mathbf{)}$ for the sides A and B, respectively. The problem of 0 violation when either N_1 or N_2 was 0, was resolved by introducing arbitral constants. R was given as 0 (indicating "no bias") if both R_A and R_B are either 0 or 1000. If only one of R_A or R_B is 0, R was given as '10' (indicating a "large bias").

   After statistics 1-7 are obtained for each of the four positions:

8. **The maximum overall ratio for the peptide: $\mathbf{R}_{max}$**

   The absolute values of R, |R|, from the four positions are compared and the highest |R| is selected as R_{max}.

4.2 Preliminary analysis

In order to examine the statistics developed earlier, preliminary data sets were obtained from the PDB. Discrimination between the data set containing amphipathic α-helices and the one without any secondary structures based on these statistics was examined.

4.2.1 Data sets

A small number of sample data was collected from PDB by manually inspecting the structures and searching the database with the "amphipathic" keyword. By using a protein structure visualization software, PyMol [3], strictly amphipathic α-helix regions were identified from six PDB entries: 1AHR (calcium-binding protein), 1BVS (holliday junction resolvase component), 1BM9 (DNA-binding protein), 1MNK (myoglobin), 2CMM (myoglobin), and 2REB (DNA binding protein).

The lengths of these amphipathic α-helix regions are from 12 to 28 amino acids. Another set of six protein segments that do not contain any secondary structures (α-helix or β-strand) was also collected from five PDB entries: 1AYN (rhinovirus coat protein), 1HQM (RNA polymerase), 1OIT (kinase), 1PK4 (hydrolase), and 1PKK (hydrolase). The lengths of all the six protein segments are 26 amino acids. These two data sets were called "positive" (containing six amphipathic α-helix regions) and "negative" (containing six protein regions that do not have any secondary structures).

4.2.2 Results

The statistics were calculated from each of the 12 peptide sequences and summarized in Tables 4.1 and 4.2. Only the statistics obtained from the starting position that gives the R_{max} are included. Appendix F includes the statistics obtained from all of the four positions.

Even though the sample size was small, comparing Tables 4.1 and 4.2, the simple

Table 4.1: Statistical analysis of the six amphipathic α-helix regions.

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AHR	15	3	0	0	20	12	1	9	80	0.25	13	10	2	6	77	3	0	0	23	3.33	2.59
1BM9	7	1	0	0	14.3	6	1	5	86	0.17	6	4	1	3	67	2	0	0	33	2	2.47
1BVS	6	6	1	6	100	0	0	0	0	1000	6	2	0	0	33	4	1	4	67	0.5	10
1MNK	11	8	2	5	72.7	3	0	0	27	2.67	9	2	0	0	22	7	1	6	78	0.29	2.22
2CMM	10	4	0	0	40	6	0	0	60	0.67	6	5	1	3	83	1	0	0	17	5	2.01
2REB	7	6	1	5	85.7	1	0	0	14	6	6	2	0	0	33	4	0	0	67	0.5	2.48

Statistics used are explained in the section 4.1.

Table 4.2: Statistical analysis of the six non structural regions.

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AYN	15	11	2	6	73.3	4	0	0	27	2.75	11	6	0	0	55	5	1	3	45	1.2	0.83
1HQM	15	7	1	3	46.7	8	1	6	53	0.88	11	4	0	0	36	7	1	4	64	0.57	0.43
1HQM2	13	4	0	0	30.8	9	1	3	69	0.44	13	8	2	4	62	5	1	3	38	1.6	1.29
1OIT	13	11	2	7	84.6	2	0	0	15	5.5	13	6	0	0	46	7	1	4	54	0.86	1.86
1PK4	14	8	1	6	57.1	6	1	3	43	1.33	12	3	0	0	25	9	1	6	75	0.33	1.39
1PKK	15	7	1	4	46.7	8	1	4	53	0.88	11	6	1	3	55	5	1	3	45	1.2	0.31

set of statistics developed identifies the difference in amphipathicity between the two data sets. Table 4.2 shows that the distribution of hydrophobic and hydrophilic amino acids in the negative data set is approximately equal between any two possible sides as expected. Figure 4.2 shows the distribution of R_{max} (the maximum overall ratio of hydrophobic to hydrophilic amino acids) obtained from the preliminary data sets. R_{max}'s calculated from the positive samples (shown with blue bars) were larger than those of negative samples (shown with red bars). Note that as described in the section 1.4, R_{max} = 10 is the arbitrary large constant given when there is no hydrophilic amino acid in one side, indicating cases of extreme bias. These results show that larger values of R_{max} are good indicators for sequences derived from amphipathic α-helices, which have more hydrophobic amino acids in one side than the other as compared to those from non-structural regions.

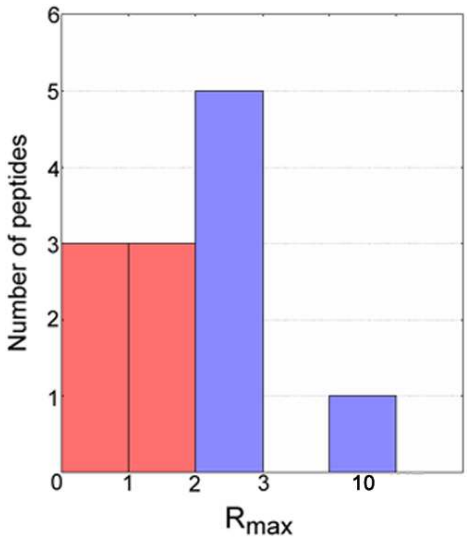


Figure 4.2: Distribution of R_{max}'s compared between positive (blue bars) and negative (red bars) data from the preliminary analysis. The average R_{max} for the negative samples is 1.02 and that for the positive samples is 2.35 (excluding cases of R_{max} = 10).

4.3 Large scale data analysis

4.3.1 Data set preparation

In order to confirm if the statistics developed based on a small sample data set were applicable for more general cases, larger data sets were prepared. 25,414 protein entries were obtained from June 2004 release of PDB. Only "monomer" (single subunit) proteins were used for this study. 4,878 "monomer" protein entries were chosen based on the absence of the "chain" information in the PDB entries (having multiple "chain" information indicates that the entry is derived from "multi-subunit" proteins). These entries were screened based on the presence of atomic coordinate information for the complete sequence of a protein; partial or fragment entries were filtered out. Out of the 1,585 complete "monomer" entries, 158 proteins were confirmed to have the complete proteins after further manually verifying each coordinate information. This final step of manual conformation was required as inconsistency frequently found in PDB entries. Some PDB entries have missing atomic coordinates information from some amino acids. Some proteins might be "dimers" (including two subunits) or "multimers" (or "polymers"; including several subunits), even if the chain information is absent from these entries, since some researchers may include such information in other parts of annotation.

Two dimer proteins from the preliminary data set (1BM9 and 1MNK) were also intentionally included in the analysis. The reason behind these exceptions was to examine if the new method developed can actually identify these peptides as amphipathic α-helices. In total, there were 160 entries that could be used to create two types of data sets: a "positive" including amphipathic α-helices and a "negative" including non-structural regions.

From these 160 PDB entries, the secondary structures were identified using the methods developed in section 3.1. 556 of 1,063 α-helices identified were ten amino acids or longer and used for the further analysis. Finally, the method described in section 3.2 was used

Table 4.3: Identification process performed in this study.

	Number of data included
PDB entries used	25,414
Monomers	4,878
Monomers except partial sequences	158
Data set used *	160
α-helices	1,063
α-helices $\geq$ 10 amino acids	556
Amphipathic α-helices identified	26

* Data sets used in the prediction analysis.

to identify amphipathic in α-helix from these 556 peptide sequences. Twenty six helices contained more than 20% of amino acids on the protein surface, and were identified as amphipathic α-helices. These 26 peptide regions were manually verified by using PyMol, a visualization software, and confirmed to have the amphipathic amino acid distribution on these helices. Table 4.3 summarizes the identification process performed in this study. Out of the 26 amphipathic α-helices, five were removed as those sequences were identical to other sequences and redundant. The total number of positive samples was, therefore, 21, and these α-helices were derived from different proteins.

A negative data set containing 21 (the same number as the positive samples) randomly selected peptides that do not have any secondary structures was prepared. Peptides with no secondary structures ranging from 13 to 20 amino acids (aa) were collected following the length distribution of α-helices included in the positive data set (from 12 to 38 aa). The reason for excluding shorter peptides was to avoid the possibility of including short secondary structure regions. As described in Chapter 3, short regions (shorter than 4 aa) that contain α-helices or β-strands were identified as non-structural regions in this study.

A set of simulated binary sequences that represents the entire theoretically possible amino acid sequence space was also prepared. In order to simplify the sequence space, a binary code (0 or 1) was used to represent two amino acid types: hydrophobic or non-

hydrophobic (hydrophilic). The length of such binary sequences was set as 15 aa based on the average length of the negative samples. The entire 32,768 binary sequences were produced and used as one data set. The objective of using this simulated data set was to examine the distributions of statistics and to compare them against the statistics obtained from natural protein data found in PDB (the positive and negative data sets). This comparison should show any difference in the theoretical and natural protein sequence spaces based on the statistics used.

4.3.2 Results and Discussion

Amphipathic α-helix prediction based on R_{max}

The statistics were calculated from each of the 42 peptide sequences of identified amphipathic α-helices in positive and negative data sets. Tables 4.4 and 4.5 list the final R_{max} values, which is the maximum overall ratio of hydrophobic amino acids between the sides A and B. Appendix G listed the detailed statistics obtained from the 42 sequences.These two tables show that R_{max} values are generally higher for the amphipathic helices than for the non-structural sequences.

Table 4.4 presents that the majority of R_{max} values obtained from amphipathic α-helices are closer to 2 or higher (the average: 1.77 excluding R_{max} = 10), whereas the R_{max} values are around 1.0 in non-0 peptides (Table 4.5: the average: 1.29 excluding R_{max} = 10). These results are consistent compared to those obtained from the preliminary analysis. Figure 4.3 compares the distributions of R_{max} between the two data sets. The distribution of R_{max} obtained from the positive data set (a) is skewed towards larger R_{max} while the negative data set distribution (b) is shifted towards smaller R_{max} values. In order to examine if the difference in the distributions is significant, Wilcoxon/Kruskal-Wallis non-parametric test was performed. The probability (P = 0.0005) indicates that the difference is highly

Table 4.4: R_{max} values obtained from the 21 amphipathic α-helix regions.

ID	R_{max}
1A8L	2.3
1A9O	1.95
1ADS	1.8
1AGX	1.95
1AH7	0.89
1AJG	2.86
1ALD	1.32
1AMP	1.66
1ANG	10[1]
1AOV	10
1AQP	10
1ARL	1.61
1AST	2.25
1BE0	10
1BEE	10
1BEO	10
1BEZ	10
1BGP	2.3
1BH0	2.12
1BIY	0
1BM9	10

[1] R_{max} = 10 is given when there is no hydrophilic amino acid in one side (A or B). As described in section 4.1 this is an arbitrary large constant indicating that there is a large bias between the two sides.

significant.

Figure 4.4 shows the result obtained from the simulated binary sequences (32,768 simulated data). The distribution in this data set appears to be basically the combination of positive and negative data sets. However, the Wilcoxon/Kruskal-Wallis non-parametric test shows that the difference between the R_{max} distributions from the simulation data set and positive data set is also highly significant (P = 0.0005).

Based on these observations for R_{max} values, the probabilities of occurrence of amphipathic helices were calculated from some ranges of R_{max}. Figure 4.7 summarizes it. Using these ranges of R_{max}, we can predict the probabilities for having an α-helix as amphipathic

Table 4.5: R_{max} values obtained from the 21 non-structure regions.

ID	R_{max}
1321	2.46
1891	0.98
1A8Q	1.1
1AC5	10[1]
1AHB	1.97
1AK9	1.39
1AMY	1.14
1APC	1.33
1AQN	1.39
1AST	0.98
1ATA	0.51
1AU9	0.52
1AYV	3.33
1AYX	1.8
1BAG	1.21
1BF2	1.27
1BG5	1.16
1BG9	1.14
1BG92	1.1
1BGO	0.41
1BGP	0.71

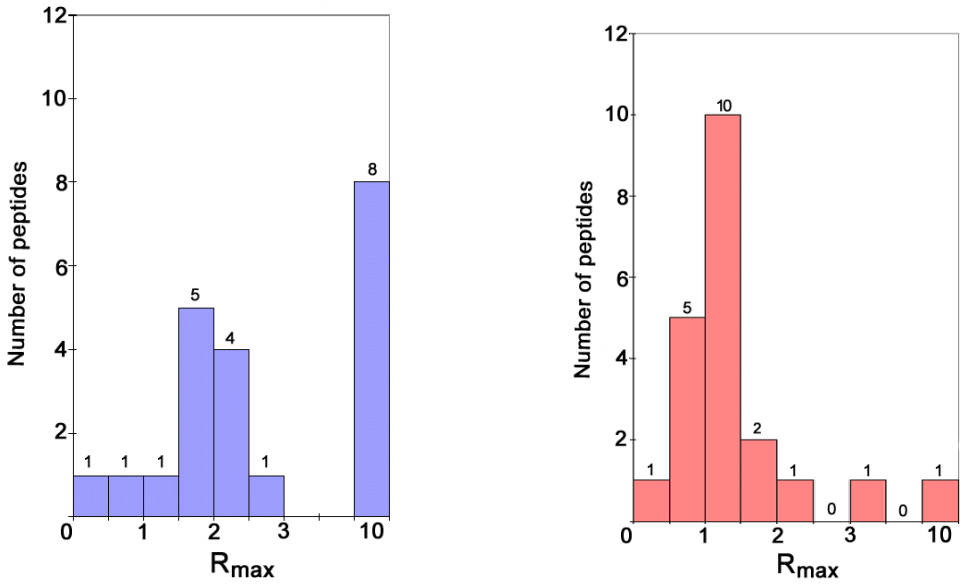
[1] R_{max} = 10 is given when there is no hydrophilic amino acid in one side (A or B). As described in section 4.1 this is an arbitrary large constant indicating that there is a large bias between the two sides.

as higher than 0.80 if $2.0 \leq R_{max}$, about 0.70 or lower if $1.5 < R_{max} < 2.0$, and lower than 0.16 if $R_{max} \leq 1.5$.

Comparative Study against Hydrophobic Moment

As described earlier, the hydrophobic moment method [5] is the only currently available method that attempts to quantify the amphipathicity of α-helices. In order to examine how this method is effective compared to our newly developed R_{max} statistics, a comparative analysis was performed.

The same data set consisting of 21 positive and 21 negative samples was used to obtain



(a) (b)

Figure 4.3: Distribution of R_{max} compared between positive (a) and negative (b) data from the large data analysis.

the "hydrophobic moment" values. The implementation available from EMBOSS [15] was used for this study. As described before (Figure 2.13), even though the hydrophobic moment method is quantitative, the final decision to identify 0 α-helix regions is largely arbitral. In this study, the maximum moment value obtained from each peptide sample was used as the index. Tables 4.6 and 4.7 list the maximum moment values obtained from the 42 peptide samples. The averages of 0 moment values are 0.41 and 0.32 from the amphipathic α-helix samples (Table 4.6) and non-structured peptide samples (Table 4.7), respectively.

Although these average values are close, Figure 4.6 shows that the maximum moment values in the positive data set are widely spread as compared to those of the negative data set. The range of the maximum moment values for the negative data set was completely overlapped with that for the positive data set. As expected, Wilcoxon/Kruskal-Wallis non-parametric test indicates that these distributions are not significantly different (P = 0.0662).

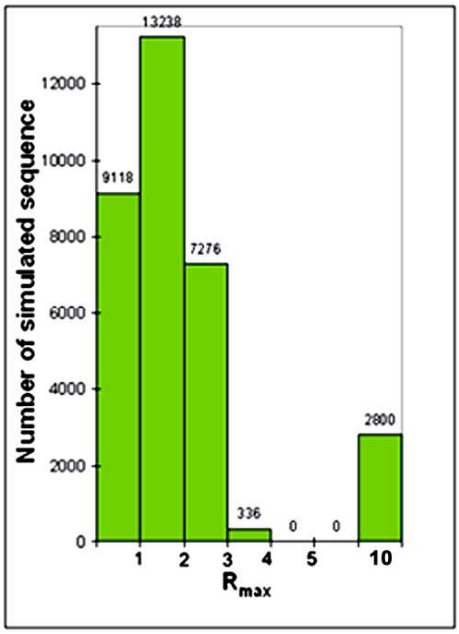


Figure 4.4: Distribution of R_{max} obtained from the simulation data set.

R_{max}	0 – 1.5	1.5 – 2.0	2.0 – 10
Prob (amphi)	0.16	0.71	0.81
Prob (non-amphi)	0.84	0.29	0.19

Figure 4.5: Probabilities of amphipathic α-helices based on R_{max} values. Prob(amphi) represents the probability of occurrence of amphipathic α-helix and Prob(non-amphi) represents the probability of non-occurrence of amphipathic α-helix.

This is very different from what we observed in the distribution of R_{max}; the R_{max} values were significantly different between positive and negative samples. These results show that it is not possible to use the maximum moment values for discriminating 0 α-helices.

Note that the negative samples show peak in the distribution of moment values (Figure 4.6(b)). This could be explained by the bias existing in the amino acid composition of protein sequences. Figure 4.7 shows the amino acid composition obtained from the complete set of Swiss-Prot protein entries. It shows that many hydrophobic amino acids (bars colored

Table 4.6: The maximum moment values obtained from the 21 amphipathic α-helix regions.

ID	Maximum moment
1A8L	0.669
1A9O	0.601
1ADS	0.292
1AGX	0.325
1AH7	0.548
1AJG	0.654
1ALD	0.381
1AMP	0.38
1ANG	0.74
1AOV	0.038
1AQP	0.441
1ARL	0.56
1AST	0.462
1BE0	0.091
1BEE	0.091
1BEO	0.531
1BEZ	0.091
1BGP	0.174
1BH0	0.49
1BIY	0.528
1BM9	0.636

in grey) are present with relatively higher frequencies in protein sequences available in the database. The negative samples in this study can be considered as the random samples from such proteins. Since the amphipathicity of an α-helix is determined by measuring the bias between two major functional groups (hydrophobic and hydrophilic), such skewed amino acid composition in negative data could generate slightly elevated moment values. This may explain the peak found in the moment value distribution from the negative samples. On the other hand, the R_{max} distribution did not show any such effect in negative samples (Figure 4.3(b)).

The moment values failed to identify some of the amphipathic helices. Figure 4.8 shows the relationships of % surface amino acids against the maximum moment values and

Table 4.7: The maximum moment values obtained from the 21 non-structure regions.

ID	Maximum moment
1321	0.536
1891	0.163
1A8Q	0.327
1AC5	0.336
1AHB	0.283
1AK9	0.302
1AMY	0.201
1APC	0.367
1AQN	0.302
1AST	0.18
1ATA	0.304
1AU9	0.362
1AYV	0.343
1AYX	0.48
1BAG	0.458
1BF2	0.454
1BG5	0.227
1BG9	0.201
1BG92	0.231
1BGO	0.324
1BGP	0.41

R_{max}. Spearman's Rho non-parametric rank correlation test showed that the correlation is not significant (P = 0.1214) for the maximum moment values (Figure 4.8(a)) whereas it is significant (P $<$ 0.0001) for R_{max} (Figure 4.8(b)).

The significant correlation of R_{max} with % surface amino acids indicates that the positive samples (blue triangles) could be predicted as amphipathic based on the higher R_{max} values. As shown in Figure 4.8(b), the majority of the positive samples have R_{max} values 1.5 or higher. On the contrary, it is not possible to predict the amphipathic α-helices from the maximum moment values. There is no significant relationship between the % surface amino acids and the maximum moment values.

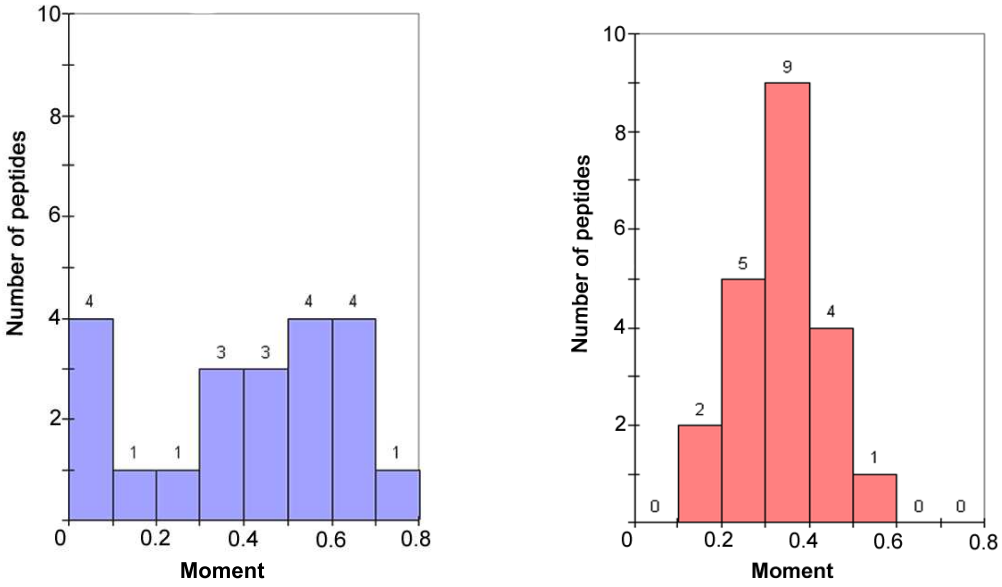


Figure 4.6: Distributions of the maximum moment values compared between positive (a) and negative (b) data. The same 42 peptide samples were used as in Figure 4.3.

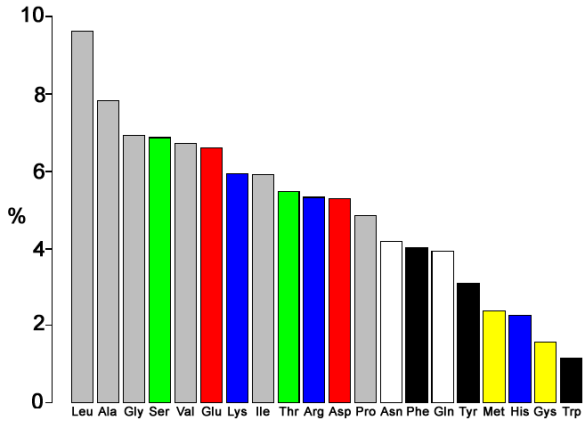
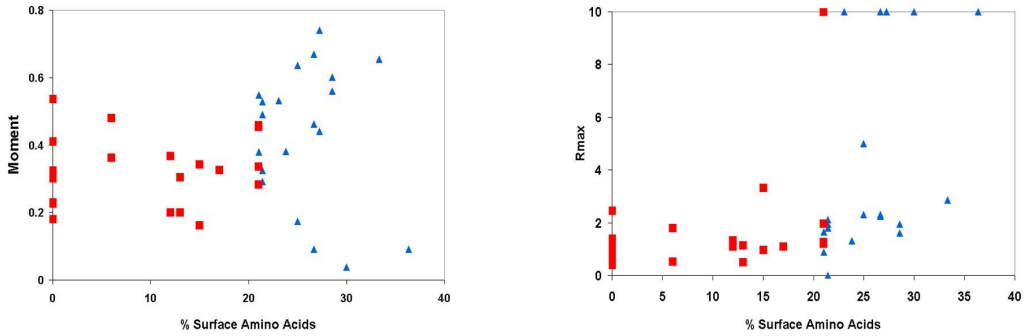


Figure 4.7: Amino acid composition obtained from 123946 protein entries in the Swiss-Prot protein database (courtesy [8]). The colored bars represent the 0 groups of amino acids: gray = aliphatic, red = acidic, green = small hydroxy, blue = basic, black = aromatic, white = amide, and yellow = sulfur.



(a) (b)

Figure 4.8: Relationship between % surface amino acids and the maximum moment values (a) and R_{max} (b). Positive samples are represented by blue triangles and negative 0 are represented by red squares. Note that higher than 20% surface amino acids from α-helix regions were used to identify 0 α-helix. The red squares with higher than 20% surface amino acids are negative samples, some of whose amino acids are at the protein surface but they are not in α-helix region. Also note that the R_{max} = 10 is the arbitrary large constant given when there is an extreme bias in 10 hydrophobic to hydrophilic ratio (see section 4.1).

Chapter 5

Conclusion and Future work

The methods for identifying secondary structures and amphipathic α-helices based on atomic coordinate information of protein structures were developed. In order to develop a prediction method for secondary structures, especially amphipathic helices, it was required to first define and then identify the secondary structures using a consistent method. The method developed identified 1,063 α-helices from 160 protein entries. 556 α-helices of length ten or longer amino acids were selected for further analysis.

A new method was developed to identify amphipathic α-helices. Amphipathic α-helices were identified by searching helices that were on the surface of the protein. This problem was solved by developing a three-level process: the cell level, atomic level, and amino acid level. Finally, the helix was identified as amphipathic if more than 20% of the amino acids are on the surface. This new method was able to detect 26 amphipathic α-helices from 160 PDB entries, all of which have not been annotated as amphipathic in the PDB database. This identification method is based only on protein structural information, and there was no such method available before.

Using the newly developed methods, a data set including both amphipathic α-helices and protein regions with no secondary structure was prepared. It enabled us to develop

a prediction method for amphipathic α-helices from primary structure information. A set of simple statistics was developed, and it was shown that the R_{max} value (maximum overall ratio of hydrophobic amino acids) can discriminate amphipathic α-helices from non-amphipathic α-helices. Using R_{max}, we can estimate probabilities of having an amphipathic α-helix based on the amino acid sequence. It can be used for the prediction.

The new identification method developed in this study was able to detect a very long α-helix that is partially amphipathic as shown in Figure 3.14(b). The currently available methods (helical wheel, helical net, and hydrophobic moment) are more likely not be able to detect this type of helices, since amphipathic structure is not well-defined in such helices. For example, the "hydrophobic moment" method indeed failed to identify 9 amphipathic α-helices. This is because the moment becomes low due to the irregularity in the distribution of hydrophobic residues in the helix.

Two PDB entries, 1MNK and 1BM9, were intentionally included when the new identification methods were applied to identify amphipathic α-helices. From the both entries, α-helices were identified including the regions previously known to be amphipathic α-helix. However, these particular regions known to be amphipathic α-helices could not be detected as amphipathic. From 1BM9, however, another α-helix was identified as amphipathic. The amphipathic α-helices expected to be identified are 13 aa long for 1BM9 and 20 aa long for 1MNK. The percentage of surface amino acids in these regions were 7.14% and 10% for the regions of 1BM9 and 1MNK, respectively. As explained before, the threshold used to identify amphipathic α-helices is to have 20% or higher amino acids at the surface of a helix region. Since we already know these helices to be amphipathic based on the visualized protein structure, the 20% threshold may need to be even lower to make the method more flexible. It should increase the probability of identifying more amphipathic helices. Of course it will also increase the false positive rate. It requires more examples and analysis. One possibility to reduce the false positive could be, for example, to incorporate both of

the length and the percentage of surface amino acids within the helix regions as part of the threshold.A larger data set needs to be used in the further study. It should allow us to refine the threshold for identifying amphipathic α-helices. And it will give us more refined scales for prediction probabilities.

The majority (or all) of the amphipathic α-helices identified in this study have not been described previously. The ultimate confirmation needs to be experimentally done. However, these methods will help researchers to identify more amphipathic α-helix candidates and their further studies on protein functions.

When we do not have protein structure information, secondary structures need to be also predicted based on amino acid sequences. As described in Chapter 1, there are many prediction methods available for this purpose. Such methods need to be incorporated with the prediction method developed in this study to make the method usable for more general sequence analysis. It is also possible that the same or extended set of the statistics developed in this study is used for the entire prediction including secondary structures.

The amphipathic property is not related only to α-helices, but β-strands/sheet are also present in the protein structure and also have an important biological function associated to it. It is useful if we can extend the method for identifying and predicting both of amphipathic α-helices and β-strands/sheets. A larger data analysis should provide more insights on such applications of the strategy developed in this study.

Since detecting amphipathic α-helices is important for examining protein structures, such information can be used to improve other bioinformatics tools, as multiple alignment and protein classification. Such application possibility can be further explored in the future.

Bibliography

[1] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000. website: http://www.rcsb.org/pdb/.

[2] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, Inc., New York, NY, 1999.

[3] W. L. DeLano. The PyMOL Molecular Graphics System, 2002. website: http://www.pymol.org.

[4] P. Dunnill. The use of helical net-diagrams to represent protein structures. *Biophysical Journal*, 8(7):865–75, 1968.

[5] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger. The helical hydrophobic moment: A measure of the amphipathicity of a helix. *Nature*, 299:371–374, 1982.

[6] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger. The helical hydrophobic moment detects periodicity in protein hydrophobicity. *Biophysics*, 81:140–144, 1984.

[7] R. M. Epand. *The Amphipathic Helix*. CRC Press, Inc., Boca Raton, FL, 1993.

[8] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, and A. Bairoch. Expasy: the proteomics server for in-depth protein knowl-

edge and analysis. *Nucleic Acids Research*, 31:3784–3788, 2003. website: http://bo.expasy.org/sprot/relnotes/relstat.html.

[9] V. Ingram. Mit biology hypertextbook. website: http://web.mit.edu/esgbio/www/lm/proteins/aa/aminoacids.html.

[10] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton. Procheck: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26:283–291, 1993.

[11] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell. *Molecular Cell Biology*. W. H. Freeman and Company, New York, NY, 2000.

[12] A. L. Morris, M. W. MacArthur, E. L. Hutchinson, and J. M. Thornton. Stereochemical quality of protein structure coordinates. *PROTEINS: Structure, Function, and Genetics*, 12:345–364, 1992.

[13] D. C. Ophardt. Virtual chembook. website: http://www.elmhurst.edu/ chm/vchembook/index.html.

[14] G. A. Petsko and D. Ringe. *Protein Structure and Function*. Sinauer Associates, Inc., Sunderland, MA, 2004.

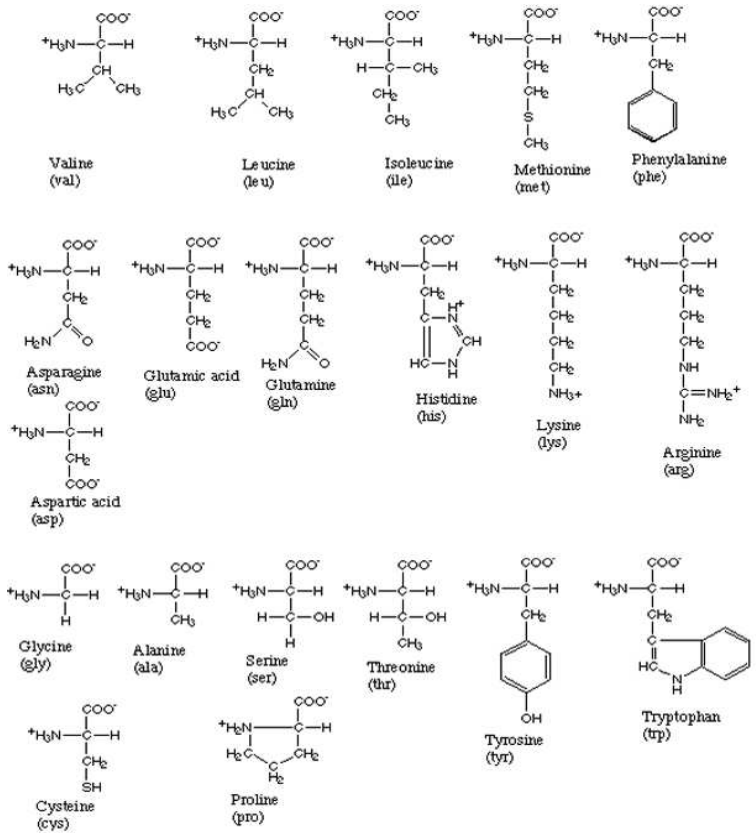
[15] P. Rice, I. Longden, and A. Bleasby. Emboss: The european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277, 2000. website: http://biocore.unl.edu/EMBOSS/index.html.

[16] M. Shiffer and A. B. Edmundson. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophysical Journal*, 7:121–135, 1967.

[17] D. Voet and J. G. Voet. *Biochemistry*. John Wiley and Sons, Inc., Hoboken, NJ, 2004.

Appendix A

Twenty amino acids and their Chemical structures (courtesy [9])



Appendix B

An example PDB entry (1AJG: a myoglobin)

```
HEADER    OXYGEN TRANSPORT                        02-MAY-97   1AJG
TITLE     CARBONMONOXY MYOGLOBIN AT 40 K
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MYOGLOBIN;
COMPND   3 CHAIN: NULL;
COMPND   4 BIOLOGICAL_UNIT: MONOMER;
COMPND   5 OTHER_DETAILS: CARBONMONOXY MYOGLOBIN, CO LIGAND BOUND TO
COMPND   6 FE OF THE HEME, HEME BOUND TO NE2 OF HIS 93
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: PHYSETER CATODON;
SOURCE   3 ORGANISM_COMMON: SPERM WHALE
KEYWDS    OXYGEN TRANSPORT, RESPIRATORY PROTEIN, HEME
EXPDTA    X-RAY DIFFRACTION
AUTHOR    T.Y.TENG,V.SRAJER,K.MOFFAT
REVDAT   1   12-NOV-97 1AJG    0
JRNL        AUTH   T.Y.TENG,V.SRAJER,K.MOFFAT
 . . .
JRNL        REFN   ASTM NSBIEW  US ISSN 1072-8368                2024
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   T.Y.TENG,V.SRAJER,K.MOFFAT
 . . .
REMARK   1  REFN   ASTM BICHAW  US ISSN 0006-2960                0033
REMARK   2
REMARK   2 RESOLUTION. 1.7  ANGSTROMS.
REMARK   3
 . . .
REMARK 200 METHOD USED TO DETERMINE THE STRUCTURE: DIFFERENCE FOURIER
REMARK 200 SOFTWARE USED: X-PLOR
REMARK 200 STARTING MODEL: PDB ENTRY 1MBO
 . . .
REMARK 900
REMARK 900 RELATED ENTRIES
REMARK 900 THIS ENTRY IS RELATED TO PDB ENTRY 1AJH.
```

```
DBREF  1AJG      1   153  SWS    P02185   MYG_PHYCA        1    153
SEQRES   1    153  VAL LEU SER GLU GLY GLU TRP GLN LEU VAL LEU HIS VAL
SEQRES   2    153  TRP ALA LYS VAL GLU ALA ASP VAL ALA GLY HIS GLY GLN
* . * *
SEQRES  11    153  MET ASN LYS ALA LEU GLU LEU PHE ARG LYS ASP ILE ALA
SEQRES  12    153  ALA LYS TYR LYS GLU LEU GLY TYR GLN GLY
HET    HEM    154       43     PROTOPORPHYRIN IX CONTAINS FE(II)
HET    CMO    155        2
HET    SO4    156        5
HET    SO4    157        5
HETNAM     HEM PROTOPORPHYRIN IX CONTAINING FE
HETNAM     CMO CARBON MONOXIDE
HETNAM     SO4 SULFATE ION
HETSYN     HEM HEME
FORMUL   2  HEM    C34 H32 N4 O4 FE1
FORMUL   3  CMO    C1 O1
FORMUL   4  SO4    2(O4 S1 2-)
FORMUL   5  HOH   *183(H2 O1)
HELIX    1   A SER      3  GLU     18  1                                   16
HELIX    2   B ASP     20  SER     35  2                                   16
 * * *
HELIX    7   G PRO    100  ARG    118  7                                   19
HELIX    8   H GLY    124  LEU    149  8                                   26
LINK         FE   HEM   154                 NE2 HIS    93
LINK         FE   HEM   154                 C   CMO   155
CRYST1   63.430   30.430   34.120  90.00 105.67  90.00 P 1 21 1      2
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.015765  0.000000  0.004423        0.00000
SCALE2      0.000000  0.032862  0.000000        0.00000
SCALE3      0.000000  0.000000  0.030440        0.00000
ATOM      1  N   VAL     1      -3.710  15.394  13.509  1.00 21.43           N
ATOM      2  CA  VAL     1      -3.436  15.765  14.923  1.00 20.18           C
 * * *
ATOM   1240  O   GLY   153       4.598  24.472  -4.001  1.00 32.34           O
ATOM   1241  OXT GLY   153       4.780  25.435  -5.991  1.00 34.99           O
TER    1242      GLY   153
HETATM 1243 FE   HEM   154      14.555  27.882   4.668  1.00  7.74          FE
HETATM 1244  CHA HEM   154      15.863  31.036   4.935  1.00  6.95           C
HETATM 1245  CHB HEM   154      13.872  27.863   8.020  1.00  6.49           C
 * * *
HETATM 1479  O   HOH   530      25.100  27.692  17.628  1.00 49.74           O
HETATM 1480  O   HOH   531      -4.077  16.611  10.497  1.00 39.50           O
CONECT  764  762  763 1243
 * * *
CONECT 1296 1293
CONECT 1297 1293
MASTER      213    0    4    8    0    0    0    6 1479    1   57   12
END
```

Appendix C

Table C.1: Ranges of torsion angles used to determine α-helix and β-strands.

Secondary structures	Torsion angle ranges (ϕ, ψ)[1]
α-helix	(-130, -10) - (-120, 20)
	(-120, -30) - (-110, 30)
	(-110, -40) - (-100, 30)
	(-100, -60) - (-80, 30)
	(-80, -60) - (-70, 20)
	(-70, -70) - (-60, 0)
	(-60, -70) - (-50, -10)
	(-50, -70) - (-40, -20)
	(-40, -70) - (-30, -30)
β-sheet	(-176, -110) - (-160, 180)
	(-160, 120) - (-150, 180)
	(-150, 110) - (-140, 180)
	(-140, 100) - (-130, 180)
	(-130, 90) - (-120, 180)
	(-120, 100) - (-100, 180)
	(-100, 90) - (-70, 180)
	(-70, 100) - (-60, 170)
	(-60, 110) - (-50, 150)
	(-50, 110) - (-40, 140)

[1] See the section 3.1 for the description how these ranges were obtained.

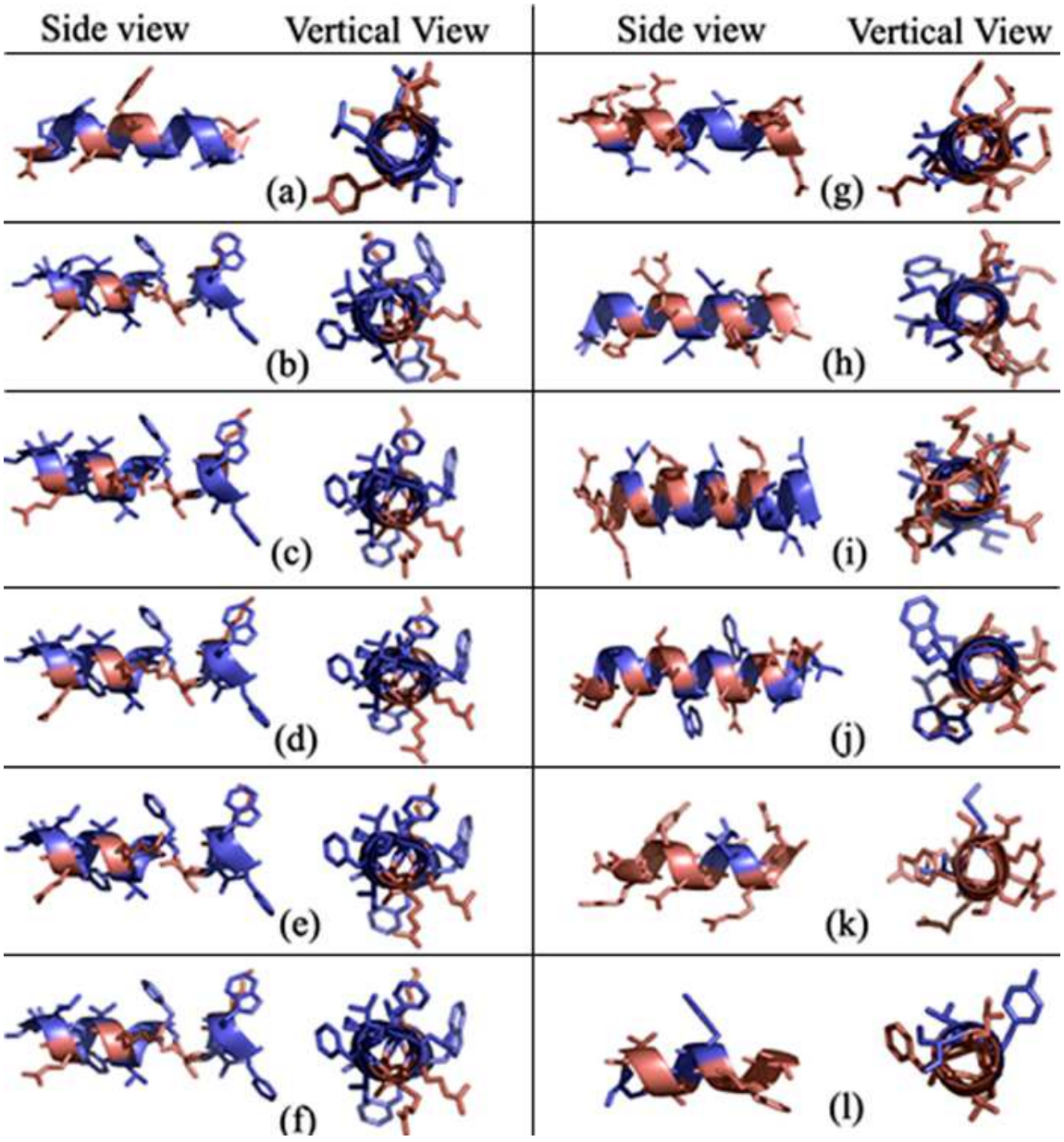
Appendix D

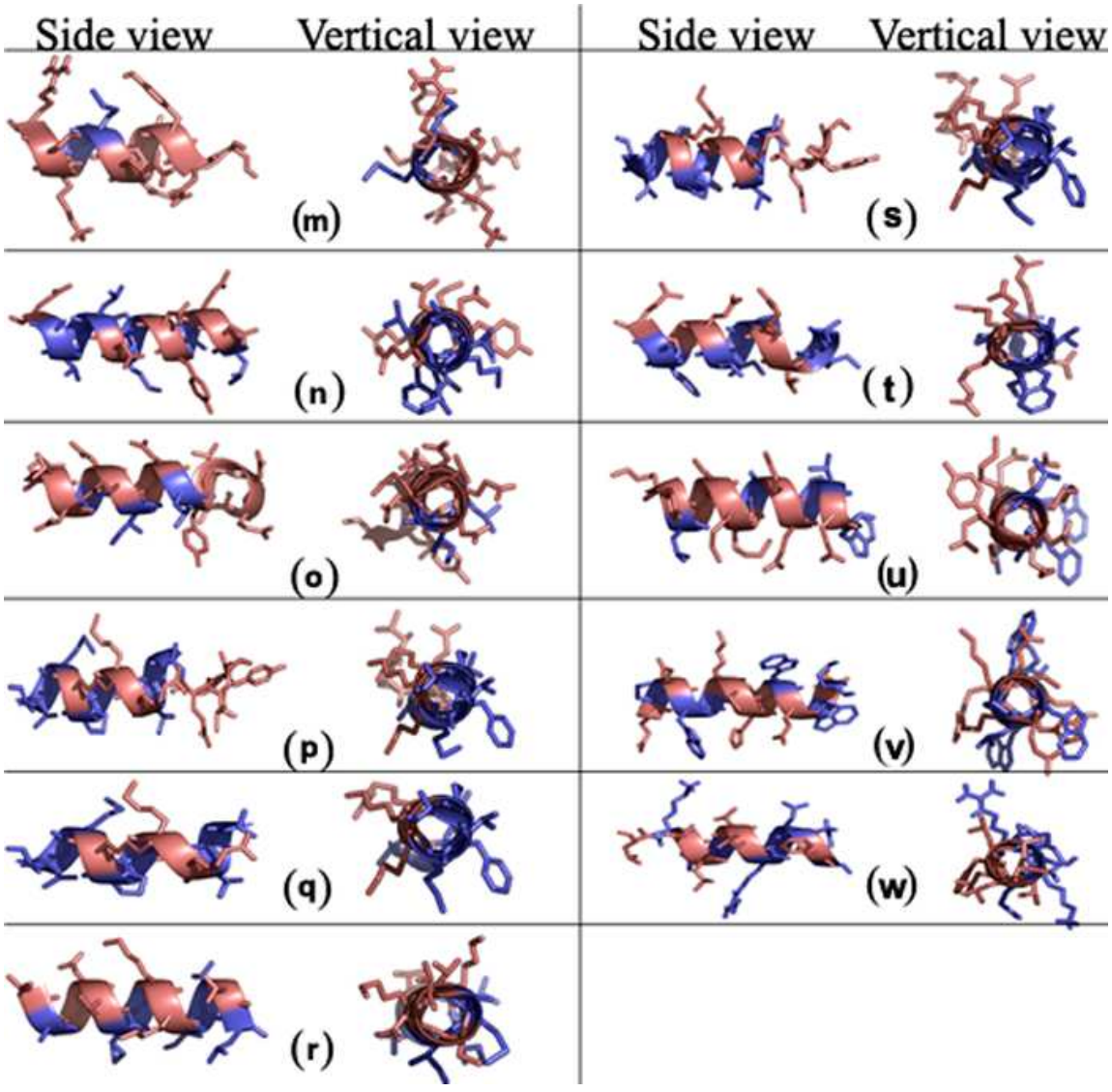
Table D.1: Atomic radii used in the study.

Atoms	Radius (Å)
Carbon (C)	0.767
Nitrogen (N)	0.702
Oxygen (O)	0.659
Sulfur (S)	1.052

Appendix E

Twenty three amphipathic α-helices identified in this study α-helices are derived from: (a)1AGX (bacterial amidohydrolase), (b)1A9O pentosyltransferase), (c)1A9P (pentosyltransferase), (d)1A9Q (pentosyltransferase), (e)1A9R (pentosyltransferase), (f)1A9T (transferase), (g)1ADS (oxidoreductase), (h)1AJH (myoglobin), (i)1ALD (lyase), (j)1AMP (hydrolase), (k)1ANG (hydrolase), (l)1AOV (transferin), (m)1AQP (hydrolase), (n)1ARL (carboxypeplidase), (o)1AST (hydrolase), (p)1BE0(dehalogenase), (q)1BEE(dehalogenase), (r)1BEO (fungal toxic elicitor), (s)1BEZ (dehalogenase), (t)1BGP (oxidoreductase), (u)1BH0(synthetic hormone), (v)1BIY (iron-binding protein), and (w)1BM9 (DNA-binding protein). Three more identified α-helices are listed in Figure 3.14.


Side view
Vertical View
Side view
Vertical View
(a)
(g)
(b)
(h)
(c)
(i)
(d)
(j)
(e)
(k)
(f)
(l)


Side view
Vertical view
Side view
Vertical view
(m)
(n)
(o)
(p)
(q)
(r)
(s)
(t)
(u)
(v)
(w)

Appendix F

Statistical analysis of the six amphipathic α-helix regions from all four positions.

Table F.1: Statistics of six amphipathic α-helix regions at position 1

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AHR	15	6	1	3	40	9	2	4	60	0.67	13	7	2	3	54	6	1	4	46	1.17	0.56
1BM9	8	2	0	0	25	6	1	3	75	0.33	5	3	0	0	60	2	0	0	40	1.5	1.51
1BVS	7	5	0	0	71	2	0	0	29	2.5	5	3	0	0	60	2	0	0	40	1.5	0.51
1MNK	11	3	0	0	27	8	1	6	73	0.38	9	7	2	4	78	2	0	0	22	3.5	2.22
2CMM	9	5	1	3	56	4	0	0	44	1.25	7	4	0	0	57	3	0	0	43	1.33	0.06
2REB	8	4	0	0	50	4	0	0	50	1	5	4	1	4	80	1	0	0	20	4	1.39

Table F.2: Statistics of six amphipathic α-helix regions at position 2

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AHR	15	3	0	0	20	12	1	9	80	0.25	13	10	2	6	77	3	0	0	23	3.33	2.59
1BM9	7	1	0	0	14	6	1	5	86	0.17	6	4	1	3	67	2	0	0	33	2	2.47
1BVS	6	6	1	6	100	0	0	0	0	1000	6	2	0	0	33	4	1	4	67	0.5	10
1MNK	11	4	1	3	36	7	2	3	64	0.57	9	6	1	4	67	3	0	0	33	2	1.26
2CMM	8	7	2	4	88	1	0	0	13	7	8	2	0	0	25	6	1	3	75	0.33	3.05
2REB	7	3	1	3	43	4	1	3	57	0.75	6	5	1	4	83	1	0	0	17	5	1.9

Table F.3: Statistics of six amphipathic α-helix regions at position 3

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AHR	15	9	2	3	60	6	1	3	40	1.5	13	4	0	0	31	9	2	5	69	0.44	1.23
1BM9	7	4	1	3	57	3	0	0	43	1.33	6	1	0	0	17	5	1	4	83	0.2	1.89
1BVS	7	5	0	0	71	2	0	0	29	2.5	5	3	0	0	60	2	0	0	40	1.5	0.51
1MNK	11	8	2	5	73	3	0	0	27	2.67	9	2	0	0	22	7	1	6	78	0.29	2.22
2CMM	9	4	0	0	44	5	0	0	56	0.8	7	5	1	4	71	2	0	0	29	2.5	1.14
2REB	7	6	1	5	86	1	0	0	14	6	6	2	0	0	33	4	0	0	67	0.5	2.48

Table F.4: Statistics of six amphipathic α-helix regions at position 4

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AHR	16	11	2	7	69	5	1	3	31	2.2	12	2	0	0	17	10	1	7	83	0.2	2.4
1BM9	8	4	1	3	50	4	0	0	50	1	5	1	0	0	20	4	1	4	80	0.25	1.39
1BVS	7	3	0	0	43	4	1	4	57	0.75	5	5	1	5	100	0	0	0	0	1000	10
1MNK	12	6	1	3	50	6	0	0	50	1	8	4	1	3	50	4	0	0	50	1	0
2CMM	10	4	0	0	40	6	0	0	60	0.67	6	5	1	3	83	1	0	0	17	5	2.01
2REB	8	5	1	3	63	3	0	0	38	1.67	5	3	1	3	60	2	0	0	40	1.5	0.11

Table F.5: Statistics of six non-structural regions at position 1

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AYN	13	8	1	5	62	5	1	3	38	1.6	13	9	2	3	69	4	0	0	31	2.25	0.34
1HQM	13	5	0	0	38	8	1	4	62	0.63	13	6	1	3	46	7	1	5	54	0.86	0.31
1HQM2	13	4	0	0	31	9	1	3	69	0.44	13	8	2	4	62	5	1	3	38	1.6	1.29
1OIT	13	11	2	7	85	2	0	0	15	5.5	13	6	0	0	46	7	1	4	54	0.86	1.86
1PK4	13	5	0	0	38	8	1	6	62	0.63	13	6	0	0	46	7	0	0	54	0.86	0.31
1PKK	13	6	1	3	46	7	1	4	54	0.86	13	7	1	4	54	6	1	3	46	1.17	0.31

Table F.6: Statistics of six non-structural regions at position 2

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AYN	14	8	2	3	57	6	1	4	43	1.33	12	9	2	5	75	3	1	3	25	3	0.81
1HQM	14	4	0	0	29	10	1	7	71	0.4	12	7	1	3	58	5	0	0	42	1.4	1.25
1HQM2	14	8	1	3	57	6	0	0	43	1.33	12	4	0	0	33	8	1	4	67	0.5	0.98
1OIT	14	9	2	4	64	5	1	3	36	1.8	12	8	1	6	67	4	0	0	33	2	0.11
1PK4	14	8	1	6	57	6	1	3	43	1.33	12	3	0	0	25	9	1	6	75	0.33	1.39
1PKK	14	7	1	4	50	7	1	4	50	1	12	6	1	3	50	6	1	3	50	1	0

Table F.7: Statistics of six non-structural regions at position 3

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AYN	15	10	2	3	67	5	0	0	33	2	11	7	1	4	64	4	1	3	36	1.75	0.13
1HQM	15	7	1	3	47	8	1	6	53	0.88	11	4	0	0	36	7	1	4	64	0.57	0.43
1HQM2	15	8	2	4	53	7	1	3	47	1.14	11	4	0	0	36	7	1	3	64	0.57	0.69
1OIT	15	8	2	3	53	7	1	4	47	1.14	11	9	2	6	82	2	0	0	18	4.5	1.37
1PK4	15	6	0	0	40	9	2	3	60	0.67	11	5	0	0	45	6	1	5	55	0.83	0.21
1PKK	15	7	1	4	47	8	1	4	53	0.88	11	6	1	3	55	5	1	3	45	1.2	0.31

Table F.8: Statistics of six non-structural regions at position 4

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1AYN	15	11	2	6	73	4	0	0	27	2.75	11	6	0	0	55	5	1	3	45	1.2	0.83
1HQM	15	8	1	3	53	7	0	0	47	1.14	11	3	0	0	27	8	1	6	73	0.38	1.1
1HQM2	15	6	0	0	40	9	1	4	60	0.67	11	6	0	0	55	5	0	0	45	1.2	0.58
1OIT	15	10	1	6	67	5	0	0	33	2	11	7	1	3	64	4	1	3	36	1.75	0.13
1PK4	15	5	0	0	33	10	1	7	67	0.5	11	6	1	5	55	5	0	0	45	1.2	0.88
1PKK	15	7	1	3	47	8	1	4	53	0.88	11	6	1	3	55	5	1	3	45	1.2	0.31

Appendix G

Results for the statistical analysis of the amphipathic and non-structural regions. See the section 4.1 for the explanation of each statistics.

Table G.1: Statistical analysis of the amphipathic α-helix regions.

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
1A8L4	9	3	0	0	33.33	6	1	3	66.67	0.5	6	5	1	5	83.33	1	0	0	16.67	5	2.3
1A9O3	8	7	1	6	87.5	1	0	0	12.5	7	6	3	0	0	50	3	0	0	50	1	1.95
1ADS2	8	2	0	0	25	6	1	4	75	0.33	6	4	1	3	66.67	2	0	0	33.33	2	1.8
1AGX4	8	7	1	7	87.5	1	0	0	12.5	7	6	3	0	0	50	3	0	0	50	1	1.95
1AH79	20	9	1	3	45	11	1	4	55	0.82	18	12	2	3	66.67	6	0	0	33.33	2	0.89
1AJG2	8	7	2	4	87.5	1	0	0	12.5	7	7	2	0	0	28.57	5	1	3	71.43	0.4	2.86
1ALD8	12	9	2	6	75	3	0	0	25	3	9	4	0	0	44.44	5	0	0	55.56	0.8	1.32
1AMP3	11	4	0	0	36.36	7	1	3	63.64	0.57	8	6	1	4	75	2	0	0	25	3	1.66
1ANG2	6	0	0	0	0	6	1	6	100	0	5	4	1	4	80	1	0	0	20	4	10
1AOV17	6	5	1	3	83.33	1	0	0	16.67	5	4	4	1	4	100	0	0	0	0	1000	10
1AQP2	6	0	0	0	0	6	1	6	100	0	5	4	1	4	80	1	0	0	20	4	10
1ARL1	8	4	1	4	50	4	1	3	50	1	6	5	1	4	83.33	1	0	0	16.67	5	1.61
1AST5	8	1	0	0	12.5	7	1	5	87.5	0.14	7	4	0	0	57.14	3	0	0	42.86	1.33	2.25
1BE09	9	3	0	0	33.33	6	2	3	66.67	0.5	6	6	1	6	100	0	0	0	0	1000	10
1BEE9	6	6	1	6	100	0	0	0	0	1000	5	2	0	0	40	3	1	3	60	0.67	10
1BEO4	7	7	1	7	100	0	0	0	0	1000	6	0	0	0	0	6	1	6	100	0	10
1BEZ9	9	3	0	0	33.33	6	2	3	66.67	0.5	6	6	1	6	100	0	0	0	0	1000	10
1BGP12	6	5	1	4	83.33	1	0	0	16.67	5	6	2	0	0	33.33	4	1	3	66.67	0.5	2.3
1BH01	8	5	1	4	62.5	3	0	0	37.5	1.67	6	1	0	0	16.67	5	1	5	83.33	0.2	2.12
1BIY1	8	4	0	0	50	4	0	0	50	1	6	3	0	0	50	3	0	0	50	1	0
1BM92	7	6	1	4	85.71	1	0	0	14.29	6	5	0	0	0	0	5	1	5	100	0	10

Table G.2: Statistical analysis of the non-structural regions.

ID	A side										B side										R_{max}
	N_A	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_A	N_B	N_1	C_1	L_1	P_1	N_2	C_2	L_2	P_2	R_B	
132L	10	7	1	4	70	3	0	0	30	2.33	6	1	0	0	17	5	1	3	83	0.2	2.46
189L	7	4	0	0	57	3	0	0	43	1.33	6	2	0	0	33	4	0	0	67	0.5	0.98
1A8Q	10	5	1	4	50	5	1	4	50	1	8	6	2	3	75	2	0	0	25	3	1.1
1AC5	8	4	0	0	50	4	0	0	50	1	6	0	0	0	0	6	1	6	100	0	10
1AHB	8	1	0	0	13	7	2	4	88	0.14	6	3	0	0	50	3	0	0	50	1	1.97
1AK9	9	8	1	6	89	1	0	0	11	8	6	4	1	3	67	2	0	0	33	2	1.39
1AMY	9	5	1	3	56	4	0	0	44	1.25	7	2	0	0	29	5	1	5	71	0.4	1.14
1APC	9	5	1	3	56	4	0	0	44	1.25	8	2	0	0	25	6	1	3	75	0.33	1.33
1AQN	9	8	1	6	89	1	0	0	11	8	6	4	1	3	67	2	0	0	33	2	1.39
1AST	7	3	0	0	43	4	0	0	57	0.75	6	4	0	0	67	2	0	0	33	2	0.98
1ATA	8	4	0	0	50	4	0	0	50	1	8	5	1	4	63	3	0	0	38	1.67	0.51
1AU9	10	4	1	3	40	6	1	5	60	0.67	7	2	0	0	29	5	0	0	71	0.4	0.52
1AYV	8	7	2	4	88	1	0	0	13	7	5	1	0	0	20	4	1	4	80	0.25	3.33
1AYX	9	6	0	0	67	3	0	0	33	2	8	2	0	0	25	6	1	5	75	0.33	1.8
1BAG2	8	5	0	0	63	3	0	0	38	1.67	6	2	0	0	33	4	0	0	67	0.5	1.21
1BF2	8	1	0	0	13	7	1	5	88	0.14	6	2	0	0	33	4	0	0	67	0.5	1.27
1BG5	10	9	2	5	90	1	0	0	10	9	6	4	1	4	67	2	0	0	33	2	1.16
1BG92	9	5	1	3	56	4	0	0	44	1.25	7	2	0	0	29	5	1	5	71	0.4	1.14
1BG9	10	8	1	7	80	2	0	0	20	4	7	4	1	4	57	3	0	0	43	1.33	1.1
1BGO	8	4	0	0	50	4	1	3	50	1	5	3	1	3	60	2	0	0	40	1.5	0.41
1BGP	10	4	0	0	40	6	1	3	60	0.67	8	2	0	0	25	6	1	5	75	0.33	0.71