DePaul University
UNIVERSITY LIBRARIES

DePaul University
Digital Commons@DePaul

College of Liberal Arts & Social Sciences Theses and Dissertations

College of Liberal Arts and Social Sciences

8-2011

Analysis of protein secondary structure via the discrete wavelet transform

Timothy E. Vanderleest
DePaul University, timvanderleest@gmail.com

Follow this and additional works at: https://via.library.depaul.edu/etd

Recommended Citation

Vanderleest, Timothy E., "Analysis of protein secondary structure via the discrete wavelet transform" (2011). *College of Liberal Arts & Social Sciences Theses and Dissertations*. 95.
https://via.library.depaul.edu/etd/95

This Thesis is brought to you for free and open access by the College of Liberal Arts and Social Sciences at Digital Commons@DePaul. It has been accepted for inclusion in College of Liberal Arts & Social Sciences Theses and Dissertations by an authorized administrator of Digital Commons@DePaul. For more information, please contact digitalservices@depaul.edu.

ANALYSIS OF PROTEIN SECONDARY STRUCTURE VIA THE DISCRETE WAVELET TRANSFORM

A Thesis

Presented in

Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

August, 2011

BY

Timothy Vanderleest

PHYSICS DEPARTMENT

College of Liberal Arts and Sciences

DePaul University

Chicago, Illinois

TABLE OF CONTENTS

LIST OF FIGURES **4**

ABSTRACT **6**

CHAPTER 1 Protein Folding **7**
- 1.1 Introduction 7
- 1.2 Protein Structure 9
  - 1.2.1 Amino acids 9
  - 1.2.2 The Polypeptide Chain 12
  - 1.2.3 The Three-Dimensional Structure of Proteins 14
  - 1.2.4 Alpha Helices and Beta Sheets 16
- 1.3 Protein Physics 20
  - 1.3.1 The Hydrophobic effect 20
  - 1.3.2 Molecular Dynamics and Forces 22
- 1.4 Models for Prediction of Secondary Structure 25

CHAPTER 2 Data Selection **29**
- 2.1 The Protein Data Bank 29
  - 2.1.1 The PDB file 30
  - 2.1.2 Protein Structure Determination 32
- 2.2 Protein Structural Classification 37
  - 2.2.1 The SCOP database 37
- 2.3 Filtering the Data 38
- 2.4 Data Characteristics 42

CHAPTER 3 The Discrete Wavelet Transform **46**
- 3.1 Introduction 46
- 3.2 The Haar Wavelet 47
  - 3.2.1 Example Calculation 51
  - 3.2.2 Edge Effects 52
- 3.3 The Daubechies Wavelets 54
- 3.4 Wavelet Reconstruction 57

TABLE OF CONTENTS – *Continued*

CHAPTER 4 Results 61
4.1 Introduction 61
4.2 Methods and Evaluation Measures 63
4.2.1 Prediction Evaluation Measures 67
4.3 Per-Residue Evaluation by Wavelet Enveloping 71
4.3.1 Threshold Optimization 76
4.4 Scale-Scale Measure 85

CHAPTER 5 Conclusion 91
5.1 Discussion of Results 92
5.2 Future Work 94
5.3 Summary 95

CHAPTER 6 Acknowledgements 96

APPENDIX A Success Measurement Program 97

APPENDIX B Wavelet Enveloping Prediction Program 100

APPENDIX C Secondary Structure Plotting Program 105

LIST OF FIGURES

1.1 The twenty standard amino acid structural formulas. 11
1.2 Diagram of the L and D isomer forms of amino acids. 12
1.3 Formation of a peptide bond between two amino acids. 13
1.4 Diagram of the main-chain angles ϕ and ψ. 14
1.5 Ramachandran plot for L-Ala residues. 15
1.6 The four levels of protein structure. 16
1.7 Diagram of the alpha helix structure. 17
1.8 Diagram of beta sheet structure. 19
1.9 Profile of the van der Waals interaction potential. 24

2.1 Example of some the important records of the PDB file 1A9N. . . . 31
2.2 Basic diagram of X-ray crystallography process. 34
2.3 Example of PDB data issue: non-standard residues. 40
2.4 Example of PDB data issue: secondary structure outside of chain length. 41
2.5 Histogram of α-helix lengths and β-strand lengths in the $\alpha + \beta$ data set. 43
2.6 Histogram of the α-helix lengths and β-strand lengths in the α/β data set. 44
2.7 Histogram of the chain lengths of the 6939 $\alpha + \beta$ chains and of the 8572 α/β chains. 45

3.1 Haar's scaling function $\varphi_{u,w}$ and wavelet function $\psi_{u,w}$. 48
3.2 Diagram breaking down a step function and wavelet function into two step functions of half the scale. 49
3.3 Diagram equating single step functions into sums of larger scale step and wavelet functions. 50
3.4 Multiscale decomposition of an arbitrary signal with the Haar wavelet transform. 53
3.5 The Daubechies 4 basic building block or scaling function $\varphi(r)$. . . . 56
3.6 The Daubechies wavelet function $\psi(r-1)$. 56
3.7 Diagram of the wavelet reconstruction of the original signal from the detail and approximation coefficients. 58
3.8 Diagram showing how to reconstruct a detail or approximation signal. 58
3.9 The hydrophobicity signal S of a protein and the reconstructed details at the first 6 scales. 60

LIST OF FIGURES – *Continued*

4.1 Flowchart outline of the main processing steps in our analysis. . . . 64
4.2 Hydrophobicity detail at four scales of PDB ID 1KJK. 67
4.3 Using the enveloping technique we analyze a protein using the three different hydrophobicity scales. 73
4.4 Distributions of the evaluation measures for the KD scale, H+S evaluation, with 1σ threshold. 75
4.5 Plot of B/P verses the threshold height for the 59 proteins with helix and strand Prevalence between 20-30%. 78
4.6 Three examples of proteins from the $\alpha + \beta$ database with high values of MCC. 82
4.7 Three examples of proteins from the α/β database with moderate to low values of MCC. 83
4.8 Scatter plot showing how the performance of our technique varied over proteins of different lengths. 85
4.9 Results using the scale-scale measure on 3 proteins from the $\alpha + \beta$ database with significantly different Prevalence levels and good levels of correlation. 87
4.10 Results using the scale-scale measure on 3 proteins from the $\alpha + \beta$ database with correlations closer to the database average. 88
4.11 Distribution of MCC values for the H+S evaluation using the scale-scale measure. 90

ABSTRACT

This project develops a secondary structure prediction approach that uses the discrete wavelet transform. In order to use the wavelet technique, we convert the primary amino acid sequence of the protein to a numerical signal using the hydrophobic tendencies associated with the amino acids. The data used in this project consists of both $\alpha + \beta$ and α/β proteins coming from the Structural Classification of Proteins (SCOP) protein database. This data provides both protein primary sequences and secondary structure locations. In total, 13,435 individual proteins and nearly 15,511 unique protein subunits are analyzed. We use three different experimentally determined hydrophobicity scales for comparison. A control data set is formed by creating 200 realizations of each protein, each realization being a random permutation of the proteins amino acid sequence. The realizations are subjected to the same analysis as the parent protein. Our analysis involves examining the correlation between locations of significant hydrophobicity fluctuations and secondary structure, where significance is determined by comparison to the control data set. Our focus is on using the first and second scales of the wavelet detail but we also construct a scale-scale measure that combines these scales to detect secondary structure. Using standard performance measures, like the Matthews correlation coefficient (MCC) and the accuracy (Q), we find that our method does show promise at being a useful tool for predicting the locations of secondary structures in protein given just the amino acid sequence.

CHAPTER 1

Protein Folding

1.1 Introduction

One of the most important problems in molecular biology today is the prediction of the structure and ultimately the function of proteins from their amino acid sequence. This problem only continues to grow in importance as the number of new protein sequences with unknown structure grows at an increasing rate. Currently the only accurate ways of determining protein structure are experimental techniques such as X-ray crystallography or NMR spectroscopy, but these methods are both expensive and time consuming. Thus a method of accurately predicting protein structure has been a top priority for many biologists, chemists, and physicists for decades.

The Human Genome Project (HGP), completed in 2003, identified all of the approximately 20-25,000 genes in human DNA. This is just one of many genomes that has been fully sequenced in the past couple decades which combine to make up millions of known gene sequences. Genes carry the information for making all of the proteins required by organisms, thus the estimated number of protein sequences is also in the millions. Despite all of this sequence information the number of known protein structures is only in the tens of thousands (there are approximately 70 thousand in the Protein Data Bank). This imbalance is one factor that is driving the effort to predict protein structure.

The importance of proteins to biology cannot be overemphasized. Proteins support every aspect of biological activity. They perform vital structural, transport, enzymatic, and regulatory functions in the cell. Defects in the structure of proteins can result in many different diseases and even cancer. For example, one protein that is associated to many different types of human cancer is called p53. Normal p53 functions as a tumor suppressor by regulating the cell cycle and has been called "the guardian of the genome [1]." Mutations in the p53 protein alters it's ability to regulate the cell and this leads to tumors and cancer. Studies have shown that different mutations in p53 lead to different types of cancer such as bladder, colon, esophagus, liver, leukemias and lymphomas, lung, breast, brain, ovary, and sarcoma [2]. An understanding of the link between protein structure and function could help in determining the role protein mutations have on tumor formation and in manipulating protein activity for cures.

The basis for the effort to predict protein structure comes from the famous hypothesis put forward in 1972 by the Nobel Prize laureate Christian Anfinsen. Anfinsen postulated that the three dimensional structure of protein is dictated by the "totality of interatomic interactions and hence by the amino acid sequence, in a given environment [3]." This hypothesis implies that a protein sequence along with characteristics of the environment (e.g. temperature and pH) are sufficient information for deducing the unique structure known as the *native* state or conformation. Thus far essentially all globular proteins studied appear to agree with Anfinsen's hypothesis. The challenge now is in determining an algorithm that takes the protein sequence as input and outputs the the total three dimensional structure.

While prediction of the overall three-dimensional structure is the main objective, the first step in this endeavor is the prediction of secondary structure elements such as alpha helices and beta sheets (§ 1.2.4). This reduces the complex three-dimensional problem into a greatly simplified one-dimensional problem, a mapping from amino acid sequence to the secondary structure identity of each residue along the chain. Secondary structure prediction is a hot area of research with over a hundred different techniques published, most of which are either based on statistics, knowledge of physical or chemical principles, or some hybrid method. Despite the reduction in complexity in going to the one-dimensional problem, secondary structure is still quite difficult to predict accurately.

The first section of this chapter begins with an overview of protein composition and structure. The second section discusses some of the important physical principles involved in protein folding and includes a description of the molecular dynamics approach to simulating the folding process. An important concept in this section is the hydrophobic effect which is considered to be the main driving force behind protein folding and a critical component to this project. Lastly, this chapter ends with a description of the general types of secondary structure prediction methods including three popular examples. However the approach taken in this thesis which will be discussed later in chapter 4 is different in many ways from these general methods.

1.2 Protein Structure

Proteins are very diverse macromolecules varying in size, composition, structure, and function. Before one can approach the problem of structure prediction it is first necessary to have a basic understanding of the chemistry and structural organization of proteins. This section covers amino acids, the polypeptide chain, and the three dimensional structure of proteins.

1.2.1 Amino acids

Amino acids are the building blocks of proteins. Despite the great diversity of functionality in proteins most are composed of just 20 common amino acids. Each of these 20 common or standard amino acids have a name, a three letter abbreviation and a one letter symbol (see Table 1.1 below) for convenience in presenting protein sequence information.

Amino acids are composed of four groups bonded to a carbon atom known as the alpha carbon (C_α): the carboxyl group (COO^-), the amino group (NH_3^+), a

Table 1.1: Amino acid names, abbreviations, and symbols

Name	Abbreviation	Symbol	Name	Abbreviation	Symbol
Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartic Acid	Asp	D	Phenylalanine	Phe	F
Cysteine	Cys	C	Proline	Pro	P
Glutamic Acid	Glu	E	Serine	Ser	S
Glutamine	Gln	Q	Threonine	Thr	T
Glycine	Gly	G	Tyrosine	Tyr	Y
Histidine	His	H	Tryptophan	Trp	W
Isoleucine	Ile	I	Valine	Val	V

hydrogen atom (H), and the R group (see Fig. 1.1). What distinguishes the 20 common amino acids is their R groups, also known as side chains. These side chains vary in structure, size, electric charge, and influence how the amino acid reacts to water. Figure 1.1 displays the structural formulas of the amino acids and groups them into some of the major categories according to physical properties. Seven of the amino acids are categorized as having non-polar aliphatic R groups and consist mainly of straight or branching hydrocarbon chains. Five of the amino acids are classified as having polar and uncharged R groups. The R groups of the three aromatic amino acids contain planar ring systems and are relatively non-polar. Lastly, there are three amino acids with positively charged R groups and two with negatively charged R groups. One amino acid that stands out from the rest is proline. It has a side chain that covalently bonds to the nitrogen atom of its amino group. At a point later on in the chapter (§ 1.3.1) we discuss the hydrophobic effect which is based on the interactions between polar (water attracting) and non-polar (water repelling) amino acids.

Based on the absolute configuration of the four groups around the alpha carbon,

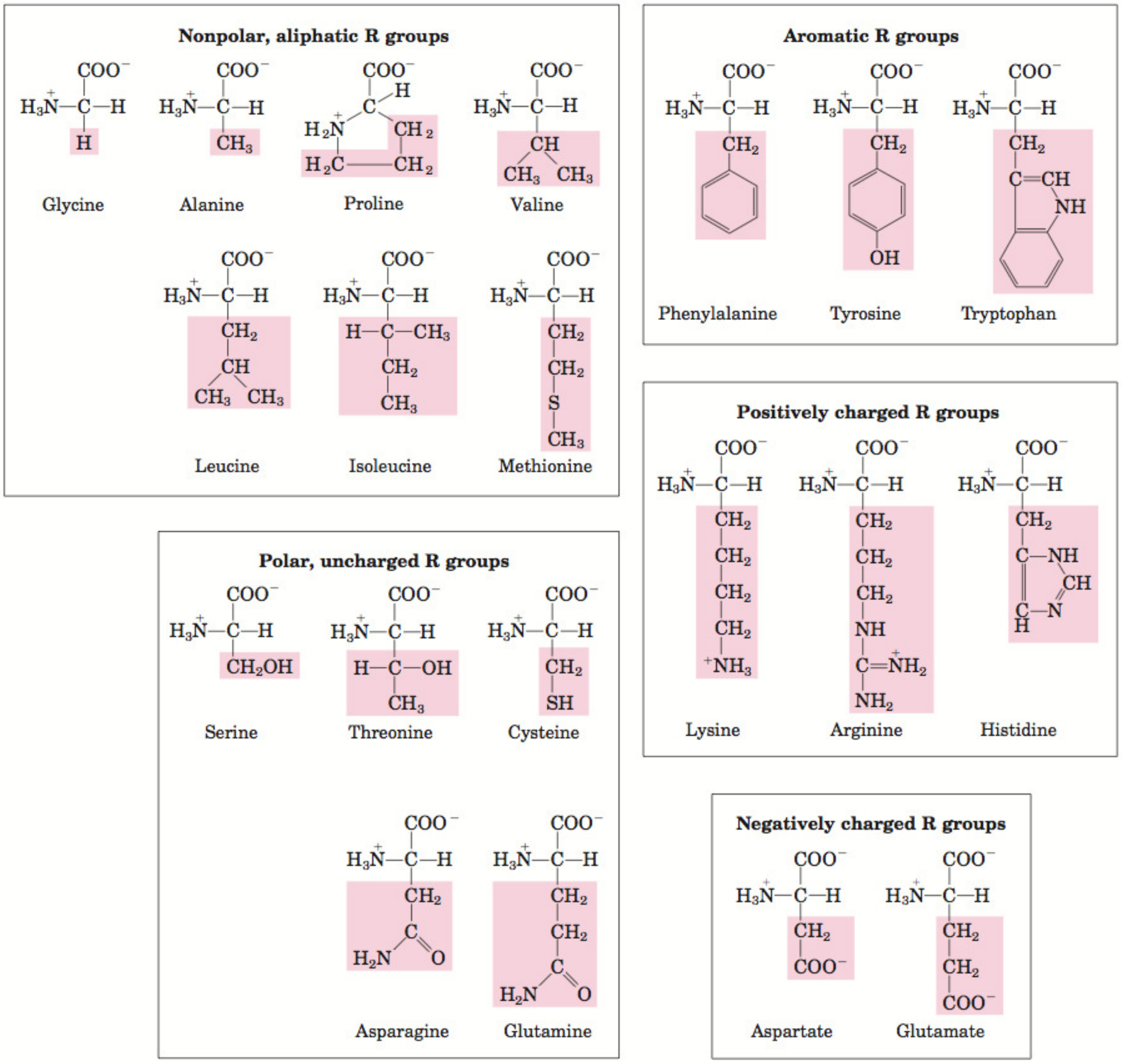

Nonpolar, aliphatic R groups
Glycine
Alanine
Proline
Valine
Leucine
Isoleucine
Methionine
Aromatic R groups
Phenylalanine
Tyrosine
Tryptophan
Positively charged R groups
Lysine
Arginine
Histidine
Polar, uncharged R groups
Serine
Threonine
Cysteine
Asparagine
Glutamine
Negatively charged R groups
Aspartate
Glutamate

Figure 1.1: The twenty standard amino acid structural formulas. The R group which distinguishes the amino acids are highlighted [4].

amino acids can be classified into two categories: D and L isomers. The difference between the two is that they are mirror images of each other (see Fig. 1.2). Proteins are built up of only L amino acids for it is only these that are encoded by the genes. This is important because the formation of stable substructures in proteins generally requires the amino acid components to be of the same isomer type [4].

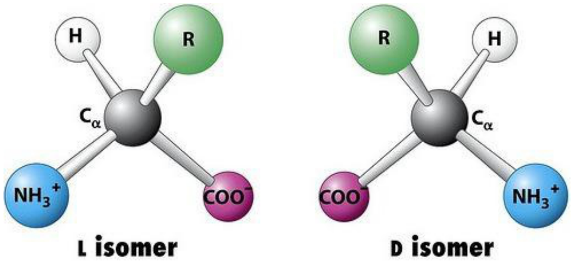


Figure 1.2: Diagram of the L and D isomer forms of amino acids. In protein they are virtually always found in the L form [5].

Most proteins are composed only of the 20 standard amino acids and these are called simple proteins. There are some however, termed conjugated proteins, that contain permanent chemical groups in addition to amino acids. The non-amino acid parts of these proteins are called prosthetic groups. Conjugated proteins are classified according to the type of prosthetic group attached. Some examples are lipoproteins, glycoproteins, phosphoproteins, and metalloproteins. In this project only simple proteins were used in our data set.

1.2.2 The Polypeptide Chain

Proteins are synthesized by the formation of peptide bonds between amino acids (see Fig. 1.3). In the peptide bond the amino group of one amino acid couples with the carboxyl group of the other. Two amino acids join together to form a dipeptide, three form a tripeptide, four a tetrapeptide, and so on. When these bonds are formed a molecule of water is released and consequently the amino acids in the chain are called *residues*. The term peptide is usually used to refer to short sequences of residues whereas the term polypeptide applies to longer chains of residues. The lengths of protein chains vary greatly with some shorter than a hundred residues and some exceeding a few thousand.

Every protein that is not cyclic has a first residue and a last residue. The amino group of the first residue is known as the N-terminus and the carboxyl group of

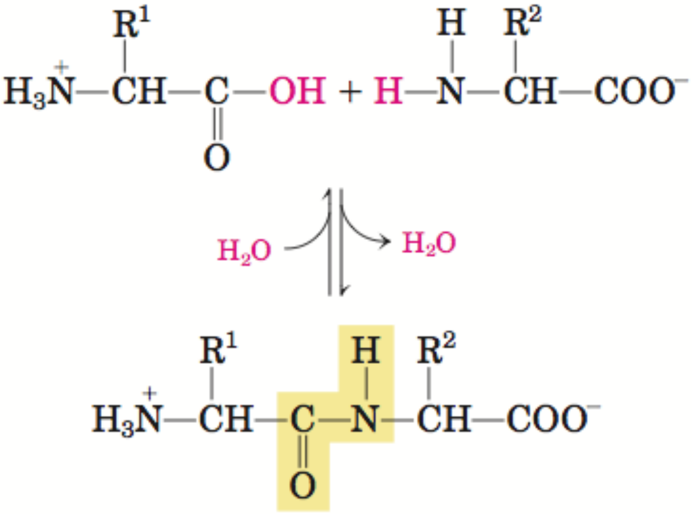

R^1
H
R^2
$H_3\overset{+}{N}$—CH—C—OH + H—N—CH—COO⁻
O
H_2O
H_2O
R^1
H
R^2
$H_3\overset{+}{N}$—CH—C—N—CH—COO⁻
O

Figure 1.3: Formation of a peptide bond (shaded) between two amino acids. The carboxyl group of one amino acid reacts with the amino group of the other resulting in the condensation of a water molecule and a strong peptide bond [4].

the last residue is the C-terminus. When protein amino acid sequences are listed they read from left to right starting with the N terminal reside and ending with the C-terminal residue. This order is also the sequence in which proteins are synthesized by ribosomes.

The polypeptide conformation refers to the curve in three-dimensional space that the back-bone traces out. The back-bone can be thought of as a chain consisting of flat rigid peptide units that are connected by the C_α atoms of the amino acids. The C_α atom has two single bonds of which both are able to rotate and this provides flexibility to the chain. The torsion angle of the C_α-N bond is designated by ϕ and the torsion angle of the C_α-C bond by ψ (see Fig. 1.4). To specify the full conformation of a protein these two angles are needed for each of these bonds along the chain.

The Ramachandran plot is a convenient way of looking at the likelihood of the

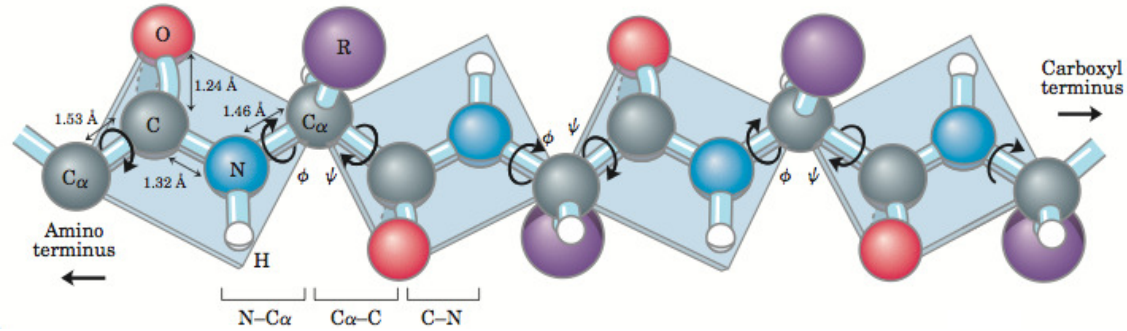


Figure 1.4: The orientation of any two residues in a peptide bond can be defined by the angles ϕ and ψ which constitute the degrees of freedom in a polypeptide chain [4].

various conformations of a peptide bond (see Fig. 1.5). It serves as a contour plot over the two dimensional conformation space defined by ψ and ϕ in which the contour levels represent the measure of how favorable a conformation is. Regions of the plot that are unfavorable or disallowed are due to steric interactions i.e. a result of the fact that two atoms of a molecule can not occupy the same space. The Ramachandran plots of most of the residues look almost identical except for glycine and proline. Glycine, which has a single hydrogen atom for its side chain, is less sterically restricted and therefore the Ramachandran plot shows a much broader range of allowed conformations. Proline on the other hand is greatly restricted due to its cyclic side chain.

1.2.3 The Three-Dimensional Structure of Proteins

There are four different levels of structural organization in proteins arranged hierarchically (see Fig. 1.6). The first level corresponding to the smallest scale is known as the *primary structure*. The primary structure is simply the sequence or linear order of amino acids from the N-terminal residue to the C-terminal residue. Proteins are defined by this primary sequence information and all subsequent levels of structure (secondary, tertiary, and quaternary) rely on it.

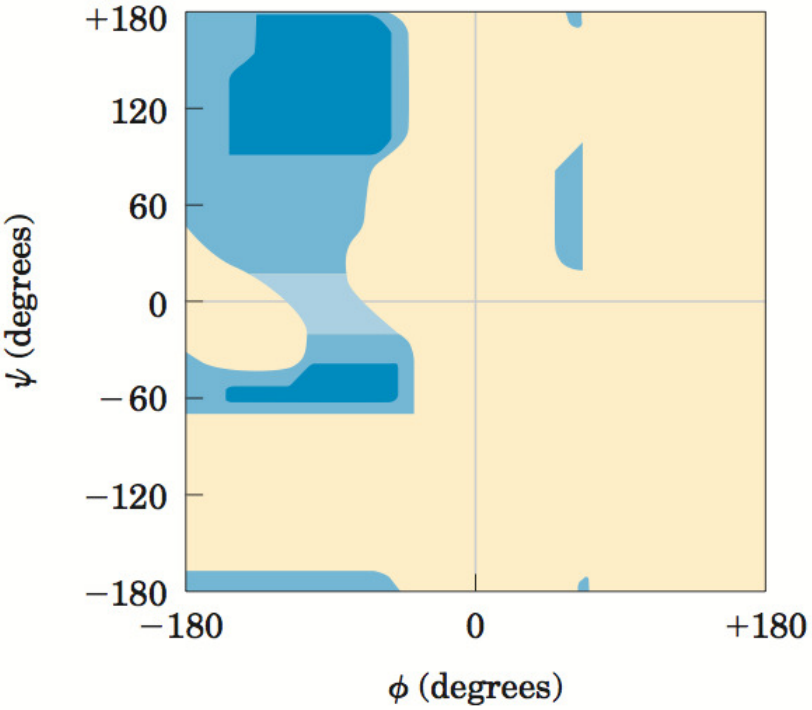


Figure 1.5: Ramachandran plot for L-Ala residues. The conformation of peptides are defined by the values of the angles ϕ and ψ. The darkest level of shading corresponds to angles that are fully allowed, the second level are allowed but less likely due to unfavorable atomic contacts, the third level is permissible but even less likely, and all other angles rarely occur [4].

The next level of protein structure, corresponding to a larger scale, is called *secondary structure*. Secondary structure refers to the local conformation of a few or many neighboring residues. Two major types of secondary structure that are found often in proteins are α-helices and β-sheets (§ 1.2.4). These are regular repeating structures. Regions of the chain that do not repeat but take on sort of a random appearance are referred to as coil. In the effort to predict protein structure the first step is to accurately predict helix and sheet structures.

The third level of protein structure is called *tertiary structure*. This is the global three-dimensional shape of the polypeptide backbone. The two major structural

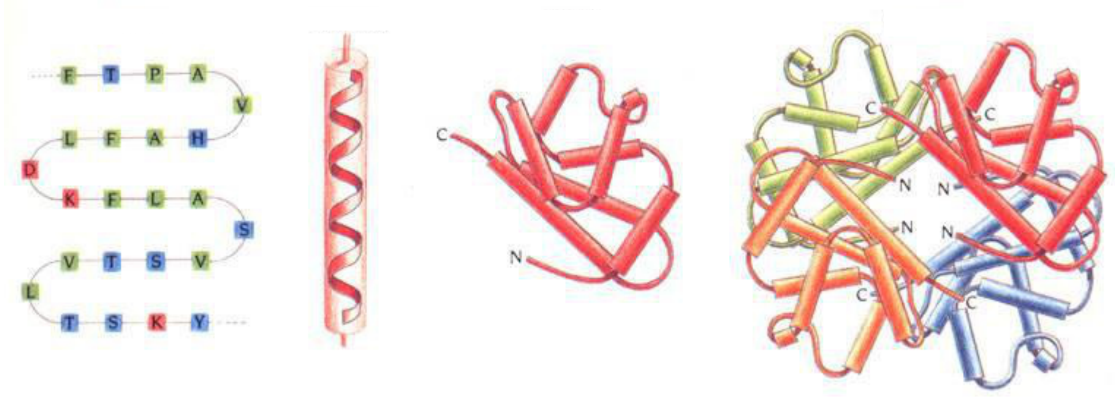


Figure 1.6: The four levels of protein structure (from left to right): primary, secondary, tertiary, and quaternary [6].

classes of proteins are globular and fibrous proteins. Fibrous proteins have an elongated shape and are dominated by helices and sheet secondary structure. Globular proteins, which are the focus of the structure prediction effort, have a spherical-like shape with parts of the polypeptide chain clustered in an interior region.

The fourth and last level of protein structure, called *quaternary structure*, applies to multi-subunit (or multi-chain) proteins. Many proteins actually consist of assemblies of more than one polypeptide chain connected by non-covalent bonding. These subunits are often identical but can be quite different in amino acid sequence and length. Quaternary structure is the arrangement of the subunits that form the whole protein.

1.2.4 Alpha Helices and Beta Sheets

In 1951 Linus Pauling, Robert Corey, and Herman Branson [7] predicted that the polypeptide chain could take on a helical structure which they called the α helix (see Fig. 1.7). Their prediction was based on x-ray studies from William Astbury [8] who, in the 1930s, found a regular structure in the protein of hair and porcupine quills that repeats every 5.15 to 5.2 Å. This distance corresponded to the regular separation distance between winds of the helix.

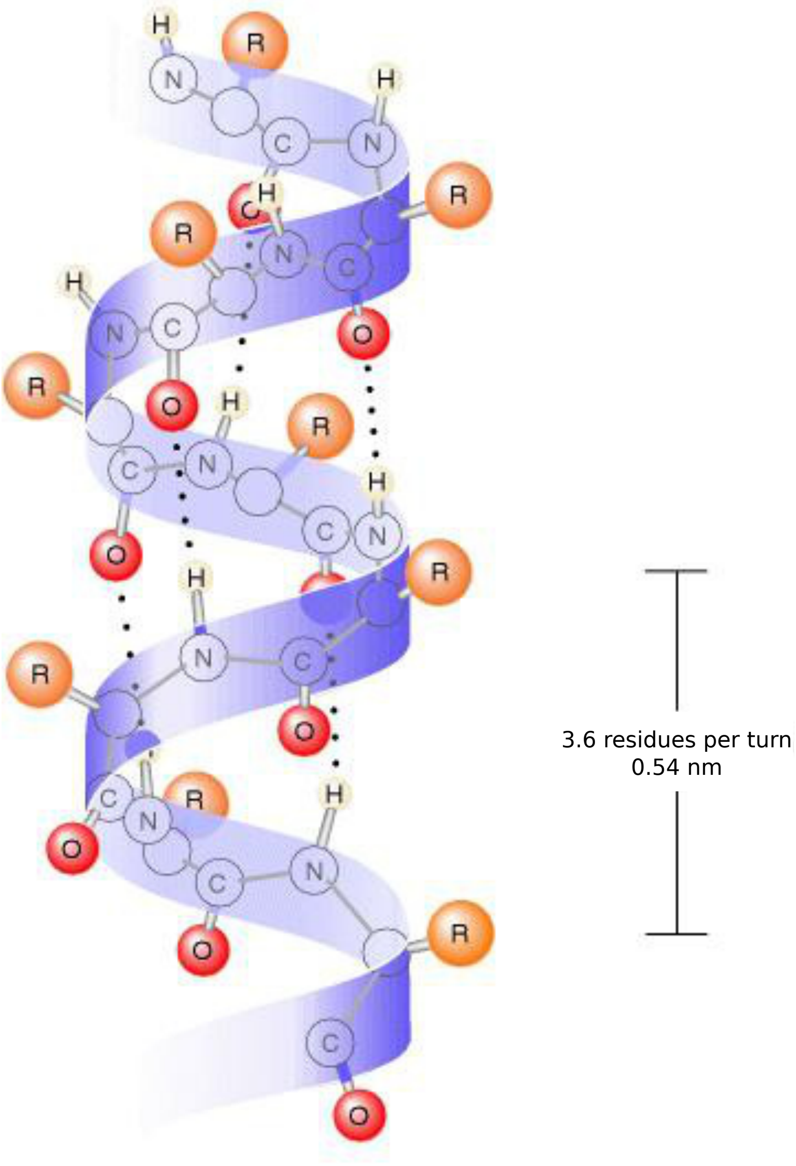


Figure 1.7: The alpha helix structure repeats every 3.6 residues which amounts to a distance along the helical axis of 0.54 nm. The dotted lines represent hydrogen bonds between the carboxyl and amino groups of every fourth residue. The open circles in the diagram are the alpha carbon atoms, C_α.

Alpha helices are found in proteins when a stretch of residues all have approximately the same bond angles ϕ and ψ of $-60°$ and $-50°$ respectively. These angles correspond to the bottom left region of the Ramachandran plot (see Fig. 1.5). Alpha helices vary in length from four or five to 40 residues with the average being around 10 [6]. With 3.6 residues and 5.2 Å per winding, a 10 residue α helix would be a little under 3 full rotations and have a length of about 15 Å.

The side chains of helices project outward from their central axis and therefore do not interfere with the structure except for Proline whose side chain bonds to its main chain N atom. For this reason Proline is rarely found in helices but when it does it usually produces a bend in the helical axis. A common location for helices to be found is on the protein globular surface with one side facing the core of the protein and the other exposed to solution. In these helices it is often found that the side facing the interior contains more non-polar (water repelling) side chains and the exposed side more polar (water attracting) leading to a periodicity of polarity in the sequence of about 3 to 4 residues. These are called amphipathic alpha helices.

The alpha helix is just one type of helix structure possible in proteins. Helix structures result from hydrogen bonds between the C=O group of one residue and the H–N group of a neighboring residue along the polypeptide backbone. When this bond is with the adjacent or second residue you have a 2_7-helix and for bonding with the third, fourth, or fifth residue you get 3_{10}, 4_{13}(also called α), and 5_{16}(called π)-helix respectively. The most abundant helix is the alpha helix but 3_{10}-helices are sometimes found as short fragments. In addition to different bond locations helices can also can be oriented right-handed or left-handed based on which way they wind. Left handed helixes are hardly ever observed in proteins.

The second major secondary structure found in proteins is the beta sheet which was also predicted by Pauling and Corey [9]. In this structure different regions of the chain are hydrogen bonded together side by side forming a pleated sheet (see Fig. 1.8). Each chain segment in the beta sheet conformation is called a beta strand.

It is import to note that while beta strands are local structures between neighboring residues beta sheets can be very nonlocal with strands joined from distant parts of the polypeptide chain. Strands are typically between 5 to 10 residues in length and nearly fully extended. The ϕ and ψ angles for these structures are in the top left region of the Ramachandran plot and there is much broader range of allowed angles than for alpha helices.

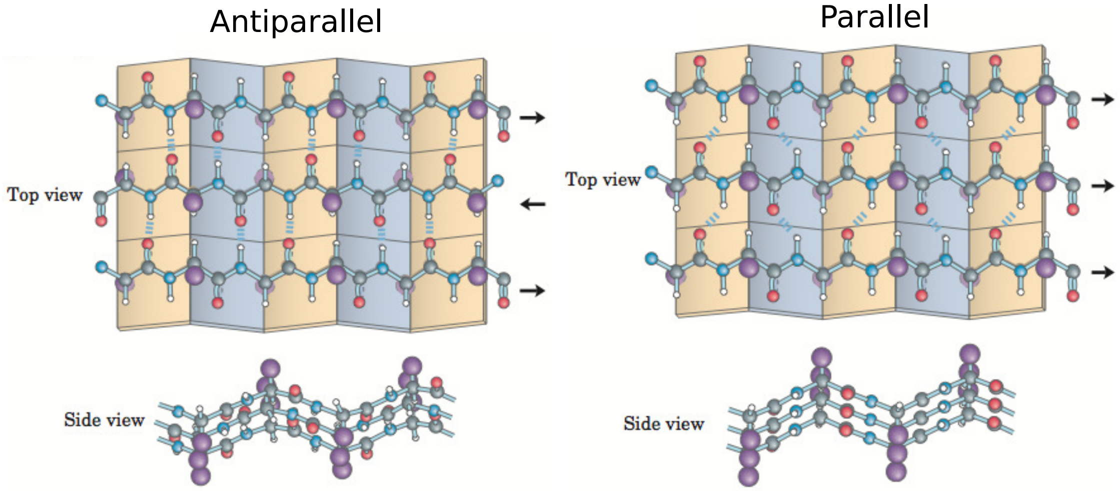


Figure 1.8: Antiparallel and parallel beta sheet conformations. The R-groups are colored purple and as shown in the side view alternate between being above and below the sheet [4].

There are two possible arrangements of two strands within a β-sheet conformation. They can be oriented parallel such that the strand's amino or carboxyl terminals point in the same direction or antiparallel if the directions are opposite. Beta sheets may consist of strands that are all parallel, all antiparallel, or a mixture of the two. The side chains of the residues protrude out normal to the surface of the sheet switching between the sides of the sheet with each consecutive residue.

Besides α-helices and β-sheets a third type of secondary structure is the turn or β-turn. This is a structure in which the polypeptide backbone folds back on itself or just changes direction. Turns are generally classified based on the number of residues that compose them which is often three or four.

1.3 Protein Physics

Initially protein folding was considered a problem that strictly belonged to the realm of biochemistry, however by the early 1990's physics based approaches had achieved many successes in explaining key aspects of protein folding [10]. One classic example is the so called Levinthal's paradox, a thought experiment that asked how a protein could fold so quickly when there are an astronomically large number of possible conformations for the protein to search through before finding the native state. Physics has provided the answer, through statistical mechanical modeling, that folding involves a funnel-shaped energy landscape. In other words proteins don't randomly search through conformation space but rather they gradually approach the minimum energy which corresponds to the native state. In this section we look at some of the forces that effect protein and the molecular dynamics approach to protein folding. Molecular dynamics or, as it is sometimes referred to, the "pure physics" approach is not the method used in this thesis. In this work we take a semi-physical approach in that we only consider the driving force behind protein structure, the hydrophobic effect.

1.3.1 The Hydrophobic effect

It is widely viewed that the hydrophobic effect is what drives protein folding and for this reason it serves as the basis of the secondary structure prediction method of this thesis. To put it simply, non-polar atoms or molecules such as hydrocarbons have a preference to reduce their contact with water. This phenomenon accounts for the insolubility of oil in water. Polar molecules, on the other hand, are attracted to water because water molecules can hydrogen bond to them. Accordingly, non-polar molecules are referred to as *hydrophobic* and polar molecules are called *hydrophilic.*

One of the major consequences of the hydrophobic effect is that non-polar residues congregate towards the core of globular proteins, effectively hiding from water molecules. On the other hand polar residues are usually found on or near the surface. Since the side chains of many amino acids are hydrophobic it is understandable that the hydrophobic effect may play a significant part in intramolecular forces.

Table 1.2: The three hydrophobicity scales used in this work

Animo acid	Kyte-Doolittle	Hopp-Woods	Engelman-Steitz
Ala	1.8	-0.5	-1.6
Arg	-4.5	3.0	12.3
Asn	-3.5	0.2	4.8
Asp	-3.5	3.0	9.2
Cys	2.5	-1.0	-2.0
Gln	-3.5	0.2	4.1
Glu	-3.5	3.0	8.2
Gly	-0.4	0.0	-1.0
His	-3.2	-0.5	3.0
Ile	4.5	-1.8	-3.1
Leu	3.8	-1.8	-2.8
Lys	-3.9	3.0	8.8
Met	1.9	-1.3	-3.4
Phe	2.8	-2.5	-3.7
Pro	-1.6	0.0	0.2
Ser	-0.8	0.3	-0.6
Thr	-0.7	-0.4	-1.2
Trp	-0.9	-3.4	-1.9
Tyr	-1.3	-2.3	0.7
Val	4.2	-1.5	-2.6

Experimental studies have been carried out on each of the standard amino acids to produce hydrophobicity scales. Three prominent scales are the Kyte-Doolittle (KD) [11], Hopp-Woods (HW) [12], and Engelman-Steitz (ES) [13] scales which are shown in Table 1.2. These scales give relative measures of how attractive or repulsive the interaction each residue experiences in an aqueous environment.

In the Kyte-Doolittle scale positive values are assigned to hydrophobic residues whereas in the Engleman-Steitz and Hopp-Woods it is just the opposite. The KD scale is a very popular scale for identifying hydrophobic regions of proteins; it has had good success in predicting transmembrane and surface-exposed regions of proteins. In this project we use these hydrophobicity scales to replace the primary amino acid sequence with a sequence of hydrophobicity values (or a hydrophobicity signal) which we then analyze for predicting secondary structure.

1.3.2 Molecular Dynamics and Forces

This section briefly describes some of the equations used in molecular dynamics (MD) along with the major forces involved. MD is important because it gives information about the folding and unfolding pathways, the native structure, and the inter-residue interactions of proteins. To start we give an overview of some of the major forces involved in MD.

Electrostatic interactions occur between the R groups of charged residues and between the NH_3^+ and COO^- groups of a protein's terminal residues. Of the standard amino acids lysine, arginine, and histidine have positively charged R groups and aspartate and glutamate have negatively charged R groups (see Fig. 1.1). As a consequence of their charges these residues are usually found near the aqueous surface of proteins where they can interact with water molecules. The energy of the interactions between any two charged atoms is described by the Coulomb potential and can be expressed by the following:

$$V_{ij} = \frac{q_i q_j}{4\pi\epsilon r_{ij}} \tag{1.1}$$

where q_i and q_j are the effective charges on the jth and ith atoms, the distance between them is r_{ij}, and ϵ is the dielectric constant.

All atoms or molecules, without net charge, are found to attract at large distances and repel when they are close together. This is known as the van der Waals interaction and it is caused by coordinated fluctuating dipoles. Electrons are mobile and at any instant they may be found to one side of a molecule resulting in a temporary dipole. These temporary dipoles induce dipoles in the neighboring molecules which results in synchronized fluctuating dipoles. The attractive force between any two dipoles of like polarization arrises because of the electrostatic force between the positive side of one molecule and the negative side of the other. The repulsive effect is due to the Pauli exclusion principle. The total potential of van der Waals interactions is approximately described by the Lennard-Jones potential of the form (see Fig. 1.9):

$$V_{ij} = E_0 \left[\left(\frac{r_0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_0}{r_{ij}} \right)^{6} \right] \tag{1.2}$$

were r_0 is the equilibrium distance between the ith and jth atoms and E_0 is the minimum energy. For different pairs of atoms r_0 and E_0 will in general have different values. Van der Waals forces are extremely weak when compared to other forces influencing protein conformation yet the large number of these close interactions can be an important force in maintaining tertiary structure.

The accuracy of MD simulations is dependent on the exactness of the potential functions used. Classically the atoms making up the molecules of a protein can be treated by Newton's laws of motion:

$$m_i \ddot{\mathbf{r}}_i = \mathbf{F}_i \qquad i = 1, 2, ..., N \qquad \ddot{\mathbf{r}}_i = \frac{d^2 \mathbf{r}_i}{dt^2} \tag{1.3}$$

where m_i is the mass of the ith of N atoms, $\mathbf{r}_i$ is its corresponding position, F_i the net force acting on the atom, and $\ddot{\mathbf{r}}_i$ the corresponding acceleration.

At the fully atomic level the forces are only potential forces however collisions and friction forces should be included to mimic collisions of the solute with the environment. This is the Langevin or Brownian dynamics treatment of the motion of particles in a fluid. With these considerations Equation 1.3 becomes a stochastic

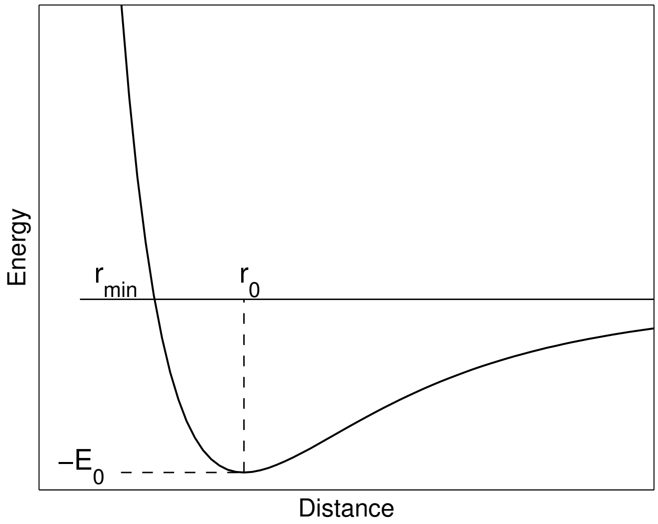


Figure 1.9: Profile of the van der Waals interaction potential.

differential equation with forces given by Equation 1.4:

$$F_i = -\nabla_{r_i} U\left(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N\right) - m_i \gamma_i \dot{\mathbf{r}}_i + \mathbf{R}_i(t) \qquad i = 1, 2, ..., N, \tag{1.4}$$

where U is the potential energy of the system which would include van der Waals and electrostatic potentials (Equations 1.1 and 1.2); γ_i and $\dot{\mathbf{r}}$ are the friction coefficient and velocity of atom i, respectively; and $\mathbf{R}_i(t)$ is the vector of random forces, of zero mean, on the atoms from the solvent environment [14].

Equations 1.3 and 1.4 along with the atom's initial coordinates and velocities must be solved numerically. There are a variety of numerical algorithms to choose from but the demand in molecular dynamics for high accuracy and low computational cost has lead to Verlet-type algorithms (the Verlet, velocity-Verlet, and the leap-frog algorithms) being the most common [14]. A couple of the advantages of Verlet-type algorithms is that they are fourth-order (the error goes as the fourth power of the integration time step Δt) and it conserves energy when there are no non-conservative forces involved. The velocity-Verlet algorithm is shown below.

Update positions in step 1:

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \dot{\mathbf{r}}(t)\Delta t + \ddot{\mathbf{r}}(t)\Delta t^2/2 \tag{1.5}$$

Update velocities in step 2:

$$\dot{\mathbf{r}}(t+\Delta t) = \dot{\mathbf{r}}(t) + [\ddot{\mathbf{r}}(t) + \ddot{\mathbf{r}}(t+\Delta t)]\Delta t/2 \tag{1.6}$$

For the numerical algorithm to be stable a time step must be chosen that is smaller than the fastest motions of the system. Hydrogen bond vibrational periods are on the order of 10 fs, thus time steps are typically chosen on the order of 1 fs. These time scales are much smaller than the time scales of helix formation or the folding of α-helical proteins which are on the order of microseconds [14].

At present MD falls short in three major respects: its ability to reproduce true protein potential energy functions, the computational power necessary to carry out micro to milli-second (very long compared to protein folding time scales) simulations, and the ability to obtain adequate sampling to characterize the folding process and analyze the data efficiently [15]. Because of the computational cost MD simulations are limited to small proteins, typically fewer than 100 residues. However, once “physics-only” or “physics-mainly” approaches do become successful there will be numerous advantages such as the ability to predict changes in conformation, to understand protein folding mechanisms and motions, and the ability to design synthetic proteins for new applications [16].

1.4 Models for Prediction of Secondary Structure

In the previous section we discussed molecular dynamics and some of the challenges that it faces in modeling the structure and folding pathways. While MD is a pure physics based method the majority of secondary structure prediction methods are based on empirical data. Of the numerous different methods that have been devised there are roughly three different categories they fall into: statistical, knowledge-based, and hybrid systems. For sake of illustration, below we take a look at two very well known statistical-based systems, the Chou-Fasman and GOR methods, and the highly successful knowledge-based system, machine learning. Today machine learning methods have reached accuracy measures of $\sim$80% (where this is the Q measure defined later in § 4.2.1).

One thing to note about secondary structure prediction methods is that they sometimes vary in what is predicted. For example, most methods only predict the two well defined secondary structure elements, α-helices and β-sheets, and everything else is considered coil. However, some advanced modern methods aim to predict other structures such as 3_{10}-helices or various types of turns.

The Chou-Fasman system [17], originally developed in the 1970's, is one of the most widely used schemes. This is a statistical based system that owes its popularity to its being easy to understand and reasonably successful. The Chou-Fasman method uses a set of proteins with known primary and secondary structure to calculate the statistical propensities of a particular residue for forming either an α-helix or a β-strand. The propensities are used to classify the residues into six classes according to their likelihood of forming an α-helix and six classes according to their likelihood of forming a β-strand. The class appointments are then used to predict the probable locations of helices and strands and finally the prediction is modified by a series of rules that constructs the final prediction. While the Chou-Fasman system is easy to understand it has a few disadvantages. For one, it isn't directly related to any chemical or physical theory and therefore doesn't give any physical explanation for the formation of secondary structure. Also, the statistics used are naive and the prediction rules are somewhat arbitrary.

The GOR (Garnier-Osguthorpe-Robson) method [18] is another statistical method first developed in the 1970's and through several improvements over the years is now in its fifth version. Unlike the Chou-Fasman method GOR is much more complex using Bayesian statistical principles and information theory. It uses a database of 267 protein chains comprised of 63,566 residues and known secondary structure to calculate what is called the information function:

$$I(S;R) = \log[P(S|R)/P(S)]. \tag{1.7}$$

The information function is defined as the logarithm of the ratio of conditional probability $P(S|R)$ where S is one of the three conformation states [helix (H), extended strand (E), or coil (C)] for the residue R (one of the 20 possible standard amino acids) and the probability $P(S)$ of the occurrence of conformation S. The state of a particular residue depends on not only the type of amino acid R but also the neighboring residues in the sequence. Thus GOR uses a centered window of 17 residues – eight nearest residues on each side – and assumes that the information function is the sum of information from single residues and residue pairs over the 17 residue window. The database is used to find pair frequencies in the different possible states S allowing the program to predict probabilities of the conformation of a unknown sequence.

Knowledge-based methods of prediction, which have been some of the most successful, use existing solved protein structures to learn or extract knowledge that can be applied to new proteins whose structure is unknown. Machine learning methods use computers to automate the process of extracting knowledge and learning structural relationships between objects. They have been applied to the prediction of all levels of protein structure (secondary, tertiary, and quaternary) including features such as binding sites, functional sites, and transmembrane helices [19]. A few of the major types of machine learning methods include neural networks, support vector machines, and hidden Markov models (HMMs). For secondary structure prediction the learning goal is to map the input sequence of amino acids to the output sequence of features. Some of the advantages of machine learning is that it can exploit physico-chemical knowledge and that the rules it generates are more comprehensible than statistical methods.

One thing that has been a great benefit to the protein structure prediction community is the Critical Assessment of Methods of Protein Structure Prediction (CASP) experiment [20]. The primary goal of of CASP is to establish the capabilities and limitations of current sequence based prediction methods, to ascertain where progress is happening, and to determine where the field is being held back. The way the experiment works is first a new protein sequence is collected whose structure has yet to be determined experimentally. The sequence is then made accessible to prediction teams who have a specific deadline to turn in their predictions and it is sent to registered CAFASP (Fully Automated) servers who are given 48 hours to reply with their predictions. The accuracy of the predictions are measured against the experimentally determined structure which is followed by a meeting to discuss the results. The CASP experiment has taken place every two years since 1994 and more than 100 international research groups participate.

This thesis pursues a relatively new and rather unique approach to protein secondary structure prediction and it has a couple of important advantages. For one thing it is not based on statistical measurements or knowledge gained from an outside group of proteins. Rather, this method is based partially on physical theory via the hydrophobic effect and thus may lead to important physical insights. Moreover this approach is capable of handling large data sets quickly and proteins of any length. In chapter 3 we discuss the main tool involved in our analysis, the discrete wavelet transform (DWT), and chapter 4 gets into the details of how we use wavelets to make predictions of secondary structure.

CHAPTER 2

Data Selection

The data used in this project consists of both protein amino acid sequences and experimentally determined secondary structure. This type of data is what is found in a protein structure database. The leading repository of protein structures is the Protein Data Bank (PDB) which is the topic of the first section of this chapter. Included in that section is an overview of the major experimental techniques used in protein structure determination. Although the ultimate source of our data is the PDB we access our files through a database that organizes proteins according to structural comparison, namely the Structural Classification of Proteins (SCOP) database. The second section of this chapter discusses the hierarchically organized SCOP database and the two protein classes we analyze in this project: $\alpha + \beta$ and α/β proteins. In the final two sections we cover issues with some of the data that led us to filter our database and we show some graphs that reveal some of the characteristics of the proteins such as chain lengths secondary structure lengths.

2.1 The Protein Data Bank

The Protein Data Bank (PDB) first started at the Brookhaven National Laboratory in 1971 and at that time had only seven protein structures. Over the years the database has grown at an exponential rate and now has over 60,000 structures that are made publicly available over the Web to the global community at no cost. In 1998, to deal with the rapid expansion of the database, the management of PDB was moved to the Research Collaboratory for Structural Bioinformatics (RSCB) which is an effort involving scientists at Rutgers University, the San Diego Supercomputing

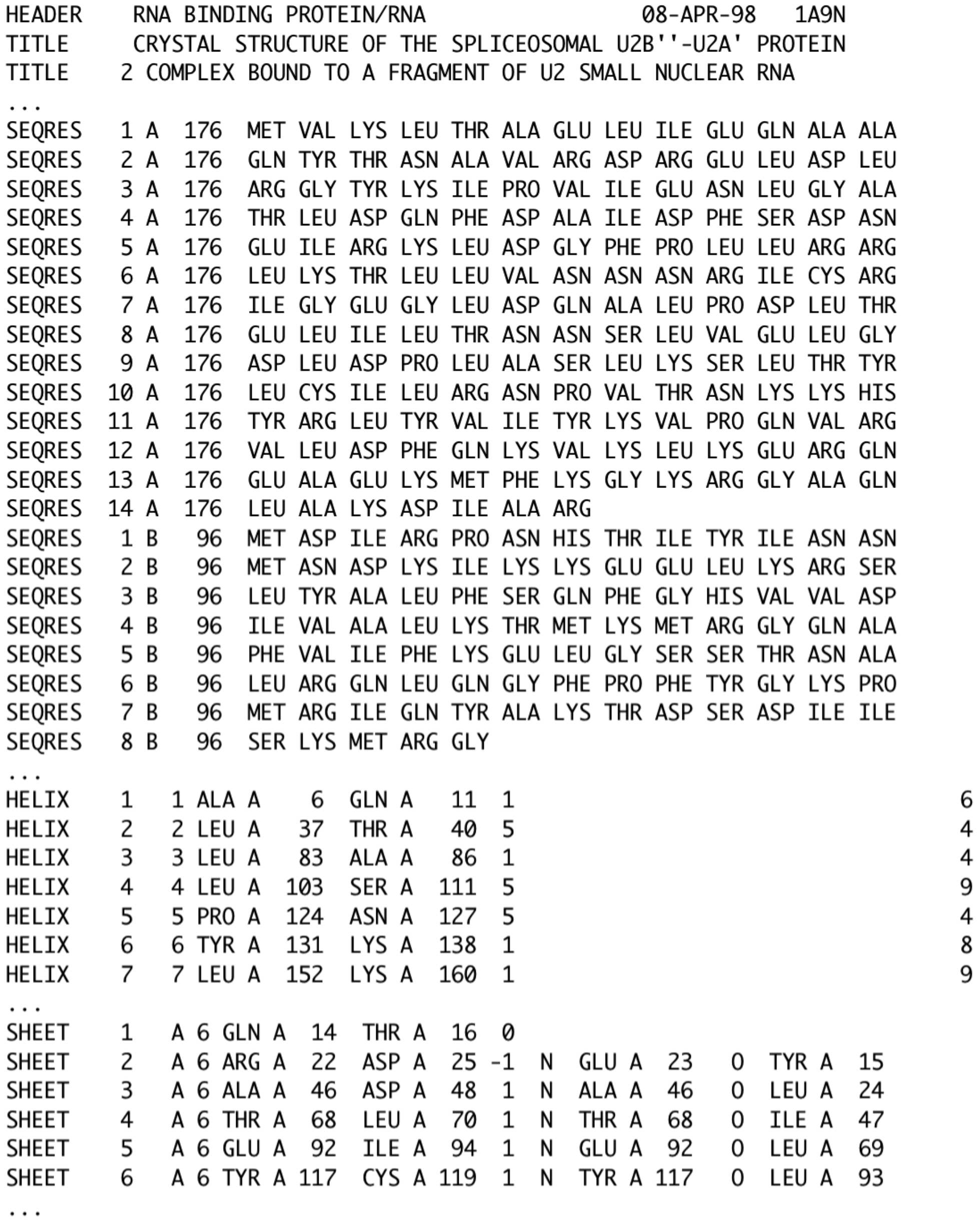
Center, and the National Institute of Standards and Technology.

PDB can be accessed on the Web (http://www.pdb.org/) where users can submit, search, and retrieve structural data. Over the Web structural biologists can deposit their structural data through ADIT (AutoDep Input Tool) which also checks the data format and provides a diagnostic report evaluating the correctness of newly deposited data. The structure validation process checks things like bond distances, bond angles, atom labels, torsion angles, and packing quality. The validation tools have been an important service for structural biologists as there had been a number of seriously flawed structures deposited into the PDB before their initiation.

2.1.1 The PDB file

The primary purpose of PDB is to store and make available its database of macromolecular structures. The data for each individual macromolecule is stored in ASCII formatted files which are man and machine readable. Figure 2.1 displays just a few of the types of records contained in a PDB file including the types used for this project (these files are typically hundreds of lines long so a lot is left out in the figure). Each line in the PDB entry consists of 80 columns, the first six columns of which identify the line. The first line in any entry is the HEADER line which gives the the PDB ID code, classification, and the date of deposition. The PDB ID code is a four-character identifier of which the first character is always a number between 1 and 9. After the header one or more TITLE lines give a description of the experiment. The records important to this thesis that contain information on the primary sequence and the secondary structure are identified by the names SEQRES, HELIX, and SHEET.

The SEQRES records lists the residues that make up the protein backbone from the N-terminal residue to the C-terminal residue. For cyclic peptides, a residue is arbitrarily assigned as the N-terminus. The residues, thirteen per line, are labeled by their three letter abbreviations (§ 1.2.1) separated by spaces. If a given protein



```
HEADER    RNA BINDING PROTEIN/RNA                 08-APR-98   1A9N
TITLE     CRYSTAL STRUCTURE OF THE SPLICEOSOMAL U2B''-U2A' PROTEIN
TITLE    2 COMPLEX BOUND TO A FRAGMENT OF U2 SMALL NUCLEAR RNA
...
SEQRES   1 A  176  MET VAL LYS LEU THR ALA GLU LEU ILE GLU GLN ALA ALA
SEQRES   2 A  176  GLN TYR THR ASN ALA VAL ARG ASP ARG GLU LEU ASP LEU
SEQRES   3 A  176  ARG GLY TYR LYS ILE PRO VAL ILE GLU ASN LEU GLY ALA
SEQRES   4 A  176  THR LEU ASP GLN PHE ASP ALA ILE ASP PHE SER ASP ASN
SEQRES   5 A  176  GLU ILE ARG LYS LEU ASP GLY PHE PRO LEU LEU ARG ARG
SEQRES   6 A  176  LEU LYS THR LEU LEU VAL ASN ASN ASN ARG ILE CYS ARG
SEQRES   7 A  176  ILE GLY GLU GLY LEU ASP GLN ALA LEU PRO ASP LEU THR
SEQRES   8 A  176  GLU LEU ILE LEU THR ASN ASN SER LEU VAL GLU LEU GLY
SEQRES   9 A  176  ASP LEU ASP PRO LEU ALA SER LEU LYS SER LEU THR TYR
SEQRES  10 A  176  LEU CYS ILE LEU ARG ASN PRO VAL THR ASN LYS LYS HIS
SEQRES  11 A  176  TYR ARG LEU TYR VAL ILE TYR LYS VAL PRO GLN VAL ARG
SEQRES  12 A  176  VAL LEU ASP PHE GLN LYS VAL LYS LEU LYS GLU ARG GLN
SEQRES  13 A  176  GLU ALA GLU LYS MET PHE LYS GLY LYS ARG GLY ALA GLN
SEQRES  14 A  176  LEU ALA LYS ASP ILE ALA ARG
SEQRES   1 B   96  MET ASP ILE ARG PRO ASN HIS THR ILE TYR ILE ASN ASN
SEQRES   2 B   96  MET ASN ASP LYS ILE LYS LYS GLU GLU LEU LYS ARG SER
SEQRES   3 B   96  LEU TYR ALA LEU PHE SER GLN PHE GLY HIS VAL VAL ASP
SEQRES   4 B   96  ILE VAL ALA LEU LYS THR MET LYS MET ARG GLY GLN ALA
SEQRES   5 B   96  PHE VAL ILE PHE LYS GLU LEU GLY SER SER THR ASN ALA
SEQRES   6 B   96  LEU ARG GLN LEU GLN GLY PHE PRO PHE TYR GLY LYS PRO
SEQRES   7 B   96  MET ARG ILE GLN TYR ALA LYS THR ASP SER ASP ILE ILE
SEQRES   8 B   96  SER LYS MET ARG GLY
...
HELIX    1   1 ALA A    6  GLN A   11  1                                  6
HELIX    2   2 LEU A   37  THR A   40  5                                  4
HELIX    3   3 LEU A   83  ALA A   86  1                                  4
HELIX    4   4 LEU A  103  SER A  111  5                                  9
HELIX    5   5 PRO A  124  ASN A  127  5                                  4
HELIX    6   6 TYR A  131  LYS A  138  1                                  8
HELIX    7   7 LEU A  152  LYS A  160  1                                  9
...
SHEET    1   A 6 GLN A  14  THR A  16  0
SHEET    2   A 6 ARG A  22  ASP A  25 -1  N  GLU A  23   O  TYR A  15
SHEET    3   A 6 ALA A  46  ASP A  48  1  N  ALA A  46   O  LEU A  24
SHEET    4   A 6 THR A  68  LEU A  70  1  N  THR A  68   O  ILE A  47
SHEET    5   A 6 GLU A  92  ILE A  94  1  N  GLU A  92   O  LEU A  69
SHEET    6   A 6 TYR A 117  CYS A 119  1  N  TYR A 117   O  LEU A  93
...
```

Figure 2.1: Example of some the important records of the PDB file 1A9N.

contains multiple chains each chain is identified by a different character in the third field from the left. For example, the PDB entry displayed in Fig. 2.1 has one chain identified by “A” with 176 residues and a second chain identified by “B” with 96 residues.

The HELIX and SHEET records are used to identify the positions of helices and sheets in the molecule. In both types of records there are many different fields with different information related to these structures. The helix records indicate the chain the helix belongs to, the starting and ending positions, the types of the starting and ending residues, the class of the helix (e.g. right-handed alpha, right-handed 310, left-handed pi, etc.), and there is a space allotted for comments about the helix. SHEET records, in addition to identifying starting and ending sequence numbers, also indicate the “sense” of each strand in a sheet with respect to the previous strand i.e. the number 0 indicates the first strand, 1 if the strand is parallel to the previous strand, and -1 if it is anti-parallel.

Other types of important records in a PDB file include the ATOM records which give the three-dimensional atomic coordinates of the standard residues, HET records identify non-standard groups (heterogens), FORMUL records give chemical formula of non-standard groups, LINK records identify inter-residue bonds, etc. The data in a PDB file can be used by molecular viewing software that read in 3-D coordinates and displays various representations of the molecule.

2.1.2 Protein Structure Determination

The effort of secondary structure prediction relies on accurately measured experimental data. This section gives an overview of the two most common experimental methods for determining the structures of proteins: X-ray crystallography and NMR spectroscopy. For both of these methods additional molecular structure information is used to create the final atomic model. For instance, amino acid sequence, bond lengths, certain bond angles, and stereochemical information is often already known. Scientists can then build a model that is consistent with both the expected geometry and the experimental data.

As of the beginning of 2010 there were 61,280 protein, peptide, and virus structures deposited into the Protein Data Bank. Of this number 57,298 entries (93.5 %) were determined by X-ray crystallography, 8,449 (13.7 %) by NMR, 295 (0.5 %) by electron microscopy, and 170 (0.3 %) by hybrid or other methods.

As the fastest and oldest technique, X-ray Crystallography is responsible for an overwhelming majority of the known structures in the PDB and the majority of the new entries. In 1912 Max von Laue showed that X-rays were diffracted by crystals earning himself the Nobel Prize in Physics in 1914 [21]. In the same year Lawrence Bragg along with his father William discovered that it is possible to calculate the positions of atoms within a crystal based on the pattern of spots produced on photographic plates when X-rays are incident upon the crystal (see Fig. 2.2) and earned themselves the Nobel Prize in Physics as well in the following year [22]. The spots on the plates represented points of maximum intensity of the scattered X-rays. The angle θ of scattering is given by Bragg's law:

$$n\lambda = 2d\sin\theta \qquad (\text{where } n = 1, 2, 3, ...) \tag{2.1}$$

where n is the order of the maxima, λ is the wavelength of the X-rays, and d is the spacing between atoms. Bragg's law allows one to determine the atomic structure of a crystal based on the intensity pattern of the scattered beam.

In protein crystallography Bragg's law is extended into three dimensions in what are known as the Laue equations:

$$\begin{aligned} a(\cos\alpha_i - \cos\alpha_r) =& h\lambda \qquad (\text{where } h = 1, 2, 3, ...) \\ b(\cos\beta_i - \cos\beta_r) =& k\lambda \qquad (\text{where } k = 1, 2, 3, ...) \\ c(\cos\gamma_i - \cos\gamma_r) =& l\lambda \qquad (\text{where } l = 1, 2, 3, ...) \end{aligned} \tag{2.2}$$

where a, b, and c represent the spacing in each of the three dimensions and α, β, and γ represent their respective incident (subscript i) and scattering (subscript r) angles.

The end result of X-ray diffraction experiments is an electron density map. Since electrons are tightly localized around the nuclei this electron density map can be used to make a good approximation of atomic positions within a molecule. One big advantage of using crystals is that you can have large numbers of molecules oriented in the same direction resulting in greater intensity of the scattered signal.

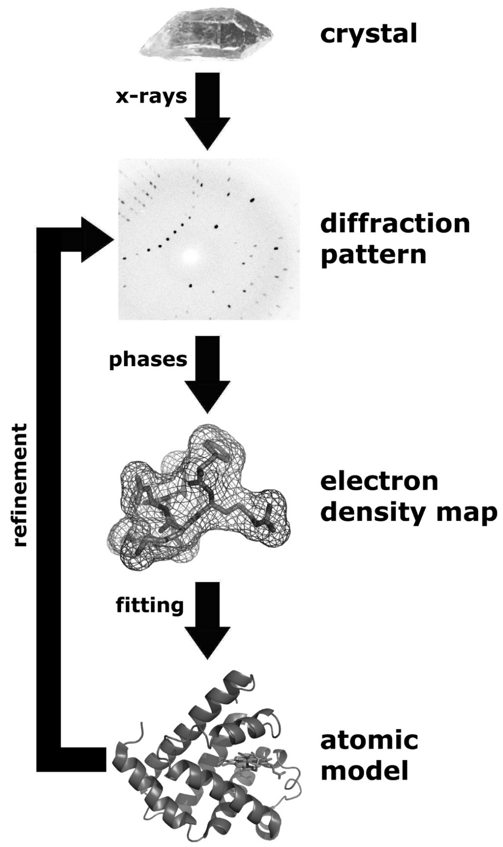

crystal
x-rays
diffraction pattern
phases
electron density map
fitting
atomic model
refinement

Figure 2.2: Basic diagram of X-ray crystallography process.

One of the most time consuming steps in protein crystallography is in the production of protein crystals. Crystallization requires the formation of stable crystals with sufficient long-range order and dimensions along each axis of more than 0.1 mm. To undergo crystallization a purified protein is slowly precipitated from an aqueous solution. Individual protein molecules take on a regular orientation by lining up in a series of “unit cells.” The quality of the structures determined from X-ray diffraction is only as good as the quality of the crystals that they are acquired from.

The second major tool for determining protein structure is NMR (Nuclear Magnetic Resonance) spectroscopy. This technique makes use of the properties of atomic nuclei to ascertain the quantities and locations of atoms in a polypeptide chain. It has a particular advantage over X-ray crystallography in cases where the proteins are difficult to crystallize. One drawback of NMR is that it is not effective for the more lengthy proteins. Proteins of around 150 or fewer residues are no problem but proteins from 150 to 300 residues require advanced strategies.

NMR relies on the spin property of atomic nuclei. Nuclear spin, represented by the symbol $\mathbf{I}$, is the vector sum of the angular momentum of all nucleons in a nucleus. Since nucleons are fermions which obey the Pauli exclusion principle they pair up in such a way to give net spins of zero or integer multiples of 1/2 dependent on the properties of valence nucleons. Nuclei that have no spin ($\mathbf{I} = 0$) are not affected by an external magnetic field making them NMR “silent” i.e. they cannot be detected. The proton nucleus (hydrogen atom without any neutrons) is the most important spin 1/2 nucleus in NMR and it also conveniently has a high natural occurrence in proteins. Two other atoms that are also important in protein structure are carbon and nitrogen, however in their most abundant natural isotopes (^{12}C and ^{14}N) they are NMR silent because of their even-even numbers of protons and neutrons which pair up to give zero spin. To overcome this issue advanced techniques in molecular biology have made it possible to enrich proteins with the half-spinned ^{13}C and ^{15}N isotopes.

The NMR experiment is performed by placing the protein samples in a uniform magnetic field. The spin 1/2 nuclei, in the presence of the external magnetic field, become polarized with the two possible orientations of either parallel or anti-parallel to the external field. These two states have an energy difference with the anti-parallel configuration being at a slightly higher energy. Radio frequency pulses are irradiated upon the sample in order to tip the ensemble of spins by 90 degrees resulting in spin precession at the Larmor frequency about the external magnetic field axis. The observable signal in all NMR experiments is the current induced in a detector coil by the spin precession of nuclei at the Larmor frequency. The signals are influenced by several different parameters such as the chemical shift, spin-spin coupling constants, spin lattice relaxation time, peak intensity, and the nuclear Overhauser effect. These parameters are effected by the proximity of atoms with other atoms.

The experimental data from NMR provides a list of atoms that are nearby in space (not just nearby in sequence) and this information gives indication of secondary structure and helps to assemble individual regions into the correct overall structure. Computations are performed which optimize the fit between the coordinates and the experimental data and this together with stereochemical restrictions produces the resultant set of atomic coordinates. Since the protein samples are not crystallized but rather are in solution, the proteins are free to move about, thus NMR experiments typically yield 15 to 20 similar structures that are consistent with the experimental data and stereochemical restraints.

The question has been asked as to how we know whether the structure results we get from experiments such as X-ray crystallography or NMR spectroscopy are relevant to the native structure *in vivo* (inside the cell). One particular experiment performed on several proteins seems to indicate that the structures are the same. In diffusing substrate into crystalized samples, the same enzymatic activity or function has been observed in the crystallized protein as is found *in vivo* [23]. In some cases comparison studies are done in which the same proteins are prepared in different crystal forms or different packing patterns and the crystallography results have only minor, tenths of an Å, differences in atomic coordinates. Furthermore, comparison between crystal structures and NMR structures of the same protein give very good agreement.

2.2 Protein Structural Classification

Proteins are often classified according to structural similarities and there are some databases whose objective is to organize proteins according to such a classification system. These databases often take the raw data from PDB. Two such databases that organize proteins into a hierarchy of structural similarity are CATH (Class, Architecture, Topology, Homology) and SCOP (Structural Classification of Proteins). The source of data used for this thesis is the SCOP database.

2.2.1 The SCOP database

SCOP's objective is to provide a comprehensive description of the structural and evolutionary relationships between all proteins in the PDB database [24]. One major difference between SCOP and CATH is that the SCOP classification is done manually by visual inspection and comparison of structures whereas CATH uses a combination of automatic and manual analysis.

The broadest level of categorizing proteins (i.e. the top of the hierarchy) is referred to as the *class* and is based on the proportion of residues adopting helical or strand conformation. Proteins within the SCOP database are currently divided into 11 different classes. The major classes include all α proteins which are composed almost entirely of α-helices, all β proteins which are composed almost entirely of β-sheets, α/β proteins with both α-helices and β-sheets often close together, $\alpha + \beta$ proteins with largely segregated α-helices and β-sheets, and lastly multi-domain proteins which consist of domains belonging to different classes. Other classes include membrane, small proteins, coiled coil, low resolution, peptides, and designed proteins.

After the class is chosen the major three levels of the SCOP hierarchy are the *fold*, *superfamily*, and *family*. Proteins with the same major secondary structures arranged in the same way and the same topological connections are considered to be part of the same fold category. Proteins within the same fold may have differences in the size or conformation of turns or in peripheral secondary structure elements. It isn't necessarily the case that proteins within the same fold have a common evolutionary origin, it could be that a certain packing arrangement is favored by the physics and chemistry of protein folding.

The fold category is further broken down into the next level in the hierarchy, the superfamily. Proteins within a superfamily are similar in their structural and functional features but do not necessarily have a significant resemblance in residue sequence. It is probable that members of a superfamily have some common evolutionary origin. The last level of the hierarchy is the family. Proteins of the same family have a clear evolutionary relationship. Usually there is high sequence identity in addition to structural and functional similarity.

2.3 Filtering the Data

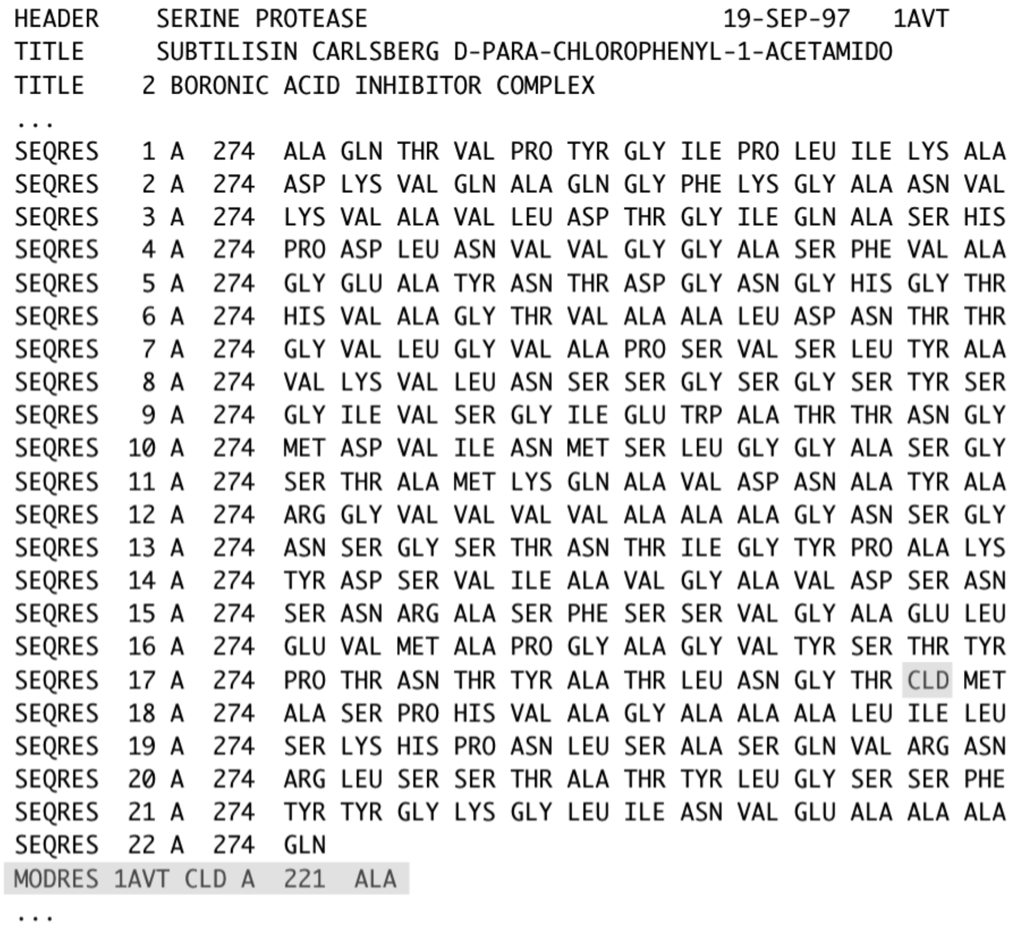
The PDB file format is archaic in nature and is full of inconsistencies making it problematic for a computer to read. In particular, files deposited before 1995 contained frequent exceptions and variations in labeling, numbering, and formatting [25]. This section highlights the major issues encountered that caused us to filter the data. Of the many different classes of proteins in the SCOP database discussed above, two were selected four our analysis: $\alpha + \beta$ and α/β. Initially our $\alpha + \beta$ and α/β data sets consisted of 9728 and 10961 proteins respectively. The issues described below include non-standard residues, secondary structure elements that were located outside the designated protein length, proteins that consisted of no secondary structure, and lastly secondary structure endpoints that occurred before their initial points.

One issue was non-standard residue abbreviations. This was not an error in files just a complication we were not prepared to deal with. PDB contains more than just simple proteins and in cases where there are non-standard residues there are different codes used in the SEQRES records of the PDB file. These codes are then defined in MODRES records (see Fig. 2.3) which describe the modifications and include correlations with standard residues. Instead of altering our computer code to replace these modified residues with standard ones these data were left out. In our $\alpha + \beta$ data set there were 1964 of these files bringing us to a new total of 7764 files for analysis. The α/β data set had 1692 files with non-standard residue names resulting in a new total of 9269 proteins.

Another issue we came across was an apparent error in the numbering of secondary structure positions. We found a surprisingly large number of cases in which the positions of secondary structure were outside the position of the final C-terminal residue or even before the first N-terminal residue (see Fig. 2.4). In some of these cases it appeared that the secondary structure positions may have just been shifted by a constant amount. One possible reason for some of these instances in multi-chain proteins is that counters may not have been reset after individual chains. However, the majority of the cases were found in single chain proteins. In the $\alpha + \beta$ data set there were 1812 PDB files with this issue. In the α/β data set there were 1732 files. All the files with these errors were also filtered from our data set leaving the $\alpha + \beta$ and α/β data sets with 5952 and 7537 protein files respectively.

The third issue was PDB files that contained no α-helix or β-sheet records whatsoever. Lack of secondary structure would not necessarily be a problem but considering that both classes are defined as having both α-helices and β-sheets the concern is that secondary structure data may have been lost. We found 17 of these files in the $\alpha + \beta$ set and 21 in the α/β set. Again these files were removed to yield totals of 5935 $\alpha + \beta$ proteins and 7516 α/β proteins.

Lastly, we found files that had secondary structure elements with endpoints located before their initial points (i.e. negative lengths). This again is an apparent

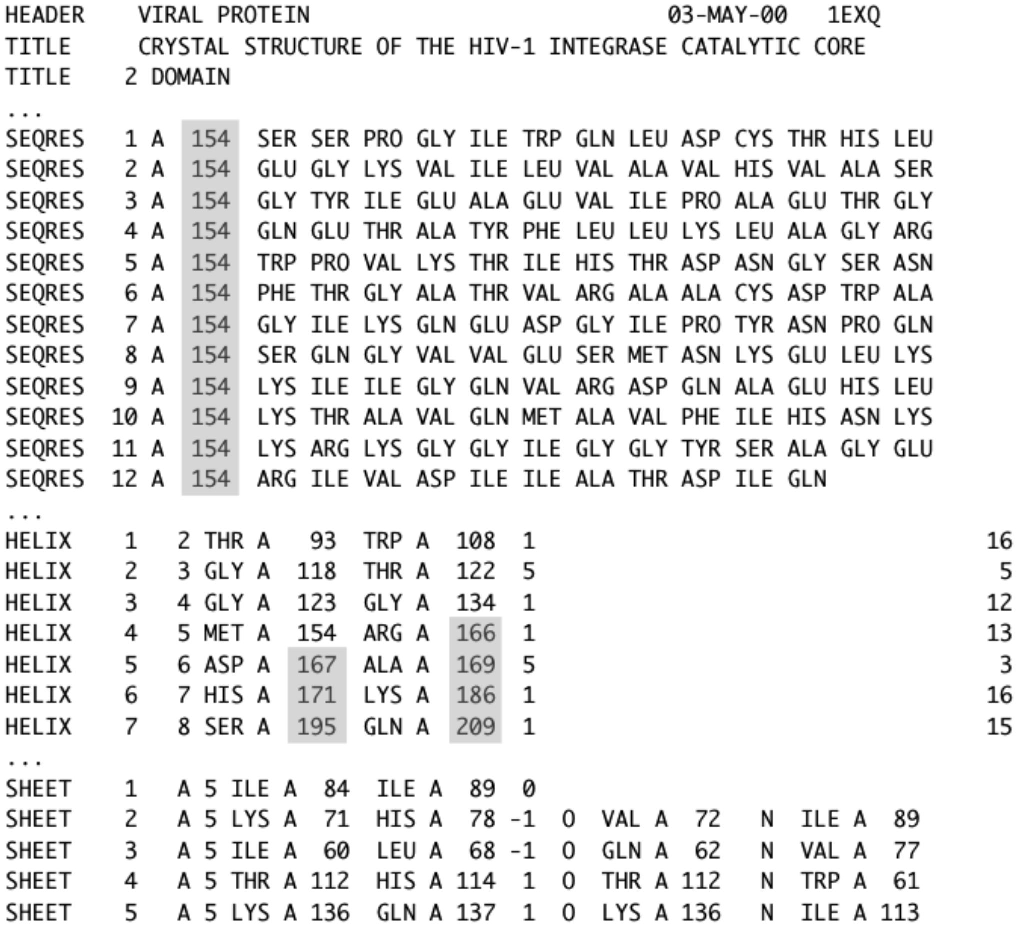


```
HEADER    SERINE PROTEASE                         19-SEP-97   1AVT
TITLE     SUBTILISIN CARLSBERG D-PARA-CHLOROPHENYL-1-ACETAMIDO
TITLE    2 BORONIC ACID INHIBITOR COMPLEX
...
SEQRES   1 A  274  ALA GLN THR VAL PRO TYR GLY ILE PRO LEU ILE LYS ALA
SEQRES   2 A  274  ASP LYS VAL GLN ALA GLN GLY PHE LYS GLY ALA ASN VAL
SEQRES   3 A  274  LYS VAL ALA VAL LEU ASP THR GLY ILE GLN ALA SER HIS
SEQRES   4 A  274  PRO ASP LEU ASN VAL VAL GLY GLY ALA SER PHE VAL ALA
SEQRES   5 A  274  GLY GLU ALA TYR ASN THR ASP GLY ASN GLY HIS GLY THR
SEQRES   6 A  274  HIS VAL ALA GLY THR VAL ALA ALA LEU ASP ASN THR THR
SEQRES   7 A  274  GLY VAL LEU GLY VAL ALA PRO SER VAL SER LEU TYR ALA
SEQRES   8 A  274  VAL LYS VAL LEU ASN SER SER GLY SER GLY SER TYR SER
SEQRES   9 A  274  GLY ILE VAL SER GLY ILE GLU TRP ALA THR THR ASN GLY
SEQRES  10 A  274  MET ASP VAL ILE ASN MET SER LEU GLY GLY ALA SER GLY
SEQRES  11 A  274  SER THR ALA MET LYS GLN ALA VAL ASP ASN ALA TYR ALA
SEQRES  12 A  274  ARG GLY VAL VAL VAL VAL ALA ALA ALA GLY ASN SER GLY
SEQRES  13 A  274  ASN SER GLY SER THR ASN THR ILE GLY TYR PRO ALA LYS
SEQRES  14 A  274  TYR ASP SER VAL ILE ALA VAL GLY ALA VAL ASP SER ASN
SEQRES  15 A  274  SER ASN ARG ALA SER PHE SER SER VAL GLY ALA GLU LEU
SEQRES  16 A  274  GLU VAL MET ALA PRO GLY ALA GLY VAL TYR SER THR TYR
SEQRES  17 A  274  PRO THR ASN THR TYR ALA THR LEU ASN GLY THR CLD MET
SEQRES  18 A  274  ALA SER PRO HIS VAL ALA GLY ALA ALA ALA LEU ILE LEU
SEQRES  19 A  274  SER LYS HIS PRO ASN LEU SER ALA SER GLN VAL ARG ASN
SEQRES  20 A  274  ARG LEU SER SER THR ALA THR TYR LEU GLY SER SER PHE
SEQRES  21 A  274  TYR TYR GLY LYS GLY LEU ILE ASN VAL GLU ALA ALA ALA
SEQRES  22 A  274  GLN
MODRES 1AVT CLD A  221  ALA
...
```

Figure 2.3: One example of a file (PDB ID code: 1AVT) from the α/β data set which has the non-standard residue identified as CLD (highlighted). The MODRES record (also highlighted) indicates the closest standard residue replacement would be ALA.

error and these files were also filtered out. The $\alpha + \beta$ set contained 8 of these files whereas the α/β set had 16. After filtering out all of these PDB files the grand total number of proteins for processing were 5927 $\alpha + \beta$ proteins and 7500 α/β proteins.

While the numbers above indicate the total number of proteins, some of these proteins contain multiple chains. For this project we decided to include these multi-chain proteins in our data set. Many of these proteins contained one or more exactly



```
HEADER    VIRAL PROTEIN                           03-MAY-00   1EXQ
TITLE     CRYSTAL STRUCTURE OF THE HIV-1 INTEGRASE CATALYTIC CORE
TITLE    2 DOMAIN
...
SEQRES   1 A  154  SER SER PRO GLY ILE TRP GLN LEU ASP CYS THR HIS LEU
SEQRES   2 A  154  GLU GLY LYS VAL ILE LEU VAL ALA VAL HIS VAL ALA SER
SEQRES   3 A  154  GLY TYR ILE GLU ALA GLU VAL ILE PRO ALA GLU THR GLY
SEQRES   4 A  154  GLN GLU THR ALA TYR PHE LEU LEU LYS LEU ALA GLY ARG
SEQRES   5 A  154  TRP PRO VAL LYS THR ILE HIS THR ASP ASN GLY SER ASN
SEQRES   6 A  154  PHE THR GLY ALA THR VAL ARG ALA ALA CYS ASP TRP ALA
SEQRES   7 A  154  GLY ILE LYS GLN GLU ASP GLY ILE PRO TYR ASN PRO GLN
SEQRES   8 A  154  SER GLN GLY VAL VAL GLU SER MET ASN LYS GLU LEU LYS
SEQRES   9 A  154  LYS ILE ILE GLY GLN VAL ARG ASP GLN ALA GLU HIS LEU
SEQRES  10 A  154  LYS THR ALA VAL GLN MET ALA VAL PHE ILE HIS ASN LYS
SEQRES  11 A  154  LYS ARG LYS GLY GLY ILE GLY GLY TYR SER ALA GLY GLU
SEQRES  12 A  154  ARG ILE VAL ASP ILE ILE ALA THR ASP ILE GLN
...
HELIX    1   2 THR A   93  TRP A  108  1                                  16
HELIX    2   3 GLY A  118  THR A  122  5                                   5
HELIX    3   4 GLY A  123  GLY A  134  1                                  12
HELIX    4   5 MET A  154  ARG A  166  1                                  13
HELIX    5   6 ASP A  167  ALA A  169  5                                   3
HELIX    6   7 HIS A  171  LYS A  186  1                                  16
HELIX    7   8 SER A  195  GLN A  209  1                                  15
...
SHEET    1   A 5 ILE A  84  ILE A  89  0
SHEET    2   A 5 LYS A  71  HIS A  78 -1  O  VAL A  72   N  ILE A  89
SHEET    3   A 5 ILE A  60  LEU A  68 -1  O  GLN A  62   N  VAL A  77
SHEET    4   A 5 THR A 112  HIS A 114  1  O  THR A 112   N  TRP A  61
SHEET    5   A 5 LYS A 136  GLN A 137  1  O  LYS A 136   N  ILE A 113
```

Figure 2.4: One example of a file (PDB ID code: 1EXQ) which has helix structures (highlighted) extending out to a chain position of 209 when the length of the chain is only 154 residues.

identical chains so to prevent duplicates the processing was limited to all the unique chains of each protein. The advantage of including multi-chain proteins was that it greatly increased the number protein chains analyzed. The number of unique chains processed in the $\alpha + \beta$ and α/β data sets were 6939 and 8572 respectively.

2.4 Data Characteristics

This final section provides some information on the characteristics of the proteins from the filtered data used in our analysis. Since this project is on secondary structure prediction one thing of interest would be a knowledge of the lengths of these structures. Figures 2.5 and 2.6 display length distributions of helices and strands for the $\alpha + \beta$ and α/β data sets. One thing that becomes immediately apparent from these figures is that the distributions look very similar amongst the two different classes of proteins. We see that helices have a much broader length distribution compared to strands. Both helix distributions peak around lengths of 5 or 6 residues but also have another significant peak at 3 residues corresponding to short 3_{10}-helix fragments. Besides the peaks, both helix distributions are quite steady out to lengths of about 14 and then broadly decay with a small amount exceeding 25 residues. Strands on the other hand have a more narrow length distribution. For both classes there is a negligible amount of strands that exceed 15 residues in length and they both peak between 3 and 6.

One final characteristic that we looked at was the distribution of chain lengths for both data sets (see Fig. 2.7). Amongst the two classes these distributions had some differences. The $\alpha + \beta$ set strongly peaked around 130 residues and then steadily decreased to negligible counts after about 600 residues. The α/β set is a broad almost symmetric distribution with a peak centered at around 300 residues and which, ignoring a few fluctuations, also drops off around 600.

The filtered $\alpha + \beta$ and α/β data sets discussed above comprise a total of 15511 proteins chains for our analysis. One of the advantages of our technique is its ability to process large numbers of proteins fairly quickly. For example, most of our computer programs, written in the Matlab language, can process the 15511 chains in less than an hour. In the analysis (chapter 4) $\alpha + \beta$ and α/β data sets are processed separately so the two can be compared.

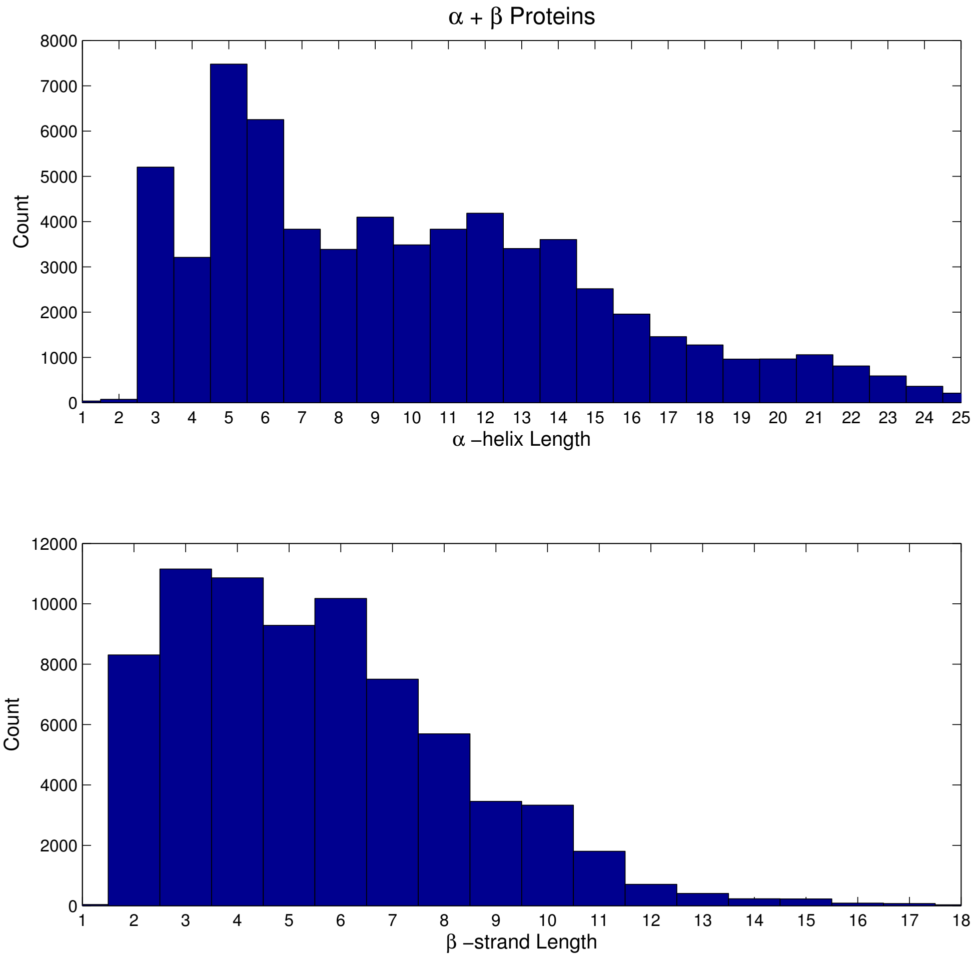


Figure 2.5: Histogram of the counts of α-helix lengths (top) and β-strand lengths (bottom) in the $\alpha + \beta$ data set.

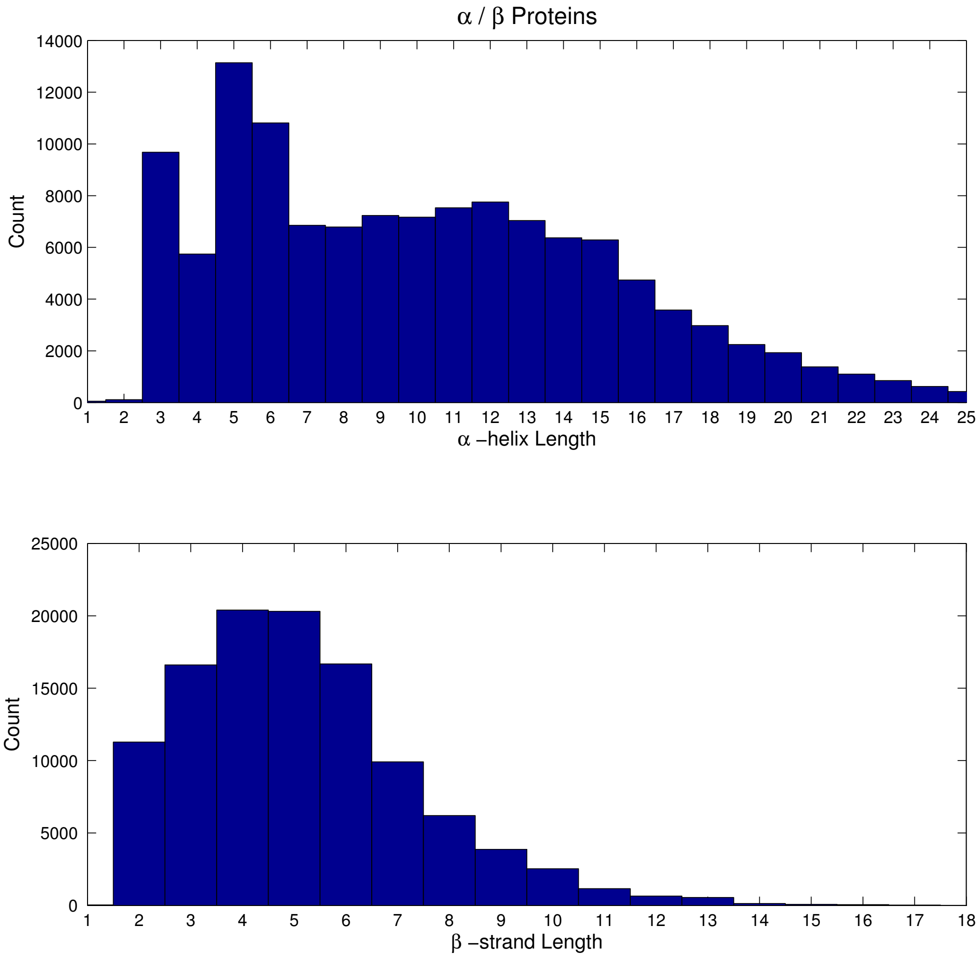


Figure 2.6: Histogram of the counts of α-helix lengths (top) and β-strand lengths (bottom) in the α/β data set.

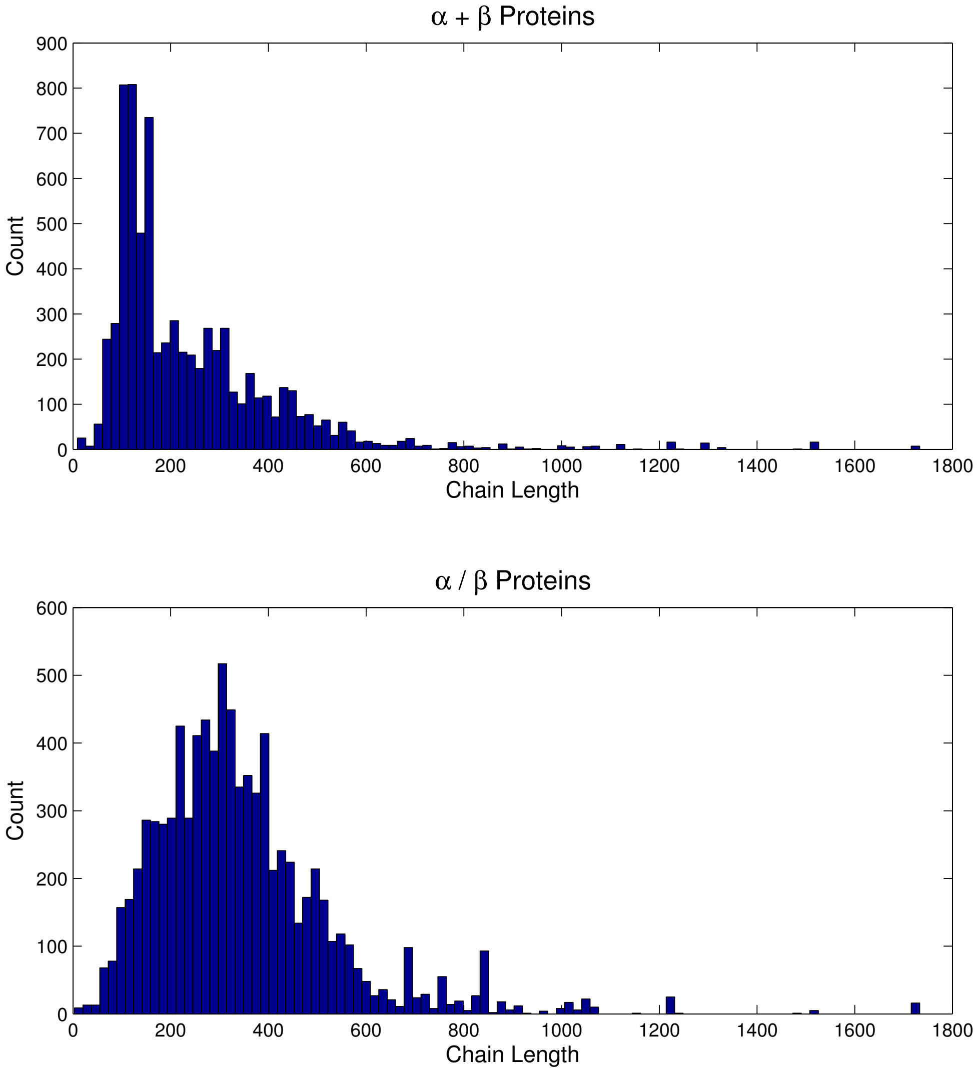


Figure 2.7: Histogram of the chain lengths of the 6939 $\alpha + \beta$ chains (top) and of the 8572 α/β chains.

CHAPTER 3

The Discrete Wavelet Transform

3.1 Introduction

This chapter focuses on the major analysis technique used in this project, the discrete wavelet transform (DWT). Wavelets are a mathematical tool used to analyze signals. They have been applied to many different problems in engineering, computer science and scientific research including image processing, heart rate and ECG (electrocardiogram) analyses, data compression, and communication applications.

The signals that we analyze in this project are the arrays of hydrophobicity values of the residues in a protein chain ordered from the N-terminal residue to the C-terminal residue. The position of the amino acid residue, which we will designate here by r_i, can be considered the discrete independent variable. The values of the dependent variable $s_i = f(r_i)$ represents the magnitude of the hydrophobicity that corresponds to the residue at position r_i.

On the topic of signal analysis the Fourier transform and Fourier series are perhaps the most popular tools. However, wavelets have some significant advantages over Fourier analysis. The Fourier transform is a transform from the space (or time) domain to the frequency domain or vice versa. It is very good at giving the frequency components of a signal but it does not give localized frequency content. For example, the Fourier transform of a sine wave in the time domain gives a delta function in the frequency domain. Wavelets solve this problem by providing both time- and frequency-domain analysis or time-frequency analysis. A method known as windowed Fourier transform fixes the time-frequency localization issue nevertheless the wavelet transform is better at “zooming in” on short-lived high frequencies [26].

Moreover, the windowed Fourier transform requires that the user choose a window size making it, in a sense, arbitrary.

Another important feature of wavelets is known as multi-resolution analysis (MRA). With multi-resolution analysis a complicated function is broken down into several simpler functions that can be studied separately. In situations where a function has slowly varying and quickly varying segments MRA makes it possible to focus on one particular resolution on its own. Since proteins structure is known to be caused by both local and non-local interactions amongst residues along the polypeptide chain MRA should be a convenient analysis tool.

3.2 The Haar Wavelet

While the main type of wavelet used in this project are Daubecies wavelets, this first section covers the much easier to conceptualize Haar wavelet. The Haar wavelet is the oldest and most simple type of wavelet dating back to 1910 [27] (although at that time the term "wavelet" was not used) when the first paper was published by Alfréd Haar. The mathematical functions that define the Haar wavelets are the scaling function, also called the basic step function $\varphi(r)$ given by

$$\varphi_{u,w}(r) = \begin{cases} 1 & \text{if } u \le r < w, \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

and the wavelet function $\psi(r)$ given by

$$\psi_{u,w}(r) = \begin{cases} 1 & \text{if } u \le r < v, \\ -1 & \text{if } v \le r < w, \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

where $v = (u + w)/2$ is the midpoint of the wavelet function over the interval $[u, w)$ (see Fig. 3.1 shows both). The basic step function, through dilations and translations, is the generator of the wavelet functions. As we'll see, the wavelet transform is a process of changing basis from scaling functions to coarser scaling and wavelet functions.

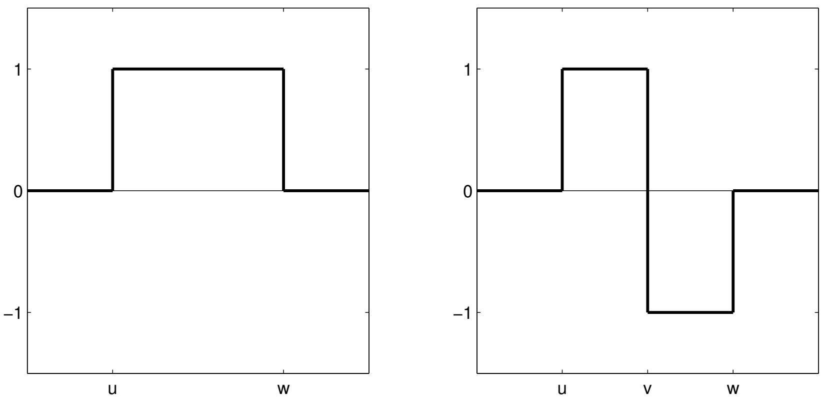


Figure 3.1: Haar's scaling function $\varphi_{u,w}$ (left) and wavelet function $\psi_{u,w}$ (right) defined over the interval $[u, w)$ where $v = (u + w)/2$.

Adding together two single step functions produces one larger step function of twice the scale:

$$\varphi_{r_0,r_2} = \varphi_{r_0,r_1} + \varphi_{r_1,r_2}. \tag{3.3}$$

Likewise taking the difference between the same two smaller steps gives a wavelet function of twice the scale (see Fig. 3.2):

$$\psi_{r_0,r_2} = \varphi_{r_0,r_1} - \varphi_{r_1,r_2}. \tag{3.4}$$

It is easy to show, through adding and subtracting Eqs. 3.3 and 3.4, that each short step can be expressed in terms of a scaling function and a wavelet function of twice the scale (see Fig. 3.3):

$$\frac{1}{2}\left(\varphi_{r_0,r_2} + \psi_{r_0,r_2}\right) = \varphi_{r_0,r_1}, \tag{3.5}$$

$$\frac{1}{2}\left(\varphi_{r_0,r_2} - \psi_{r_0,r_2}\right) = \varphi_{r_1,r_2}. \tag{3.6}$$

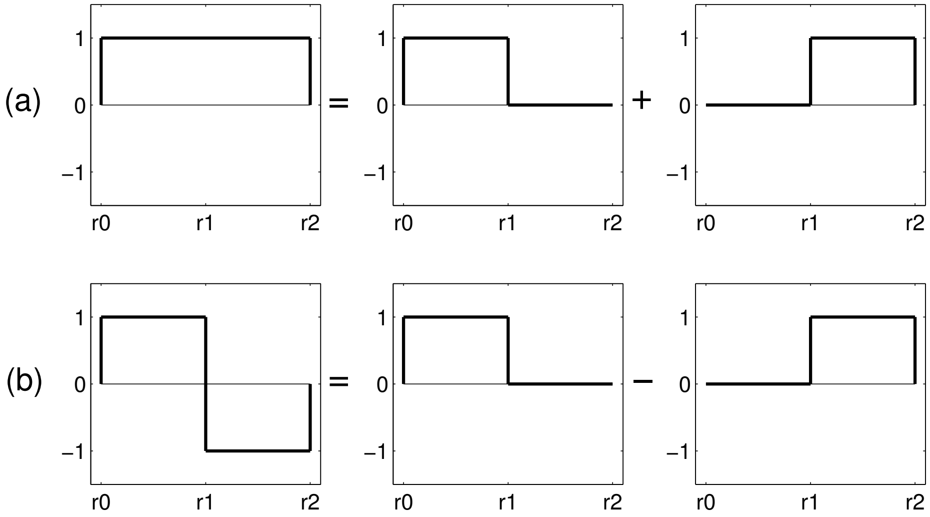


Figure 3.2: (a) The scaling function on the interval $[r_0, r_2)$ can be broken into two single steps, one after the other. (b) The wavelet function on the same interval is equivalent to the difference between the same two single steps.

Equations 3.5 and 3.6 provide a relationship between narrower basic step functions and dilated step functions and wavelet functions. This relationship is the basis for the Haar wavelet transform.

Representing the signal of data as a mathematical function is the first step in applying wavelet analysis. The signal is represented in terms of scaling functions which serve as an orthonormal set of basis functions. In a set of n data points (r_i, s_i) where $s_i = f(r_i)$ is the height at abscissa r_i the simple function or step function is denoted by $\tilde{f}$:

$$\begin{aligned}\tilde{f} =& s_0 \cdot \varphi_{r_0,r_1} + s_1 \cdot \varphi_{r_1,r_2} + \cdots + s_{n-1} \cdot \varphi_{r_{n-1},r_n} \\ =& \sum_{i=0}^{n-1} s_i \cdot \varphi_{r_i,r_{i+1}}.\end{aligned} \tag{3.7}$$

We are now equipped to perform the Haar wavelet transform or decomposition (carried out below in Eqs. 3.8 through 3.11). For a signal with two data points

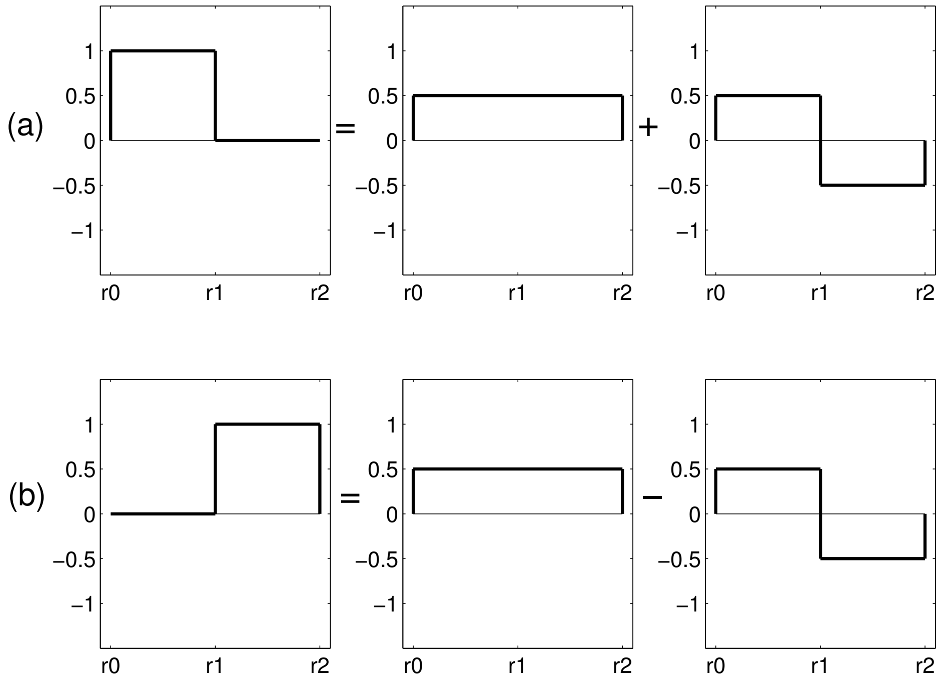


Figure 3.3: (a) A single step over the interval $[r_0, r_1)$ is equivalent to half of the larger scale step function $\frac{1}{2}\varphi_{r_0,r_2}$ plus half the wavelet function $\frac{1}{2}\psi_{r_0,r_2}$. (b) A step over the interval $[r_1, r_2)$ is equivalent to the difference between the same two functions.

(s_0, s_1) we first construct the simple signal function given by Eq. 3.7. Next we substitute in the relations given by Eq. 3.5 and 3.6 and finally we combine the like terms:

$$\begin{aligned}
\tilde{f} &= s_0 \cdot \varphi_{r_0,r_1} + s_1 \cdot \varphi_{r_1,r_2} && (3.8)\\
&= s_0 \cdot \frac{1}{2}\left(\varphi_{r_0,r_2} + \psi_{r_0,r_2}\right) + s_1 \cdot \frac{1}{2}\left(\varphi_{r_0,r_2} - \psi_{r_0,r_2}\right), && (3.9)\\
&= \frac{s_0 + s_1}{2} \cdot \varphi_{r_0,r_2} + \frac{s_0 - s_1}{2} \cdot \psi_{r_0,r_2}, && (3.10)\\
&= a_0 \cdot \varphi_{r_0,r_2} + c_0 \cdot \psi_{r_0,r_2}. && (3.11)
\end{aligned}$$

The result is a more coarse scaling function with coefficient $a_0 = (s_0 + s_1)/2$, also called *approximation coefficient*, representing the localized *average* of the signal over the interval $[r_0, r_2)$. The second term in Eq. 3.11 is the wavelet function with the coefficient $c_0 = (s_0 - s_1)/2$, also called *detail coefficient*, representing the *change* or fluctuation in the signal over the same interval. Since the wavelet function, as defined (Eq. 3.2), makes a jump from 1 to -1 the jump in the signal is given by $c_0 \cdot (-2)$.

In effect the wavelet transform is a change of basis from scaling functions to multi-scale wavelet functions. In the transformation no information is lost and there is an inverse transform that brings back the original signal.

3.2.1 Example Calculation

For purpose of illustration we perform the Haar wavelet transform on a signal with some real numbers. Take the signal $\vec{s} = (5, -1, 7, 3, 1, 1, 2, 4)$, which for example could represent the hydrophobicity of a peptide with eight residues. We begin by expressing the signal in terms of scaling function coefficients $\vec{\mathbf{a}}$,

$$\begin{aligned} \vec{s}^{(j_{max})} =& \vec{s}^{(3)} = (5, -1, 7, 3, 1, 1, 2, 4) \\ =& \vec{\mathbf{a}}^{(3)} = (a_0^{(3)}, a_1^{(3)}, a_2^{(3)}, a_3^{(3)}, a_4^{(3)}, a_5^{(3)}, a_6^{(3)}, a_7^{(3)}), \end{aligned} \tag{3.12}$$

where j_{max} is given by the number of points in the signal $N = 2^{j_{max}}$ (in this case $N = 8$, thus $j_{max} = 3$). For the original signal the resolution of the signal j is equal to j_{max} and after each pass of the transform it decreases by one corresponding to an increase in scale (see Fig. 3.4 shows the signal at each scale).

For the first pass of the decomposition ($j = j_{max} - 1 = 2$) we carry out the procedure from Eq. 3.8 through 3.11 on each pair of approximation coefficients e.g. $(a_0^{(3)}, a_1^{(3)})$, $(a_2^{(3)}, a_3^{(3)})$, etc. The following is the first level decomposition:

$$\begin{aligned} \vec{\mathbf{a}}^{(2)} &= \left(\frac{5 + (-1)}{2}, \frac{7 + 3}{2}, \frac{1 + 1}{2}, \frac{2 + 4}{2} \right) = (2, 5, 1, 3), \\ \vec{\mathbf{c}}^{(2)} &= \left(\frac{5 - (-1)}{2}, \frac{7 - 3}{2}, \frac{1 - 1}{2}, \frac{2 - 4}{2} \right) = (3, 2, 0, -1). \end{aligned} \tag{3.13}$$

The first pass gives us 4 approximation coefficients and 4 detail coefficients, each half the length of the original signal. These coefficients tell us that over the region $[r_0, r_2)$ the average is 2 and there is a jump of $3 \cdot (-2) = -6$ (3 is the coefficient of the wavelet function which jumps by -2), over the region $[r_2, r_4)$ the average is 5 and there is a jump by $2 \cdot (-2) = -4$, and so on. The approximation coefficients are a smoothed version of the signal whereas the detail coefficients captures the fluctuations from the average.

For the second pass ($j = 1$) we decompose the approximation of the function from the first pass:

$$\begin{aligned} \vec{\mathbf{a}}^{(1)} &= \left(\frac{2+5}{2}, \frac{1+3}{2}\right) = (3.5, 2), \\ \vec{\mathbf{c}}^{(1)} &= \left(\frac{2-5}{2}, \frac{1-3}{2}\right) = (-1.5, -1). \end{aligned} \tag{3.14}$$

This process can be repeated until the largest possible scale has been reached ($j = 0$)

$$\begin{aligned} \vec{\mathbf{a}}^{(0)} &= \left(\frac{3.5+2}{2}\right) = (2.75), \\ \vec{\mathbf{c}}^{(0)} &= \left(\frac{3.5-2}{2}\right) = (0.75), \end{aligned} \tag{3.15}$$

where 2.75 represents the average of the entire signal and there is a jump of $0.75 \cdot (-2) = -1.5$ at the halfway point.

3.2.2 Edge Effects

One requirement in using a DWT is that the signal be an integer power of two in length. Since very rarely can the length of a signal be controlled (e.g. protein chain lengths) there are several different methods to extend a signal to allow for the wavelet decomposition. Any method of signal extension introduces border distortions but due to the localization of wavelets this only affects a few coefficients near the end of the signal. Some of the popular methods described below are zero-padding, symmetrization, and periodic extension.

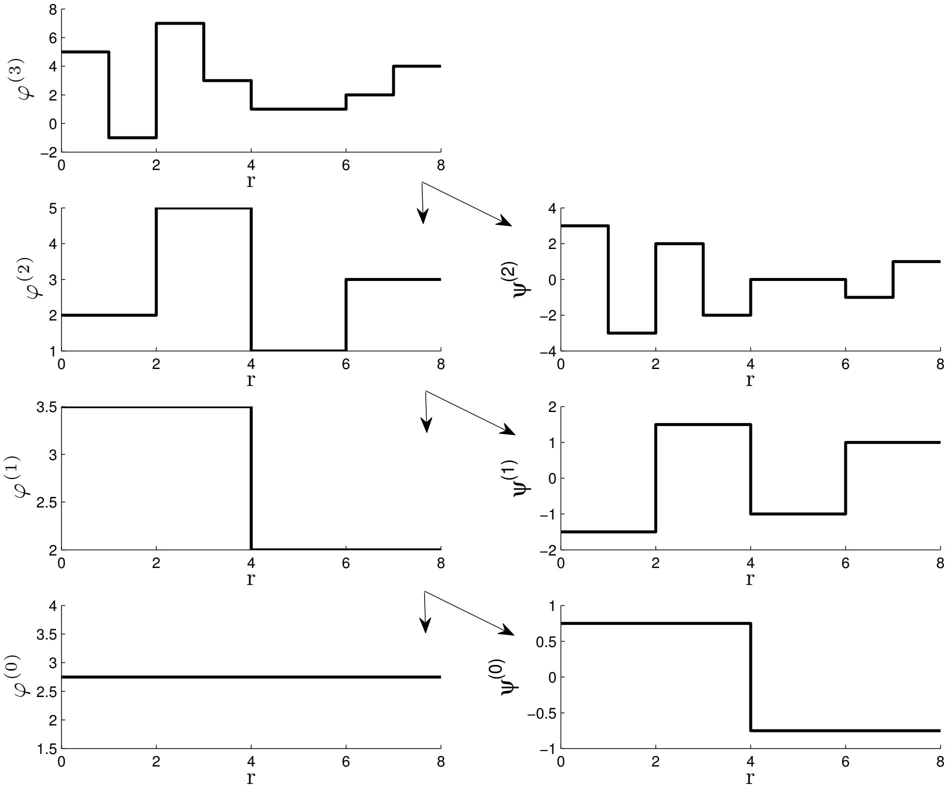


Figure 3.4: Multiscale decomposition of an arbitrary signal with the Haar wavelet transform.

- Zero-Padding, as the name suggests, means that you just extend the array with zeros to the next highest integer power of two. For example, if we had a protein chain with 311 residues then we would extend it's hydrophobicity signal with zeros out to 512 or 2^9. Zero-padding is considered to be the least accurate method for signal extension, however the advantage of zero-padding is that it is extremely easy and the fastest computationally.

- Periodic extension is a popular method in which you treat the data as if it were periodic. After the end of the signal the beginning of the signal is repeated until the next integer power of two is reached. Periodic extensions are also fast and yet there is less distortion compared to zero-padding.

- Symmetric extension involves lengthening the end of a signal by its mirror reflection. For example, if a signal ended with 0, 1, 2, 3 then the extension would begin with 3, 2, 1, 0. Both symmetric and periodic extensions are a better choice than zero padding because they use values and slopes that are similar to those in the data.

3.3 The Daubechies Wavelets

The wavelets used in this project are the Daubechies D4 (or 4 Tap because as we'll see it involves 4 coefficients) wavelets developed by Ingrid Daubechies in the 1990's [26]. The major difference between the Daubechies and the Haar wavelets is that the Daubechies wavelets do not have jump discontinuities and as such represent signals in frequency or scale space with better localization. The theory behind these wavelets is much more intricate than that of the Haar wavelets, nevertheless, all of the main concepts carry over from the previous section.

Like the Haar system the Daubechies system has a scaling function or building blocks, although they cannot be expressed in terms of elementary functions (e.g. sines, exponentials, polynomials, etc.). Rather they are defined by a set of initial conditions

$$\begin{aligned}\varphi(0) =&0,\\ \varphi(1) =&\frac{1+\sqrt{3}}{2},\\ \varphi(2) =&\frac{1-\sqrt{3}}{2},\\ \varphi(3) =&0,\end{aligned} \tag{3.16}$$

and the recurrence relation

$$\begin{aligned}\varphi(r) = \frac{1+\sqrt{3}}{4}\varphi(2r) + \frac{3+\sqrt{3}}{4}\varphi(2r-1) \\ + \frac{3-\sqrt{3}}{4}\varphi(2r-2) + \frac{1-\sqrt{3}}{4}\varphi(2r-3),\end{aligned} \tag{3.17}$$

where φ is also zero outside the interval from 0 to 3. For convenience we denote the coefficients of Eq. 3.17 with the abbreviations h_0, h_1, h_2 and, h_3:

$$h_0 = \frac{1+\sqrt{3}}{4}, \qquad h_1 = \frac{3+\sqrt{3}}{4}, \qquad h_2 = \frac{3-\sqrt{3}}{4}, \qquad h_3 = \frac{1-\sqrt{3}}{4}, \tag{3.18}$$

so that Eq. 3.17 can be expressed as an inner product with the h's:

$$\varphi(r) = h_0 \cdot \varphi(2r) + h_1 \cdot \varphi(2r-1) + h_2 \cdot \varphi(2r-2) + h_3 \cdot \varphi(2r-3). \tag{3.19}$$

Equation 3.19 along with the initial conditions 3.16 is the generator of the building block function (φ is shown in Fig. 3.5). The values of φ in between the initial values are found starting with half-integer values of r (e.g. 1/2, 3/2, 5/2) and then proceeding, using the half-integer values, to find the quarter-integer values and so on.

The associated wavelet function ψ is expressed in terms of the basic building block φ by the following recurrence relationship:

$$\psi(r) = -h_0 \cdot \varphi(2r-1) + h_1 \cdot \varphi(2r) - h_2 \cdot \varphi(2r+1) + h_3 \cdot \varphi(2r+2). \tag{3.20}$$

Due to the boundaries on the scaling function $\varphi(r) = 0$ if $r \leq 0$ or $3 \leq r$, it follows that $\psi(r) = 0$ if $r \leq -1$ or $2 \leq r$. Equation 3.20 along with the initial conditions on φ generates the the wavelet function (shown in Fig. 3.6).

The Daubechies wavelet transform algorithm emerges from the recursion relations (Eq. 3.17 and 3.20). Substituting $r/2$ for r in both recursion relations and translating ψ by one unit to the right gives:

$$\varphi([r/2]) = h_0 \cdot \varphi(r) + h_1 \cdot \varphi(r-1) + h_2 \cdot \varphi(r-2) + h_3 \cdot \varphi(r-3), \tag{3.21}$$

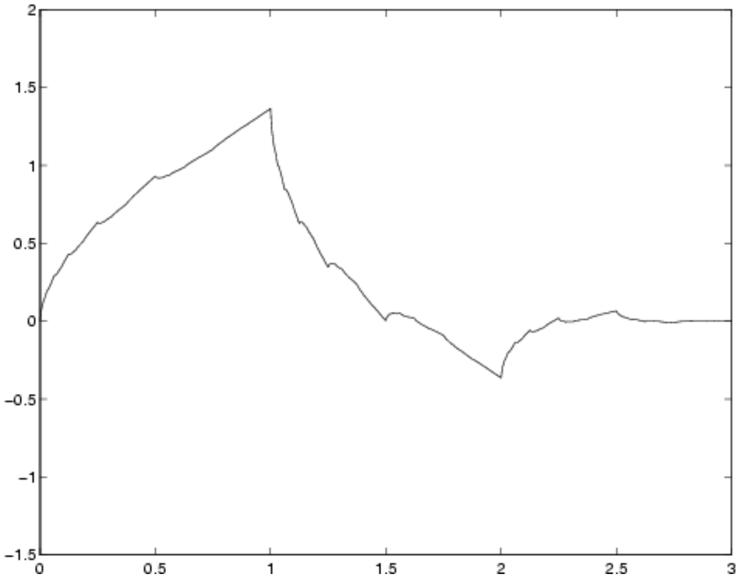


Figure 3.5: The Daubechies 4 basic building block or scaling function $\varphi(r)$.

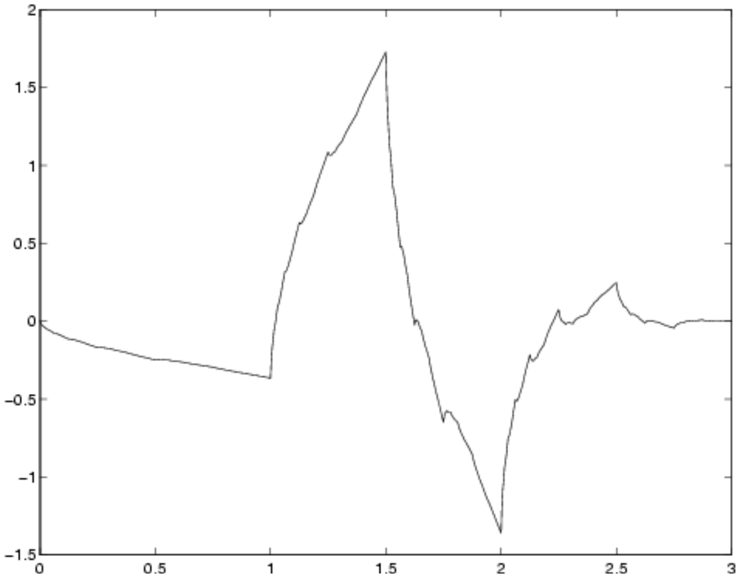


Figure 3.6: The Daubechies wavelet function $\psi(r-1)$.

$$\psi([r/2]-1) = h_3 \cdot \varphi(r) - h_2 \cdot \varphi(r-1) + h_1 \cdot \varphi(r-2) - h_0 \cdot \varphi(r-3). \tag{3.22}$$

Equations 3.21 and 3.22 show how four consecutive scaling functions can be transformed to give both a scaling function, $\varphi([r/2])$, and wavelet function, $\psi([r/2]-1)$, of two times the scale. The calculation of a single scale wavelet transform is demonstrated by the following matrix equation:

$$\begin{pmatrix} a_0^{(j-1)} \\ c_0^{(j-1)} \\ a_1^{(j-1)} \\ c_1^{(j-1)} \\ a_2^{(j-1)} \\ c_2^{(j-1)} \\ \vdots \end{pmatrix} = \frac{1}{2} \begin{pmatrix} h_0 & h_1 & h_2 & h_3 & 0 & 0 & \cdots \\ h_3 & -h_2 & h_1 & -h_0 & 0 & 0 & \cdots \\ 0 & 0 & h_0 & h_1 & h_2 & h_3 & \cdots \\ 0 & 0 & h_3 & -h_2 & h_1 & -h_0 & \cdots \\ 0 & 0 & 0 & 0 & h_0 & h_1 & \cdots \\ 0 & 0 & 0 & 0 & h_3 & -h_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \times \begin{pmatrix} a_0^{(j)} \\ a_1^{(j)} \\ a_2^{(j)} \\ a_3^{(j)} \\ a_4^{(j)} \\ a_5^{(j)} \\ \vdots \end{pmatrix}. \tag{3.23}$$

The result of the matrix multiplication gives the next higher scale $(j-1)$ approximation and detail coefficients arranged in an alternating fashion in a column array. To perform the next scale decomposition the approximation coefficients, $\vec{\mathbf{a}}^{(j-1)}$, are separated from detail coefficients and the same matrix operation is carried out again to produce $\vec{\mathbf{a}}^{(j-2)}$ and $\vec{\mathbf{c}}^{(j-2)}$. This process can be continued until there is just one approximation coefficient $a^{(0)}$.

3.4 Wavelet Reconstruction

Thus far we've discussed signal decomposition into detail and approximation coefficients (here we label them as cD and cA) and we've seen that these decomposed signals contain fewer and fewer coefficients as we move to higher scale. The first scale detail coefficients, for example, are half in number as the original signal. This loss in resolution is a problem for localized analysis and that's where the topic of wavelet reconstruction comes in. The process of reconstruction is used to synthesize signals from detail and approximation coefficients. We could for example retrieve the original signal from a reconstruction of the same coefficients attained by the decomposition process (see Fig. 3.7). However, our purpose is to reconstruct the details and approximations so that they have the same length as the original signal.

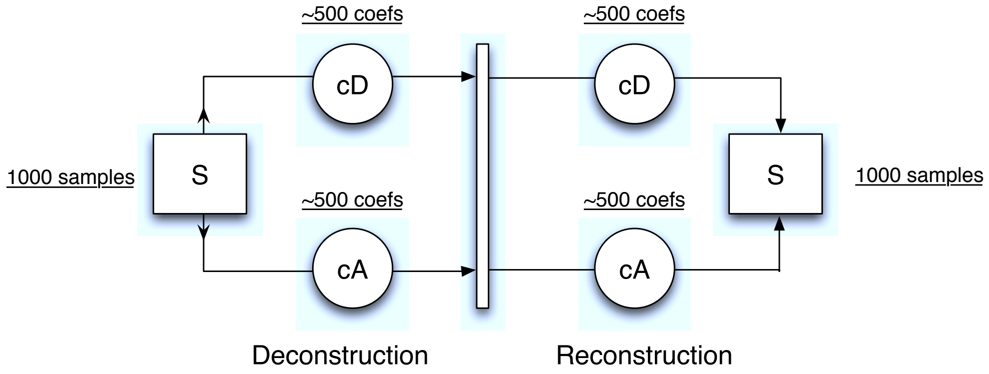

~500 coefs
cD
~500 coefs
cD
1000 samples
S
~500 coefs
~500 coefs
S
1000 samples
cA
cA
Deconstruction
Reconstruction

Figure 3.7: Diagram of the wavelet reconstruction of the original signal from the detail (cD) and approximation (cA) coefficients.

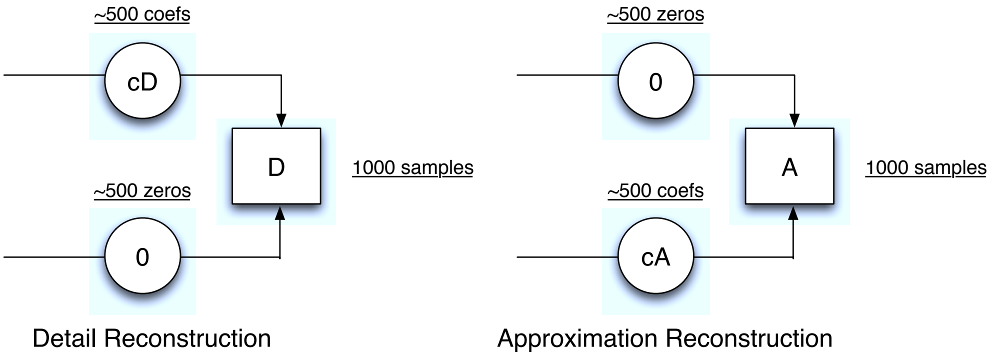

~500 coefs
cD
D
1000 samples
~500 zeros
0
Detail Reconstruction
~500 zeros
0
A
1000 samples
~500 coefs
cA
Approximation Reconstruction

Figure 3.8: Diagram showing how a detail signal D is reconstructed from the detail coefficients (cD) and an array of zeros in place of the approximation coefficients. Likewise the approximation A can be reconstructed from the approximation coefficients (cA) and an array of zeros in place of the detail coefficients.

The reconstruction or synthesis process involves the inverse wavelet transform. The inverse wavelet transform is performed by the inverse of equation 3.23 which is possible because the matrix of h's is orthogonal and thus its inverse is its transpose [28]. As we said earlier, information is not lost in the transform process and this is what makes possible the reconstruction of signals. The inverse transform takes as inputs the approximation and detail coefficients and produces a signal characterized by them. In order to reconstruct a detail signal D you perform the inverse transform using the detail coefficients as one input and in place of the approximation coefficients your second input is an array of zeros (see Fig. 3.8). The same procedure is used to produce an approximation signal A except that detail coefficients are instead replaced with zeros. The result is that A and D have the same number points as the original signal S and in fact $S = A + D$.

Wavelet reconstruction can be used on multi-level transformed signals also to reconstruct the detail each level of the decomposition. Figure 3.9 is an example that shows the reconstructed detail of a signal at six different scales. Remember that the detail represents the fluctuation in the signal from the average of the signal. With each pass of the wavelet transform (moving downward in Fig. 3.9) we are looking at coarser fluctuations corresponding to larger scale regions of the signal. In this project we use functions that are part of the Matlab Wavelet Toolbox for both decomposition and reconstruction.

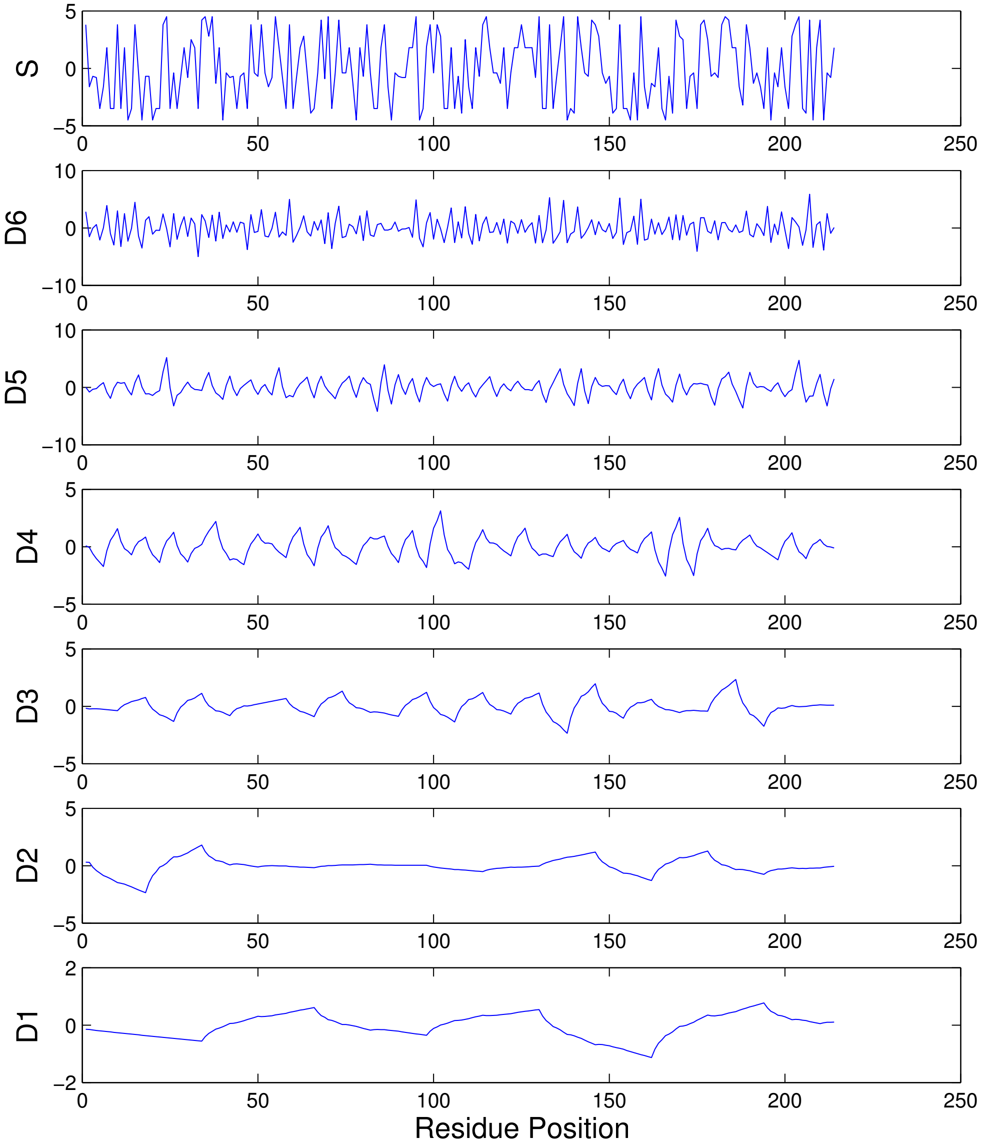


Figure 3.9: The hydrophobicity signal S of a protein (top plot) and the reconstructed details at the first 6 scales.

CHAPTER 4

Results

4.1 Introduction

As discussed in chapter 1 the three dimensional structure of protein is determined by the sequence of amino acids that form the polypeptide chain. Encoded in this sequence is the information necessary to produce the native structure but due to the complex relationship between sequence and structure, the code has not yet been deciphered. There are two main approaches to solving this problem. One approach is the pure physics based approach or molecular dynamics which involves computing the folding trajectory using atomic force fields. However, at this time these direct calculation methods are computationally expensive and limited to short proteins. The other general approach is to use a knowledge based algorithm. With this type of approach prediction of an unknown protein's structure is made based on statistical probabilities derived from a set of known proteins with similar sequence. The downside of the knowledge based approaches are that they, for the most part, disregard the fundamental forces involved in the folding and they rely on large sets of similar proteins with solved structures.

In recent years some researchers have taken another approach that uses statistical analysis to detect periodicities or other implicit order in the amino acid sequence. Pattini and Cerutti [29] have used a wavelet approach to analyze the protein's hydrophobicity profile in order to predict the presence of alpha helices. The wavelet transform has been used to find functional similarity of proteins by Wen *et al.* [30]. Giuliani *et al.* [31] have used various signal analysis methods including wavelets, singular value decomposition, and recurrence quantification analysis to search for sequence-structure relationships.

In this project the approach we take for finding sequence-structure relationships is wavelet based. The discrete wavelet transform (DWT), covered in chapter 3, is a technique that is able to capture multi-scale features in signals. In order to use this technique it is necessary to represent the protein sequence as a numerical signal that contains some physical significance in terms of protein structure. We choose hydrophobicity scales (discussed in § 1.3.1) because of the general understanding that hydrophobic forces play a dominant role in the native structures of proteins.

With wavelets there are two ways of analyzing a signal. One way is to use the approximation coefficients which give a smoothed or coarse grained version of the signal. In this type of analysis one can look at the average value of the signal over quasilocalized regions. This approach was taken by Pando *et al.* [32] where they used coarse grained versions of the hydrophobicity to detect secondary structure. In this project we adopt the other approach which is to look at the wavelet detail at different scales which represents the fluctuations from the average of the hydrophobicity. Our hypothesis is that there may be correlation between locations of secondary structure and significant fluctuations in hydrophobicity.

One point that gives some validation to this hypothesis comes from observed hydrophobic patterns in secondary structure. It has been found that helices and strands near the globular protein surface tend to follow alternating patterns of hydrophobic and hydrophilic residues because of the orientations of the R-groups within these structures [33]. In particular helixes, with their arrangement of 3.6 residues per turn, tend to follow the pattern 0011011 or 0010011 (where 1 represents hydrophobic and 0 hydrophilic or vice versa) such that the side of the helix facing the solvent is hydrophilic and the side facing the interior is hydrophobic. The R-groups of beta strands, on the other hand, alternate with each successive residue between the different sides of the sheet and thus the hydrophobicity patterns of these structures near the globular surface tend to follow the pattern 010101. These are the types of patterns that we would expect wavelet detail to be very good at detecting, particularly at the first or second wavelet scales. However, this line of reasoning would suggest that secondary structure near the globular core, which tend to consist mainly of hydrophobic residues, would not be detected by wavelet detail.

The proteins that we analyze are comprised of the two different classes $\alpha+\beta$ and α/β proteins. As discussed in chapter 2, $\alpha+\beta$ proteins contain mainly segregated α-helices and β-sheets whereas these structures in α/β proteins occur close together sometimes alternating between the two. Initially we obtained a total of 20,689 PDB files but due to a number of apparent errors and complications (see § 2.3) we were forced to filter the data. The filtering process left us with 5935 $\alpha+\beta$ proteins consisting of 6939 unique chains and 7500 α/β proteins consisting of 8572 unique protein chains. These two sets of proteins were analyzed separately so that comparisons could be made between the results of the two classes. To evaluate our methods against an entire data set we calculate a number of performance measures (explained below in § 4.2.1) and look at distributions of these values.

This research is intended to lay a foundation and establish tools for finding protein sequence-structure relationships. We expect that further work with these techniques will lead to improved results.

4.2 Methods and Evaluation Measures

This section gives a description of the main method of our approach in detecting the locations of secondary structure. Figure 4.1 shows a flow chart of all the essential steps of the process. The flowchart begins at the top with the raw structural data of a protein taken from a PDB file. The primary amino acid sequence data from the PDB file is extracted and the sequence is converted into a hydrophobicity signal using one of the three hydrophobicity scales (described in § 1.3.1). Those three scales again are the Kyte-Doolittle (KD), Hopp-Woods (HW), and Engelman-Steitz (ES) scales. We tested each of these three scales with our technique to compare the results and to determine whether any one of them is particularly better to use with this approach.

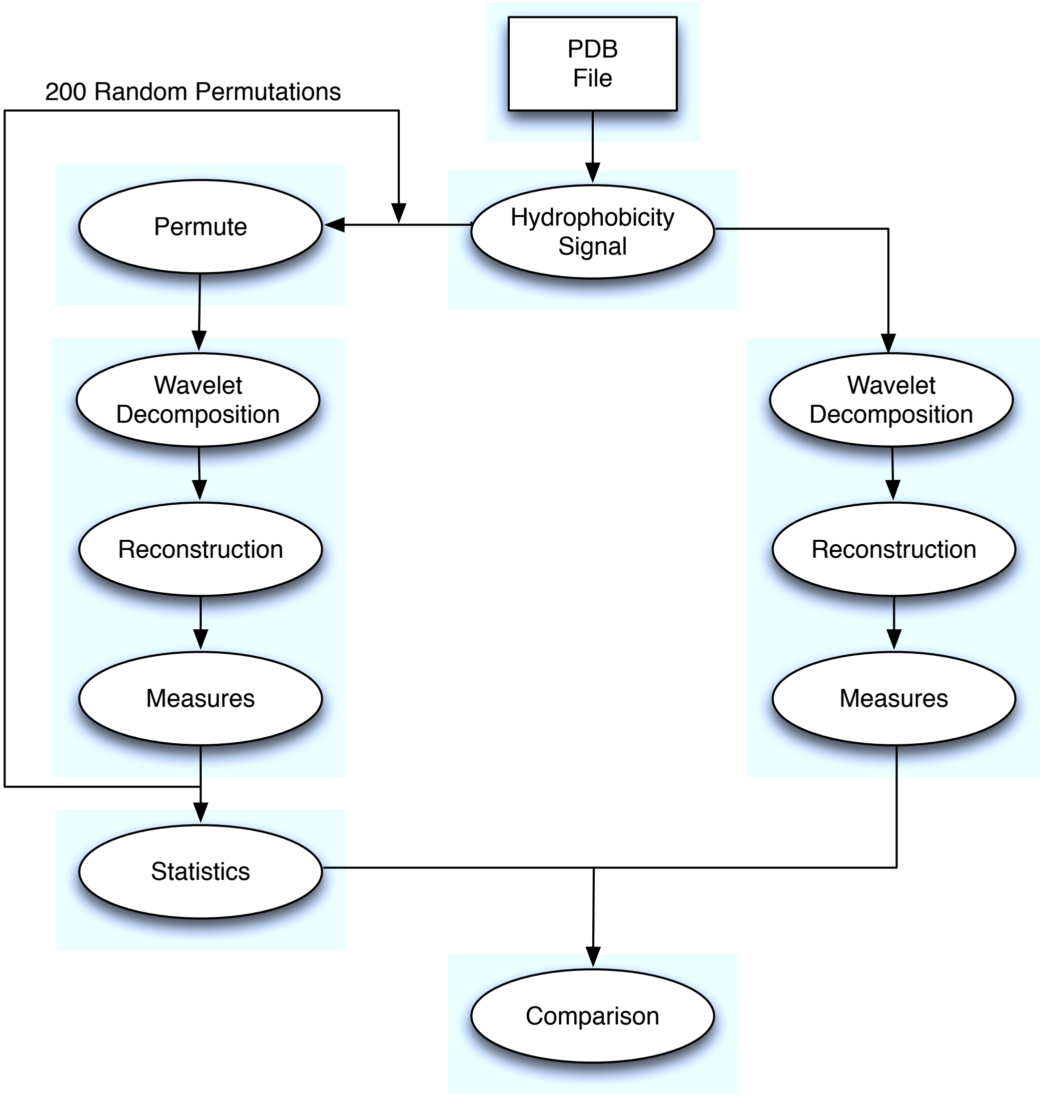

PDB
File
200 Random Permutations
Permute
Hydrophobicity
Signal
Wavelet
Decomposition
Reconstruction
Measures
Wavelet
Decomposition
Reconstruction
Measures
Statistics
Comparison

Figure 4.1: Flowchart outline of the main processing steps in our analysis.

After the protein primary sequence has been converted to a hydrophobicity signal the flowchart breaks off into two paths: the path to the right corresponds to the real signal of the protein, the path to the left corresponds to a set of randomized versions or *realizations* of the protein for comparison. Let us first go through the processing of the real protein sequence. The first step in the process is wavelet decomposition by way of the Daubechies wavelets. As discussed in the previous chapter (see § 3.2.2) protein chains are rarely integer power of two in length and thus must be extended. The extension inevitably causes distortion near the edges but this only affects a few coefficients near the end. We tested out a few different extension methods that are built into the Matlab Wavelet Toolbox and found no significant differences in the results thus we went the default symmetrization method. This method extends the signal with values that mirror the end of the signal. The fact that we found no differences in the various extension methods is a result of the localization property of the wavelet.

The wavelet transform provides the detail and approximation coefficients which both have a reduced resolution compared the the original signal. For local analysis it is important to to retain the resolution of the original signal by the process of wavelet reconstruction (as described in § 3.4). The reconstruction process takes the detail coefficients and synthesizes a detail signal with the same number of points as the original hydrophobicity signal yet without adding any new information.

Up to this point we've described the processing involved for the real protein hydrophobicity signal. In order to determine the significant features in this signal we first must have a way of defining significance. Our approach is to compare the detail of the real protein to the details of a set of 200 random realizations of the protein. Each realization consists of the same amino acid occurrence just arranged differently. This tests the hypothesis that not only which amino acids, but also where they are located in the protein is important to structure. In effect we take the primary sequence and randomly permute it 200 times carrying out the same process on each realization as was performed on the actual protein sequence. This set of realizations forms what we call the control data which is used to generate a threshold for determining significance. There are two basic types of thresholds that we could generate. One type would be a threshold that varies across the length of the protein due to the local variation of the control data. In this case the threshold at each residue location is computed as the average of the control data at that location plus or minus one standard deviation (1σ) of the control data at that residue location. The other type of threshold is a flat global threshold which would be computed by taking the average of the localized threshold. For all the analysis in this chapter we adopt the flat threshold but there is very little difference between the two. Another issue that we discuss later in the chapter is whether the 1σ threshold is an appropriate level for identifying secondary structure. It must be remembered that the 1σ threshold is not calculated from a random distribution but a distribution that has the same amino acid frequencies of the parent protein. Thus the 1σ threshold is relatively robust. In a slightly different approach Pando *et al* [32] showed that the 1σ threshold was a valid cutoff.

Figure 4.2 shows the reconstructed detail at the first four scales ($j = j_{max}-1$, $j = j_{max} - 2$, etc.) along with the flat 1σ threshold obtained from the 200 realizations. At the bottom of each plot are also shown the actual locations of secondary structure as given from the PDB file with black lines representing helix structures and red lines representing strands. As the figure shows there are wavelet structures that exceed the threshold at each of the four scales. However, the first scale seems to correspond the best with the secondary structure. The strongest detail peaks occur across almost the entire span of the single helix. There are also salient peaks that seem to match up well with the three strand structures. In the next section of this chapter we test out a technique using the *envelope* of the wavelet detail to perform a two-state secondary structure prediction of each residue in a protein. First, we must have a way to quantify the performance of a two-state prediction.

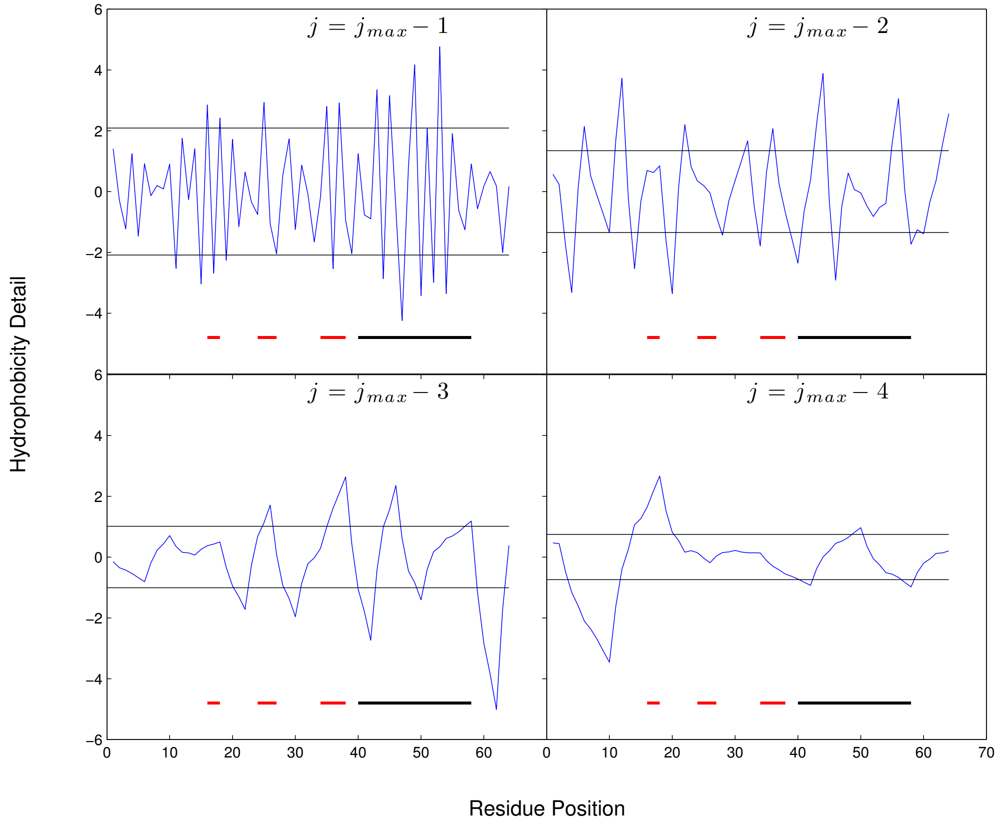


Figure 4.2: The blue signals are the detail at the first four scales for the protein 1KJK (PDB ID). The threshold produced by the 200 random realizations is colored black. The locations of the known secondary structure, as given in the PDB file, is plotted at the bottom of each window where the color black indicates helix and red indicates strand. Particularly at the the first scale we see clear correspondence between the helix and significant detail.

4.2.1 Prediction Evaluation Measures

In order to determine how well a secondary structure prediction method works there must be a system to measure performance relative to the known structure in the database. There are a number of different performance measures, each with their own strengths and weaknesses. The measures that we will consider here include the sensitivity S_n, specificity S_p, accuracy Q, and the Matthews correlation coefficient MCC and are some of the most commonly used measures in protein secondary structure prediction.

As a general example consider an amino acid sequence of length N. The structural data that we take from the PDB for comparison is the secondary structure assignments $\mathbf{D} = d_1, d_2, ..., d_N$ where each element d_i corresponds to the accepted structural state of a residue. For simplicity consider the dichotomous case of two alternative classes such as helix versus non-helix. The structural data array values d_i would then take on the binary values 1 for residues with helices and 0's for all other residues. Now our prediction algorithm or model outputs a prediction $\mathbf{M} = m_1, m_2, ..., m_N$ where the values m_i will be 1 for the prediction of a helix, which in our case corresponds to a wavelet structure exceeding the threshold, and 0 for the prediction of a non helix, corresponding to a wavelet structure below the threshold. The important question is how to compare $\mathbf{M}$ with $\mathbf{D}$ to best evaluate the performance of the prediction $\mathbf{M}$.

With both $\mathbf{M}$ and $\mathbf{D}$ binary, their comparison can be summarized by the following four numbers:

- T_p (True Positive) is the number of times d_i and m_i are both 1 or a structure is correctly predicted,

- T_n (True Negative) is the number of times d_i and m_i are both 0 or there is correct prediction of no structure,

- F_p (False Positive) is the number of times d_i is 0 and m_i is 1 or a structure is predicted where there is none,

- F_n (False Negative) is the number of times d_i is 1 and m_i is 0 or a structure exists where it is not predicted,

- The sum the four numbers satisfies: $T_p + T_n + F_p + F_n = N$.

These four numbers are sometimes arranged into what is known as a contingency table or confusion matrix (see Table 4.1 below).

Table 4.1: The Contingency Table shows the four numbers that are used to evaluate the performance of a prediction.

		PREDICTED		
		yes	no	Total
ACTUAL	yes	T_p	F_n	$T_p + F_n$
	no	F_p	T_n	$F_p + T_n$
	Total	$T_p + F_p$	$F_n + T_n$	N

Prediction evaluation measures are designed to give a single number for the comparison between **M** and **D**. The single number measures are constructed from some combination of T_p, T_n, F_p, and F_n. Whenever an overall measure is constructed from local information, there is a loss of information. Nevertheless, the overall measure provides a clearer picture of the validity of the method. Two such measures, the sensitivity (S_n) and the specificity (S_p), are defined as follows

$$S_n = \frac{T_p}{T_p + F_n}, \tag{4.1}$$

$$S_p = \frac{T_p}{T_p + F_p}. \tag{4.2}$$

The sensitivity is a measure of the proportion of correctly predicted structures over the total number of structures. The specificity is the proportion of correctly predicted structures to all predicted structures. It is trivial to see that both these measures have maximum values of 1 and are also both zero if there are no True Positives. Both of these measures however are biased because they ignore the performance of handling negative cases. Sensitivity relates only the first column of the contingency table and specificity only the first row, neither of them takes into account the number of True Negatives T_n. These values do provide useful information, they just fail to give the full picture. Sometimes what is actually looked at is the average of the two values or alternatively their geometric mean $G = \sqrt{S_n S_p}$. G has a value of 1 for a perfect prediction and in the case of no positive predictions has the value 0. However, as G is just a combination of the sensitivity and specificity, it also ignores the performance of handling negative cases.

Due to its intuitive ease of understanding a more common measure for evaluating performance is the Q statistic (also called the accuracy or success rate) defined as

$$Q = \frac{T_p + T_n}{N}. \tag{4.3}$$

Q is the proportion of all correct cases to the total number of cases N and is often presented as a percentage. While Q is an important measure it tells nothing about how the false cases F_p and F_n are divided, information that could indicate whether the algorithm is biased toward too many predictions or too few. We address biasing later in the chapter.

A more complex measure that is commonly used in bioinformatics is the Matthews correlation coefficient (MCC) defined by

$$MCC = \frac{T_p T_n - F_p F_n}{\sqrt{(T_n + F_n)(T_n + F_p)(T_p + F_n)(T_p + F_p)}}. \tag{4.4}$$

The importance of the MCC is that it gives a value of 1 when there is complete association, 0 when there is no association, and -1 when there is complete negative association. The MCC is typically a much smaller value than Q but in many cases gives a much better representation of how “good” the prediction is.

While the above mentioned measures are typically used to gauge how well **M** performs there are two additional measures that will prove important for making adjustments to the prediction threshold. It is straightforward to see that lowering the threshold will naturally lead to greater numbers of positive predictions (wavelet structures above the threshold) and lowering the threshold would do the opposite. This raises the question of whether the 1σ threshold is appropriate in terms of the number of positive predictions our model makes. In other words, do we tend to predict more or less secondary structure than what is known to be? This question can be answered in terms of the following two measures:

$$\text{Prevalence } = P = \frac{T_p + F_n}{N} \tag{4.5}$$

$$\text{Bias } = B = \frac{T_p + F_p}{N} \tag{4.6}$$

The first quantity P represents the Prevalence of the positive cases, i.e., the proportion of the protein chain that contains secondary structure. Clearly the number of known structures correctly predicted (T_p) plus the number of known structures not predicted (F_n) sums up to the total number of experimentally known structure. Prevalence does not depend on the prediction algorithm but rather is a property of the protein and it will vary from protein to protein. Bias on the other hand is the tendency of the algorithm to output positive predictions and is defined as the total number of predictions over the total number of cases. The result of increasing or decreasing the threshold will affect the Bias. A common rule of thumb for most models is to parametrize them in such a way that $B \approx P$ or $B/P \approx 1$ (i.e. such that the number of predicted structures approximately equals the number of known structures). A good prediction will not only meet this criteria but will also have a good correlation. Note that setting $B = P$ is equivalent to setting $F_n = F_p$. The Matlab program that we used to compute the above mentioned evaluation measures is given in Appendix A.

4.3 Per-Residue Evaluation by Wavelet Enveloping

Much of our initial research involved evaluating how well wavelet structures matched with regions of secondary structure rather than predicting the state of each residue individually. When a wavelet structure peaked in a region of secondary structure it was counted as a True Positive, in a region of no secondary structure a False

Positive, etc. This type of analysis ignored starting and ending positions and the span of secondary structure. The approach we take here seeks to overcome these issues by using not the rapidly fluctuating detail but rather its envelope as the determining factor of significance (i.e. the predictor). In this section we test a wavelet enveloping technique with the per-residue type evaluation of performance. This means that each residue is determined to be either T_p, T_n, F_p, or F_n and the four numbers sum up to the number of residues in the chain N.

The wavelet envelope is generated by linearly interpolating between consecutive maxima in the reconstructed detail for the top envelope and consecutive minima for the bottom envelope. Using the detail envelope instead of the detail itself results in much broader regions of significance with lengths more like what would be expected of secondary structure. Figure 4.3 shows the results of this technique for the three different hydrophobicity scales on a protein from the α+β database. Most noticeably using the KD scale there is clear similarity between the known structure **D** and the prediction **M**. For this case (with the KD scale) the majority of all four helices are predicted well although there are some breaks and overlap. The salient wavelet structures also appear to correlate with the β-strands in many cases. Appendix B lists the Matlab program for performing this technique on a single protein.

A problem with our technique at this point is that we have not determined a way to distinguish between helix structures and strands. This means that performance evaluation can be done in a couple of ways. One way would be to just test how well the technique predicts helices. In this case we ignore all known strand segments when constructing the **D** array. For the examples in Fig. 4.3 this would mean that all of the predicted structures that line up with strands would be counted as false positives. The other way would be to just test how well the technique predicts helices and strands without distinguishing between the two (a two-state prediction not three). In this way of evaluation a positive prediction that matched with either a strand or a helix would count as a True Positive. Instead of picking one route or

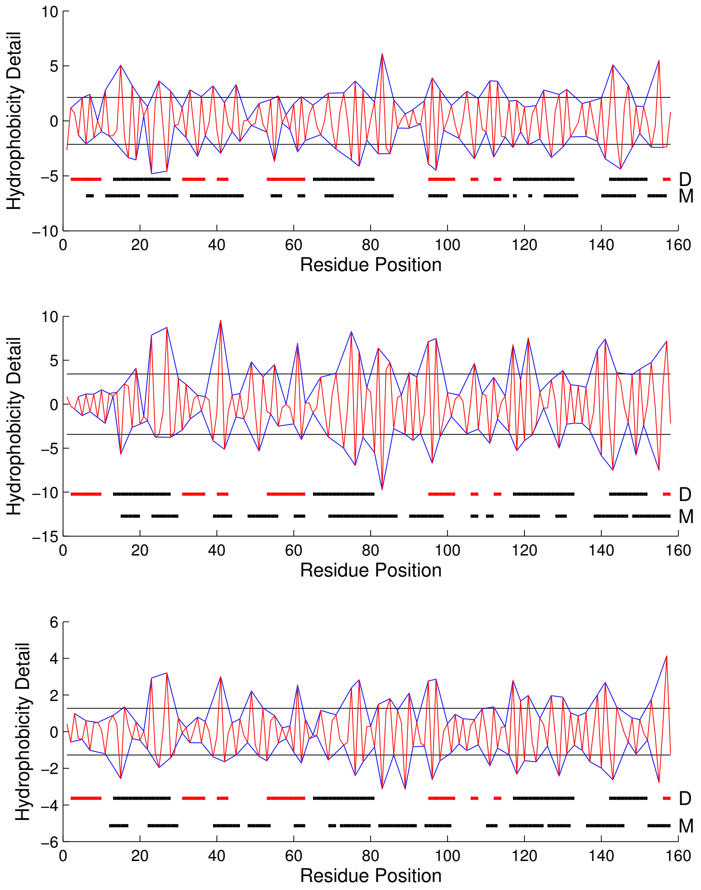


Figure 4.3: Using the enveloping technique we analyze the $\alpha+\beta$ protein 1F9Y using the three hydrophobicity scales: KD (top), ES (middle), and HW (bottom). The red and blue curves are the first scale reconstructed detail and its envelope respectively. For all three cases the threshold (black) is set to 1σ and displayed at the bottom of the window is the actual secondary structure **D** (where black lines are helices and red are strands) and what is predicted from the salient features of the envelope **M**. Here the KD scale performs the best with $Q = 0.61$ and $MCC = 0.18$. All three cases are evaluated against helices and strands without distinguishing the two.

the other we perform both a helix only evaluation and a general helix and strand structure evaluation. A strand only evaluation would be another route however in all the cases we looked at this correlation was poor thus we didn't pursue that test.

Above we looked at some of the plots of the enveloping technique for an individual protein. Analyzing one or two proteins is not enough to really say how well a technique works in general thus in this section we show some results across our two databases of just over 15,500 chains. Table 4.2 shows the results of enveloping the first wavelet scale ($j = j_{max} - 1$) detail over the entire $\alpha + \beta$ dataset of 6939 chains. The first three columns identify some of the inputs and parameters of the algorithm. The first column, H-scale, identifies which hydrophobicity scale was used. The second column identifies which type of secondary structure we are evaluating our prediction against: H for helices, and H+S for helices and strands. The third column shows the threshold parameters that we chose which are all 1σ for now. All of the following columns correspond to prediction evaluation measures. These values are the average across the whole $\alpha + \beta$ set.

Table 4.2: Results for first scale detail with enveloping on the $\alpha + \beta$ database.

h-scale	struc	t.h.	S_n	S_p	Q	MCC	B/P
KD	H	1σ	0.6784	0.4156	0.5234	0.1005	2.1222
KD	H+S	1σ	0.6678	0.6451	0.5869	0.1361	1.0764
HW	H	1σ	0.6335	0.4161	0.5284	0.0937	1.9748
HW	H+S	1σ	0.6001	0.6237	0.5487	0.0740	1.0034
ES	H	1σ	0.6351	0.4114	0.5233	0.0809	1.9828
ES	H+S	1σ	0.5919	0.6103	0.5348	0.0400	1.0132

There is one pattern in particular that stands out in Table 4.2. We see that the average Bias over Prevalence ratio B/P is close to 1 for evaluation against helices and strands whereas it is around 2 for evaluation against only helices. A B/P value of 1 indicates that the algorithm is not biased (i.e., it is predicting the correct amount of structure) whereas a value of 2 indicates that the algorithm is biased towards predicting twice as much structure as is truly occurring. The pattern is not something that should be surprising. For a given $\mathbf{M}$, if $\mathbf{D}$ includes more structure (in this case additional strands S) then there will be a higher Prevalence, thus a smaller B/P ratio. The 1σ threshold on the average seems to be appropriate for the H+S evaluation at least in terms of Bias. However, if this approach is best suited for predicting helices only we are doing it a disservice in our choice of threshold. For helix only prediction, raising the threshold would reduce the number of positive predictions, thus lowering the Bias B and the B/P ratio.

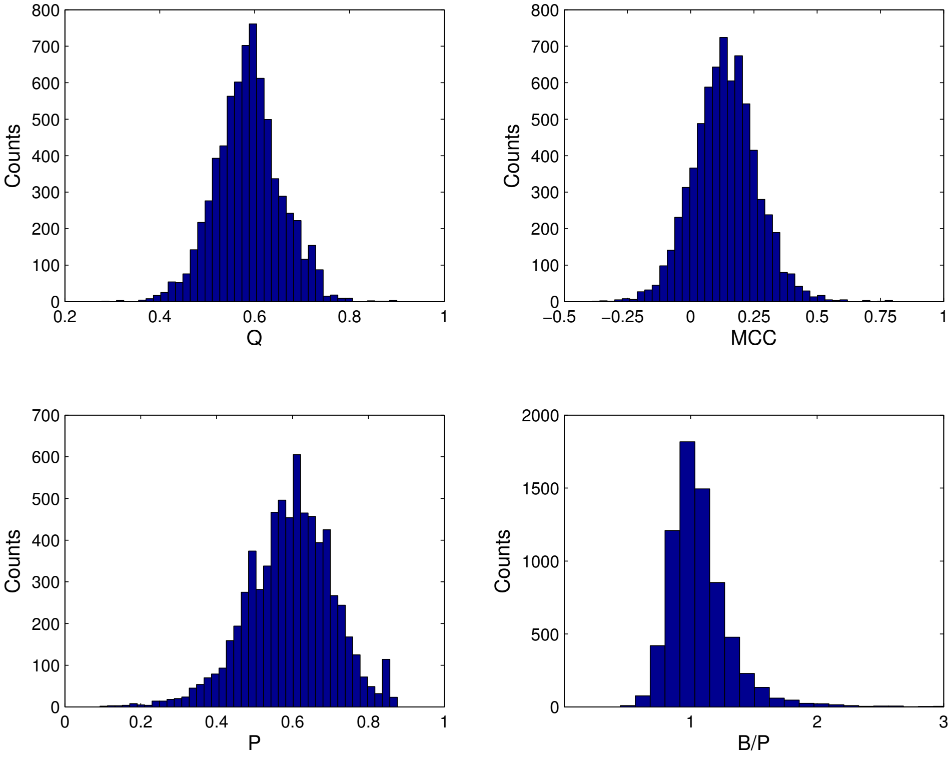


Figure 4.4: Distributions of the evaluation measures for the KD scale, H+S evaluation, with 1σ threshold (second row of Table 4.2).

The data in Table 4.2 are the average values over a very large database of protein chains. A much better picture of what's happening can be provided by a distribution of the results. Figure 4.4 shows distributions of Q, MCC, P, and B/P for the second trial (second row) in Table 4.2. These distributions show that there are wide ranges of values in performance. The Q accuracy goes from 40% to around 80% and MCC values, while on the average show a small but positive correlation, go from about -0.25 to 0.5. The next point of interest is the distributions of P and B/P. The Prevalence P, remember, is the proportion of the protein chain that is occupied by secondary structure. The Prevalence distribution as shown in Fig. 4.4 corresponds to helices and strands and is characteristic of the protein chains in the $\alpha+\beta$ data set (it doesn't change over the H+S trials). As the figure shows, P varies from around 0.2 to over 0.8, thus some proteins have very little secondary structure and some are almost composed entirely of helices and strands. Looking at the distribution of B/P we see a distribution that is broader than we would like. While the average of 1.0764 is an acceptably "unbiased" value (i.e., $B/P \approx 1$), there are significant numbers of chains that are both under-biased by less than $B/P = 0.5$ and some over-biased by more than $B/P = 2$. These results suggest that the 1σ threshold may not be universally suitable over protein chains of varying Prevalence.

For the rest of the chapter we will use the KD hydrophobicity scale in our analysis. Table 4.2 showed that there are no significant differences between the three scales, however the KD scale performed slightly better on the average.

4.3.1 Threshold Optimization

The above observations lead us to take a different approach at evaluating the performance of our model. Instead of using a universal threshold we divide the protein chains up according to levels of Prevalence and then individually determine the appropriate threshold that gives an unbiased prediction for each level. Note that this step makes a leap from a pure prediction based model to a model that uses known information of secondary structure Prevalence in order to set an appropriate threshold. Although the following results are not a pure prediction they still provide valuable information about our approach. Using an unbiased threshold only forces the model to output the correct amount of structure, the correlation is still determined by how well the predicted structure matches with known structure. However, when we come to it will be necessary to discuss the effect this threshold optimizing approach has on the Q accuracy and why the MCC is a more appropriate measure.

Figure 4.5 demonstrates how we determined an acceptable threshold for a given set of proteins with similar P. The plot shows the B/P ratio verses the threshold height for 59 protein chains with Prevalence between 20 and 30%. As we would expect, increasing the threshold height results in a decreased Bias which ultimately goes to zero as the threshold overcomes the greatest detail peak. We found that proteins of similar Prevalence also had similar curves on this plot. The point where these curves intersect with $B/P = 1$ gives the threshold for an unbiased prediction. Thus we took the distribution of points that intersected with $B/P = 1$ and used the average value as the threshold for the set of proteins. In the figure we also show 1σ error bars to give a sense of how the intersections varied.

Tables 4.3 and 4.4 shows the results for the $\alpha+\beta$ and α/β databases respectively. Both tables display results for the helix and strand evaluation (H+S) and a helix only (H) evaluation for the first two wavelet scales. The abbreviation "t.h." stands for threshold. The proteins were broken up into 8 divisions based on P (note that we used slightly different divisions for the H evaluation due to the overall smaller values of P). The numbers of chains in each division greatly varied as shown in the tables. Also notice that we have not included the S_n and S_p measures in these tables. This was done because we found they provided no interesting information on performance. The constraint that $B/P \approx 1$ results in $S_n \approx S_p$ and furthermore we found that they expectedly increased with increasing P as the number of allowed True Positives increased. The measures that we did use were Q and MCC although even Q has some issues we discuss next.

The Q measure is often used to gauge the performance of prediction models but we found it to be less valuable in our situation due to our treatment of the threshold.

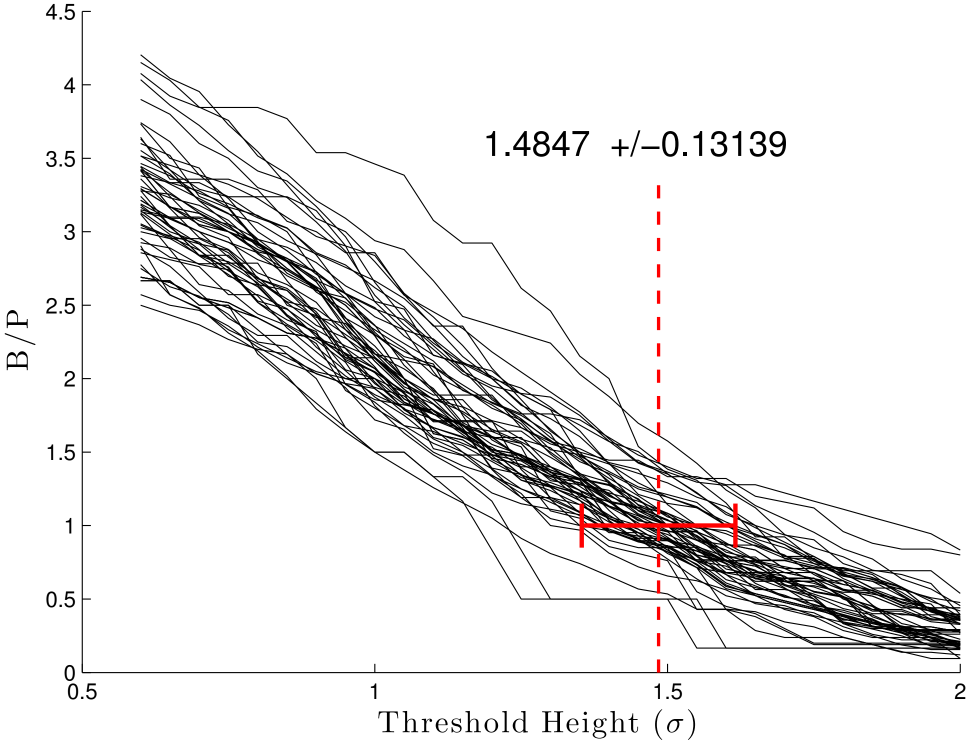


Figure 4.5: Plot of B/P verses the threshold height for the 59 proteins with helix and strand Prevalence between 20-30% ($0.2 < P \leq 0.3$) from the $\alpha + \beta$ database. The threshold height for this set which gives minimal biasing ($B/P \approx 1$) was found to be 1.4847 ± 0.13139. This data corresponds to the second row of Table 4.3 for H+S evaluation.

In either of the tables if we look across the range of different Prevalence divisions the average Q starts out high for small P, it decreases around the 40-50% P range, and then increases back up for high P. At first appearance it would seem our method works very well on proteins with either very low or very high Prevalence, however, adjusting the threshold for an unbiased prediction can significantly increase accuracy even when it is not a good prediction. For example, if a protein of 100 residues has 10 residues in the helix state and we predict exactly 10 residues in the helix state but all of them are False Positives then our totals are $T_p = 0$, $T_n = 80$, $F_p = 10$,

Table 4.3: $\alpha + \beta$ database results for both helix and strand (H+S) and helix only (H) evaluation at the first two wavelet scales. The KD hydrophobicity scale is used and proteins are grouped according to similar Prevalence so as to reduce biasing.

H + S		$j = j_{max} - 1$				$j = j_{max} - 2$			
P	chains	t.h.	Q	MCC	B/P	t.h.	Q	MCC	B/P
$0.0 - 0.2$	24	1.78σ	0.742	0.050	1.078	2.45σ	0.741	0.034	1.051
$0.2 - 0.3$	59	1.48σ	0.656	0.122	1.021	2.06σ	0.608	-0.005	1.031
$0.3 - 0.4$	258	1.30σ	0.600	0.146	1.034	1.76σ	0.567	0.067	1.013
$0.4 - 0.5$	1082	1.20σ	0.558	0.113	1.002	1.52σ	0.549	0.094	0.996
$0.5 - 0.6$	2050	1.07σ	0.569	0.126	1.012	1.34σ	0.545	0.081	0.995
$0.6 - 0.7$	2379	0.95σ	0.604	0.132	1.011	1.16σ	0.576	0.075	0.997
$0.7 - 0.8$	876	0.83σ	0.665	0.138	0.997	1.00σ	0.639	0.080	0.988
$0.8 - 1.0$	211	0.68σ	0.775	0.168	1.002	0.81σ	0.745	0.063	0.997
Total Ave.		1.02σ	0.601	0.130	1.010	1.27σ	0.580	0.080	0.997

H		$j = j_{max} - 1$				$j = j_{max} - 2$			
P	chains	t.h.	Q	MCC	B/P	t.h.	Q	MCC	B/P
$0.0 - 0.1$	197	2.17σ	0.892	0.003	1.060	2.85σ	0.895	-0.002	1.006
$0.1 - 0.2$	695	1.74σ	0.736	0.044	1.061	2.34σ	0.730	0.025	1.059
$0.2 - 0.3$	1361	1.54σ	0.650	0.087	1.028	2.00σ	0.639	0.062	1.030
$0.3 - 0.4$	1692	1.38σ	0.582	0.089	1.023	1.77σ	0.577	0.075	1.013
$0.4 - 0.5$	1725	1.26σ	0.555	0.100	1.005	1.58σ	0.559	0.108	1.002
$0.5 - 0.6$	589	1.12σ	0.559	0.114	1.008	1.36σ	0.545	0.087	1.001
$0.6 - 0.7$	402	1.03σ	0.613	0.155	0.996	1.12σ	0.602	0.139	0.970
$0.7 - 1.0$	278	0.80σ	0.678	0.124	0.989	0.94σ	0.684	0.114	1.022
Total Ave.		1.37σ	0.617	0.092	1.021	1.74σ	0.613	0.081	1.015

and $F_n = 10$ and we achieve a Q accuracy of 80%. While 80% of the time we are correct we fail to predict all 10 helix-state residues (False Negatives) and worse we falsely predict 10 helix-state residues (False Positives). The MCC however gives a more honest evaluation of the prediction. For this example the MCC would give a value of -0.111 which indicates a negative association or a prediction that is worse than random. There are a couple examples in Tables 4.3 and 4.4 of accuracy above 60% yet negative MCC. The MCC is used often in bioinformatics along with Q to give extra support for the evaluation. To get a sense of the MCC measure a

Table 4.4: α/β database results for both helix and strand (H+S) and helix only (H) evaluation at the first two wavelet scales. The KD hydrophobicity scale is used and proteins are grouped according to similar Prevalence so as to reduce biasing.

H+S		$j = j_{max} - 1$				$j = j_{max} - 2$			
P	chains	t.h.	Q	MCC	B/P	t.h.	Q	MCC	B/P
$0.0 - 0.2$	18	1.80σ	0.723	-0.007	1.050	2.13σ	0.739	0.031	1.056
$0.2 - 0.3$	78	1.48σ	0.671	0.174	1.079	1.93σ	0.619	0.024	1.037
$0.3 - 0.4$	141	1.35σ	0.591	0.124	1.053	1.69σ	0.578	0.089	1.045
$0.4 - 0.5$	645	1.20σ	0.559	0.116	1.007	1.51σ	0.547	0.090	1.003
$0.5 - 0.6$	2645	1.07σ	0.564	0.116	1.002	1.35σ	0.546	0.080	0.999
$0.6 - 0.7$	3828	0.96σ	0.591	0.104	1.003	1.19σ	0.581	0.082	0.999
$0.7 - 0.8$	1178	0.84σ	0.649	0.109	1.006	1.04σ	0.633	0.077	0.998
$0.8 - 1.0$	39	0.66σ	0.686	0.041	0.945	0.72σ	0.767	0.215	0.995
Total Ave.		1.01σ	0.590	0.110	1.005	1.26σ	0.573	0.085	1.014
H		$j = j_{max} - 1$				$j = j_{max} - 2$			
P	chains	t.h.	Q	MCC	B/P	t.h.	Q	MCC	B/P
$0.0 - 0.1$	92	2.20σ	0.911	0.057	1.057	2.90σ	0.913	0.014	0.984
$0.1 - 0.2$	186	1.76σ	0.731	0.018	1.043	2.33σ	0.721	-0.019	1.024
$0.2 - 0.3$	676	1.53σ	0.630	0.052	1.035	1.99σ	0.642	0.080	1.024
$0.3 - 0.4$	2048	1.38σ	0.579	0.084	1.008	1.76σ	0.574	0.076	1.014
$0.4 - 0.5$	3277	1.24σ	0.550	0.092	1.012	1.58σ	0.546	0.083	1.004
$0.5 - 0.6$	1867	1.11σ	0.550	0.092	1.012	1.41σ	0.543	0.081	1.005
$0.6 - 0.7$	356	0.98σ	0.579	0.097	1.000	1.25σ	0.579	0.098	1.002
$0.7 - 1.0$	70	0.71σ	0.675	0.130	0.987	0.91σ	0.683	0.136	1.006
Total Ave.		1.27σ	0.576	0.082	1.001	1.62σ	0.569	0.079	1.008

few examples are shown in Fig. 4.6 of good predictions and in Fig. 4.7 some lower values.

In light of these considerations of Q and MCC the averaged results may not seem impressive. On average we are seeing a positive correlation but it is mostly weak with the highest averages around 0.17 for H+S evaluation at the first wavelet scale. Table 4.5 gives a better picture of the distribution of results we're getting for the different tests in Tables 4.3 and 4.4. For all 8 tests there are very few that exceed an MCC of 0.7. If we look at the data in just the first column we see that there are

112 (1.6%) with an MCC between 0.4 and 0.5, 381 (5.4%) between 0.3 and 0.4, and 1274 (18%) between 0.2 and 0.3. For these proteins, which make up approximately 25% of the database, the predictions have good to moderate correlations. There are 2263 (32.6%) proteins that have an MCC between 0.1 and 0.2 which are better than a random predictions but certainly not great. These results indicate that significant fluctuations in hydrophobicity do occur more frequently in regions of secondary structure than in random coil.

Table 4.5: Counts of proteins chains at different levels of MCC for the data from the 8 different tests of Tables 4.3 and 4.4 .

	$j = j_{max} - 1$				$j = j_{max} - 2$			
	$\alpha + \beta$		α/β		$\alpha + \beta$		α/β	
MCC	H+S	H	H+S	H	H+S	H	H+S	H
over 0.8	0	1	0	0	0	2	0	1
0.7 to 0.8	2	4	1	1	1	6	3	7
0.6 to 0.7	12	9	6	2	5	18	0	3
0.5 to 0.6	35	19	6	17	23	61	17	17
0.4 to 0.5	112	75	33	32	92	149	39	61
0.3 to 0.4	381	198	187	150	337	347	183	280
0.2 to 0.3	1274	800	1178	934	677	882	851	977
0.1 to 0.2	2263	2275	3314	2668	1682	1518	2429	2281
0.0 to 0.1	1880	1994	2797	2929	2321	1811	3264	2824
-0.1 to 0.0	773	1185	885	1500	1246	1347	1463	1465
-0.2 to -0.1	166	312	143	298	432	560	276	486
-0.3 to -0.2	34	53	20	29	94	189	36	148
-0.4 to -0.3	5	13	0	11	24	41	11	20
-0.5 to -0.4	1	0	1	0	5	8	0	2
under -0.5	1	1	1	1	0	0	0	0

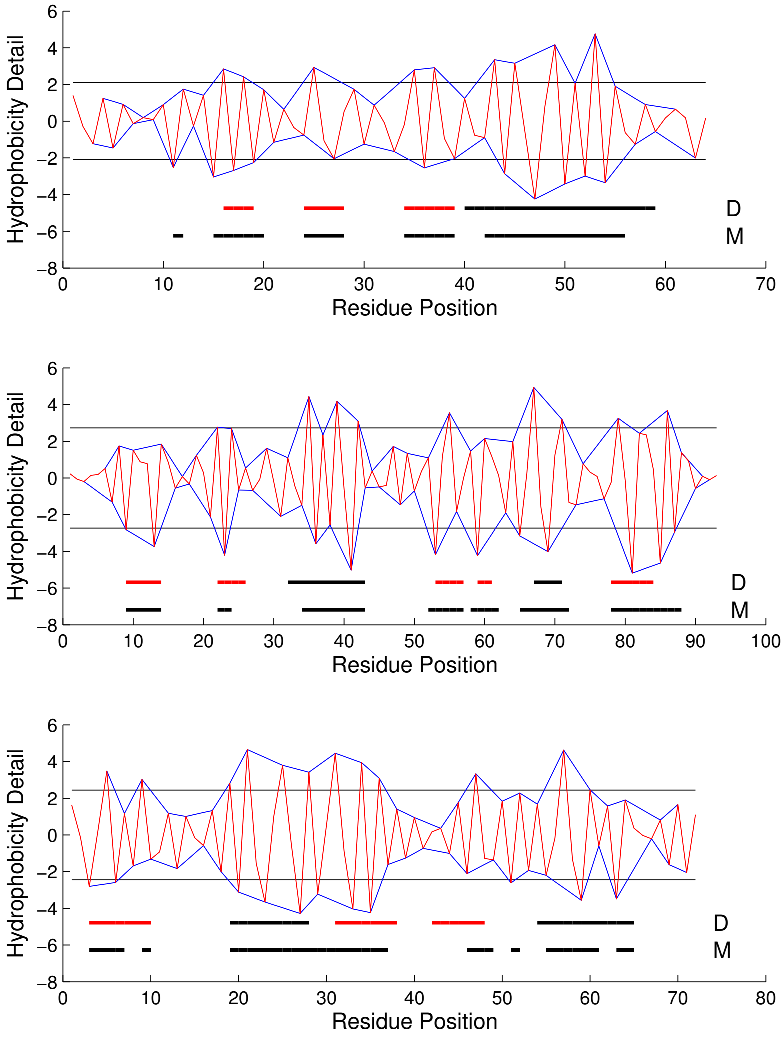


Figure 4.6: Three examples of proteins from the $\alpha + \beta$ database with high values of MCC when evaluated against helices and strands (not distinguishing the two). The top shows protein 1KJK (PDB ID) for which $Q = 0.875$ and $MCC = 0.751$ when the threshold is set to 1.00σ. The middle protein is 1WGD for which $Q = 0.849$ and $MCC = 0.698$ when the threshold is set to 1.30σ. The bottom shows protein 1AFJ for which $Q = 0.792$ and $MCC = 0.590$ when the threshold is set to 1.20σ. For all three plots the KD scale is used and $j = j_{max} - 1$.

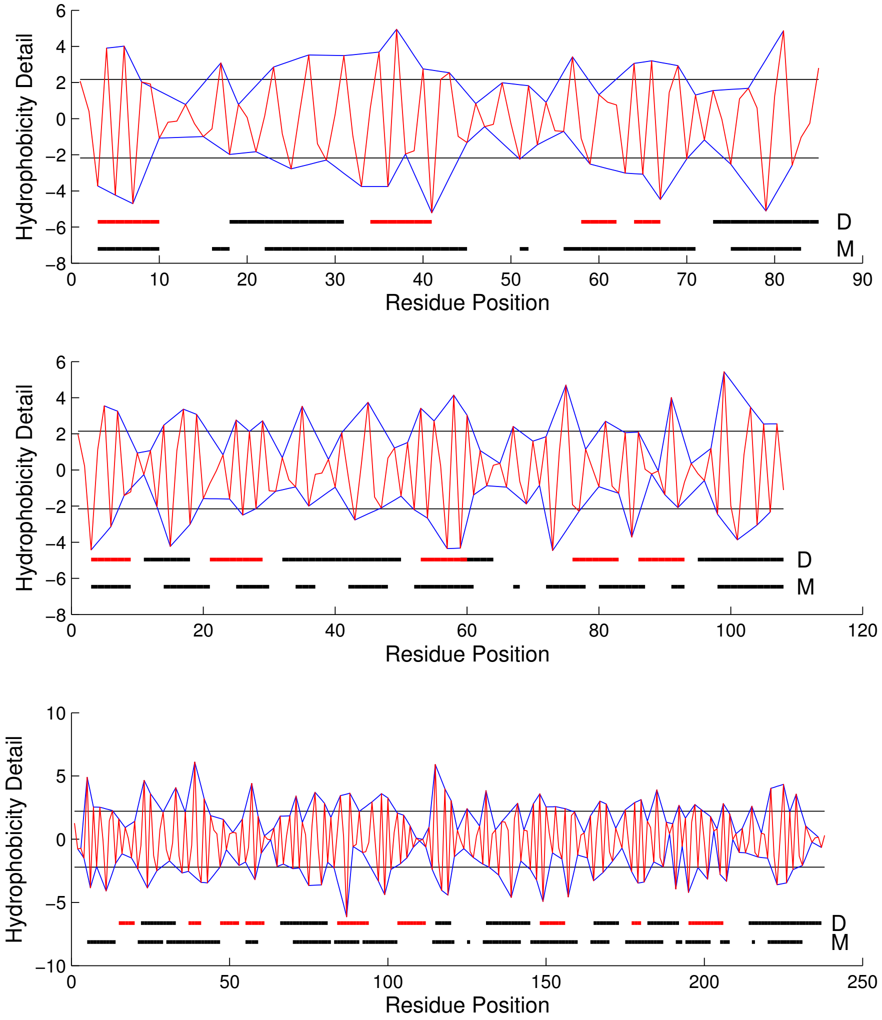


Figure 4.7: Three examples of proteins from the α/β database with moderate to low values of MCC when evaluated against helices and strands (not distinguishing the two). The top shows protein 1FO5 for which $Q = 0.694$ and $MCC = 0.383$ when the threshold is set to 1.00σ. The middle protein is 2TIR for which $Q = 0.620$ and $MCC = 0.198$ when the threshold is set to 1.00σ. The bottom shows protein 1ECP for which $Q = 0.567$ and $MCC = 0.081$ when the threshold is set to 1.05σ. For all three plots the KD scale is used and $j = j_{max} - 1$.

One last issue of interest is in whether there is anything that we can say is characteristically different about the proteins which yield good results with our technique and those that don't. In the introduction to this chapter we mentioned a theoretical basis for why fluctuations in hydrophobicity would tend to be found to occur in secondary structure residing near the protein globular surface. Working from this reasoning one might expect that our model might perform better on shorter proteins which should have a greater proportion of their chain exposed to the surface and thus more "detectable" secondary structure. To test this assumption we look at a scatter plot of MCC verses the protein chain length. This type of plot for each of the eight different tests we performed above looked qualitatively the same. Figure 4.8 shows the results for the $\alpha + \beta$ H+S evaluation at the first wavelet scale. While the data does show that our best performing predictions happen for the shortest proteins we also see that some of the worst performance also happens for short chains. The variation in the values of MCC essentially grows as we look at shorter chain proteins.

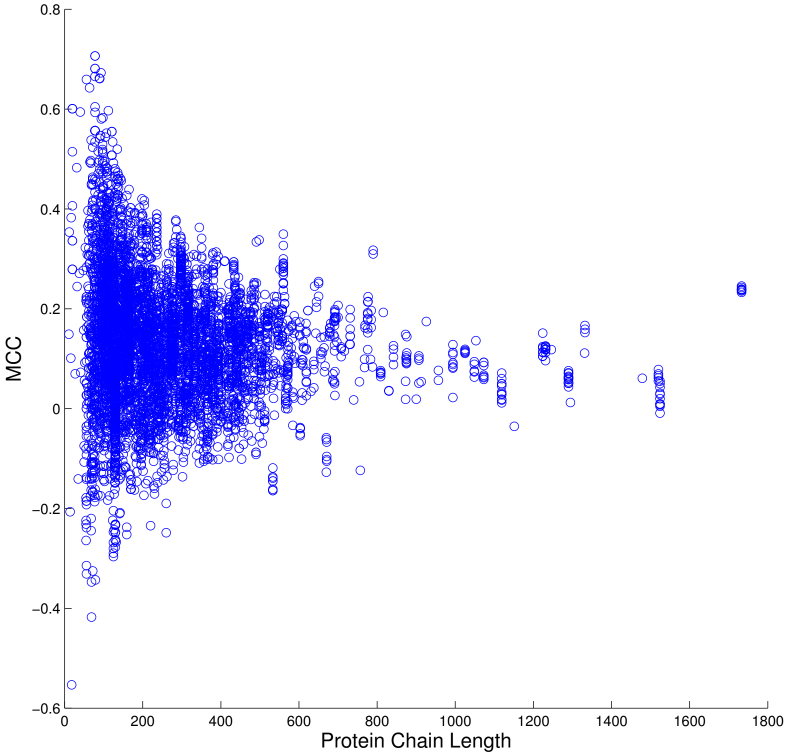


Figure 4.8: Scatter plot showing how the performance of our technique varied over proteins of different lengths. These are the results for the $\alpha + \beta$ proteins at scale $j = j_{max} - 1$ and hydrophobicity scale KD.

4.4 Scale-Scale Measure

When a wavelet decomposition is performed information is sorted according to scale and different scales are typically examined separately. In the previous section, for example, we analyzed the first and second wavelet scales separately. On average we found positive correlations for both of these scales. Another reasonable approach would be to consider the signal of a measure that combines two or more scales because a single scale by itself may miss structure that another scale captures. Here we construct a measure that combines two adjacent scales defined by

$$SS_j = \frac{D_{j,l}^2 D_{j-1,l}^2}{\langle D_{j,l}^2 \rangle \langle D_{j-1,l}^2 \rangle} \tag{4.7}$$

where $D_{j,l}$ is the reconstructed detail coefficient at scale j, l is the residue index number which goes from 1 to N, and the bracket notation "$\langle\rangle$" signifies an average over the entire signal. This measure will yield especially high significance in locations where there is high significance at both scales and especially low significance in locations in where both scales have low significance. Physically this measure may be giving us information about hierarchical structure, i.e., structure at one scale that is related to structure at another scale. As we did in the last section we again use the enveloping technique here to capture the regions of significance as the predictor of secondary structure. With the squaring of the coefficients we essentially mirror the signal onto the positive axis and only one threshold is necessary. The 1σ threshold is defined here as 1σ from the zero point (instead of from the average of the control data as we did in the previous section).

As before, we divided the proteins into groups of similar Prevalence and found the appropriate threshold to satisfy the condition $B/P \approx 1$. Tables 4.6 and 4.7 show the results for the $\alpha+\beta$ and α/β databases. These results are for evaluation against helices and strands only. Figure 4.9 shows the results of the technique on three individual proteins, of varying levels of Prevalence, with fairly good correlations. Figure 4.9 shows the results of the technique on three individual proteins with MCC values closer to the database average.

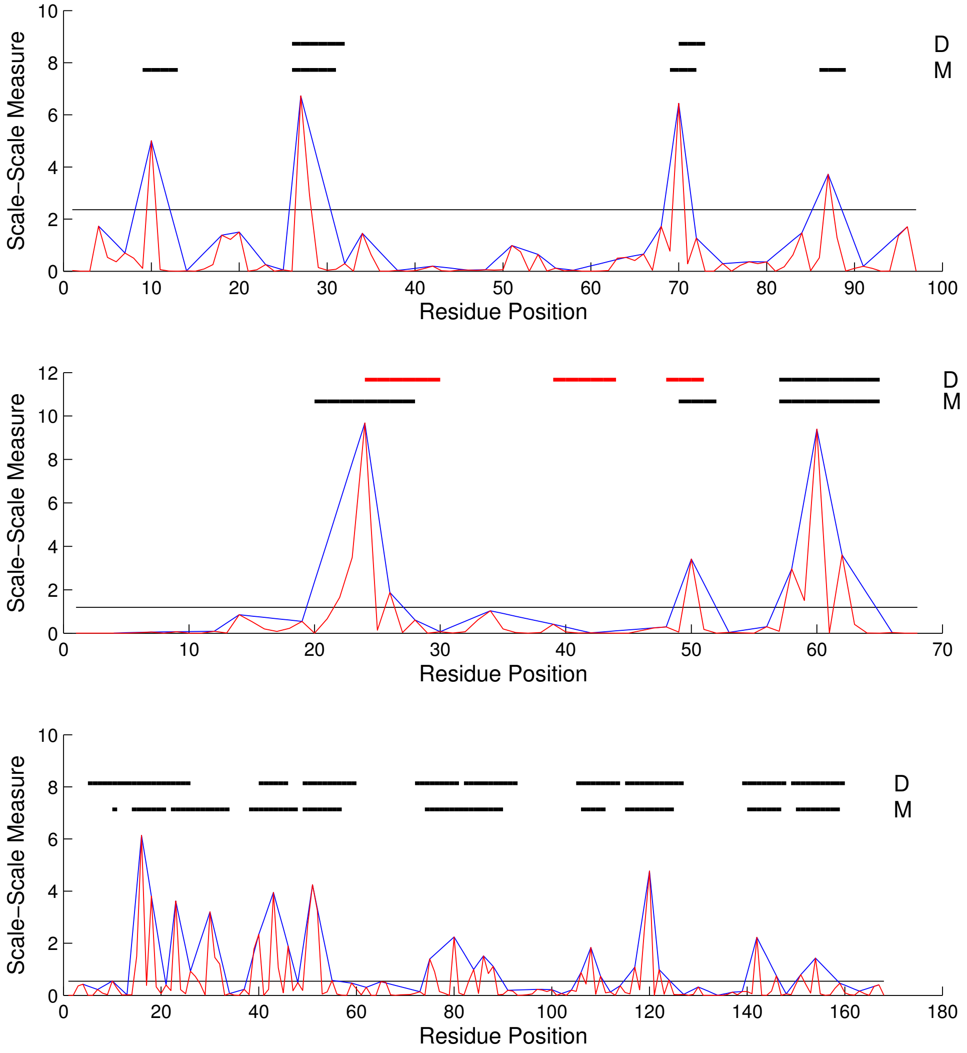


Figure 4.9: Results using the scale-scale measure on 3 proteins from the $\alpha + \beta$ database with significantly different Prevalence levels and good levels of correlation. Top: protein 1ROE for which $MCC = 0.55$. Middle: protein 1HRJ for which $MCC = 0.55$. Bottom: protein 1MX2 for which $MCC = 0.53$.

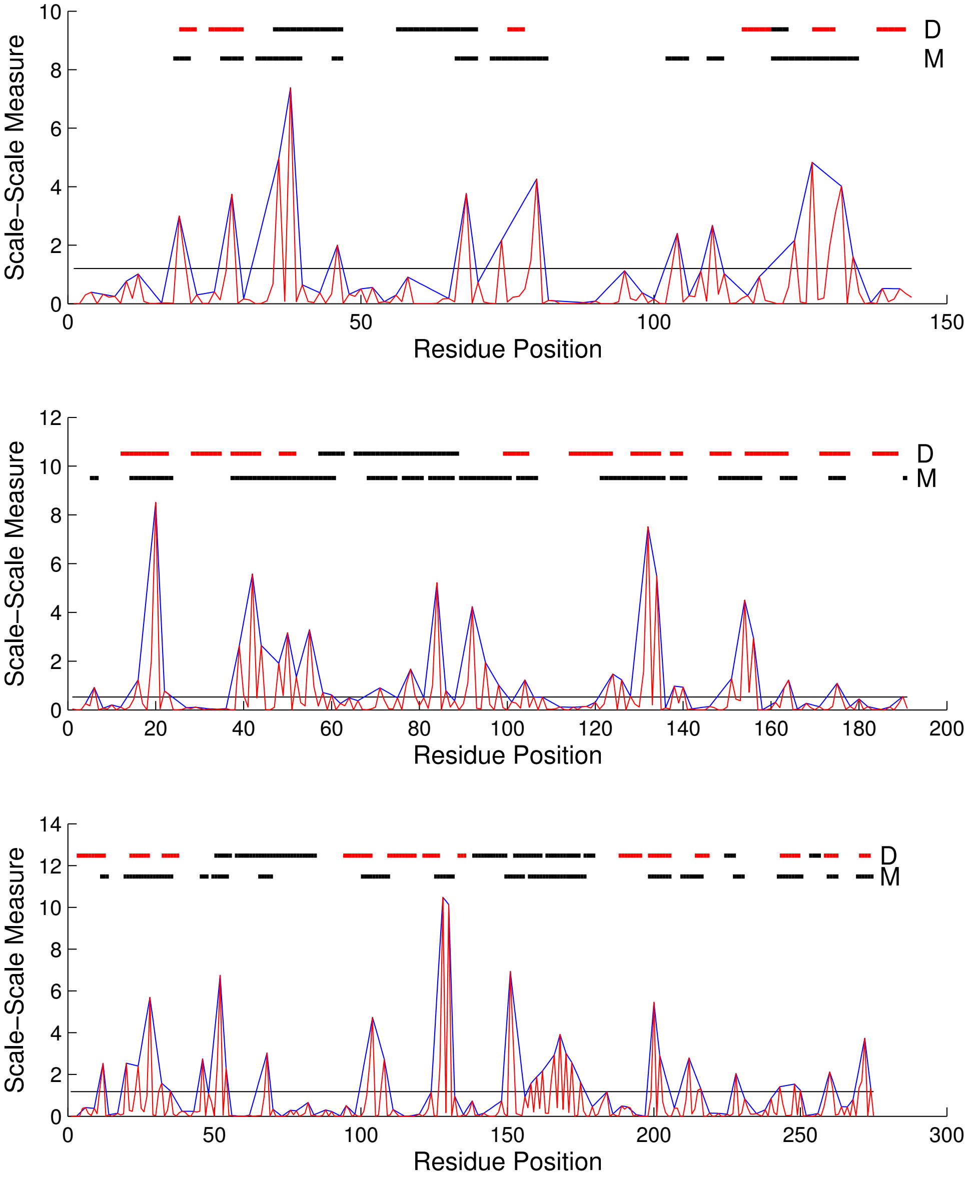


Figure 4.10: Results using the scale-scale measure on 3 proteins from the $\alpha + \beta$ database with correlations closer to the database average. Top: protein 1LIT for which $MCC = 0.20$. Middle: protein 1K8I for which $MCC = 0.18$. Bottom: protein 1AO7 for which $MCC = 0.12$.

Table 4.6: $\alpha + \beta$ database results for both helix and strand evaluation using the scale-scale measure at the first two wavelet scales. The KD hydrophobicity scale is used and proteins are grouped according to similar Prevalence so as to reduce biasing.

P	chains	t.h.	Q	MCC	B/P
$0.0 - 0.2$	24	2.39σ	0.761	0.081	0.973
$0.2 - 0.3$	59	1.52σ	0.647	0.0892	1.004
$0.3 - 0.4$	258	1.09σ	0.586	0.103	1.000
$0.4 - 0.5$	1082	0.77σ	0.563	0.124	0.991
$0.5 - 0.6$	2050	0.57σ	0.559	0.107	0.998
$0.6 - 0.7$	2379	0.42σ	0.589	0.107	0.994
$0.7 - 0.8$	876	0.27σ	0.6567	0.115	0.996
$0.8 - 1.0$	211	0.17σ	0.744	0.0658	0.993
Weighted Ave.			0.590	0.109	0.995

Table 4.7: α/β database results for both helix and strand evaluation using the scale-scale measure at the first two wavelet scales. The KD hydrophobicity scale is used and proteins are grouped according to similar Prevalence so as to reduce biasing.

P	chains	t.h.	Q	MCC	B/P
$0.0 - 0.2$	18	1.98σ	0.754	0.093	1.015
$0.2 - 0.3$	78	1.58σ	0.663	0.122	0.989
$0.3 - 0.4$	141	1.13σ	0.602	0.131	0.967
$0.4 - 0.5$	645	0.78σ	0.557	0.111	1.006
$0.5 - 0.6$	2645	0.57σ	0.557	0.101	1.000
$0.6 - 0.7$	3828	0.42σ	0.585	0.091	1.0013
$0.7 - 0.8$	1178	0.30σ	0.647	0.109	0.999
$0.8 - 1.0$	39	0.13σ	0.723	0.145	0.945
Weighted Ave.			0.585	0.099	1.000

Figure 4.11 shows the distributions of the MCC from Tables 4.6 and 4.7. The distributions show similar results to the single scale results. For the $\alpha + \beta$ proteins around 24% of the database have correlations greater than 0.2 and over 35% are higher than 0.15. The performance is slightly lower for the α/β proteins with around

16% of the database having correlations greater than 0.2 and over 30% higher than 0.15.

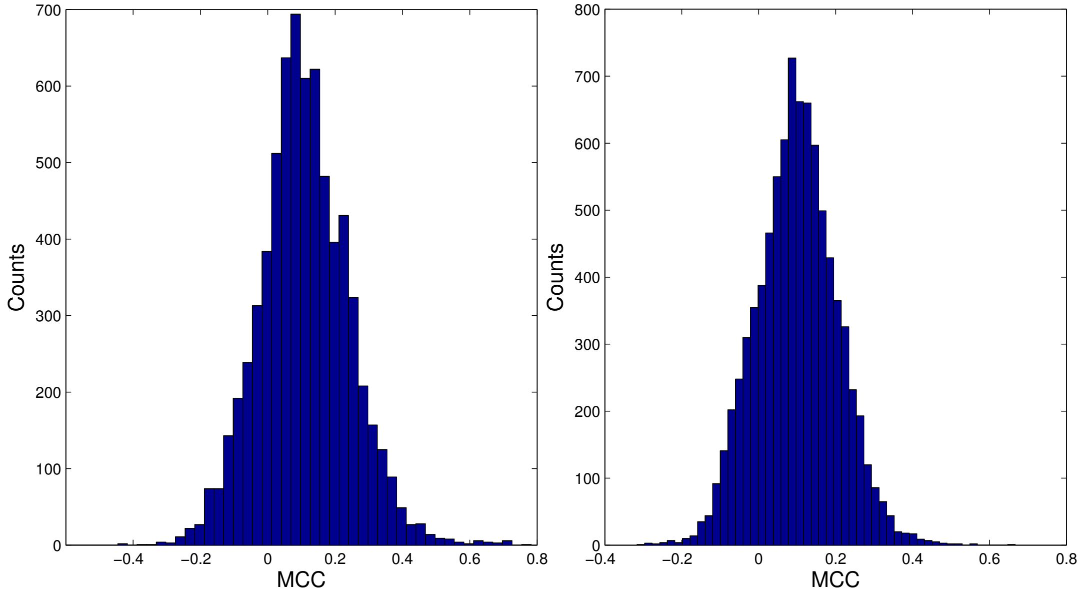


Figure 4.11: Distribution of MCC values for the H+S evaluation using the scale-scale measure. Left: the $\alpha + \beta$ database. Right: the α/β database.

In closing this chapter we emphasis again that this work is still in a preliminary stage. We expect that further work with these techniques will lead to improved results. Lastly, it may be beneficial to future work to focus on the analysis of a smaller data set.

CHAPTER 5

Conclusion

Protein secondary structure prediction is an extremely complex problem that is yet to be fully solved. This work is intended to establish a set of tools for use in the study of protein structure. The two protein classes that we studied were very diverse consisting of hundreds of different protein families. One thing that may benefit future research would be to focus on a smaller amount of data.

The approach that we took here utilized the discrete wavelet transform (DWT) to analyze the protein's hydrophobicity "signal." We tested the hypothesis that locations in the protein's sequence which exhibit significant wavelet detail (i.e., fluctuations in hydrophobicity) correspond to locations of secondary structure. For each protein we generated a control data set which consisted of 200 identically processed signals from permutations of the protein's primary sequence. The purpose of the control data was to produce a threshold for determining the significant regions of the real protein. Initially our aim was to use this approach to predict the secondary structure state of each residue in the protein chain. However, for reasons discussed below, the large data sets of 6939 $\alpha+\beta$ and 8572 α/β protein chains revealed that a universal threshold could not be found to give relatively "unbiased" predictions over all. For that reason our work shifted from a pure prediction to a project that used knowledge of the amount of secondary structure in order to adjust the threshold. This chapter will briefly summarize these and other issues from our analysis and suggest areas for future development.

5.1 Discussion of Results

There are many advantages that this approach has. One important advantage is that it incorporates physics into the prediction via the hydrophobic effect. This is a downside of many of the knowledge based methods which do not give physical intuition into why a certain structure forms. The other advantage our approach has over the knowledge based methods is that it does not require a set of learning proteins from which to gain information. While the "pure physics" based molecular dynamics methods are the ultimate aim they are difficult, computationally expensive, and at this time limited to short proteins. Our approach is capable of quickly handling large data sets and protein chains of any length. Lastly, representing the protein as a signal opens the door to a number of different statistical measures and signal analysis techniques.

Two important concepts that were discussed in chapter 4 with this type of approach are Prevalence P and Bias B (§ 4.2.1). Prevalence is the proportion of the protein chain length that consists of secondary structure – a property of the protein having nothing to do with the prediction. Bias on the other hand is the proportion of the chain length that is predicted to be in the secondary structure state and this is very much dependent on the height chosen for the threshold. Ideally a prediction model will be unbiased, which is to say that $B/P \approx 1$ and often models are parametrized in this way. Initially we postulated that a threshold in the 1σ range of the control data would be appropriate in terms of giving an unbiased prediction. What we found in our databases was that the proteins had a widely varying P ranging from less than 20% to over 80%. The 1σ threshold for proteins of low P would result in predictions of far too much structure (overbiased) whereas proteins of high P would result in the prediction of not enough structure (underbiased). Thus, we took a different route in our evaluation process. Instead of a pure prediction we used information about secondary structure Prevalence to divide the proteins up and evaluate our model on groups of proteins with similar P. For each group of proteins with similar P we carried out our performance evaluation using a threshold that would give the least amount of biasing for the whole group. While not a pure prediction, this is still a valid way to evaluate whether significant wavelet features correlate with secondary structure.

Another important issue that has not yet been resolved using our model is the ability to distinguish between helix and strand structures. However, there were many cases in which we found good correlation with both strands and helices. Thus in our analysis we performed different types of performance evaluation: we tested our prediction agains both structures without distinguishing the two, and we tested it against just helices. It may have been valuable to perform a strand only evaluation also but in all the proteins we individually inspected there seemed to be good correlation with helices.

We tested out three different hydrophobicity scales with our approach and found that the results amongst them didn't differ greatly. The wavelet decomposition seemed to reveal similar areas of significance for each of the scales. The Kyte-Doolittle (KD) scale on the average gave slightly better performance and for that reason we carried out most of our analysis using that scale.

Overall our results showed positive Matthews correlation coefficients (MCC) for the majority of the proteins in both data sets. The wavelet enveloping technique proved to be a convenient tool for identifying regions of secondary structure. We found that on average results didn't differ much between the $\alpha + \beta$ and α/β proteins. Furthermore, the average results didn't differ much for proteins of varying secondary structure Prevalence. For our analysis with the single wavelet scales we found our best results for helix and strand evaluation (H+S) at the first scale detail, $j = j_{max} - 1$. For the H+S evaluation the second wavelet scale, $j = j_{max} - 2$, gave noticeably worse results. Helix only evaluation (H) in general showed worse performance than the evaluation of both structures combined. While for all of this data the averages were low and positive the distributions were fairly broad. For example, approximately 25 % of the $\alpha + \beta$ proteins with the H+S evaluation at the first scale had MCC values greater or equal to 0.2 while the average was around 0.12. The Scale-Scale measure also gave on average positive correlations over both data sets. The averages were low, near 0.1, but again the distributions were broad with 24 % of the $\alpha + \beta$ and 16 % of the α/β proteins having MCC over 0.2. The Q accuracy ranged from about 60 to as high as 90 %. Given the size and diversity of the data set we analyzed these performance values indicate that this has promise.

5.2 Future Work

There are a number of areas of future work that may provide more information and better prediction results. First, it may be more beneficial to study a smaller data set of similar proteins (perhaps a group of proteins from within a SCOP family or superfamily) and do more visual analysis of individual proteins. Much further work can be done in the localized analysis of secondary structure detection. Here we've looked at the first and second scales of the wavelet detail and a measure that combines these scales. There may be a more suitable measure and perhaps information from the wavelet approximations may be important. Furthermore, these techniques may provide information on protein tertiary structure. One thing that we did not cover here is the use of the DWT to uncover global information about the proteins. With wavelet analysis one can look at a power spectrum which is analogous to the Fourier power spectrum. Through such global techniques it may be possible to reveal global truths such as protein classification information or secondary structure Prevalence.

5.3 Summary

The results that we found support the conclusion that wavelet analysis of the hydrophobicity signal of protein reveals secondary structure information. With that being said there may be limitations to the general approach in what it can detect. Secondary structure near the globular core may not exhibit significant fluctuations in hydrophobicity at any scale and therefore could be invisible with this type of analysis. This work has laid the foundation for a potentially rich analysis paradigm. The positive correlations we see in detecting secondary structures over as diverse a database as the SCOP $\alpha + \beta$ and α/β data sets is strong evidence that there is a great deal of validity in this approach. We trust future work will bear this out.

CHAPTER 6

Acknowledgements

First and foremost I would like to thank my advisor Dr. Jesus Pando for all his help, guidance, and the many hours of editing required of my writing. I would also like to thank the DePaul physics department for making this project possible. I thank my parents for the encouragement, the meals, and the roof over my head. And last but certainly not least I thank God for giving me the focus when I needed it most.

APPENDIX A

Success Measurement Program

The following Matlab function was used for evaluating the performance of a two-state (1's and 0's) prediction vector **M** against the two-state known structure vector **D**. Where m_i and d_i are both 1 a True Positive is counted and where both 0 a True Negative. Where m_i is 1 and d_i is 0 a False Positive is counted and for the reverse case a False Negative is counted. The algorithm uses these four numbers to calculate a number of performance measures such as the Matthews correlation coefficient (MCC), Q accuracy, Bias B, and Prevalence P.

```
function [Sn,Sp,Q,MCC,B,P] = PerResEvaluation(D,M)

% PERRESEVALUATION evaluates performance of a binary prediction M given the
% actual data D.
%       Inputs:
%           D       binary array of secondary structure data where 1
%                   represents a structure and 0 represents non—structure.
%           M       binary array of predicted structure where 1 represents
%                   the prediction of a structure and 0 the prediction of a
%                   non—structure.
%       Outputs:
%           Sn      sensitivity
%           Sp      specificity
%           Q       accuracy
%           MCC     matthew's correlation coefficient
%           B       bias
%           P       prevalence
% Written by Tim Vanderleest
% Last updated 6/19/11

if length(D) ≠ length(M)
    error('Two input data should have the same length.');
end

N = length(D);  % The number of prediction cases

% True positives (Tp), true negatives (Tn), false positives (Fp), and false
% negatives (Fn) initialized to zero:
Tp=0; Tn=0; Fp=0; Fn=0;
% Loop over the each case
for i=1:N
    if D(i)
        if M(i)
            Tp = Tp + 1;
        else
            Fn = Fn + 1;
        end
    else
        if M(i)
            Fp = Fp + 1;
        else
            Tn = Tn + 1;
        end
    end
end
%% * Measures
Sn = Tp/(Tp + Fn);          % Sensitivity
Sp = Tp/(Tp + Fp);          % Specificity
% Matthews Correlation Coefficient
denom = sqrt((Tn + Fn)*(Tn + Fp)*(Tp + Fn)*(Tp + Fp));
```

```
if denom == 0
    MCC = (Tp*Tn - Fp*Fn);
else
    MCC = (Tp*Tn - Fp*Fn)/denom;
end
Q = (Tp + Tn)/N;            % Q statistic
P = (Tp + Fn)/N;            % Prevalence
B = (Tp + Fp)/N;            % Bias
end
```

APPENDIX B

Wavelet Enveloping Prediction Program

The following Matlab function performs the wavelet enveloping technique on a single protein and compares it to a set of equivalently processed randomized sequences in order to predict secondary structure. This function does not convert the protein sequence to a hydrophobicity signal nor does it extract secondary structure data directly from PDB files. These steps were preprocessed and saved into a convenient format to be imported into this function. The amino acid hydrophobicity numbers were saved to a data file in columns starting with the N-terminal residue. If there were multiple chains each unique chain would be stored its own column. The secondary structure that was extracted from the PDB was also stored in a special format that this function was designed to read. The secondary structure data was stored in 4 columns of integer values. Each line corresponded to a secondary structure element. The first column identifies whether the structure is a helix or a strand and which chain the structure belongs to. A number 1 indicates a helix from the first chain, 2 indicates a helix from the second chain, etc. Likewise a number 51 indicates a strand from the first chain, 52 a strand from the second chain and so forth (we did find any proteins with more than 50 unique chains). The second column indicated the length of the chain. The third and fourth columns indicated the starting and ending positions of the secondary structure element. This information may make reading the code easier. For plotting the secondary structure we created a separate function which is included in Appendix C.

```
function wenvsingle(hscale,pdbID,chainnum,ssc,tholdfactor)
```

```
% WENVSINGLE compares the real protein wavelet envelope to random
% realizations for evaluating the prediction of secondary structures on a
% single protein chain.
%
%   Inputs:
%       hscale      hydrophobicity scale: 'KD', 'ES', or 'HW'.
%       pdbID       PDB Id number (e.g. '9RSA').
%       chainnum    chain number of multichain protein, e.g. 1, 2, etc.
%       ssc         choice of secondary structure to evaluate:
%                   1 for helix, 2 for strand, and 3 for both.
%       tholdfactor factor for varying the height of the threshold
%
%   Outputs: (Plot and the following values are written to screen)
%       Sn          sensitivity
%       Sp          specificity
%       Q           accuracy
%       MCC         matthew's correlation coefficient
%       B/P         bias/prevalence
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Written by Timothy Vanderleest
% Last updated 7/10/11

%% * Parameters
numrealz = 200; % Number of realizations
wtype = 'db2'; % Wavelet type set to Daubechies 4 Tap wavelets
%% * Import both hydrophobicity signal and secondary structure data
hydrodata = importdata(strcat('../KDhydroslash/',hscale,pdbID,'.dat'));
ssdata = importdata(strcat('../KDhydroslash/','SS',pdbID,'.dat'));
%% * Gather SS (secondary structure) data and the chain length
hnum = histc(ssdata(:,1),chainnum); % # of helix structures in chain.
snum = histc(ssdata(:,1),chainnum + 50);% # of strand structures in chain.
helix = zeros(hnum,2); strand = zeros(snum,2); % Construct SS arrays.
if hnum == 0 && snum == 0,return;end % If no SS in chain exit function.
```

```
% The following loop gathers the length of the chain, and the starting and
% ending points of helices and strands from the ssdata array.
j=1; k=1; lenflag = true;
for i=1:size(ssdata,1)
    if ssdata(i,1)==chainnum
        if lenflag, lenres = ssdata(i,2); lenflag = false;end
        helix(j,1)=ssdata(i,3); helix(j,2)=ssdata(i,4); j=j+1;
    end
    if ssdata(i,1) == chainnum + 50
        if lenflag, lenres = ssdata(i,2); lenflag = false;end
        strand(k,1)=ssdata(i,3); strand(k,2)=ssdata(i,4); k=k+1;
    end
end
%% * Construct D matrix (3 rows corresponding to 3 separate D's)
% The 1st row is helices only, 2nd is strands only, and the 3rd is both.
D = zeros(3,lenres);
for k=1:hnum
    for i=helix(k,1):helix(k,2)
        D(1,i) = true;
    end
end
D(3,:) = D(1,:);
for k=1:snum
    for i=strand(k,1):strand(k,2)
        D(2,i) = true;
        D(3,i) = true;
    end
end
%% * Process real chain data
hydrochain = hydrodata(:,chainnum)';    % Selects chain given by chainnum.
hydrochain(lenres+1:end) = [];          % Remove excess zeros.
[C_real,L_real]=wavedec(hydrochain,4,wtype); % Decompose into 4 scales.
real1 = wrcoef('d',C_real,L_real,wtype,1); % First 4 scales of the detail
real2 = wrcoef('d',C_real,L_real,wtype,2); % reconstructed. Note, not all
```

```
real3 = wrcoef('d',C_real,L_real,wtype,3); % of these scales were used.
real4 = wrcoef('d',C_real,L_real,wtype,4); %
realmeas = real1;    % realmeas is the measure of the detail we analyze
                     % (in this case just the first scale detail).
r = 1:lenres;
[up,down] = envelope(r,realmeas);
% The above envelope function creates both uppper and lower envelope of a
% fluctuating signal. Function written by Lei Wang and downloaded from:
% http://www.mathworks.com/matlabcentral/fileexchange/3142-envelope1-1
%% * Process random chain data with 'numrealz' realizations
controldata = zeros(numrealz,lenres);    % control data matrix for storing
                                         % each realization.
realz=zeros(1,lenres);  % initialize single realization array.
for k=1:numrealz
    n=randperm(lenres); % n is a random permutation array of indices
    for i=1:lenres
        realz(i)=hydrochain(n(i));
    end
    [C_rand,L_rand]=wavedec(realz,4,wtype);
    rand1 = wrcoef('d',C_rand,L_rand,wtype,1);
    rand2 = wrcoef('d',C_rand,L_rand,wtype,2);
    rand3 = wrcoef('d',C_rand,L_rand,wtype,3);
    rand4 = wrcoef('d',C_rand,L_rand,wtype,4);
    randmeas = rand1;
    controldata(k,:) = randmeas;
end
%% * Statistics of random data for Threshold and determination of M
rup = mean(controldata) + tholdfactor*std(controldata);
rdown = mean(controldata) - tholdfactor*std(controldata);
rup = mean(rup)*ones(1,lenres); % create globally flat upper threshold.
rdown = mean(rdown)*ones(1,lenres); % create globally flat lower threshold.
M = zeros(1,lenres);
for i=1:lenres
    if up(i) > rup(i) || down(i) < rdown(i)
```

```
        M(i) = true; % If over threshold positive prediction
    else
        M(i) = false; % If under threshold negative prediction
    end
end
[Sn,Sp,Q,MCC,B,P] = PerResEvaluation(D(ssc,:),M); % evaluation
fprintf('Sn= %f, Sp= %f, Q= %f, MCC= %f, B/P = %f \n',Sn,Sp,Q,MCC,B/P);

%% * Plotting
hold on
plot(r,rup,'k')         % Plot upper threshold from control data (black).
plot(r,rdown,'k')       % Plot lower threshold from control data (black).
plot(r,realmeas,'r')    % Plot detail from real protein (red).
plot(r,up,'b')          % Plot upper envelope from real protein (blue).
plot(r,down,'b')        % Plot lower envelope from real protein (blue).
ssplotter(D(1,:),min(down)-.5,'k',2,'D')    % Plot D helix segments (black)
ssplotter(D(2,:),min(down)-.5,'r',2,'none') % Plot D strand segments (red)
ssplotter(M,min(down)-2,'k',2,'M')          % Plot prediction M (black)
xlabel('Residue Position','FontSize',12)
ylabel('Hydrophobicity Detail','FontSize',12)
hold off
end
```

APPENDIX C

Secondary Structure Plotting Program

This short function is used for plotting secondary structure and is used by the Matlab program listed in Appendix B.

```
function ssplotter(M,height,color,width,name)
% SSPLOTTER plots the secondary structure (can be used
% to plot either M or D).
%   Inputs:
%       M   Array of predictions
%       height      plotting height (location on the graph).
%       color       plotting color, e.g. 'r' (red), 'b' (blue), etc.
%       width       plotting line width.
%       name        name the secondary structure (or not by 'none').


if name == 'none'
    noname = true;
else
    noname = false;
end

for i = 1:length(M)
    if M(i) % for each residue where M(i) is true we plot a horizontal line.
        plot([i i+1],[height height],color,'Linewidth',width)
    end
end
```

```
if ¬ noname
    text(length(M)+2,height,name,'FontSize',12)
end


end
```

REFERENCES

[1] D.P. Lane. Cancer - p53, guardian of the genome. *Nature*, 358:15–16, 1992.

[2] M. Hollstein, D. Sidransky, B. Vogelstein, and C.C. Harris. p53 mutations in human cancers. *Science, New Series*, 253:49–53, 1991.

[3] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(96):223–230, 1973.

[4] A.L. Lehninger, M.M. Cox, and D.L. Nelson. *Principles of Biochemistry.* W.H. Freeman, New York, 2004.

[5] J.M. Berg, J.L. Tymoczko, and L. Stryer. *Biochemistry. 5th edition.* W.H. Freeman, New York, 2002.

[6] C. Branden and J. Tooze. *Introduction to Protein Structure.* Garland Publishing, Inc., New York and London, 1991.

[7] L. Pauling, R.B. Corey, and H.R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 37, 1951.

[8] W.T. Astbury and A. Street. X-ray analysis of the structure of hair, wool and related fibres. 1. general. *Trans. R. Soc. Lond.*, A230, 1931.

[9] L. Pauling and R.B. Corey. The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci. USA*, 37, 1951.

[10] E. Shakhnovich. Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *Chemical Reviews*, 106(5):1559–1588, 2006.

[11] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157, 1982.

[12] T. P. Hopp and K. R. Woods. A computer program for predicting protein antigenic determinants. *Molecular Immunology*, 20, 1983.

[13] D. M. Engelman, T. A. Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, 15, 1986.

[14] H.A. Scheraga, M. Khalili, and A. Liwo. Protein-folding dynamics: Overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.*, 58:57–83, 2007.

[15] P.L. Freddolino, C.B. Harrison, Y. Liu, and K. Schulten. Challenges in protein-folding simulations. *Nature Physics*, 6, 2010.

[16] K.A. Dill, S.B. Ozkan, T.R. Weikl, J.D. Chodera, and V.A. Voelz. The protein folding problem: when will it be solved? *Current Opinion in Structural Biology*, 17:342–346, 2007.

[17] P.Y. Chou and G.D. Fasman. Prediction of protein conformation. *Biochemistry*, 13(2), 1974.

[18] J. Garnier, D.J. Osguthorpe, and B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, 120(1), 1978.

[19] J. Cheng, A.N. Tegge, and P. Baldi. Machine learning methods for protein structure prediction. *IEEE Reviews In Biomedical Engineering*, 1, 2008.

[20] J. Moult, J.T. Pedersen, R. Judson, and K. Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins*, 23:ii–iv, 1995.

[21] W. Friedrich, P. Knipping, and M. von Laue. Concerning the detection of x-ray interferences. *Nobel Lecture, Physics*, 1915.

[22] W.L. Bragg. The diffraction of x-rays by crystals. *Nobel Lecture, Physics*, 1922.

[23] D. Whitford. *Introduction to Protein Science*. Oxford, New York, 2004.

[24] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. Scop: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[25] Andreas Baxevanis and B.F. Francis Ouellette, editors. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Hoboken, 2005.

[26] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.

[27] A. Haar. Zur theorie der orthogonalen funktionensysteme. *Math. Ann.*, 69, 1910.

[28] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in Fortran: The Art of Scientific Computing*. Cambridge University Press, second edition, 1992.

[29] L. Pattini and S. Cerutti. Hydrophobicity analysis of protein primary structures to identify helical regions. *Methods of Information in Medicine*, 43, 2004.

[30] Z.N. Wen, K.L. Wang, M.L. Li, F.S. Nie, and Y. Yang. Analyzing functional similarity of protein sequences with discrete wavelet transform. *Comptuational Biology and Chemistry*, 29, 2005.

[31] A. Giuliani, R. Benigni, J.P. Abilut, C.L. Webber, P. Sirabella, and A. Colosimo. Nonlinear signal analysis methods in the elucidation of protein sequence-structure relationships. *Chem. Rev.*, 102, 2002.

[32] J. Pando, L. Sands, and S.E. Shaheen. Detection of protein secondary structures vie the discrete wavelet transform. *Physical Review*, 80, 2009.

[33] J.M. Bujnicki. *Prediction of Protein Structures, Functions, and Interactions.* John Wiley and Sons, West Sussex, 2009.