University of Alberta

Towards Improving Accuracy of Protein Content Prediction for Low Homology Sequences

by

Leila Homaeian ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Electrical and Computer Engineering

Edmonton, Alberta
Fall 2006

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-22285-0
Our file Notre référence
ISBN: 978-0-494-22285-0

Abstract

One of the main challenges in biological research is to determine protein conformation (3-D structure) and function(s), which in turn has several applications in novel drug discovery and disease treatment. Proteins have a lower level structure, called secondary structure, which contributes to protein conformation. The focus of this research is to predict the content of the secondary structure of protein sequences. Protein secondary structure content prediction can be utilized to predict protein conformation. Current alignment based approaches are applicable to the content prediction problem in case high homology is present between proteins with unknown structure and those with known structure. At the same time, growing number of proteins with unknown structure versus proteins with known structure becomes one of the main challenges facing the content prediction problem. In the course of this thesis, novel approaches with respect to feature space and prediction method are proposed to improve the accuracy of protein secondary structure content prediction for low homology protein sequences.

Acknowledgments

I would like to express my deep gratitude to my supervisor Dr. Lukasz Kurgan. Without his invaluable support and insight, completion of M.Sc. degree would have been impossible for me. Also, I would like to express my grateful appreciation to my dear husband, Reza. His unsparing support made every step throughout my M.Sc. program easy. I would like to thank our research group members too, especially Kanaka Kedarisetti, Mandana Rahbari, Rafal Rak, and Wojciech Stach. Their discussions provided me with useful feedback. Also, I would like to thank the members of my thesis committee. A special thanks to Dr. Carol A. Boliek whose warm friendship inspired me to finish my degree. I would like to thank Dr. Davood Rafiei as well. Getting involved in his research project motivated me to pursue my M.Sc. degree.

Dedication

I would like to dedicate this thesis to my dear parents whose deep unconditional love always inspires me to explore the unknown.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Expansion	**Abbreviation**
Molecular Weight	*MolW*
Isoelectric Point	*pI*
Composition Vector	*CV*
Composition Moment Vector	*CMV*
R-Groups	*RG*
Exchange Groups	*XG*
Hydrophobicity Groups	*HG*
Electronic Groups	*EG*
Chemical Groups	*CG*
Other Groups	*OG*
Fauchere Hydrophobicity Index	*FH*
Eisenhaber. Hydrophobicity Index	*EH*
Average Side Chain Mass	*M*
Hydropathy Index	*Hp*
Amino Acid	AA
Multiple Linear Regression	MLR
Neural Network	NN
10-Fold Cross Validation	10CV

Chapter 1 Introduction

Proteins are the primary molecules that contribute to cell structures. They are involved in all cellular activities [20]. One of the major challenges in biological research is to determine the conformation or three-dimensional form of a protein, since a protein's functions and properties are direct results of its conformation. The three-dimensional structure prediction is fundamental in research for novel drug discovery and disease treatment. There are sophisticated experimental methods to determine the conformation of a protein [20] such as X-Ray Crystallography and Nuclear Magnetic Resonance (NMR) Spectroscopy. However, these methods are expensive, labor intensive and in some cases, they take too much time [7]. Moreover they require complex analysis and they cannot be applied to any protein. For example, for some proteins a time-consuming trial and error process is needed to crystallize the proteins and it could take decades or even years [20]. NMR analysis is only applicable to short proteins with less than 200 amino acids (for information on protein length see Chapter 2). Therefore there has been an increasing interest in the computational approaches for the protein structure prediction [31].

Another issue that makes protein structure prediction more challenging is the growing number of proteins with unknown structure versus the ones for which the structure is known. Therefore the chance of having proteins with unknown structures that are not similar to proteins with known structure, increases. This questions the functionality of the many alignment-based approaches, which attempt to find proteins homologous (sequence homology is to be defined in section 2.3) to a query protein to determine its structure. The

NCBI[1] Proteins database contains over 2.2 million proteins. The SWISS-PROT[2] has over 200,000 sequence entries. PDB[3], which is the only database that includes three-dimensional protein structures, contains only about 33,000 proteins. On average about 75,000 new proteins were added to NCBI last year, among those the structure for about 500 proteins were inserted into PDB.

Proteins have a lower level of structure, referred to as *secondary structure,* which mainly contributes to proteins' conformation. Secondary structure consists of three major components: *Helix*, *Strand* and *Coil. Protein secondary structure prediction* is an intermediate step towards protein tertiary structure prediction. This project specifically aims at predicting the amount of secondary structure components or, more specifically, *secondary structure content* of a given protein. Since the content of each secondary structure component is a real number between zero and one, this research focuses on 'prediction' rather than 'classification'. There are two main computational prediction methods which were used in the past with respect to the protein secondary structure content prediction task: Neural Networks and Regression [15] [24] [26] [30] [33] [42] [43] [44]. Neural Networks are undesirable due to the complexity of the model, which results in long learning time, and lack of interpretability. In this project novel prediction methods that aim at improving the accuracy of regression-based prediction are investigated. To this end, four goals are followed:

1. The first goal is to investigate whether specialized (with respect to specific subsets of data) regression-based methods can improve the protein secondary structure content results. In this case the general set of proteins is divided into a number of

[1] National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/
[2] Protein Knowledgebase http://us.expasy.org/sprot/
[3] Protein Data Bank http://www.rcsb.org/pdb/

subsets based on different criteria, and a regression-based model is derived for each subset.

2. A protein needs to be encoded as a fixed-length vector to be fed into the content prediction method. In this project an aggregation of sequence representation methods used in related works is employed. The second goal is to perform feature selection and investigate the possibility of improving the content prediction accuracy.
3. The third goal is to design a new protein sequence representation that provides more comprehensive information about each protein. This will improve the accuracy of protein secondary structure content prediction.
4. Finally the fourth goal is to investigate how different base functions might change the result of regression-based approaches. Quadratic, cubic, Fourier transform coefficients, and exponential functions are among those tested.

We note that prior studies in content prediction did not address the above directions, i.e. splitting the protein data, performing feature selection, improving sequence representation and investigating different regression base functions.

A low-homology subset of Protein Data Bank called PDBSelect25[4] [11] is used as a dataset to test the proposed approaches in each goal. The low-homology means that proteins in this set are dissimilar which makes the prediction task more challenging. PDBSelect25 contains over 2,000 proteins and the homology level among the protein sequences is less than 30% (for more information see section 2.3).

The rest of the thesis is organized as the follows. Chapter 2 presents the background knowledge on proteins and their structure. It also describes the sequence representation.

[4] http://bioinfo.tg.fh-giessen.de/pdbselect/

Chapter 3 reviews the related work. Formal definition of the problem addressed in this thesis is given in Chapter 4. Chapter 5 introduces datasets and data preprocessing procedure. Chapter 6 presents a detailed definition of project goals and discusses the results achieved for each goal. Moreover, it presents an experimental comparison between our approach and state-of-the-art techniques. Chapter 7 summarizes the thesis and describes directions for future work.

Chapter 2 Protein Structure

In order to derive a protein secondary structure content prediction model, it is important to first understand protein structure. This chapter provides the necessary background knowledge and is organized as follows. First a short description of protein building blocks is presented, and the hierarchical structure of proteins is introduced. Next, the notion of homologous proteins is defined. Finally the protein sequence representation, which is utilized in this project, is described.

2.1 Amino Acids

There are 20 *Amino Acids (AA),* which are the basic structural building units of proteins. Table A 1 lists the AAs along with their 1 and 3-letter representations. All AAs (also called residues) share the same general structure, as depicted in Figure 2.1[5], but their side chains (also called *R* groups) are different, which gives each AA its unique set of chemical properties. The α carbon atom (C_α) of AAs, which is adjacent to the carboxyl group (*COOH*), is bonded to four different chemical groups: an amino (NH_2) group, a carboxyl (*COOH*) group, a hydrogen (*H*) atom, and a variable *R* group.

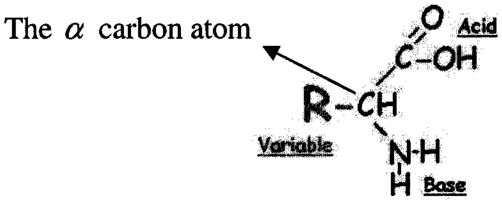

Figure 2.1. Amino acids

[5] Picture Source: http://www.rothamsted.bbsrc.ac.uk/notebook/courses/guide/aa.htm

The peptide bond is formed by a reaction between the amino group of one AA and the carboxyl group of anther AA (Figure 2.2)[5]. A short chain of AAs (at most 20-30 ones) linked by peptide bonds is called *peptide*. *Polypeptides* are longer than peptides and they can have as many as 4,000 AAs. A polypeptide folded into a three- dimensional molecule is referred to as a *protein*.

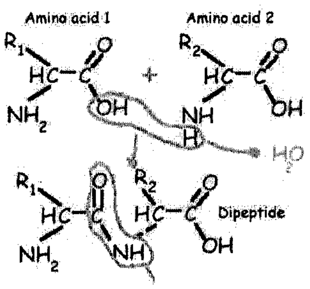

Figure 2.2. Peptide bond

2.2 Hierarchical Structure of Proteins

Four hierarchical levels of structure usually describe the shape of proteins. The first level is called the *primary structure* that is the linear sequence of AAs connected through peptide bonds, e.g. *PCSAFEFHCLSGECIHSSWRCDGGPDCKDKSDEENCA*. The length of a protein sequence is defined as the number of AAs composing its primary structure. The second level of protein structure is the *secondary structure*. Along a protein sequence, some AAs interact with each other locally and form different spatial arrangements. The Dictionary of Secondary Structures of Proteins annotates each AA as belonging to one of eight secondary structure types [33]: H (α Helix), G (3-Helix or 3_{10} Helix), I (5-Helix or π-Helix), B (residue in isolated β-bridge), E (Extended Strand), T (Hydrogen Bond Turn), S (Bend), and "_" (any other structure). Typically the above eight secondary structure types are reduced to three groups [6]: helix (which includes types H, G and I), strand (which includes types E and B), and coil (which includes T, S and the others). Helix is represented as *H* and has a spiral shape. Strand is represented as *E* and is plain-shaped, and coil is represented as *C* and has any shape other than that of helix and strand. The coil structure serves as a 'connector' between helix and strand structures. An example protein sequence and its secondary structure are shown below:

PCSAFEFHCLSGECIHSSWRCDGGPDCKDKSDEENCA
CCCCCCEECCCCCEECHHHCCCCCCCCCCCHHHCCCC

Helices and strands are arranged when stable hydrogen bonds are formed between AAs. These secondary structure components, i.e. helices and strands are periodical. In contrast coils are irregular. Figure 2.3[6] shows an example of conformation of the three secondary structure components.

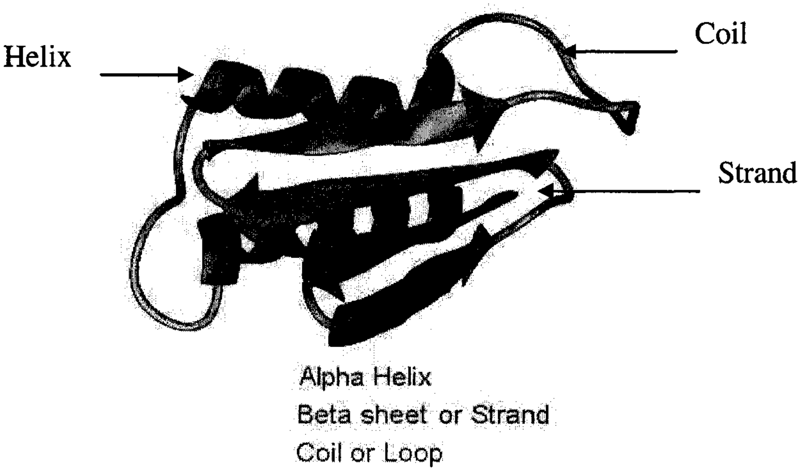

Figure 2.3. Secondary structure components

[6] Picture Source: http://xray.bmc.uu.se/Courses/PT/Project/Projects2002/Victoria_final_Figures.htm

Helical Conformation

An α-helix is formed when a polypeptide chain arranges into a regular spiral or helical conformation [20]. In this structure the peptide bonds are formed between carbonyl oxygen or the $C = O$ of nth residue and the amid hydrogen or *N-H* of (n+4)th residue. Therefore the α-helix has four residues per turn. The other two helical conformations, 3_{10} – helix and π-helix are relatively rare in proteins, with three and five residues per turn, respectively.

Strand Conformation

β-Strands are usually five to eight residues long [20]. Backbone atoms in two strands connect through hydrogen bonds and from a *sheet*. If the hydrogen-bonded strands run in the same direction the resulting sheet is called *parallel sheet*. If the hydrogen-bonded strands run in opposite directions the resulting sheet is called *anti-parallel sheet*.

Coil Conformation

Since coils are non-repetitive irregular structures, they are not easily described by structure [20][28]. Among the tree main types of coils, turns assume few defined structures; while bends or loops are irregularly shaped. Coils connect other two secondary structure components together and without them a protein would be loosely packed.

Table 1 shows the frequencies of AAs in proteins disregarding the secondary structure, as well as AA frequencies in each of helix, strand and coil structures[7]. The AAs commonly observed in each of the secondary structure components are highlighted. AAs A and L are common in helix, AA V is common is strand, and AAs G, N and P are frequently observed in coil, when compared to other AAs.

[7] This table was derived based on PDBSelect90 dataset. For more information about this dataset see Chapter 5.

Table 1. AAs Frequency in proteins and secondary structure components

AA	Frequency in Proteins (%)	Frequency in Helix (%)	Frequency in Strand (%)	Frequency in Coil (%)
A	8.02	**46.60**	16.30	37.10
C	1.59	26.30	30.90	42.80
D	5.77	29.50	11.10	59.40
E	6.59	45.60	14.30	40.10
F	3.88	33.80	31.60	34.60
G	7.57	15.40	13.30	**71.30**
H	2.42	28.70	19.50	51.80
I	5.41	35.30	36.50	28.20
K	6.02	37.40	17.20	45.40
L	8.71	**43.40**	24.60	32.00
M	2.21	39.40	21.30	39.30
N	4.31	25.50	12.90	**61.50**
P	4.69	16.70	9.72	**73.60**
Q	3.77	40.80	17.40	41.80
R	4.92	39.40	18.90	41.70
S	6.44	25.30	18.40	56.30
T	5.69	24.30	27.40	48.30
V	7.04	29.20	**41.00**	29.90
W	1.41	35.40	31.30	33.30
Y	3.52	31.60	33.00	35.40

Structural Class

The first definition of protein structural classes was officially recognized in 1976 [17]. Four structural classes of globular proteins are usually distinguished:

1. all-α class, which includes proteins with only small content of strands,
2. all-β class with proteins with only small content of helices,
3. α / β class with proteins that include both helices and strands, where strands are mostly parallel
4. $\alpha + \beta$ class, which includes proteins with both helices and strands, where strands are mostly anti-parallel

Several definitions of structural classes were developed in 1980's and redefined multiple times since then. However, the main differences among different definitions were in the thresholds used to define content of strands for all-α proteins, and content of helices for all-β proteins. Sometimes α / β and $\alpha + \beta$ classes are combined into a single $\alpha\beta$ class.

Tertiary Structure

The third level of a protein structure, referred to as *tertiary structure* is the overall conformation or the three dimensional shape of the protein. This level relies on the number, the size and the arrangement of secondary structure components [20]. Knowing the conformation of a protein is the key to understanding its properties and functions.

Quaternary Structure

Some proteins are built from multiple polypeptide *chains*. Each chain is an independent functional subunit of the protein. Proteins with multiple chains have a fourth level of structure called the *quaternary structure,* which describes the number and relative position of each chain [20]. Quaternary structure prediction deals with protein to protein interaction which is out of the scope of this thesis.

Figure 2.4 shows how protein secondary structure *content* prediction is used for protein secondary *structure* prediction. The shaded box, as mentioned earlier, shows the focus of this project.

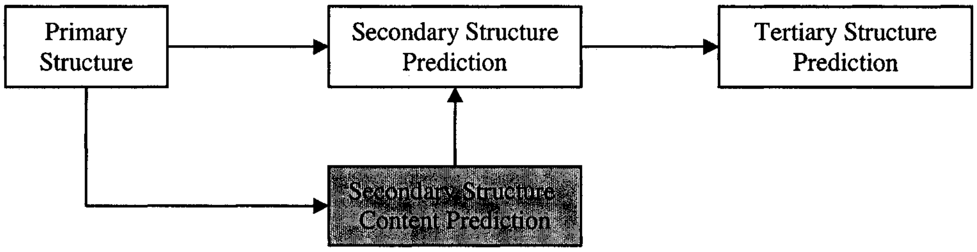
Primary Structure
Secondary Structure Prediction
Tertiary Structure Prediction
Secondary Structure Content Prediction

Figure 2.4. Relationship among prediction of different protein structure levels

2.3 Homologous Proteins

Protein homology is defined as the percentage of AAs in a protein sequence that are identical after aligning the sequence with other sequences in a protein dataset (gaps between AAs may be introduced to facilitate alignment). Proteins with similar functionality often (but not necessarily) have homologous AA sequences that match important functional domains. This fact has been used widely in Protein Science where a protein database is queried to find protein sequences with known structure (functions) that exhibit similarity to a protein sequence with unknown structure (functions). Information about similar proteins provides revealing insights into the structure and function of the query protein. This approach is also used to predict the secondary structure content of newly discovered proteins, but it can be successful only provided that a query protein has homologous peers in the database, i.e. at least 40% homology is present. In this thesis we concentrate on content prediction for low-homology proteins, in which case alignment cannot be successfully performed. This is an important problem to consider since with a growing number of proteins with unknown structures versus those with known structures, the chance of having proteins with unknown structure that are not homologous to proteins with known structure, increases (for more information see Chapter 1).

2.4 Protein Sequence Representation

The prediction of secondary protein content is usually performed with an intermediate step, in which the primary sequence is converted into its feature (also called predictor) space representation. The existing protein secondary structure content prediction methods use a limited set of features to describe the primary sequence [15] [26] [24] [43] [45] [30] [33], while other methods, such as those for prediction of protein structure or function [12] [26] [35] [7] [36] [41], use a more diverse and larger number of features. In this project the protein properties employed for the above purposes are aggregated. The reasoning behind it is to provide the proposed content prediction methods with more comprehensive information about the underlying sequences. In addition, a new set of attributes is proposed and tested. All features and their original applications are summarized in Table 2 (Table A 1 shows the corresponding indices and their values).

The following sub-sections divide the attributes into a number of sub-groups and explain them in details.

2.4.1 Index-based Attributes

This set includes the following features: molecular weight, average isoelectric point, auto-correlation functions and the five features that are proposed in this thesis. As mentioned earlier, all features are computed using the index values in Table A 1. In the following definitions *N* represents the length of the protein, i.e. number of AAs, for which the following attributes are being computed.

The **molecular weight,** *MolW,* of a protein sequence is the result of adding up the molecular weight *MolW*$_i$ (residue average) values of its residues plus the mass of a water molecule ($MolW_{H_2O}$) that is approximately 18 daltons.

$$MolW = MolW_{H_2O} + \sum_{i=1}^{N} MolW_i \qquad (1)$$

Table 2. Features used to encode protein sequences and their applications

Feature	Application	Reference(s)
Protein sequence length	Protein content and function prediction	[12][26] [35]
Average molecular weight		
Average isoelectric point		
Composition vector	Protein structure and content prediction	[15][26] [33] [42][43] [44][45]
First and second order composition moment vector	Protein content prediction	[33]
R groups	Protein structure and content prediction	[41]
Exchange groups	Protein family and structure prediction	[36][41]
Hydrophobicity groups	Protein function prediction, structural and functional relationships	[12][35] [20]
Electronic groups	Protein structure prediction	[7]
Chemical groups		
Other groups	Protein function prediction, structural and functional relationships	[12][35]
Auto-correlation functions based on $FH_{i,}$,EH_i and M_i (see Table A 1)	Protein content prediction	[24][43] [45]
Auto-correlation functions based on $Hp_{i,}$ (see Table A 1)	Protein content prediction	This project
Average Hydrophobicities based on $FH_{i,}$ and EH_i indices	Protein content prediction	This project
Sum of Hydrophobicities based on $FH_{i,}$ and EH_i indices	Protein content prediction	This project
Sum of average Hydrophobicities of each three consecutive residues $FH_{i,}$ and EH_i indices	Protein content prediction	This project
Cumulative indices based on $FH_{i,}$,EH_i indices	Protein content prediction	This project

The **average isoelectric point,** *pI,* of a protein sequence is computed based on the average isoelectric point pI_i values of its residues. The *pI* value shows the pH[8] at which a molecule carries no net electric charge and thus it is immobile in an electric field [28]

$$pI = \frac{1}{N}\sum_{i=1}^{N} pI_i \qquad (2)$$

An order *n* **auto-correlation function,** A_n^a (Equation 3), is computed by summing up the products of a_i indices (the index the function is based on) of every pair of residues

[8] The pH of a solution is determined by the relative concentration of acids and bases [28]

separated by n residues. The sum is normalized based on the number of residue pairs. In this project two AA hydrophobicity indices are used: the Fauchere (*FH*) [24] and the Eisenberg (*EH*) [4] indices. Table A 1 shows the index values for each AA. Following the published research, six auto-correlation functions are used based on *FH*, *EH* and *M* [24], i.e. $n = [1...6]$ (Equation 3) [24], and nine based on *Hp* i.e. $n = [1...9]$ (Equation 3) [14]. Hydropathy indices (*Hp*) were first proposed in [14]. They were used to identify the hydrophilic and hydrophobic regions of a protein as determined by X-ray crystallography. *M* indices are the relative side-chain mass for each residue [24].

The following equation defines auto-correlation functions where a_i is *FH*, *EH*, *Hp* or *M*:

$$A_n^a = \frac{1}{N-n}\sum_{i=1}^{N-n} a_i a_{i+n} \qquad (3)$$

The **five attributes,** which are proposed in this thesis, are also categorized as index-based features. The first one an auto-correlation function based on *Hp* index. Equations **(4)**, **(5)**, and **(7)** show the sum, average and cumulative sum of hydrophobicity indices respectively. Equation **(6)** shows the sum of hydrophobicity index averages over each three consecutive AA where b_i is *FH* or *EH*.

$$H_{sum}^b = \sum_{i=1}^{N} b_i \qquad (4)$$

$$H_{avr}^b = \frac{\sum_{i=1}^{N} b_i}{N} \qquad (5)$$

$$H^{b}_{sum3} = \sum_{i=1}^{N-3}(\sum_{j=i}^{i+3} b_j)/3 \qquad (6)$$

$$HCum_n^b = \frac{\sum_{i=1}^{N-n}\left(\sum_{j=1}^{i} b_j\right)\times\left(\sum_{j=1}^{i+n} b_j\right)}{N-n} \qquad (7)$$

We computed six cumulative density functions, for n=[1...6] in (7).

2.4.2 Composition Vector and Composition Moment Vector

Composition Vector (*CV*) is defined as the composition percentage of each residue in the primary sequence. Unlike composition vector, Composition Moment Vector (*CMV*) takes into account the position of each residue in the sequence. The following equation shows how a composition moment vector of order k is computed for each AA (i=[1...20]):

$$CMV_i^k = \frac{\sum_{j=1}^{x_i} n^k{}_{ij}}{\prod_{d=1}^{k}(N-k)} \qquad (8)$$

n_{ij} represents the j^{th} position of the i^{th} AA, x_i is the frequency of i^{th} AA, and N is the length (number of AAs) of the protein. In this project orders zero to two are used ($k \in \{0, 1, 2\}$). Note that the zeroth order reduces to the composition vector (CV). The composition vector was used extensively for both protein structure and content prediction, while composition moment vector was recently proposed for the protein content prediction task [33].

2.4.3 Property Groups

Property groups classify the AAs into groups related to specific properties of individual AAs or an entire protein molecule. The properties that are considered in this thesis are summarized in Table 3. The composition of each group, which is normalized with regard to the sequence length, gives a real-number attribute. A short description of each group is given below, which is followed by an example of how the attributes based on property groups are computed.

Table 3. Property based AA groups

Groups	Subgroups	AAs	Groups	Subgroups	AAs
R groups	Nonpolar aliphatic Polar uncharged Positively charged Negative Aromatic	AVLIMG SPTCNQ KHR DE FYW	Hydrophobicity groups	Hydrophobic Hydrophilic basic Hydrophilic acidic Hydrophilic polar with uncharged side chain	VLIMAFPWYCG KHR DE STNQ
Exchange groups	(A) (C) (D) (E) (F) (G)	C AGPST DENQ KHR ILMV FYW	Electronic groups	Electron donor Weak electron donor Electron acceptor Weak electron acceptor Neutral Special AA	DEPA VLI KNR FYMTQ GHWS C
Other groups	Charged Polar Aromatic Small	DEKHRVLI DEKHRNTQSYW FHWY AGST	Other groups	Tiny Bulky Polar uncharged	AG FHWYR NQ

Hydrophobicity group: Hydrophilic AAs are water-soluble with ionized or polar side chains. Usually they are located at the surface of a water-soluble protein. In contrast, hydrophobic AAs are slightly soluble or insoluble. They avoid aqueous environments and are usually found in interior parts of a protein [20].

R groups: This classification is based on molecular weigh (*MolW*), hydropathy index (*Hp*) and Isoelectric point (*pI*) of AAs [41]. Hydropathy index combines hydrophobic and hydrophilic tendencies.

Electronic groups: AAs are classified based on their tendency to accept or donate electrons [7]. Note that AA C has special properties and it is grouped by itself.

Other groups: Despite the overlaps between groups, each group is considered as a separate attribute. The molecular weights of tiny, small and bulky AAs are less than 80 daltons between 80 daltons and 101 daltons, and more than 120 daltons, respectively [12].

Chemical groups: Table 4 shows a detailed chemical composition of AAs [7]. A composition of each chemical group, which is normalized with regard to the protein length, gives a real-valued attribute.

Table 4. Chemical groups for AAs

AA	Associated chemical groups	AA	Associated chemical groups
A	CH CO NH CH_3	M	CH CO NH CH_2 CH_2 S CH_3
C	CH CO NH CH_2 SH	N	CH CO NH CH_2 CO C NH_2
D	CH CO NH CH_2 CO COO^-	P	CHRING CO NHRING CH_2RING CH_2RING CH_2RING
E	CH CO NH CH_2 CH_2 CO COO^-	Q	CH CO NH CH_2 CH_2 CO C NH_2
F	CH CO NH CH_2 CAROM CHAROM CHAROM CHAROM CHAROM CHAROM	R	CH CO NH CH_2 CH_2 CH_2 NH C NH_2 NH_2^+
G	CH_2 CO NH	S	CH CO NH CH_2 OH
H	CH CO NH CH_2 CAROM CHAROM N CHAROM NH	T	CH CO NH CH CH_3 OH
I	CH CO NH CH_2 CH CH_3 CH_3	V	CH CO NH CH CH_3 CH_3
K	CH CO NH CH_2 CH_2 CH_2 CH_2 NH_3^+	W	CH CO NH CH_2 CAROM CAROM CAROM NH CHAROM CHAROM CHAROM CHAROM CHAROM
L	CH CO NH CH_2 CH CH_3 CH_3	Y	CH CO NH CH_2 CAROM CHAROM CHAROM CHAROM CHAROM CAROM OH

Exchange groups: Unlike other groupings that are based on a priori knowledge about proteins, exchange groups are supported by statistical studies. A mutation point is an exchange of one AA with one another due to natural selection [3]. Exchange groups are groups of AAs based on accepted point mutation [3]. In other words, this grouping shows conservative replacements through revolution [36].

Here, we illustrate how the attributes are computed based on property groups, through an example. Consider the following sequence with length N=37:

PCSAFEFHCLSGECIHSSWRCDGGPDCKDKSDEENCA

AAs AVLIMG (underlined below) constitute the Nonpolar Aliphatic subgroup:

PCSAFEFHCLSGECIHSSWRCDGGPDCKDKSDEENCA

Therefore the corresponding attribute value is: 7/37. Similarly, the Positively Charged group is composed of AAs KHR (underlined below), and gives an attribute value of 5/37:

PCSAFEFHCLSGECIHSSWRCDGGPDCKDKSDEENCA

The chemical group CH_3 is found once in AAs A, M, T, and twice in AAs I, L, V (underlined below). The attribute value for this group is 6/37:

PCSAFEFHCLSGECIHSSWRCDGGPDCKDKSDEENCA

To summarize, this chapter introduced amino acids, which are protein building units, and protein structure. It also talked about different attributes used in this thesis for protein sequence representation. The next chapter reviews related work.

Chapter 3 Related Work

Section 2.2 presented protein's secondary structure representation using eight secondary structure components (*8-state representation*). These 8 states can be reduced into three types (*3-state representation*). Although the protein secondary structure content prediction task can be performed for the former case [1][21][2][23], in this research project we focus on the *3-state* problem.

The following sub-section reviews the related works in chronological order (prediction approaches proposed for the 8-state representation case were excluded). Next, evaluation procedures and accuracy measures that were considered in literature to evaluate the content prediction models are described. Finally, based on these evaluation criteria, a comparison of prior works is given.

3.1 Overview of Prior Works

The related works are divided into two groups: those that assume the secondary structural classes of query proteins are known, and the ones that do not utilize this a priori information.

3.1.1 Primary Structure-based Approaches

Approaches in this category use the primary structure of proteins *only*. The first secondary content prediction effort was undertaken in 1973 [15] where a Multiple Linear Regression (MLR) model was used to predict the content based on the composition vector. The model was trained on a dataset of 18 proteins; two sub-chains of two training samples were used to perform an independent test. The authors indicated that some residues in their dataset, which were in the border between different secondary structure regions, were assigned to multiple structures. Therefore the sum of the secondary structure contents might not be equal to one for some sequences. Moreover, the authors stated that the dataset includes at least two types of errors that might affect the quality of the prediction: AAs N and D, and Q with E, were in some cases confused. Secondly, they pointed out uncertainties in assigning secondary structure components to different regions in a sequence. This issue arose when comparing the eight sequences which were common between this paper and four other recent papers [29] [19] [25] [32]. The prediction was performed for four types of secondary structures: Helix, Strand, Turn and Coil. To represent sequences, the authors considered sums of up to five AA compositions that were highly (positive or negative) correlated with the four secondary structure types. This resulted in eight sums. Also absence or presence of the Heme group[9] and the positive and negative correlation sums for each structure were combined in an MLR model and the final regression coefficients were determined using the training set. To address the problem with the confusion between N with D and Q with E, the exact same procedure was followed to derive another model assuming that the aforementioned AAs are not distinguished, i.e. N and D were counted as one residue and so were Q and E.

It was not until 1992 when another content prediction approach was proposed [26]. This method utilized composition vector, molecular weight and absence or presence of bound Heme group to represent protein sequences [26]. The sequences were then fed into a tandem composed of two neural networks. The first neural network performed

[9] Heme group reflects the bias in the protein set to globins and cytochromes. It correlates positively with helix and negatively with strand [6]

memorization since it had more adjustable parameters than the learning examples. The second network was used to determine when the first neural network was generalizing.

In the next approach, two analytic vector decomposition techniques were developed to predict the content [5]. As the following formulation shows, in the first one, the AA composition (CV^q) of a query protein was represented as a linear combination of the three secondary structure components' AA compositions (CV_i, i=1 for helix; i=2 for strand; i=3 for coil) derived from a training set:

$$CV^q = \sum_{i=1}^{3} x_i CV_i \text{ such that } \sum_{i=1}^{3} x_i = 1 \text{ and } 0 \leq x_i \leq 1 \qquad \textbf{(9)}$$

The coefficients x_i are called Barycentric Coordinates. They show the portion of each secondary structure component in the secondary structure of query protein CV^q. The three basis vectors CV_i span a subspace in a 20-dimensional space. The barycentric coordinates describe the relative influence of this span on a vector CV^q in the subspace. If the subspace does not include CV^q, the following optimization criterion finds the closest solution:

$$\min(CV^q - \sum_{i=1}^{3} x_i CV_i)^2 \qquad \textbf{(10)}$$

The second technique took into account compositional couplings between any two AAs. From the mathematical viewpoint, the second moment matrix describing the AA compositional variations of secondary structural types in training dataset is computed as:

$$M_i = \sum_{s=1}^{S} (CV_i - CV_i^s)^T (CV_i - CV_i^s) \qquad \textbf{(11)}$$

where S is the number of sequences in training set, and CV_i^s is the AA composition vector for i^{th} secondary structure element of s^{th} protein.

To summarize, in the latter paper the goal is to minimize the following summation:

$$\{\sum_{i=1}^{3} x_i f(CV^q, CV_i)\}^2 \quad \textbf{(12)}$$

where function f is computed in these steps: First CV^q- CV_i is transformed from the AA fraction space into the eigenvector space of the second moment matrix. The resulting components were scaled by the square roots of the corresponding eigenvalues M_i. Finally the resulting vector was transformed back to the AA fraction space.

The authors studied the two proposed vector decomposition techniques on four protein datasets that included sequences measured at 1.8, 2.0, 2.5 and 3.0 Angstrom, which result in different qualities. At the worst 3.0 Angstrom resolution, terminal residues of secondary structure components are not clearly recognized, whereas at the best resolution level, the side-chain conformations can be reliably recognized.

In 1998, auto-correlation functions were employed for the content prediction task for the first time[10] [43]. The auto-correlation functions, which were based on the Fauchere hydrophobicity index (see section 2.4.1 for more information), along with composition vector were used to represent proteins that were fed into an MLR model later. The authors reported that using ten and four auto-correlation functions led to the least content prediction errors for helix and strand respectively.

[10] Auto-correlation functions based on hydrophobicity index were first used in 1987 to study the structures of amphipathic α helix (An amphipathic molecule contains both hydrophobic and hydrophilic regions)

In 2001 Li and Pan utilized the auto-correlation function idea proposed by Zahng et al. [43]. In this case six auto-correlation functions were based on Fauchere Hydrophobicity index and six based on side-chain mass indices (see M_i in Table A 1). These attributes along with composition vector were used in this work and prediction was performed using MLR model.

In 2004, Pilizota et al. analyzed the AA composition vector by performing feature selection to choose a subset of features leading to the least prediction error for the prediction of helix and strand contents [30]. For I = [1...20], I denotes the index of a corresponding AA, they built the MLR models with I components. Then for each model, the sum of squared correlation coefficients between each component and the target (i.e. helix or strand content) was computed. Among the models with I components, only the model with the highest sum was kept. Therefore, at last 20 models were tested and the authors reported that the models with nine and five predictors led to the best prediction results in terms of accuracy for helix and strand content precition, respectively.

In our prior contribution to protein content prediction, attribute indices 1-95, which are given in Table A 2 in the Appendix, together with the 400 element *dipeptide* composition vector were used to represent protein sequences and to design an MLR model [16]. Dipeptide is a pair of consecutive AAs. Two attribute selection procedures were used to independently design MLR prediction models for helix and strand. In the first one each feature was ranked according to its absolute correlation coefficient with the predicted target; the higher the correlation the more important the feature. The second attribute selection approach assigned importance to features based on the magnitude of their regression coefficients. Each selection method started with the complete set of attributes and proceeded by eliminating five least important features at each step. For both helix and strand content prediction, errors using four attribute subsets were reported:

1. The subset with the five most significant attributes
2. The subset with the lowest prediction error,
3. The subset with the best relative ratio between error and number of features.
4. The subset with the smallest number of attributes that does not result in a substantial drop in prediction accuracy.

In 2005, a Neural Network (NN) based approach was published. It used composition moment vector to represent protein sequences [33]. Two separate NNs were trained for helix and strand content prediction tasks. The authors compared the prediction results to sequence representation using composition vector and they concluded that composition moment vector was more informative and led to better results.

3.1.2 Secondary Structural Class-based Approaches

The a priori knowledge of secondary structural classes of proteins is recognized to improve content prediction. However this information is tightly related to content, i.e. to learn the structural class, the content and secondary structure must be known. The prediction of the structural class is a difficult task, and state-of-the-art methods achieve about 60% accuracy [17] [40]. Therefore, although we review two works that have been published in this category, we will not compare these results with ours since they require information, i.e. secondary structural class, that cannot be obtained from the protein primary sequence accurately *only*.

An MLR model for content prediction that requires knowledge of structural classes was proposed in 1998 [44]. In this paper, the secondary structural classes of proteins were incorporated by defining the following pairs of predictors (attributes): (1, 0) for the mainly-α class, (0, 1) for the mainly-β class, and (0, 0) for the $\alpha\beta$ class [44]. Based on a previous study the authors used different sets of AAs in determining different protein secondary structures, i.e. AAs A, E, M and L are associated with helices, while F, I, V, and Y are usually associated with strands. The other two predictors for helix (strand) content prediction were the sum of squares and the sum of coupling terms of helix (strand) main former amino acids. Moreover the composition vector, sequence length and squared sequence length were used to represent sequences. The protein secondary structural class assignment was performed according to the quantitative criteria given in Table 5 [44]:

Table 5. Secondary structural class assignment by Zhang et al. [44]

Structural Class	α-Helix content	β-Strand content
mainly-α	$\geq 40\%$	$\leq 5\%$
mainly-β	$\leq 5\%$	$\geq 40\%$
$\alpha\beta$	$\geq 15\%$	$\geq 15\%$

In 2001, another MLR-based method was proposed [44]. The main difference between this and the previous method was that separate MLRs were trained for α, β and $\alpha\beta$ structural classes [45]. Protein sequences were represented by the composition vector and ten auto-correlation functions based on the Fauchere Hydrophobicity index. Feature selection was performed for each MLR model; i.e. the attributes were sorted according to their linear correlation coefficients with the target. The procedure started with the most highly correlated feature and added one feature at a time, therefore resulting in 30 different models. The regression equation leading to the least prediction error was used and reported. The assignment of structural classes for the dataset used by Zhang et al.

[45] was based on protein three-dimensional coordinates (besides the secondary structure content) [27]. Elaboration on the assignment procedure is beyond the scope of this thesis. Here, we only report the secondary structure content thresholds for each class:

Table 6. Secondary structural class assignment by Zhang et al. [45]

Structural Class	α -Helix content	β -Strand content	Parallel Strand Content
Mainly- α	$> 60\%$	$< 5\%$	—
Mainly- β	$< 5\%$	$> 40\%$	—
α / β	—	—	Taken into account
$\alpha + \beta$	$\geq 15\%$ and $\leq 55\%$	$\geq 10\%$ and $\leq 45\%$	Taken into account

3.2 Comparison of Prior Works

3.2.1 Evaluation Procedures

The protein content prediction models have been evaluated through *re-substitution* (also called *self-consistency*) and *cross validation* statistical tests. The former measure evaluates a prediction model on the same dataset used for training (in-sample test), whereas in the latter case the training set is divided into *k* folds and training step is performed for *k-1* folds and the resulting model is a test on the remaining fold (out-of-sample test). This procedure is repeated *k* times and the average prediction error is reported. For *k* equal to the number of training samples, the test is called a *jackknife* test. The re-substitution test is not reliable, since it is biased due to testing the model on sequences that were used during the learning phase. Since early papers reported this test, it is still being reported for comparative purposes, although it is generally understood that these results are inflated. In protein content prediction, the test of choice that is reported in all contributions is the jackknife test. In some publications an *independent* test is performed instead of (or besides) the jackknife test. In such a case, the model is examined on a dataset different from the training one. The *mean absolute error* **(13)** is used to measure the accuracy of a content prediction model for helix and strand separately:

$$e = \frac{\sum_{i=1}^{S} \left| y_i^{obs} - y_i^{pred} \right|}{S} \qquad (13)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{S} (e - \left| y_i^{obs} - y_i^{pred} \right|)^2}{S-1}} \qquad (14)$$

where S is the number of sequences in the dataset for which the prediction is performed, y_i^{obs} and y_i^{pred} are the observed (true) and predicted secondary structure contents for sequence i, respectively. The standard deviation of the prediction error, σ (Equation **14**), is also reported.

3.2.2 Discussion

Table 7 and Table 8 show the model evaluation procedures and the accuracy measures employed in each of the eight and two approaches reviewed in sections 3.1.1 and 3.1.2, respectively. The homology level shows the level of sequence similarity among proteins in the training dataset.

The tables show that the composition vector was used in almost every contribution. Thus, we conclude that it provides useful information with respect to the protein's secondary structure. However this information is not sufficient to achieve high content prediction accuracies. Adding auto-correlation functions led to a relatively significant drop in the prediction error [43]. Li and Pan added auto-correlation functions based on a new index, *M,* (see section 2.4.1) and proposed a simple MLR model with a good prediction accuracy [24]. In this project we compare our content prediction results with Li and Pan's [24] and Zhang et al.'s [45] results (see Chapter 6). The former work is the most recent and the best representative MLR model among the works published in this area. Since it has the highest prediction accuracy on the lowest homology dataset, it establishes a solid baseline to compare our work with. Although the latter work [45] utilizes secondary structural class information, we implemented and compared this recent MLR-based method with our approach assuming that the a priori knowledge of structural class is not provided. Another reason for comparing with this method is that our prior work showed that the attribute selection procedure suggested by Zhang et al. could lead to simpler yet accurate prediction models [16].

The NN proposed by Ruan et al. [33] distinguishes pure helices (1642 sequences with zero strand content) and pure strands (405 sequences with zero helix content) with high quality and in this regard it is superior to MLR models. However this could be due to a high homology level between the training and test sets. The NN is fed with composition moments vectors of orders one and two, and the authors conclude that prediction results for pure helices, pure strands and mixed structures are mostly better than the case, in which sequences are encoded using the composition vector. The only case where the composition vector provides more information is when the NN is trained for helix content prediction and the network is tested on a pure helix dataset.

To summarize, this chapter reviewed related work in chronological order. It introduced the content prediction evaluation procedure exploited in literature. Based on the evaluation criteria, which is the mean absolute error between predicted and true content values, the related works were compared with each other. The next chapter elaborates on problem definition and reviews project goals.

Table 7. Comparison among content prediction approaches which don't exploit a priori knowledge of structural class

Method	Feature Set	Training Set		Independent Test					Re-substitution Test				Cross Validation Test[1]				Ref.
		Size	Homology Level	Data-set Size	Helix		Strand		Helix		Strand		Helix		Strand		
					e	σ	e	σ	e	σ	e	σ	e	σ	e	σ	
MLR	*CV*	18	--	--	--	--	--	--	--	--	--	--	7.1%	--	6.9%	--	[15]
NN	*CV, MolW,* Heme	104	High[2]	15	5%	3.4	5.6%	4.9	4.1%	4.5	4.1%	3.4	--	--	--	--	[26]
AVDT	*CV*	166	<=35%	--	--	--	--	--	12.6%	10.3	10.8%	9.1	14.3%	11.7	12.3%	10.4	[5]
MLR	CV, A^{FH}	261	< 35%	347	9.9%	--	8.3%	--	7.7%	0.059	7.3%	0.057	8.7%	0.067	8.1%	0.065	[43]
MLR	CV, A^{FH}, A^{M}	704	< 30%	--	--	--	--	--	8.4%	0.07	7.7%	0.06	8.8%	0.073	8.1%	0.066	[24]
MLR	*CV*	317	< 35%	158	11.19%	--	10.75%	--	10.5%		8.99%		10.82%		9.19%		[30]
NN	CMV^1, CMV^2	9159	High	1642[3]	14.5%	--	0.0%	--	6.5%	--	6%	--	--	--	--	--	[33]
				405[4]	0.0%	--	13.4%	--	6.5%	--	6%	--	--	--	--	--	
		1707	<25%	--	--	--	--	--	--	--	--	--	12.6%	0.096	11.9%	0.099	
MLR	See legend	5834	--	--	--	--	--	--	--	--	--	--	11.28%	0.102	8.67%	0.006	[16]

1: All the tests are jackknife tests except for the ones by [16] which is 10-fold cross validation (10CV). 2: The homology level was high among the sequences in training set but low between the independent test dataset and the training set 3: Pure Helix 4: Pure Strand

Table 8. Comparison among content prediction approaches which exploit a priori knowledge of structural class

Method	Feature Set	Training Set		Independent Test					Re-substitution Test				Jackknife Test				Ref.
		Size	Homology Level	Data-set Size	Helix		Strand		Helix		Strand		Helix		Strand		
					e	σ	e	σ	e	σ	e	σ	e	σ	e	σ	
MLR	CV, N	120	< 25%	--	--	--	--	--	4.0%	0.039	3.5%	0.038	5.1%	0.053	4.5%	0.053	[44]
MLR	CV, A^{FH}	210	Low	143	5.5%	0.048	5.1%	0.05	5.2%	0.05	4.7%	0.047	5.8%	0.057	5.3%	0.053	[45]

Legend for Table 7 and Table 8:
--: not reported; CV: Composition Vector; CMV^k: Composition Moment Vector of order *k*; *MolW*: Molecular Weight; A^{FH}, A^{M}: Fauchere and side mass chain based Auto-correlation functions Heme: Presence/absence of heme group; N: Sequence length. For more information on feature sets see section 2.4
The features used by [16] are indices 1-95 listed in Table A 2 plus 400 dipeptide composition

Chapter 4 Problem Description

This chapter presents the formal definition of the prediction problem considered in this thesis, which is followed by an overview of project's goals.

Secondary structure content of a protein refers to the portions of the sequence in helix, strand and coil conformation normalized by the length of the sequence. Therefore, each secondary structural content is a real number between zero and one. The sum of the three contents is equal to one. As an example, we consider the primary and secondary structures of the protein introduced in section 2.2 with 37 AAs:

PCSAFEFHCLSGECIHSSWRCDGGPDCKDKSDEENCA
CCCCCCEECCCCCEECHHHCCCCCCCCCCCHHHCCCC

The helix content for this sequence is 6/37, the strand content is 4/37, and the coil content is 27/37.

4.1 Problem Definition and Proposed Framework

Given a protein's primary structure only, we attempt to predict helix and strand contents of the sequence (coil is excluded since it is an irregular structure; see section 2.2). As for any other learning problem, a training set is used to derive two separate helix and strand content prediction models. Figure 4.1 shows the general procedure employed for the protein secondary structure content prediction. First the protein sequences in the training dataset are represented as *feature vectors* (see section 2.4 for more details)[11]. Then the

[11] We note that some feature groups overlap. These redundancies were eliminated and Table A 2 shows the final non-redundant feature list.

two models are trained to predict helix and strand contents. Finally test sequences are applied to each model to compute the corresponding content values.

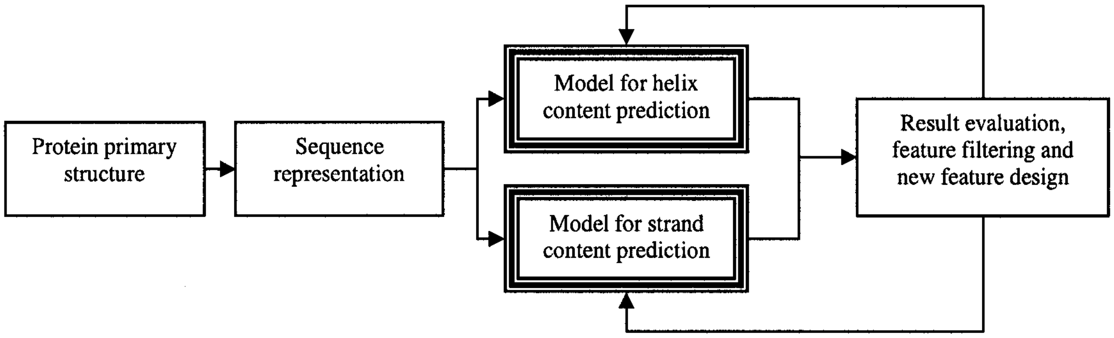
Protein primary structure
Sequence representation
Model for helix content prediction
Model for strand content prediction
Result evaluation, feature filtering and new feature design

Figure 4.1. General approach for protein secondary structure content prediction problem

Equations (15) and (16) give a formal definition of the Multiple Linear Regression model we utilized for helix and strand content (i.e. y_α and y_β) prediction:

$$y_\alpha = a_0 + \sum_{i=1}^{n} a_i f_i \qquad (15)$$

$$y_\beta = b_0 + \sum_{i=1}^{n} b_i f_i \qquad (16)$$

In these equations, *f* denotes an *n* dimensional protein feature vector. The elements of this vector are also called the regression *predictors*. The regression coefficients a_0, a_1... a_n (similarly b_0 ,b_1,...,b_n) are estimated using training data. Since helix and strand contents are real numbers between zero and one, any negative number prediction is rounded to zero, while predictions greater than one are rounded to one.

4.2 Project Goals

The thesis aims at providing an MLR-based method that improves the content prediction accuracy when compared with existing methods proposed in the related work section. To this end, four goals are investigated and we review them in this sub-section. A formal definition of these goals is provided in Chapter 6.

1.The objective is to *cluster* the proteins in the training dataset into a number of subsets, and derive separate MLR models for each of these subsets. An unseen protein is then *classified* into one of the clusters, and the corresponding MLR models are applied to predict its helix and strand contents (Figure 4.2). Since we anticipate that MLR models designed for specific data subset will perform better than a model extracted from the *entire* training data, the main issue is to minimize the classification error rate, so we minimize the number of times when a wrong MLR model is used.

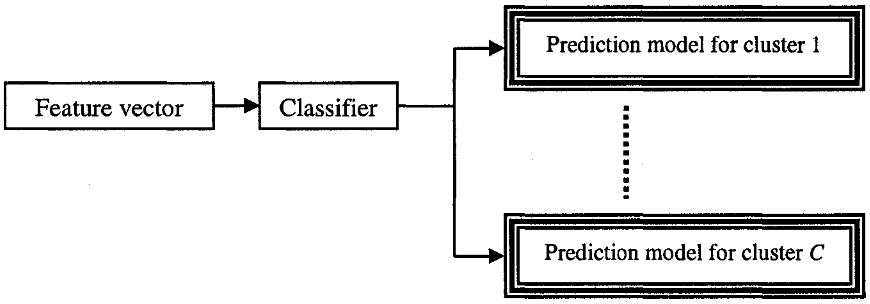
Feature vector
Classifier
Prediction model for cluster 1
Prediction model for cluster *C*

Figure 4.2. Specialized prediction content models

In another approach we divide the training set into subsets based on one of the known sequence features, i.e. length and average hydrophobicity (one at a time). In this case, there is no classification required for the test sequences. Therefore, this approach seems to be promising due to the elimination of the classification error.

2.In this project, an aggregation of protein features is used to represent protein sequences. The goal is to perform feature selection to select most informative features and eliminate unnecessary attributes with low contribution to content prediction. Also, since MLR is used, collinearities among attributes have to be eliminated. We used an attribute orthogonalization technique to remove collinearities among attributes.

3. Another goal is to design new informative attributes that can provide the model with a more comprehensive knowledge for helix and strand content prediction. Therefore, we propose to incorporate some statistical patterns, such as polypeptides that are common for specific secondary structures, to design new features.

4. The fourth goal is to compare non-linear regression models with MLR. The effect of different base functions such as quadratic, cubic, and exponential functions was investigated.

To summarize, this chapter presented the formal definition of the protein secondary structure content prediction problem studied in this thesis. It also reviewed the project goals. The next chapter talks about experimental setup in terms of datasets, data preprocessing steps and platform.

Chapter 5 Experimental Setup

This chapter describes the datasets that were used for training and testing, as well as the data preprocessing procedure. Finally, the platform that was used to run the experiments is described.

5.1 Datasets

In this project, PDBSelect25 is used as the dataset to train and test our proposed secondary structure prediction models. This dataset has sequences with less than 25% sequence homology and it is a non-redundant representative subset of PDB [11]. This dataset excludes sequences from PDB that satisfy the following criteria:

1. Contain more than 5% non-standard AAs, i.e. residues other than the 20 common AAs[12].
2. Have less than 30 AAs.
3. Are measured with a resolution greater than 3.5 Angstrom.
4. Have R-Factor greater than 30%[13].

We used PDBSelect25 version from Nov. 2004 that has 2,485 sequences. The dataset was further processed to exclude additional sequences according to the following criteria:

1. Sequences with any number of non-standard AAs.
2. Sequences with a helix fragment shorter than three residues. Such sequences are artifacts since helix fragments should be at least three to five residues long (see section 2.2).

[12] These AAs are typically termed UNK in PDB files, and they are modified residues.

[13] The R-Factor is commonly used to measure the quality of protein models obtained in X-ray crystallography.

Finally, 2,187 sequences were left in the dataset. Each sequence was represented with 140 numerical attributes, which were calculated according to feature description given in section 2.4.

As described in section 3.2.2, the highest content prediction accuracy was reported by Li and Pan [24]. To compare with their results, we obtained the dataset with 704 sequences that was used in their paper. We managed to obtain 681 sequences from the original dataset and after filtering according to the aforementioned criteria, 642 sequences were left in this dataset. The sequences that were highly homologous, i.e. with a homology level greater than 40%, to our design dataset, i.e. PDBSelect25, were removed from the Li and Pan's dataset. As a result, 384 sequences were left. This low homology dataset is referred to as LiPanLH in the rest of the document.

PDBSelect90 [11] is another dataset that was used in this project (see section 6.4.1). It contains 8,595 highly homologous proteins. This dataset was filtered according to the same criteria for PDBSelect25, which resulted in 7,544 sequences.

The former two datasets were used to evaluate quality of the proposed prediction methods while the latter dataset was used to design new attributes.

5.2 Data Preprocessing

5.2.1 Attribute Normalization

Different attributes have different ranges of values. For example, molecular weight and length have larger values than other attributes by an order of magnitude, and thus could easily outweigh them. Therefore, each attribute is normalized to values between zero and one with respect to the distribution of the values among all the sequences. Min-max normalization method was used for each feature *j* and sequence *i* [8]:

$$NewVal_i^j = \frac{Val_i^j - Val_{\min}^j}{Val_{\max}^j - Val_{\min}^j} \qquad \textbf{(17)}$$

where $Val_{\min}^j$ and $Val_{\max}^j$ denote the minimum and maximum values of feature *j* among all sequences, respectively.

5.2.2 Attribute Collinearity

Since this research applies MLR to perform prediction, a high correlation, i.e. near-linear relation among attributes, has to be avoided. The MLR formulation from Chapter 4, i.e. equations (**15**) and (**16**), can be re-written as follows:

$$Y_\alpha = FA \qquad \textbf{(18)}$$
$$Y_\beta = FB \qquad \textbf{(19)}$$

where *F* represents the training feature matrix (with proteins in rows and features as columns), *A* and *B* are vectors of the coefficients to be estimated, and Y_α and Y_β are vectors containing true helix and strand contents for proteins in the training set, respectively. Equations (**18**) and (**19**) can be rewritten as follows [9] (*T* denotes matrix transpose):

$$F^T FA = F^T Y_\alpha \qquad \textbf{(20)}$$
$$F^T FB = F^T Y_\beta \qquad \textbf{(21)}$$

If there is strong *collinearity* between features, the determinant of matrix $F^T F$ will be close to zero. In this case the estimated coefficients may be larger than what is expected or may be negative contradicting the purpose of the corresponding feature [9]. Therefore, one more preprocessing step was undertaken prior to using the dataset to compute the MLR models. One of each two attributes that were collinear was eliminated. The attribute with a higher correlation with the target content was kept.

The PDBSelect25 and LiPanLH datasets were used to establish cut-off thresholds for the correlation coefficient value to eliminate pair wise collinear attributes. Figure 5.1, Figure 5.2, Figure 5.3 and Figure 5.4, depict how helix and strand content prediction errors change for PDBSelect25 and LiPanLH as the threshold changes from 1.00 to 0.80. Based on this experiment the cut-off points of 0.90 and 0.99 for PDBSelect25, and cut-off points of 0.85 and 0.86 for LiPanLH were selected for helix and strand, respectively. Table A 3, Table A 4, Table A 5 and Table A 6 present the attribute lists after pair wise collinearity eliminations for datasets PDBSelect25 and LiPanLH.

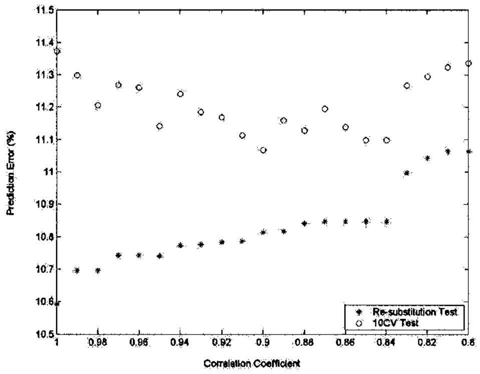

Figure 5.1. Effect of linear correlation among attributes on helix content prediction for PDBSelect25 dataset

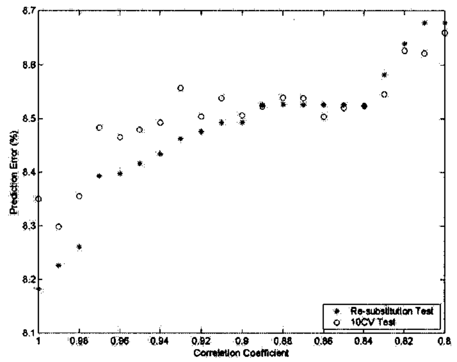

Figure 5.2. Effect of linear correlation among attributes on strand content prediction for PDBSelect25 dataset

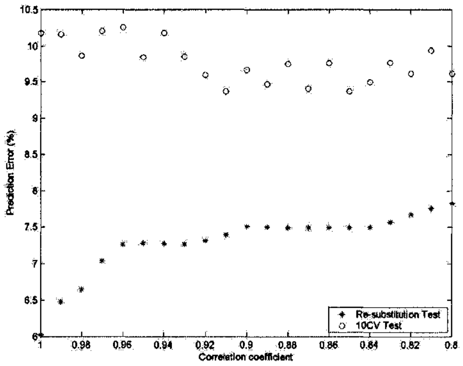

Figure 5.3. Effect of linear correlation among attributes on helix content prediction for LiPanLH dataset

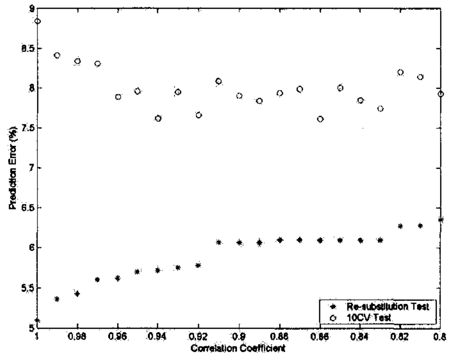

Figure 5.4. Effect of linear correlation among attributes on strand content prediction for LiPanLH dataset

5.3 Platform

The experiments were implemented in Matlab 6.5 Release 13, and Weka 3.4.5[14]. Weka is an open-source software that implements a collection of data mining algorithms and data visualization tools in Java.

To summarize, this chapter introduced the datasets PDBSelect25, PDBSelect90 and LiPanLH which are used in this project. It also talked about the data preprocessing steps, i.e. attribute normalization and attribute pair wise collinearity elimination. The next chapter presents the formal definition of project goals and experimental results compared with state-of-the-art published work.

[14] Waikato Environment for Knowledge Analysis (http://www.cs.waikato.ac.nz/ml/weka/)

Chapter 6 Experiments

This chapter is organized as follows. First the helix and strand content prediction results on PDBSelect25 using the base MLR model (introduced in section 4.1) are presented. Next the formal definitions of the project goals, which are pursued to improve the content prediction accuracy of the base MLR model, are given and the experimental results for PDBSelect25 for each goal are presented. The results are summarized and the best prediction model configurations are found and tested on the LiPanLH dataset. The methods proposed by Li and Pan [24] and Zhang et al. [45] are implemented and tested on both PDBSelect25 and LiPanLH datasets and the results are compared with results achieved by our methods.

6.1 The Base MLR Model

This section presents the prediction results when the feature sets listed in Table A 3 and Table A 4 are utilized to derive MLR models for helix and strand, respectively. According to Table 9, the base MLR predicts strand content more accurately (*e*=8.30%) than helix content (*e*=11.08%). *e* denotes the mean absolute error (Equation **(13)**).

Table 9. Content prediction errors for the base MLR model

Helix				Strand			
Re-substitution		10CV*		Re-substitution		10CV*	
e (%)	σ	*e* (%)	σ	*e* (%)	σ	*e* (%)	σ
10.67	0.09	11.08	0.09	8.23	0.07	8.30	0.07

*: 10-Fold Cross Validation

This model is used as a baseline to develop and evaluate improved MLR models. Table 10 and Table 11 show the attribute indices and the corresponding coefficients for each prediction MLR model.

Table 10. Helix prediction MLR coefficients

Feature Index	MLR Coefficient	Feature Index	MLR Coefficient	Feature Index	MLR Coefficient	Feature Index	MLR Coefficient	Feature Index	MLR Coefficient	Feature Index	MLR Coefficient
112	-259.626	81	-12.421	96	-0.196	59	0.068	134	0.194	19	27.779
79	-132.830	87	-12.260	120	-0.178	62	0.094	138	0.206	76	40.736
84	-79.533	11	-11.806	107	-0.155	60	0.099	106	0.217	92	41.321
94	-78.241	95	-11.721	132	-0.105	111	0.115	117	0.222	10	42.577
0*	-43.376	74	-9.414	140	-0.088	101	0.120	52	0.230	91	45.895
7	-28.543	20	-7.491	100	-0.080	45	0.121	24	0.248	77	52.628
82	-27.489	8	-6.097	136	-0.080	56	0.124	135	0.260	89	59.071
14	-26.994	16	-4.591	137	-0.048	49	0.136	98	0.367	15	64.739
6	-24.908	73	-4.526	110	-0.046	50	0.137	105	0.421	68	67.450
67	-21.932	86	-1.387	139	0.027	102	0.153	99	0.627	17	82.788
88	-19.555	97	-0.553	46	0.040	63	0.164	90	11.196	18	92.032
13	-19.111	133	-0.358	109	0.063	126	0.182	66	13.965	64	100.827
78	-17.984	108	-0.285	61	0.066	34	0.190	93	21.552	103	252.212

*: Index 0 refers to the MLR constant value

Table 11. Strand prediction MLR coefficients

Feature Index	MLR Coefficient	Feature Index	MLR Coefficient	Feature Index	MLR Coefficient	Feature Index	MLR Coefficient	Feature Index	MLR Coefficient	Feature Index	MLR Coefficient
71	-104.395	22	-7.703	99	-0.392	117	0.046	26	1.009	14	28.715
64	-97.856	67	-5.478	54	-0.366	100	0.062	40	1.086	65	28.912
66	-96.920	50	-1.576	105	-0.271	118	0.070	95	1.128	85	31.015
80	-73.718	52	-1.471	98	-0.240	110	0.073	41	1.142	6	31.528
75	-54.540	62	-1.328	135	-0.229	114	0.099	29	1.298	74	32.111
18	-52.703	47	-1.209	101	-0.220	96	0.100	37	1.345	12	33.044
76	-40.687	59	-1.174	119	-0.194	107	0.107	42	1.460	7	35.962
70	-40.030	49	-1.038	134	-0.178	1	0.218	39	1.615	82	39.146
8	-39.971	57	-0.927	125	-0.177	108	0.219	30	1.662	13	44.652
112	-38.766	63	-0.889	106	-0.173	111	0.268	27	1.674	69	48.380
91	-35.451	60	-0.818	115	-0.168	133	0.356	32	1.773	84	51.651
93	-33.016	61	-0.793	136	-0.118	97	0.406	87	2.634	15	51.678
77	-27.363	56	-0.755	131	-0.101	34	0.442	9	5.307	11	51.686
90	-24.754	51	-0.745	116	-0.076	28	0.620	19	6.218	103	52.185
0*	-23.270	44	-0.739	45	-0.059	35	0.763	92	7.789	79	55.068
16	-20.090	53	-0.661	138	-0.035	33	0.803	5	11.250	94	67.445
73	-15.523	46	-0.649	139	-0.026	43	0.879	81	11.754	17	78.149
89	-15.183	58	-0.646	140	-0.021	24	0.906	10	12.844	72	78.553
68	-13.930	48	-0.542	109	-0.011	31	0.911	78	18.648	21	83.854
23	-11.939	55	-0.521	137	0.014	36	0.932	20	24.557		
4	-11.554	102	-0.423	132	0.032	38	0.957	88	26.434		

*: Index 0 refers to the MLR constant value

Attributes that have positive coefficients in helix content prediction MLR and negative coefficients in strand content prediction MLR are shown in bold in Table 10 and Table 11. Attributes that have negative coefficients in helix content prediction MLR and positive coefficients in strand content prediction MLR are underlined in Table 10 and Table 11. Table 12 and Table 13 show the attributes with opposite signs in helix and strand content prediction MLRs.

Table 12. Attribute with positive and negative coefficients in helix and strand content prediction MLRs respectively

Index	Attribute
93	OG; Tiny group
91	OG; Aromatic group
90	OG; Polar group
89	OG; Charged group
77	EG; Weak electron acceptor group
76	EG; Electron acceptor group
66	RG; Positively charged group
64	RG; Nonpolar aliphatic group

Table 13. Attribute with negative and positive coefficients in helix and strand content prediction MLRs respectively

Index	Attribute
132,133,137	A_n^{Hp} , n=1, 2, 6
107,108,110	A_n^{EH} , n=3, 4, 6
96,97,100	A_n^{FH} , n=1, 2, 5
95	OG; Polar uncharged group
94	OG; Bulky group
88	CG; OH group
87	CG; NH group
84	CG; CH_3 group
82	CG; CH_2 group
81	CG; CH group
79	CG; C group
78	EG; Neutral group
74	EG; Electron donor group

It is interesting to note that the composition value for AA V (attribute with index 21) has the highest positive coefficient value in strand content prediction MLR. This is consistent with AA frequencies given in Table 1, which shows that this AA is the most frequent AA observed in strand. A composition moment vector of order one for AA A (attribute index 24) has a positive coefficient value in the helix content prediction MLR. This is also consistent with AA frequencies given in Table 1, which shows that this AA is frequently observed in helix.

6.2 Goal 1

In this section the effect of specialized MLR models on prediction accuracy is investigated. Two different approaches are pursued and presented in the following subsections, where models are created for subsets of the training data.

6.2.1 Clustering and Classification

The training set is clustered into C clusters with regard to helix, strand, and coil contents. As a result, sequences with similar content values are grouped in the same clusters. Separate MLR models for helix and content prediction are derived for each cluster c. The parameters for each model (a_{ci} for helix and b_{ci} for strand) are computed according to the following equations:

$$y_{c\alpha} = a_{c0} + \sum_{i=1}^{n} a_{ci} f_i \qquad (22)$$

$$y_{c\beta} = b_{c0} + \sum_{i=1}^{n} b_{ci} f_i \qquad (23)$$

After clusters are established, a test protein is classified into one of the clusters and the corresponding prediction model is applied. The classification is performed based on the attributes listed in Table A 3 and Table A 4 for the trand and helix content prediction, respectively. We experimented with both clustering and supervised classification algorithms. For clustering algorithms, we ran experiments with K-Means and Hierarchical clustering. For supervised classification algorithms, we ran experiments with Decision Tree, Probabilistic NN (PNN), and Discriminant Analysis (DA). A brief description of each of the clustering and classification algorithms is provided next.

- **Clustering Algorithms:**

1. K-Means Clustering Algorithm initializes clusters with K randomly chosen sequences as the cluster centers (K is a parameter defined by the user). Then it assigns each sequence to the closest cluster center. In the third step the *centroid* of each cluster (the mean of sequences within the cluster) is found and the cluster center is assigned to the cluster centroid. The algorithm iterates with refining this initial cluster assignment by re-assigning sequences to the closest centroid. This procedure is repeated until convergence or termination [18].
2. In Hierarchical Clustering Algorithms [8], initially each sequence forms a cluster that contains only one sequence, which turns out to be the centroid of that cluster. In an iterative scheme, the two most similar clusters are identified and merged to form an agglomerated cluster, reducing the number of clusters by one. Several alternatives exist for finding the distance between clusters. In the *Single Linkage* approach, this distance is equal to the distance between nearest sequences of two clusters. In the *Complete Linkage* approach, this distance is equal to the maximum distance between sequences of two clusters. In the *Average Linkage* approach, this distance is equal to the distance between the centroids of two clusters. We used single linkage approach in our experiments with hierarchical clustering algorithm.

- **Classification Algorithms:**

1. Probabilistic Neural Networks (PNNs) learn to estimate a probability density function for training data. PNN follows the optimal Bayesian classification rule [34]. The a posteriori probability is used to assign a sequence S to class c_j, among

possible classes, that maximizes $P(S|c_j)P(c_j)$. We used the radial basis function for probability density estimation.

2. Decision Tree classifiers partition training data based on attribute values. It starts with a single node (root) that includes all sequences. Then it finds an attribute that can split the sequences into subsets that reduce entropy with respect to a class attribute. It continues this procedure until the majority of sequences in tree leaves belong to one class, or there are no more attributes to split on [8]. We used the Gini index as the split criterion [8]. For a dataset D that contains S sequences from C classes, the Gini index is defined as follows:

$$Gini(D) = 1 - \sum_{i=1}^{C} P_i^2 \qquad \textbf{(24)}$$

   where P_i shows the relative frequency of class i in D.

   If D is divided into D_1 and D_2 subsets with S_1 and S_2 sequences respectively, the Gini index is defined as follows:

$$Gini(D) = \frac{S_1}{S} Gini(D_1) + \frac{S_2}{S} Gini(D_2) \qquad \textbf{(25)}$$

   The attribute that gives the smallest Gini index for D is selected to split the dataset [8].

3. Discriminant Analysis (DA) performs classification via classification functions [10]. We experimented with linear, quadratic and Mahalanobis functions. Only the results of linear functions are reported since they lead to the smallest prediction errors. The classification model includes a set of linear functions, one per each

class. For each class an MLR is derived and a test sequence is then fed into all functions to get a 'score' for each class. The sequence is assigned to the class with the highest score [10].

The experiments were performed using the following four different distance (dissimilarity) measures for K-Means and Hierarchical clustering algorithms [18] [22]:

Squared Euclidean: $d = \sum_{i=1}^{n}(X_i - Y_i)^2$

City Block: $d = \sum_{i=1}^{n}|X_i - Y_i|$

Cosine: $d = \frac{\sum_{i=1}^{n} X_i Y_i}{|X||Y|}$

Correlation $d = 1 - \frac{1}{n-1}\sum_{i=1}^{n}(\frac{X_i - \overline{X}}{s_X})(\frac{Y_i - \overline{Y}}{s_Y})$

where X and Y are feature vectors extracted from protein sequences; each vector has n attributes. $\overline{X}$ ($\overline{Y}$) and s_X (s_Y) are the mean and standard deviation for X (Y). The Squared Euclidean distance always led to the smallest prediction error during the tests. Therefore, we only report the results for this distance measure.

The hierarchical clustering algorithm created some small clusters. The number of sequences in those small clusters was less than the number of features. Thus it was not possible to continue the experiment with this clustering algorithm, i.e. MLR model could not be derived. Table 14 presents the content prediction errors when the K-Means clustering algorithm is paired with each of the classifiers; the best results are shown in bold.

As the number of clusters increases, the prediction models tend to be over-trained, i.e. the re-substitution error drops but the 10CV error increases. The DA classifier gives the best results since it does not lead to significant differences between test and train errors when compared to other classifiers. The model consisting of K-Means and DA leads to virtually the same prediction results as the base MLR does (11.08% for helix and 8.30% for strand).

Table 14. Prediction results for the K-Means algorithm

			Number of Clusters								
			2			3			4		
			Classifier			Classifier			Classifier		
Structure	Test	Measure	Tree	DA	PNN	Tree	DA	PNN	Tree	DA	PNN
HELIX	Re-substitution	*e* (%)	8.44	**11.11**	7.63	7.17	10.88	7.30	7.42	10.63	5.55
		σ	0.07	**0.10**	0.003	0.08	0.10	0.003	0.08	0.11	0.002
	10cv	*e* (%)	14.42	**11.87**	14.18	15.17	11.92	14.53	15.69	11.98	14.57
		σ	0.12	**0.11**	0.01	0.13	0.11	0.02	0.14	0.11	0.02
STRAND	Re-substitution	*e* (%)	6.67	**7.78**	7.52	5.79	7.66	7.18	5.26	7.33	7.80
		σ	0.06	**0.07**	0.06	0.05	0.08	0.07	0.06	0.08	0.07
	10cv	*e* (%)	9.55	**8.69**	9.07	10.41	8.87	10.19	10.48	8.76	10.23
		σ	0.08	**0.07**	0.07	0.09	0.08	0.08	0.10	0.08	0.08

Since the dataset is classified based on secondary structure contents, the classification problem resembles the protein structural class prediction problem (for more information on protein structural class see section 3.1.2) that is a difficult open problem for low homology datasets [17] [40]. The model configuration discussed in this section does not lead to improvement in content prediction compared to the base MLR, since sequences are classified into wrong classes and thus wrong MLRs are applied to perform the content prediction task. The following sub-section presents a different approach to derive specialized prediction models, which addresses this problem by eliminating the need to classification.

6.2.2 Grouping based on an Attribute

The objective is to group sequences based on a feature that is derived from the primary structure only. We experimented with grouping based on the sequence length and average hydrophobicity. The following two paragraphs describe the motivation behind choosing these two criteria for grouping. Equations (**22**) and (**23**) (section 6.2.1) give the formal definition of the problem studied for this goal.

° Sequence Length

Grouping based on length was reported to be useful in structural fragment classification [13]. A structural fragment is defined as the longest fragment of protein sequence corresponding to the same secondary structure. This was our motivation to derive specific MLR models for sequences of different length.

° Sequence Average Hydrophobicity

The motivation behind picking average hydrophobicity is to investigate whether there is a relation between hydrophobic tendency of a protein and its helix/strand content. Several other researchers also acknowledged the importance of the relation between hydrophobicity and the structure [45] [24].

As both Table 15 and Table 16 show, the models tend to be over-fitted when the number of groups increases; the best results are shown in bold. Grouping based on sequence length leads to better results when compared with average hydrophobicity. However, the base MLR model is superior to both approaches discussed in this sub-section.

Table 15. Prediction results when grouping sequences by length

	Helix				Strand			
# of groups	Re-substitution		10CV		Re-substitution		10CV	
	e (%)	σ	*e* (%)	σ	*e* (%)	σ	*e* (%)	σ
2	**10.31**	**0.09**	**11.19**	**0.09**	**7.36**	**0.06**	**8.27**	**0.07**
3	10.10	0.08	11.38	0.10	6.96	0.06	8.51	0.07
4	9.80	0.08	11.69	0.10	6.67	0.06	8.71	0.07
5	9.51	0.08	11.83	0.10	6.33	0.05	9.05	0.08
6	9.44	0.08	12.18	0.10	6.21	0.05	9.70	0.08
7	9.07	0.07	12.31	0.10	5.69	0.05	9.78	0.09
8	8.83	0.07	12.61	0.11	5.58	0.05	10.68	0.09
9	8.63	0.07	12.98	0.11	5.19	0.05	11.00	0.10
10	8.51	0.07	13.23	0.11	4.99	0.04	11.92	0.10

Table 16. Prediction results when grouping sequences by average hydrophobicity

	Helix				Strand			
# of groups	Re-substitution		10CV		Re-substitution		10CV	
	e (%)	σ	*e* (%)	σ	*e* (%)	σ	*e* (%)	σ
2	**10.31**	**0.09**	**11.31**	**0.10**	**7.44**	**0.06**	**8.43**	**0.07**
3	10.10	0.08	11.39	0.10	7.15	0.06	8.70	0.07
4	9.87	0.08	11.63	0.10	6.95	0.06	9.16	0.08
5	9.56	0.08	11.82	0.10	6.56	0.05	9.44	0.08
6	9.41	0.08	12.11	0.10	6.23	0.05	9.84	0.08
7	9.17	0.08	12.60	0.11	5.94	0.05	10.11	0.09
8	8.94	0.07	12.81	0.11	5.64	0.05	10.92	0.10
9	8.84	0.07	13.42	0.11	5.29	0.04	11.51	0.10
10	8.45	0.07	13.80	0.12	4.95	0.04	12.13	0.10

6.3 Goal2

In this section the effect of feature selection and feature extraction on the content prediction accuracy is investigated. One feature extraction approach and two feature selection techniques, which are specific to regression-based models, are followed.

6.3.1 Feature Selection

The objective is to choose a concise subset of the aggregated set of 140 attributes. First we present the formal definition of the employed attribute selection procedures and then we report the prediction results when each procedure was applied.

1. **Forward Selection**

This method starts with the feature that has the highest absolute correlation coefficient value with the predicted content. The second feature is added to the model such that it leads to the best two-variable model given that the first feature is already included in the model [9]. Thus the method proceeds by adding one attribute at a time provided that the attribute leads to the least Residual Sum of Squares (*RSS*) compared to other attributes not in the model. *RSS* is computed according to the following equation:

$$RSS = \sum_{i=1}^{S} (y_i^{obs} - y_i^{pred})^2 \qquad \textbf{(26)}$$

where y_i^{obs} and y_i^{pred} are the observed (true) and predicted secondary structure contents for sequence *i*, respectively. The stopping criterion is based on the *F*-statistic for testing the hypothesis that the estimated coefficient for the attribute to be added is zero [9]. Therefore, assuming that *k* features are already in the model; attribute *j* is added to the model if the following inequality is satisfied [9]:

$$FStat_j = \max(\frac{RSS_k - RSS_{k+1}}{RMS_{k+1}}) > FStat_{in} \qquad (27)$$

where RSS_k is the residual sum of squares for the current model, and RSS_{k+1} and RMS_{k+1} are the residual sum of squares and residual mean of squares, respectively if attribute j is included in the model. The value $FStat_{in} = 2$ is usually used regardless of the degree of freedom [9].

2. Backward Elimination

This method begins by fitting a model using all features. Next, it eliminates the least significant feature, which is measured by the magnitude of the *t*-statistic, at each step [9]. Given that the model includes *k* attributes; attribute *j* is considered for deletion according to the following inequality (Equation (**26**) defines *RSS*):

$$FStat_j = \min(\frac{RSS_{k-1} - RSS_k}{RMS_k}) < FStat_{out} \qquad (28)$$

where RSS_{k-1} is the residual sum of squares if feature j is eliminated from the model, and RSS_k and RMS_k are the residual sum of squares and residual mean of squares for the current model, respectively. The value $FStat_{out} = 2$ is often used [9].

The forward selection and backward elimination algorithms were applied to PDBSelect25 to select a subset of attributes in Table A 2. The selected subsets, which were utilized to encode sequences in LiPanLH to perform content prediction, are presented in Table 17.

Table 17. Selected attributes for each structure

Attribute Selection Algorithm	Structure	Number of Selected Attributes	Selected Attributes
Forward Selection	Helix	47	1 3 4 7 13 15 20 21 22 24 27 29 32 34 43 44 49 52 54 56 60 63 66 69 71 73 96 97 98 99 100 102 104 105 106 107 108 114 117 118 119 133 134 135 137 138 140
	Strand	56	4 5 13 19 21 22 24 25 27 29 31 32 33 34 39 41 42 43 45 46 47 48 49 50 52 55 56 57 58 59 60 62 63 65 71 84 92 97 98 99 101 103 104 105 106 107 108 110 113 119 120 133 134 135 136 138
Backward Elimination	Helix	89	1 2 3 5 6 7 12 13 14 19 22 24 29 30 31 32 33 34 35 36 38 39 40 41 43 44 51 53 54 55 58 61 64 65 66 68 75 76 79 80 81 82 83 87 88 89 94 96 97 98 99 100 102 103 104 105 106 107 108 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
	Strand	88	1 2 6 11 13 14 15 17 18 20 21 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 70 71 72 74 76 77 78 79 80 81 82 88 91 94 97 98 99 101 102 104 105 106 107 108 115 119 120 123 125 133 134 135 136 138

Table 18 shows the helix and strand content prediction errors when using attributes in Table A 5 and Table A 6 to derive base MLR models for helix and strand content prediction on LiPanLH (see section 5.2.2).

Table 18. Prediction results on LiPanLH using base MLR model

Helix				Strand			
Re-substitution		10CV		Re-substitution		10CV	
e (%)	σ	*e* (%)	σ	*e* (%)	σ	*e* (%)	σ
7.49	0.06	9.77	0.08	6.10	0.05	7.93	0.07

According to Table 19, applying forward selection leads to lower prediction errors for both helix and strand compared to backward elimination and the base MLR model. Moreover this attribute selection algorithm gives smaller subset of attributes.

Table 19. Prediction results on LiPanLH using the selected attributes

Attribute Selection Algorithm	Helix				Strand			
	Re-substitution		10CV		Re-substitution		10CV	
	e (%)	σ	*e* (%)	σ	*e* (%)	σ	*e* (%)	σ
Forward Selection	7.83	0.06	9.10	0.07	6.26	0.05	7.59	0.07
Backward Elimination	7.10	0.05	9.83	0.08	5.86	0.05	8.24	0.07

6.3.2 Feature Extraction

°Principal Component Analysis (PCA)

In Chapter 4 we discussed the problem of collinearity among features in an MLR model. One way to deal with this issue is to express the model based on a new set of attributes that are linear transformations of the original attributes [9]. Based on MLR formulation originally presented in Chapter 4, the following equation holds (we present the analysis only for helix since for strand the steps are identical):

$$Y_\alpha = FA \qquad \textbf{(29)}$$

where Y_α is a vector containing the true helix content for proteins in training set, matrix *F* is the feature matrix and *A* represents the vector of coefficients to be determined. Let *C* be the correlation matrix of *F*, and *E* be the orthogonal matrix of eigenvectors of *C*. The new model is formulated as follows:

$$Y_\alpha = PX \qquad \textbf{(30)}$$

where *P* = *FE*. The new *orthogonal* features are defined by columns of *P* and are called the *Principal Components*. Matrix *X* represents the coefficients in the transformed model. In principal component regression, some features can be eliminated based on the magnitude of the eigenvalues of *E*. Usually features $P_1P_2...P_k$ are deleted from the model if their corresponding eigenvalues $\lambda_1\lambda_2...\lambda_k$ are small.

The PCA was applied to PDBSelect25 (using all the 140 attributes) and the eigenvalue thresholds leading to the least prediction error were reported for each of helix and strand, i.e. 8.1776e-009 and 2.0029e-008. Next the transformation matrix *E* and thresholds were applied to build MLR prediction models on LiPanLH (Table 20). Comparing the base MLR results on LiPanLH (Table 18), i.e. e=9.77% and e=7.93% for helix and strand respectively, it is concluded that the base MLR leads to better prediction accuracy.

Table 20. Prediction results using PCA on LiPanLH

Helix				Strand			
Re-substitution		10CV		Re-substitution		10CV	
e (%)	σ	e (%)	σ	e (%)	σ	e (%)	σ
7.78	0.06	9.90	0.08	6.60	0.05	8.75	0.07

6.4 Goal3

This goal is to design new attributes that provide the MLR prediction models with more useful information about the secondary structure content. The new attributes are then added to the original feature set, which is used in section 6.1.

6.4.1 Secondary Structure-Derived Polypeptides

The objective is to find a set of polypeptides (all of the same length) that are commonly observed in each of helix and strand structures. One attribute for a sequence is the sum of frequencies of each polypeptide in the corresponding protein sequence. Therefore, considering polypeptides of the same length, *two* new attributes are derived; one for each structure. We experimented with di-peptides, tri-peptides, and tetra-peptides. PDBSelect90 (see section 5.1) was used to derive the set of polypeptides. This dataset contains sequences with high homology. The sequences highly homologous to sequences in PDBSelect25, i.e. more than 40%, were removed from the dataset. As a result 4,746 sequences were kept. This dataset is referred to as PDBSelect90LH from now on. The procedure to extract the set of polypeptides (i.e. di, tri, and tetra) follows:

1. Scan PDBSelect90LH and report the frequency of each polypeptide observed in helix across the entire dataset;
2. Normalize each frequency to the frequency of the corresponding polypeptide disregarding the secondary structure;
3. Sort polypeptides with regard to the normalized frequencies;
4. Based upon a threshold, keep the most frequent polypeptides.

Given that μ and σ' represent the mean and standard deviation of the frequencies respectively, the cut-off points considered for helix and strand polypeptides separately, are as follows:

$$\mu + k\sigma' \text{ where } k \in \{0.1, 0.2, 0.3, \ldots, 1\}$$

The sets of di-peptides and tri-peptides mainly observed in helix (strand) include di-peptides with *all* residues in helix (strand). However for tetra-peptides two different sets for each structure are considered: a set including tetra-peptides with *three* residues of the same secondary structure, and a set including tetra-peptides with *all four* residues of the same secondary structure.

Table 21 and Table 22 show the best result, for PDBSelect25, obtained by trying all cut-off points and adding the two new attributes to the original feature set for helix and strand content prediction MLRs, respectively; the best results are shown in bold.

The tables show that the lowest helix (strand) content prediction error, i.e. 10.82% (8.10%) 10CV error, is achieved by adding the tetra-peptide (with *all four* residues of the same secondary structure) based attributes. As described in section 6.1, the base MLR model gives 11.08% and 8.30% 10CV errors for helix and strand content predictions. According to a paired t-test, improvements achieved by adding the two new attributes are statistically significant at 99.75% (*t*=4.186) and 99.95% (*t*=12.947) confidence levels for helix and strand, respectively, when compared to results of the base MLR model.

Table 21. Prediction results for helix with two new attributes

Polypeptide	Number of residues of the same secondary structure	Cut-off point	Helix			
			Re-substitution		10CV	
			e (%)	σ	e (%)	σ
Tetra-peptide	3	$\mu+\sigma'$ for helix; $\mu+0.3\sigma'$ for strand	10.63	0.10	11.01	0.09
Tetra-peptide	4	$\mu+0.8\sigma'$ for helix; $\mu+0.7\sigma'$ for strand	**10.45**	**0.09**	**10.82**	**0.09**
Tri-peptide	3	$\mu+0.2\sigma'$ for helix; $\mu+0.8\sigma'$ for strand	10.50	0.09	10.91	0.09
Di-peptide	2	$\mu+0.8\sigma'$ for helix; $\mu+0.7\sigma'$ for strand	10.58	0.09	10.99	0.09

Table 22. Prediction results for strand with two new attributes

Polypeptide	Number of residues of the same secondary structure	Cut-off point	Strand			
			Re-substitution		10CV	
			e (%)	σ	e (%)	σ
Tetra-peptide	3	$\mu+0.7\sigma'$ for helix; $\mu+0.1\sigma'$ for strand	7.80	0.06	8.25	0.07
Tetra-peptide	4	$\mu+0.6\sigma'$ for helix; $\mu+0.2\sigma'$ for strand	**7.66**	**0.06**	**8.10**	**0.07**
Tri-peptide	3	$\mu+0.2\sigma'$ for helix; $\mu+0.5\sigma'$ for strand	7.73	0.06	8.16	0.07
Di-peptide	2	$\mu+0.4\sigma'$ for helix; $\mu+0.8\sigma'$ for strand	7.80	0.06	8.22	0.07

6.5 Goal4

The last goal aims at studying the effect of non-linear base functions on the accuracy of the regression based model. The functions experimented with are as follows:

1. Quadratic, Cubic and 4th order regression:

$$y_\alpha = a_0 + \sum_{j=1}^{k}\sum_{i=1}^{n} a_{ji} f_i^{\,j} \qquad (31)$$

$$y_\beta = b_0 + \sum_{j=1}^{k}\sum_{i=1}^{n} b_{ji} f_i^{\,j} \qquad (32)$$

k determines the order of the regression function, i.e. *k*=2 for quadratic, *k*=3 for cubic and *k*=4 for 4th order regression.

2. Exponential:

$$y_\alpha = (1 + \exp(a_0 + \sum_{i=1}^{n} f_i a_i))^{-1} \qquad (33)$$

$$y_\beta = (1 + \exp(b_0 + \sum_{i=1}^{n} f_i b_i))^{-1} \qquad (34)$$

3. Fourier Transform Coefficients:

$$y_\alpha = a_0 + \sum_{i=1}^{n} \sin(f_i \pi) a_{1i} + \cos(f_i \pi) a_{2i} \qquad (35)$$

$$y_\beta = b_0 + \sum_{i=1}^{n} \sin(f_i \pi) b_{1i} + \cos(f_i \pi) b_{2i} \qquad (36)$$

In the above equations, y_α and y_β are the estimated helix and strand contents respectively; f_i is the feature vector of a protein sequence with *n* elements; *a* and *b* are the vectors of coefficients to be estimated by the helix and strand content prediction models, respectively.

Table 23 shows the result for each non-linear function; the best results are shown in bold. For higher order linear regressions, i.e. quadratic, cubic and 4th order regression, increasing the order makes the model more over-fitted. The exponential based helix content prediction model converged in neither re-substitution nor 10CV tests. The model based on this function significantly deteriorates the strand content prediction accuracy. The Fourier Transform Coefficients is slightly inferior to the MLR model. Finally, it is observed that the Quadratic function is the most accurate among the tested functions. However the MLR model is superior (*e*=11.08% for helix and *e*=8.30 for strand) to all the non-linear functions.

Table 23. Effect of different base function on the regression model

Non-Linear Function	Helix				Strand			
	Re-substitution		10CV		Re-substitution		10CV	
	e (%)	σ	*e* (%)	σ	*e* (%)	σ	*e* (%)	σ
4th order regression	9.44	0.08	11.64	0.10	6.65	0.06	9.39	0.10
Cubic	9.90	0.08	11.70	0.10	6.98	0.06	8.84	0.08
Quadratic	**10.32**	**0.08**	**11.33**	**0.10**	**7.43**	**0.06**	**8.46**	**0.07**
Exponential	10.72	0.09	11.37	0.10	18.84	0.15	18.87	0.15
Fourier Transform Coefficients	10.63	0.09	11.85	0.10	7.62	0.06	8.62	0.07

6.6 Comparison with Related Work

Table 24 summarizes the best results achieved for different goals. The most successful configurations from each goal are applied on LiPanLH (which is a low homology dataset containing sequences with low homology to sequences in PDBSelect25) and Table 25 presents the outcomes. Moreover each table presents the result of Li and Pan's approach [24] and Zhang et al.'s [45] on the corresponding datasets; the best results are shown in bold. Table 26 and Table 27 show results of the t-test based significance test. It shows which of our proposed approaches lead to significantly better results than those reported by Li and Pan [24], and Zhang et al. [45]. A negative *t* value shows that the corresponding model approach is inferior to the model compared with. The significant results are shown in bold.

Table 24. Summary of best results by different goals on PDBSelect25

Method	Helix				Strand			
	Re-substitution		10CV		Re-substitution		10CV	
	e (%)	σ	*e* (%)	σ	*e* (%)	σ	*e* (%)	σ
Base MLR	10.67	0.09	11.08	0.09	8.23	0.06	8.30	0.07
Goal 1: K-Means + DA (two clusters)	11.11	0.10	11.87	0.11	7.78	0.07	8.69	0.07
Goal 1: divisions based on length (two groups)	10.31	0.09	11.19	0.09	7.36	0.06	8.27	0.07
Goal 3: Tetra-peptides of the same secondary structure	**10.45**	**0.09**	**10.82**	**0.09**	**7.66**	**0.06**	**8.10**	**0.07**
Goal 4: Quadratic regression	10.32	0.08	11.33	0.10	7.43	0.06	8.46	0.07
Li and Pan's approach [24]	11.38	0.10	11.60	0.10	8.58	0.07	8.72	0.07
Zhang et al.'s approach [45]	11.40	0.09	11.56	0.09	8.63	0.07	8.76	0.07

Table 25. Prediction results on LiPanLH dataset

Method	Helix				Strand			
	Re-substitution		10CV		Re-substitution		10CV	
	e (%)	σ	*e* (%)	σ	*e* (%)	σ	*e* (%)	σ
Base MLR	7.49	0.06	9.77	0.08	6.10	0.05	7.93	0.07
Goal 1: K-Means + DA (two clusters)	6.17	0.07	10.68	0.09	5.68	0.06	9.49	0.09
Goal 1: divisions based on length (two groups)	6.15	0.05	10.36	0.09	5.35	0.04	9.00	0.07
Goal 2: Forward selection	**7.83**	**0.06**	**9.10**	**0.07**	**6.26**	**0.05**	**7.59**	**0.07**
Goal 2: PCA	7.78	0.06	9.90	0.08	6.60	0.05	8.75	0.07
Goal 3: tetra-peptides of the same secondary structure	**7.21**	**0.06**	**9.16**	**0.07**	**5.88**	**0.05**	**7.53**	**0.06**
Goal 4: Quadratic regression	6.14	0.05	10.81	0.10	4.95	0.04	8.85	0.09
Li and Pan's approach [24]	8.29	0.07	9.10	0.08	6.85	0.06	7.71	0.07
Zhang et al.'s approach [45]	8.46	0.07	9.23	0.08	6.89	0.06	7.54	0.06
Combination of Goal 2 and new attributes from Goal 3 (CM)	**7.39**	**0.06**	**8.75**	**0.07**	**6.02**	**0.05**	**7.38**	**0.07**

Table 26. Paired t-test results on PDBSelect25

Method	Helix				Strand			
	Li and Pan's [24]		Zhang et al.'s [45]		Li and Pan's [24]		Zhang et al.'s [45]	
	Significant	Level	Significant	Level	Significant	Level	Significant	Level
Goal 3*	**Yes (*t*=4.134)**	**99.75%**	**Yes (*t*=3.916)**	**99.75%**	**Yes (*t*=9.255)**	**99.95%**	**Yes (*t*=8.437)**	**99.95%**
Base MLR	**Yes (*t*=3.238)**	**99.5%**	**Yes (*t*=3.019)**	**99%**	**Yes (*t*=6.306)**	**99.95%**	**Yes (*t*=5.839)**	**99.95%**

*: tetra-peptides of the same secondary structure

Table 27. Paired t-test results on LiPanLH

Method	Helix				Strand			
	Li and Pan's [24]		Zhang et al.'s [45]		Li and Pan's [24]		Zhang et al.'s [45]	
	Significant	Level	Significant	Level	Significant	Level	Significant	Level
Goal 3*	No (*t*=-0.642)	N/A	No (*t*=-0.708)	N/A	No (*t*=0.120)	<75%	No (*t*=-0.416)	N/A
Base MLR	No (*t*=-1.258)	N/A	No (*t*=-1.533)	N/A	No (*t*=-0.905)	N/A	No (*t*=-1.579)	N/A
Goal2: Forward Selection	No (*t*=0.287)	<75%	No (*t*=0.625)	<75%	No (*t*=0.291)	<75%	No (*t*=-0.566)	N/A
Goal2+ Goal3* (CM)	**Yes (*t*=2.088)**	**95%**	**Yes (*t*=3.559)**	**99.5%**	No (*t*= 1.716)	90%	No (*t*=0.779)	75%

*: tetra-peptides of the same secondary structure

Table 24 shows that our base MLR model is superior to the approaches proposed by Li and Pan [24], and Zhang et al. [45] on PDBSelect25. However it does not achieve better results on the LiPanLH dataset (Table 25). The new tetra-peptide based attributes proposed in Goal 3 lead to significantly better results for PDBSelect25, but according to the paired t-test they did not result in statistically significant improvement in accuracy on the LiPanLH dataset. The forward selection algorithm, which was trained on PDBSelect25, reported concise sets of attributes for helix and strand content prediction (Table 17) and according to Table 25 it results in better content prediction accuracy on LiPanLH dataset compared with both competing methods. However as Table 27 presents, the selected sets of attributes do not achieve statistically significant better prediction.

Based on these results, we used a combination of our two best prediction models, i.e. the forward attribute selection and tetra-peptide based attributes, to build a prediction model that potentially can achieve higher accuracy. The following diagram (Figure 6.1) depicts the steps to build the combined model (CM). Table 25 shows the results of the CM model on LiPanLH dataset. Table 27 shows results of t-test based significance test for the CM method. The test shows that the improvements in results are statistically significant, compared to both competing methods for helix content prediction and better (but not significantly) for the strand prediction. The more significant improvements for helices are due to the design of attributes used in this project to encode protein sequences, which are focused on describing local regions in the sequences. The beta sheets are more difficult to correctly predict since they are formed from strands that are far apart in a protein primary sequence. Helices, on the other hand, are more local structures when compared with the sheets [20].

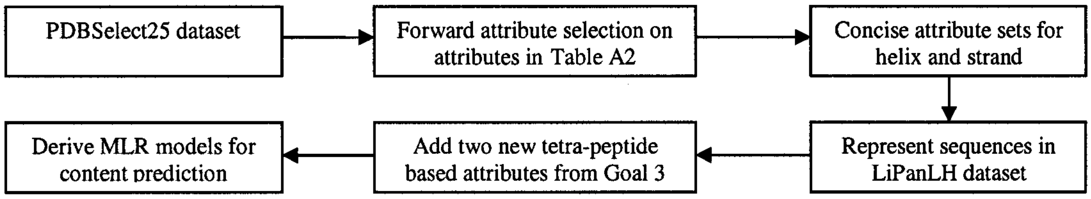
PDBSelect25 dataset
Forward attribute selection on attributes in Table A2
Concise attribute sets for helix and strand
Represent sequences in LiPanLH dataset
Add two new tetra-peptide based attributes from Goal 3
Derive MLR models for content prediction

Figure 6.1. Steps taken to build the CM model

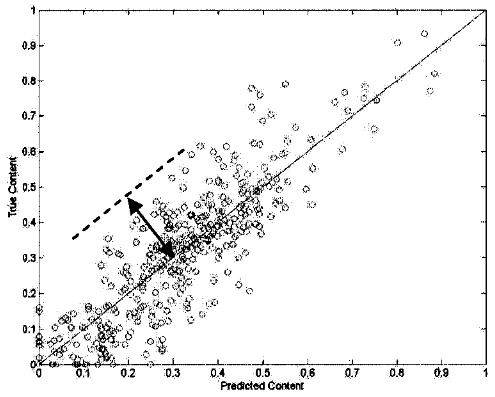

Figure 6.2. Predicted vs. true helix contents using CM model

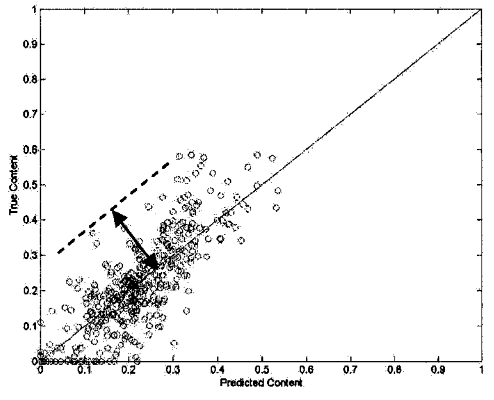

Figure 6.3. Predicted vs. true strand contents using CM model

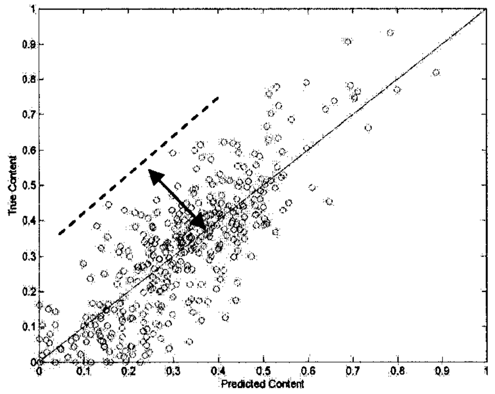

Figure 6.4. Predicted vs. true helix contents using Li & Pan's approach [24]

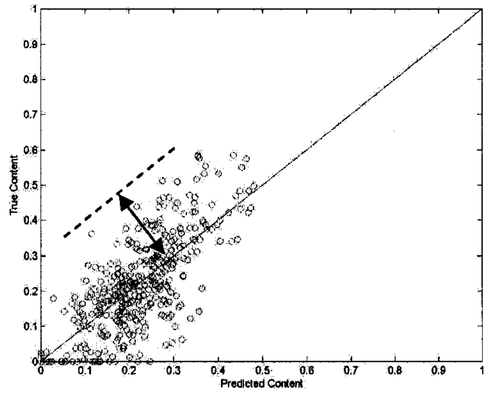

Figure 6.5. Predicted vs. true strand contents using Li & Pan's approach [24]

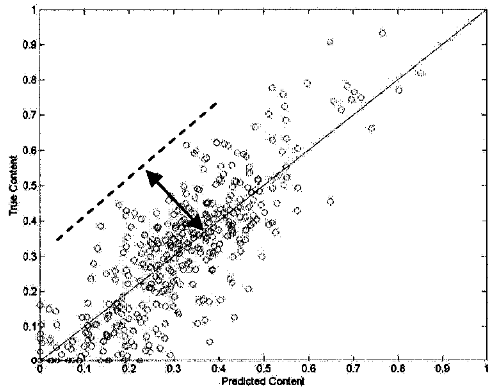

Figure 6.6. Predicted vs. true helix contents using Zhang et al.'s approach [45]

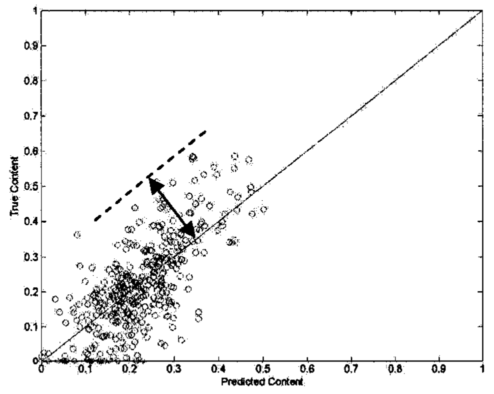

Figure 6.7. Predicted vs. true strand contents using Zhang et al.'s approach [45]

Scatter plots shown in Figure 6.2 through Figure 6.7 depict predicted versus true contents when applying CM, Li and Pan's [24], and Zhang et al.'s [45] prediction models on LiPanLH dataset. These graphs are used to visually contrast the effects of higher prediction accuracy achieved by applying CM model with the prediction obtained by using competing models. The diagonal line shows the ideal situation when 100% content prediction accuracy is achieved. As the dashed lines show, the result of CM model for helix content prediction is more centered around the diagonal, i.e. closer to the ideal case, when compared with the two other methods. However, as expected, this difference is less noticeable for strand content prediction.

To summarize, this chapter showed prediction results using the base MLR model. It also presented the formal definition of each project goal implemented to improve the prediction accuracy of the base MLR model. At the end, the best results from each goal were compared with state-of-the-art published work and a paired t-test was performed to see if the improvement in accuracy was significant. It is concluded that the CM model combining the forward feature selection and the tetra-peptide based attributes leads to significant improvement in content prediction accuracy for the LiPanLH dataset. The next chapter summarizes the thesis and discusses directions to extend this research.

Chapter 7 Summary and Conclusion

This chapter summarizes the work that was done in this thesis and draws the conclusions. Finally, it provides some directions for the feature work.

7.1 Discussion

In this project we studied the problem of protein secondary structure content prediction for low homology protein sequences. The aim was to improve the content prediction accuracy. To this end, we implemented several novel prediction systems and compared their quality with the published research. A comprehensive aggregated set of features was used to encode protein sequences. The experiments were performed on the low homology dataset PDBSelect25 [11]. At first, two base MLR models were derived for helix and strand content prediction tasks (section 6.1). Then four goals were defined and pursued to improve the results of the base MLRs. We tried specialized prediction models (Goal 1 section 6.2), attribute selection (Goal 2 section 6.3), new attribute design (Goal 3 section 6.4), and finally non-linear base functions for our regression based prediction method (Goal 4 section 6.5).

The best results from each goal were compared with the results published by Li and Pan [24] and Zhang et al. [45]. The latter two papers were identified as superior when compared with other related contributions for this problem. The paired t-test showed that targeting attributes, whether by new attribute design or attribute selection, is the best point of attack to improve the quality of protein secondary structure content prediction.

The combination of forward attribute selection algorithm and the two new quara-peptide based attributes, led to significantly better results when compared with state-of-the-art published research.

7.2 Feature work

According to our experiments, designing auto-correlation functions that target strands could lead to content prediction improvement. The challenge is to find criteria to locate possible strand occurrences in a sequence. Given that, our simulated biased experiment indicates that about 2% could be gained for both helix and content predictions. Therefore, our feature work will focus on this issue.

Appendix

Table A 1. Amino acids, their codes and their indices used to encode protein sequences

AA	3-letter Code	1-letter Code	*MolW*	*pI*	*FH*	*EH*	*M*	*Hp*
Alanine	Ala	A	71.0791	6.01	0.42	0.62	0.115	1.8
Arginine	Arg	R	156.188	10.76	-1.37	-2.53	0.777	-4.5
Asparagine	Asn	N	114.104	5.41	-0.82	-0.78	0.446	-3.5
Aspartate	Asp	D	115.0887	2.77	-1.05	-0.90	0.446	-3.5
Cysteine	Cys	C	103.1437	5.07	1.34	0.29	0.36	2.5
Glutamine	Gln	Q	128.131	5.65	-0.30	-0.85	0.55	-3.5
Glutamate	Glu	E	129.1157	3.22	-0.87	-0.74	0.55	-3.5
Glycine	Gly	G	57.0521	5.97	0.00	0.48	0.0007	-0.4
Histidine	His	H	137.1414	7.59	0.18	-0.40	0.63	-3.2
Isoleucine	Ile	I	113.16	6.02	2.46	1.38	0.13	4.5
Leucine	Leu	L	113.16	5.98	2.32	1.06	0.13	3.8
Lysine	Lys	K	128.1792	9.74	-1.35	-1.50	0.48	-3.9
Methionine	Met	M	131.1977	5.47	1.68	0.64	0.577	1.9
Phenylalanin	Phe	F	147.1772	5.48	2.44	1.19	0.7	2.8
Proline	Pro	P	97.1171	6.48	0.98	0.12	0.323	-1.6
Serine	Ser	S	87.0784	5.68	-0.05	-0.18	0.238	-0.8
Threonine	Thr	T	101.1054	5.87	0.35	-0.05	0.346	-0.7
Tryptophan	Trp	W	186.2139	5.89	3.07	0.81	1	-0.9
Tyrosine	Tyr	Y	163.1756	5.67	1.31	0.26	0.82	-1.3
Valine	Val	V	99.133	5.97	1.66	1.08	0.33	4.2

Table A 2. All the attributes used for protein representation

Feature	Abbreviation	Indices
Protein sequence length	*N*	1
Average molecular weight of the sequence	*MW*	2
Average isoelectric point of the sequence	*pI*	3
Composition vector (in alphabetical order)	*CV*	4-23
First order composition moment vector (in alphabetical order)	CMV^1	24-43
Second order Composition moment vector (in alphabetical order)	CMV^2	44-63
R-groups (*AVLIMG, SPTCNQ, KHR, DE, FYW*)	*RG*	64-68
Exchange groups (*AGPST, DENQ, ILM*)	*XG*	69-71
Hydrophobicity groups (*VLIMAFPWYCG, STNQ*)	*HG*	72-73
Electronic groups (*DEPA, LIV, KNR, FYMTQ, GHWS*)	*EG*	74-78
Chemical groups (*C, CAROM, CH,* CH_2,, CH_2RING, CH_3,, *CHAROM, CO, NH, OH*)	*CG*	79-88
Other groups (*DEKHRVLI, DEKHRNTQSYW, FHWY, AGST, AG, FHWYR, NQ*)	*OG*	89-95
Auto-correlation functions based on *FH* (*n*=[1...6])	A_n^{FH}	96-101
Sum over *FH* indices	H_{sum}^{FH}	102
Average of *FH* indices	H_{avr}^{FH}	103
Sum of *FH* averages over each three consecutive AA	H_{sum3}^{FH}	104
Auto-correlation functions based on *EH* (*n*=[1...6])	A_n^{EH}	105-110
Sum over *EH* indices	H_{sum}^{EH}	111
Average of *EH* indices	H_{avr}^{EH}	112
Sum of *EH* averages over each three consecutive AA	H_{sum3}^{EH}	113
Auto-correlation functions based on *M* (*n*=[1...6])	A_n^{M}	114-119
Cumulative density for *FH* indices	$HCum_n^{FH}$	120-125
Cumulative density for *EH* indices	$HCum_n^{EH}$	126-131
Auto-correlation functions based on *Hp* (*n*=[1...9])	A_n^{Hp}	132-140

Table A 3. Attribute list after pairwise collinearity elimination on PDBSelect25 (threshold: 0.90)

6	7	8	10	11	13	14	15	16	17	18	19	20	24	34	45	46	49
50	52	56	59	60	61	62	63	64	66	67	68	73	74	76	77	78	79
81	82	84	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
101	102	103	105	106	107	108	109	110	111	112	117	120	126	132	133	134	135
136	137	138	139														

Table A 4. Attribute list after pairwise collinearity elimination on PDBSelect25 (threshold: 0.99)

1	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	26	27	28	29	30	31	32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
76	77	78	79	80	81	82	84	85	87	88	89	90	91	92	93	94	95
96	97	98	99	100	101	102	103	105	106	107	108	109	110	111	112	114	115
116	117	118	119	125	131	132	133	134	135	136	137	138	139	140			

Table A 5. Attribute list after pairwise collinearity elimination on LiPanLH (threshold: 0.85)

2	4	7	8	10	13	14	17	18	20	21	32	45	46	49	51	55	56
59	60	61	62	63	64	66	67	68	72	73	74	76	77	78	79	81	82
83	84	86	87	88	89	91	92	93	94	95	96	97	98	99	100	101	103
105	106	107	108	109	110	111	112	114	125	131	132	133	134	135	136	137	138
139	140																

Table A 6. Attribute list after pairwise collinearity elimination on LiPanLH (threshold: 0.86)

6	9	11	12	13	15	17	22	23	36	39	40	41	44	45	47	48	50
53	54	58	64	65	66	67	68	69	73	74	75	76	77	78	79	81	82
86	87	88	89	90	93	94	95	96	97	98	99	100	101	102	103	105	106
107	108	109	110	112	113	115	120	127	132	133	134	135	136	137	138	139	140

References

[1] Chou, K.C., Using Pair-Coupled Amino Acid Composition to Predict Protein Secondary Structure Content. *Journal of Protein Chemistry,* 18(4):473-80, May 1999.

[2] Cai, Y.D., Liu, X.J. and Chou, K.C., Prediction of protein secondary structure content by artificial neural network, *Journal of Computational Chemistry.* April 30;24 (6):727-31, April 2003.

[3] Dayhoff, M.O. (ed.), *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation (Washington, D.C.), 5, 1972.

[4] Eisenberg, D., Weiss, R.M. and Trewilliger T.C., The Hydrophobic Moment Detects Priodicity in Protein Hydrophobicity, *Proceedings of National Academy of Science*, 81:1, pp 140-144, 1984.

[5] Eisenhaber, F., Imperiale, F., Argos, P. and Frommel, C., Prediction of Secondary Structural Contents of Proteins from Their Amino Acid Composition Alone, I New Analytic Vector Decomposition Methods, *Proteins*, 25:2, 157-168, 1996.

[6] Eyrich, A.V., Przybylski, D., Koh, I.Y.Y., Grana, O., Pazos, F., Valencia, F. and Rost, B., CAFASP3 in the Spotlight of EVA, *Proteins*, 53 S6, 548-560, 2003.

[7] Ganapathiraju, M.K., Klein-Seetharaman, J., Balakrishnan, N. and Reddy, R., Characterization of Protein Secondary Structure, *IEEE Signal Processing Magazine*, 78-87, May 2004.

[8] Han, J., *Data mining: Concepts and Techniques*, San Francisco: Morgan Kaufmann Publishers, 2001.

[9] Hocking, R.R., *Methods and Applications of Linear Models: Regression and Analysis of Variance,* Wiley Series in Probability and Statistics, 1996.

[10] Hill, T. and Lewicki, P., *STATISTICS Methods and Applications*, StatSoft, Tulsa, OK, 2006.

[11] Hobohm, U., and Sander, C., Enlarged representative set of protein structures, *Protein Science* 3 (1994) 522.

[12] Hobohm, U., and Sander, C., A Sequence Property Approach to Searching Protein Databases, *Journal. of Molecular Biology*, 251, 390-399, 1995.

[13] Kedarisetti, K.D., Computational Prediction of Three State Secondary Structure for Protein Structural Fragments. *M.Sc. Thesis,* University of Alberta, 2005.

[14] Kyte, J. and Doolitle, R.F., A Simple Method for Displaying the Hydropathic Character of a Protein, *Journal. of Molecular Biology*, 157, pp 105-132, 1982.

[15] Krigbaum, W. R. and Knutton, S. P., Prediction of the Amount of Secondary Structure in a Globular Protein from its Amino Acid Composition, *Proceedings of the National Academy of Science*, 70, 2809-2813, 1973.

[16] Kurgan, L.A. and Homaeian, L., Prediction of Secondary Protein Structure Content from Primary Sequence Alone - a Feature Selection Based Approach, *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition* pp. 334-345. 2005.

[17] Kurgan, L.A. and Homaeian, L., Prediction of Structural Classes for Protein Sequences and Domains- Impact of Prediction Algorithms, Sequence Representation and Homology, and Test Procedures on Accuracy, *J. of Pattern Recognition*, special issue on Bioinformatics, Elsevier, 2006 (to appear).

[18] Larose, D.T. and Hoboken, N.J., *Discovering Knowledge in Data: an Introduction to Data Mining*. Wiley-Interscience, c2005, 2005.

[19] Leberman, R., Secondary Structure of Tobacco Mosaic Virus Protein, *Journal of Molecular Biology*, Volume 55, Issue 1, 23-30, 1970.

[20] Lodish, H., Berk, A., Zipursky, S.L., Matsudaria, P., Baltimore, D. and Darnell, J.E., *Molecular Cell Biology*, 4th ed., W.H. Freeman and Company, New York, 2000.

[21] Liu, W.M. and Chou, K.C., Prediction of Protein Secondary Structure Content, *Journal of Protein Engineering*, 12:12 1041-1050, 1999.

[22] Moore, D.S., *The Basic Practice of Statistics*. New York, W. H. Freeman and Co., 2000.

[23] Lee, S., Lee, B.C. and Kim, D., Prediction of protein secondary structure content using amino acid composition and Evolutionary Information, *Proteins,* Dec 2005.

[24] Lin, Z. and Pan, X., Accurate Prediction of Protein Secondary Structural Content, *Journal of Protein Chemistry*, Vol. 20, No. 3, 2001.

[25] Lewis, P.N., and Scheraga, H.L., Predictions of Structural Homologies in Cytochrome *c* Proteins, *Archives of Biochemistry and Biophysics,* Volume 144, Issue 2, 576-583,1971.

[26] Muskal, S.M. and Kim, S-H., Predicting Protein Secondary Structure Content: a Tandem Neural Network Approach, *Journal of Molecular Biology*, 225, 713-727, 1992.

[27] Michie, A.D., Orengo, C.A. and Thornton, J.M., Analysis of Domain Structural Class Using an Automated Class Assignment Protocol, Journal *Molecular Biology,* 262, 168-185, 1996.

[28] Nelson, D. and Cox, M., *Lehninger Principles of Biochemistry Amino*, Worth Publishers, 2000.

[29] Ptitsyn, 0.B. and Finkelstein, A.V., Connexion between the Secondary and Primary Structures of Globular Proteins, *Biophysics,* 15, 785-796, 1970.

[30] Pilizota, T., Lucic, B. and Trinajstic, N., Use of Variable Selection in Modeling the Secondary Structural Content of Proteins from Their Composition of Amino Acid Residues, *Journal of Chemical Information and Computer Sciences*[15], Jan-Feb;44(1):113-21, 2004.

[31] Przybylski, D. and Rost, B., Alignments Grow, Secondary Structure Prediction Improves. *Proteins,* 1; 46(2):197-205, Feb 2002.

[32] Robson, B., and Pain, R.H., Analysis of the Code Relating Sequence to Conformation in Proteins: Possible Implications for the Mechanism of Formation of Helical Regions, *Journal of Molecular Biology,* Volume 58, Issue 1, 237-257, 1971.

[33] Ruan, J., Wang, K., Yang, J., Kurgan, L.A. and Cios, K., Highly Accurate and Consistent Method for Prediction of Helix and Strand Content from Primary Protein Sequences, *Artificial Intelligence in Medicine*, special issue on *Computational Intelligence Techniques in Bioinformatics*, accepted, 2005.

[34] Specht, D.F., Probabilistic Neural Networks, *Neural Networks*, 3, 109-118. 1990.

[35] Syed, U. and Yona, G., Using a Mixture of Probabilistic Decision Trees for Direct Prediction of Protein Function, *Proceedings of RECOMB 2003 Conference*, 224-234, 2003.

[36] Wang, J., Ma, Q., Shasha, D. and Wu, C.H., Application of Neural Networks to Biological Data Mining: a Case Study in Protein Sequence Classification, *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 305-309, 2000.

[15] Now called *Journal of Chemical Information and Modeling*

[40] Wang Z-X. and Yuan, Z, How Good is the Prediction of Protein Structural Class by the Component-Coupled Method, *Proteins*, 38, 165-175, 2000.

[41] Yang, X. and Wang, B., Weave Amino Acid Sequences for Protein Secondary Structure Prediction, *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 80-87, 2003.

[42] Zhang, C.T., Zhang, Z. and He., Z., Prediction of the Secondary Structure of Globular Proteins Based on Structural Classes, *Journal of Protein Chemistry*, 15, 775-786, 1996.

[43] Zhang, C.T., Lin, Z.S., Zhang, Z. and Yan, M., Prediction of Helix/Strand Content of Globular Proteins Based on Their Primary Sequences, *Protein Engineering*, 11:11, 971-979, 1998a.

[44] Zhang, C.T., Zhang, Z. and He, Z., Prediction of the Secondary Structure Contents of Globular Proteins based on Three Structural Classes, *Journal of Protein Chemistry*, 17, 261-272, 1998b.

[45] Zhang, Z.D., Sun, Z.R. and Zhang, C.T., A New Approach to Predict the Helix/Strand Content of Globular Proteins, *Journal of Theoretical Biology*, 208, 65-78, 2001.