University of Alberta

HYBRID MODELS FOR PROTEIN SECONDARY STRUCTURE CONTENT PREDICTION

by

Mandana Rahbari 

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 2008

Your file Votre référence
ISBN: 978-0-494-47395-5
Our file Notre référence
ISBN: 978-0-494-47395-5

To my beloved parents and my dear Dmitry.

ABSTRACT

Secondary protein structure content prediction is an important intermediate problem in establishing accurate methods for prediction of higher-level structures (2D and 3D). The content was recently applied to prediction of structural classes, folding rates and transition, enzyme proteins and types, and analysis of protein interactions. A new machine learning based protein secondary structure content prediction method named Length Adjusted Models for Improving Content prediction Accuracy (LAMICA) is introduced. The proposed learning models are trained on features extracted from two complementary secondary structure prediction methods, PSIPRED and PROFsec, and further advanced by the addition of physicochemical, energetic and conformational features. Two separate models are created: one for small protein sequences and the other for large sequences. The models are tuned by a two-stage feature selection process. The proposed method outperforms state-of-the-art prediction techniques in terms of content prediction error. In addition we demonstrate that the predicted content can be used to improve protein secondary structure and class predictions.

ACKNOWLEDGEMENT

I express my sincere gratitude to my supervisor Prof. Lukasz Kurgan for his guidance and support. In addition, I would like to thank Prof. Marek Reformat and Prof. Jack Tuszynski for their valuable comments and suggestions. Finally, I would like to thank my colleagues Ke Chen, Kanaka Kedarisetti, Wojciech Stach, Michael Kurgan, Rafal Rak and Hua Zhang for helpful discussions and feedback.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Proteins are among the most important biochemical molecules that provide invaluable structure and services to the host organism. For instance, they form the cytoskeleton (e.g., tubulin), perform signaling (src) and transporting functions (hemoglobin), implement immune responses (antibodies), etc. Proteins are natural polymers consisting of amino acid (AA) units. The number of AAs in protein molecules ranges from several (for short peptides) to thousands. Each protein, at the primary structural level, is a sequence of AAs. The secondary (2D) protein structure corresponds to a specific local, spatial arrangement of AAs caused mostly by hydrogen bonding. Most of the existing secondary protein structure data has been obtained by one of three methods: X-ray and electron crystallography based on a crystallized protein or NMR from a purified protein in solution. High expenses incurred by these two methods, inherent practical restrictions, and high numbers (millions) of unsolved proteins (with known sequences but unknown structure) result in a great demand for more time and cost-effective methods to find protein structure.

National Center for Biotechnology Information (NCBI) database contains over 5 million protein sequences (2008), while Protein Data Bank (PDB), which is the only database that contains three dimensional protein structure, contains just about 50,000 proteins, see Table 1.1.

year	#depositions
before 2000	10991
2000	2983
2001	3286
2002	3563
2003	4830
2004	5508
2005	6678
2006	7282
2007	8128
2008	2782
total	50830

Table 1.1: Statistics for PDB structures that were deposited.

A reasonable and time-efficient candidate to replace the experimental methods is a computational (in-silico) prediction of the 3D and 2D protein structure from its AA sequence. Early in-silico prediction methods were based on local properties associated with neighboring residues of a given sequence, while later research focused on prediction based on global protein properties via multiple sequence alignment. Unfortunately, despite the low cost and high computational efficiency all currently known computational prediction techniques suffer from relatively low prediction accuracy of about 80%.

The lower level of protein structure, referred to as protein secondary structure, consists of three major components: helix, strand, and coil. This level of the structure plays a major role in proteins' conformation. This project intends to predict the quantities of components (helix, coil, and strand) in a protein secondary structure. These quantities are also called secondary structure content. The prediction is based on machine learning techniques. Protein content prediction is an important intermediate problem in establishing accurate methods for prediction of higher level structures.

Protein secondary structure content prediction was tackled in the past

using two machine learning methods: Neural Networks and Regression (Krigbaum and Knutton, 1973; Lin and Pan, 2001; Muskal and Kim, 1992; Ruan *et al.*, 2005). Neural networks are characterized by long learning time and lack of interpretability. Therefore alternative regression type methods received significant attention. In this project a new protein secondary structure content prediction method, called Length Adjusted Models for Improving Content prediction Accuracy (LAMICA) based on support vector machine regression is investigated.

To this end, three goals are addressed in this work:

- to develop a method that provides lower content prediction error compared to existing methods for content prediction and gives more accurate results than content computed from secondary structure predictions.
- to verify whether the features computed from predicted protein secondary structure are useful for content prediction.
- to validate whether the predicted content can be applied to solve other practical problems.

The thesis is organized as follows. Chapter 2 presents the background knowledge about proteins and their structures. It also describes the protein sequence characterization, methods and datasets that are utilized in this project. Chapter 3 reviews the related works. The proposed solution for the problem addressed in this thesis is described in Chapter 4. In Chapter 5 the experimental comparison between our proposed method and state-of-the-art techniques is presented. Chapter 5 also discusses results achieved for the above mentioned goals. Chapter 6 summarizes the contributions and concludes the thesis.

Chapter 2

Background

2.1 Protein Secondary Structure

2.1.1 Amino acids

Polymers are large molecules made from many repeating subunits known as monomers. Twenty amino acids (monomers) are found in our body that compose the protein sequences (polymers). In each amino acid the central carbon is bonded with four different groups: Hydrogen group (−H), Carboxyl group (−COOH), amino group ($-NH_2$) and the side chain symbolized by R, see Fig 2.1. Amino acids vary in their side chain groups.

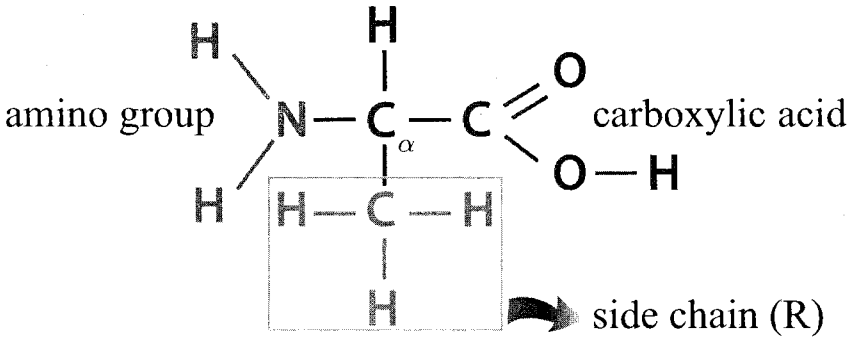

Figure 2.1: The structure of alanine amino acid, with the amino group on the left and the carboxyl group on the right.

amino acid	three-letter code	one-letter code
alanine	ala	A
arginine	arg	R
asparagine	asn	N
aspartic acid	asp	D
cysteine	cys	C
glutamine	gln	Q
glutamic acid	glu	E
glycine	gly	G
histidine	his	H
isoleucine	ile	I
leucine	leu	L
lysine	lys	K
methionine	met	M
phenylalanine	phe	F
proline	pro	P
serine	ser	S
threonine	thr	T
tryptophan	trp	W
tyrosine	tyr	Y
valine	val	V

Table 2.1: Twenty standard amino acids

The twenty amino acids are denoted by either three-letter or one-letter codes. Table 2.1 gives these notations.

To form a protein, amino acids are linked to each other by dehydration synthesis to form peptide bonds. A peptide bond is formed by the reaction of an amino group of one amino acid with the carboxyl group of another amino acid, see Fig 2.2. Each amino acid chain is known as primary amino acid structure of the peptide. A short chain of amino acids, including 20-30 amino acids,

Peptide bond	Average length	Single bond	Average length	Hydrogen bond	Average (+/−30)
C_α - C	1.53 Å	C - C	1.54 Å	O-H - - - O-H	2.80 Å
C - N	1.33 Å	C - N	1.48 Å	N-H - - - O=C	2.90 Å
N - C_α	1.46 Å	C - O	1.43 Å	O-H - - - O=C	2.80 Å

Table 2.2: The average bond lengths in a peptide chain.

linked by peptide bonds is called a peptide. Polypeptides are longer than peptides and they can include as many as 4000 amino acids. Some proteins have one polypeptide chain while others, for example hemoglobin, contain several polypeptide chains. The sequence of amino acids in each polypeptide chain or protein is unique for that protein. The primary amino acid structure (defined in Section 2.1.2) contains the information necessary for folding the peptide chain into its "native structure". If even one amino acid in the polypeptide is changed, it can sometimes result in changing protein's function. For example, sickle cell anemia is caused by a change in only one nucleotide in the DNA sequence that causes just one amino acid in one of the hemoglobin polypeptide molecules to change. As a result, the whole red blood cell ends up being deformed and unable to properly carry oxygen. A few important characteristics of bonds are given in Table 2.2.

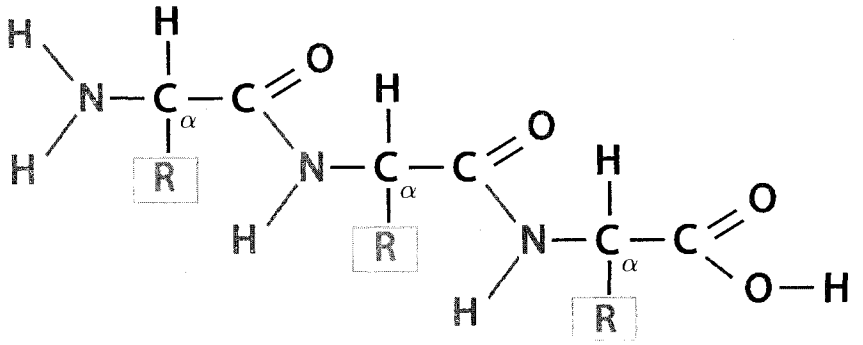

Figure 2.2: The condensation of three amino acids to form a polypeptide chain.

2.1.2 Protein Structure

A polypeptide chain can bend and twist in several ways. Most of the polypeptides fold up into an energetically stable, three dimensional molecule, shortly after being synthesized. Four distinct structural levels are defined below.

Primary Structure

The sequence of amino acids, which are connected through peptide bonds, is called the primary structure of the peptide or protein. The length of the peptide is defined as the number of the amino acids composing its primary structure. The sequence starts at the N-terminal (amino group), and ends with a C-terminal (carboxyl group).

Secondary Structure

Due to interaction between chemical groups in the amino acids, some spatially local 3D patterns frequently occur within a folded protein. These frequently occurring shapes are called protein secondary structures. The most common secondary structures are α-helix, β-sheet and β-turn.

The Dictionary of Secondary Structures of Proteins (DSSP) is a single letter coding method, which is used to describe the protein secondary structures based on hydrogen bonding patterns. DSSP annotates each amino acid as belonging to one of eight secondary structure conformations using eight single letter codes:

- G = 3_{10}-helix. (Minimum length 3 amino acids)
- H = α-helix. (Minimum length 4 amino acids)
- I = Π-helix. (Minimum length 5 amino acids)
- T = hydrogen bonded turn
- E = β sheet. (Minimum length 2 amino acids)

- B = β-bridge
- S = bend
- C/L/" " (space) = coil/loop. (For those amino acids which are not in any of the above conformations)

DSSP eight secondary structure types can be reduced to three state secondary structure by assigning H=HGI, E=EB, C=STC. Helix is represented as *H* and has a spiral shape. Strand is shown by *E* and is a plane-shaped structure and Coil, which is represented by *C*, serves as a linker between helices and strands, see Fig 2.3.

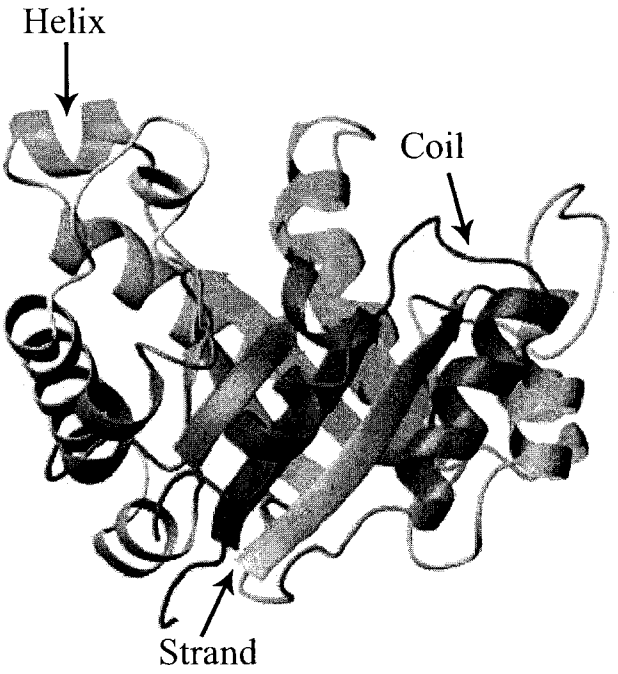

Figure 2.3: Secondary structure components of a protein.

By building models of peptides using known information about bond lengths and angles, the first elements of secondary structure, the α-helix and the β-sheet, were suggested in 1951 by Linus Pauling (Pauling *et al.*, 1951).

Helix Conformation

The helix structure is formed when a polypeptide chain arranges into spiral conformation. A repeated hydrogen bonding between amino group (N-H) of i^{th} amino acid in polypeptide chain and carboxyl group (C=O) of the amino acid n residues earlier (i-n) forms a helix (n equals four for α-helix, and three and five for 3_{10}-helix and Π-helix, respectively). 3_{10}-helix and Π-helix are relatively rare. The helix conformation is tight and there is almost no free space in this structure. The amino acid side chains are on outside and they point slightly downwards. Helices range from four to forty residues, and a typical helix is about ten amino acids long. Fig. 2.4 shows the side and top view of a helix.

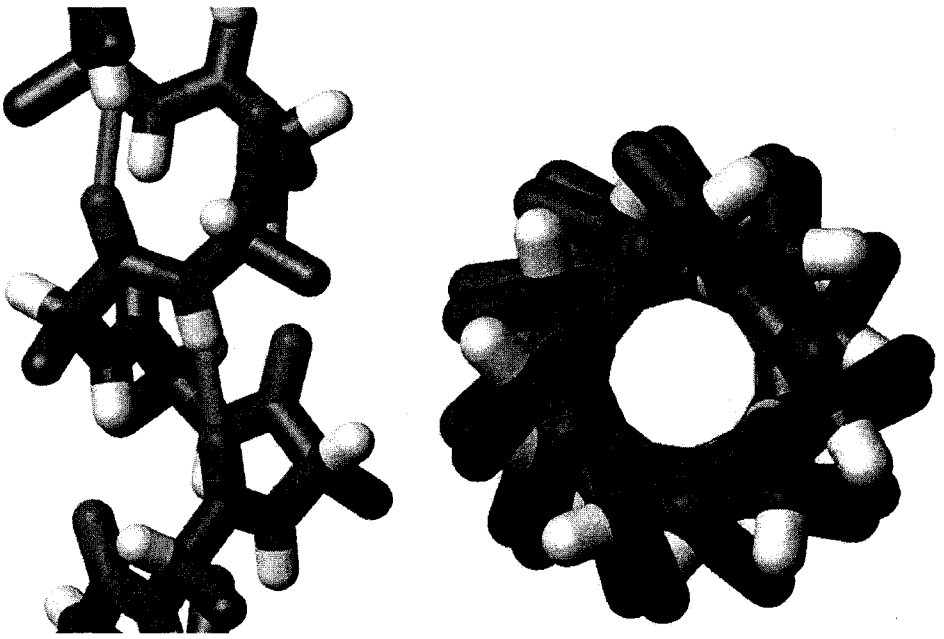

Figure 2.4: Helix. Left panel is the side view of an α-helix composed of alanine residues. Right panel shows the top view of the same α-helix.

Strand Conformation

β-strands are positioned adjacent to the other β-stands and form hydrogen bonds with their neighbors. N-H groups in the backbone of one strand form hydrogen bonds with C=O groups in the backbone of the adjacent strand. This

way a network of hydrogen bonds is established and strands form a sheet. If the strands are in the same direction with respect to the protein sequence, the resulting sheet is called *parallel sheet*, otherwise it is called *antiparallel sheet*. β-strands are typically five to ten amino acids long. The strand conformation is shown in Fig. 2.5.

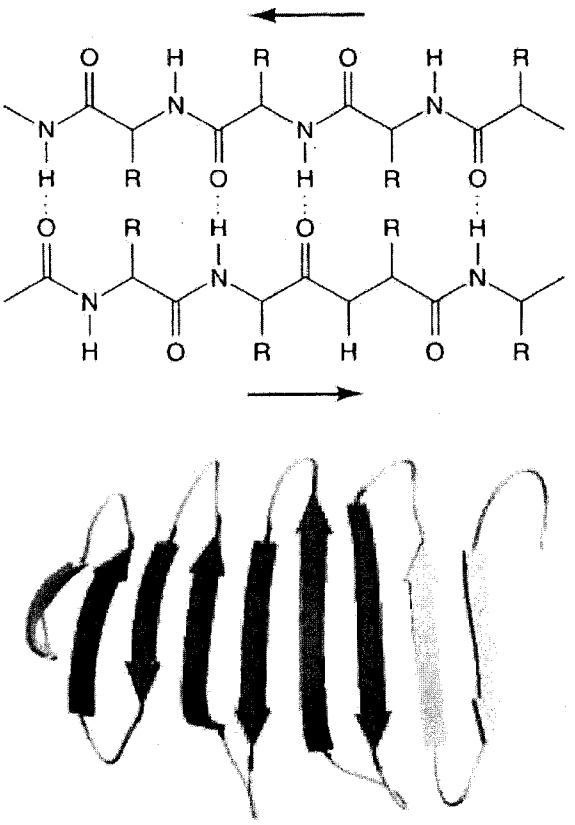

Figure 2.5: Antiparallel sheet.

Coil Conformation

Since coils are non-repetitive irregular structures, it is not easy to describe them by structure. Among three types of coil, turns seem more structurally defined than bends and loops. Coils connect (link) two other secondary structure components and without them a protein would be loosely packed.

Tertiary Structure

The secondary structure elements are usually folded into a more stable compact shape using turns and loops. Although the formation of the tertiary structure is usually driven by energy minimization, burial of the hydrophobic residues and other interactions like hydrogen bonds, ionic interactions, and disulphide bonds also stabilize the tertiary structure. This level of the protein structure depends on the number, size and the arrangement of the secondary structure elements. The information about conformation of the protein is the key for understanding its properties and functions.

Quaternary Structure

The quaternary structure is established based on interactions between multiple polypeptide chains. The individual chains in the quaternary structure are called subunits. Not all proteins have quaternary structure, since they might function as monomers. This structure is stabilized by the same interactions as the tertiary structure.

2.1.3 Classifications of protein structures

Organization of protein structures according to folding pattern imposes a very useful logical structure on the entries in the Protein Data Bank (PDB), which is the centralized, world-wide repository of protein structures (Berman *et al.*, 2000). Several databases derived from the PDB are built around classifications of protein structures. They offer useful features for exploring the protein structures, including search using keywords or sequences, navigation among similar structures at various levels of the classification hierarchy, presentation of structures, exploring the databank for structures similar to a new structure, and links to other sites. These databases include SCOP (Structural Classification Of Proteins) (Murzin *et al.*, 1995), CATH (Class, Architecture, Topology, Homologous superfamily) (Orengo *et al.*, 1997), and FSSP/DDD (Fold classification based on Structure-Structure alignment of Proteins/Dali

Domain Dictionary) (Holm and Sander, 1994, 1998).

Class	folds	superfamilies	families
All α proteins	259	459	772
All β proteins	165	331	679
α and β proteins (α/β)	141	232	736
α and β proteins ($\alpha+\beta$)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

Table 2.3: Statistics for SCOP classification.

SCOP

The SCOP database organizes protein structures in a hierarchy according to evolutionary origin and structural similarity. The lowest level of the SCOP hierarchy includes individual domains, extracted from the PDB entries. Domains are grouped into families of homologues, for which the similarities in structure, sequence, and sometimes function imply a common evolutionary origin. Families containing proteins of similar structure and function, but for which the evidence for evolutionary relationship is suggestive but not compelling, form superfamilies. Superfamilies that share a common folding topology, for at least a large central portion of the structure, are grouped as folds. Finally, each fold group falls into one of the general classes. The major classes in SCOP are all α , all β , $\alpha+\beta$, α/β , and miscellaneous 'small proteins', which often have a small amount of helices and strands and which are held together by disulphide bridges. All α structural class includes proteins with only small content of strands, while all β structural class includes proteins with only small content of helices. Both $\alpha+\beta$ and α/β classes include proteins with helices and strands. $\alpha+\beta$ contains the proteins with mostly antiparallel

Class	helix content (%)	strand content (%)
All-α	> 15	< 10
All-β	< 15	> 10
$\alpha\beta$	> 15	> 10
Irregular	< 15	< 10

Table 2.4: Defined criteria by (Eisenhaber *et al.*, 1996) for assignment of structural classes based on the content of the corresponding secondary structure.

strands, while α/β includes proteins with mostly parallel strands (Nakashima *et al.*, 1986). Sometimes $\alpha+\beta$ and α/β classes are combined into a single $\alpha\beta$ class (Eisenhaber *et al.*, 1996).

The SCOP release of September 2007 contained 34494 PDB entries, split into 97178 domains. The distribution of entries at different levels of the hierarchy is shown in Table 2.3.

Nakashima's Structural Class Assignment

This method, introduced by Nakashima et. al(Nakashima *et al.*, 1986) considers five structural classes (Levitt and Chothia, 1976) (α, β, $\alpha + \beta$, α/β and "irregular"), and analyzes the relation of protein's structural class and its amino acid composition.

In this method a 20-dimensional space is used to represent a protein via its amino acid composition. The frequencies of the 20 amino acids form the 20 coordinates of a point representing a protein in this 20-dimensional space. The distribution of proteins in composition space is investigated and some criteria of assigning the structural class to a protein are defined, which were later adapted by Eisenhaber (Eisenhaber *et al.*, 1996) for the case of four structural classes. Eisenhaber assumes that structural class is defined via secondary structure content according to Table 2.4.

2.1.4 Protein Sequence Characteristics

Secondary structure content prediction is usually performed using an intermediate step, in which the protein primary structure is encoded into its feature space representation. Existing content prediction methods use a limited number of features to represent the protein sequence when compared to other structure prediction techniques. In this work protein features employed for content and structure prediction are aggregated to provide a more comprehensive feature space representation of the protein sequences. In this section, the employed features are explained in detail.

Index-base Attributes

The index-based properties used in this work include molecular weight, isoelectric point, hydrophobicity indices and 48 other index-based diverse amino properties analyzed by Tomii and Kenehisa (Tomii and Kanehisa, 1996).

The *molecular weight* (MolW) of a protein sequence is the result of adding up of its amino acids' molecular weight values plus the weight of a water molecule ($\text{MolW}_{\text{H}_2\text{O}}$), which is approximately 18 daltons (Da).

$$\text{MolW} = \text{MolW}_{\text{H}_2\text{O}} + \sum_{i=1}^{N} \text{MolW}_i, \tag{2.1}$$

where N denotes the length of the protein sequence. Table 2.5 shows the value for MolW.

The amino acid *isoelectric point* property (pI) shows the pH at which a molecule carries no electric charge and therefore it is stationary in an electric field (Nelson and Cox, 2000). Isoelectric point indices are shown in Table 2.5.

The chemical properties of the amino acid side chain give the amino acid its character. Depending on the polarity of the side chain an amino acid can be hydrophobic or hydrophilic, which play an important role in protein structure and protein-protein interaction. Hydrophobic and hydrophilic molecules are also called nonpolar and polar molecules, respectively. Hydrophilic molecules

are charge-polarized and capable of establishing hydrogen bonds, which enables them to dissolve in water. Hydrophobic molecules are nonpolar and prefer neutral molecules and nonpolar solvents. In hydrophobic environment, these molecules cluster together and form micelles. In this work we are using three hydrophobicity indices by ((Engelman *et al.*, 1989) (EnH), (Eisenberg *et al.*, 1984) (EH) and (Fauchere and Pliska, 1983) (FH)), see Table 2.5.

It has been shown that similarities in physical, chemical, energetic and conformational properties enable amino acids to conserve their specific ideal environments and spatial positions in the folded conformation of proteins (Prabhakaran and Ponnuswamy, 1979). Tomii and Kenehisa (Tomii and Kenehisa, 1966) analyzed 48 diverse amino acid properties, which are shown to be involved in proteins stability. In this work, five out of above mentioned 48 amino acid index-based properties are used to build the model, i.e., average medium contact (Mc), coil tendency (P_c), helix tendency (P_h), turn tendency (P_t), and average power to be at helix N-terminal (Pn). They are shown in Table 2.5.

Property groups

Property groups classify the AAs into groups related to specific properties of individual AAs or entire protein molecule. The properties that are considered in this thesis are summarized in Table 2.6. 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.

Hydrophobicity group: AAs can be categorized into different hydrophobicity groups (Lodish *et al.*, 2000) according to their hydrophilic and hydrophobic character. These characteristics vary depending on the polarity of the side chain. Hydrophilic amino acids with polar side chain are located in the surface of a water soluble protein. Hydrophobic amino acids stay away from aqueous environment and are slightly soluble or insoluble. They are found in interior parts of a protein. Hydrophobicity groups are shown in Table 2.6.

R groups: In this group AAs are categorized based on their MolW, pI and hydropathy index (Hp) (Yang and Wang, 2003). Hp index combines hy-

AA	MolW	pI	EnH	EH	FH	Mc	P_c	P_h	P_t	Pn
A	71.0791	6.01	1.6	0.62	0.42	2.11	0.71	1.42	0.66	1.59
C	103.1437	5.07	2	0.29	1.34	1.88	1.19	0.7	1.19	0.33
D	115.0887	2.77	-9.2	-0.9	-1.05	1.8	1.21	1.01	1.46	0.53
E	129.1157	3.22	-8.2	-0.74	-0.87	2.09	0.84	1.51	0.74	1.45
F	147.1772	5.48	3.7	1.19	2.44	1.98	0.71	1.13	0.6	1.14
G	57.0521	5.97	1	0.48	0	1.53	1.52	0.57	1.56	0.53
H	137.1414	7.59	-3	-0.4	0.18	1.98	1.07	1.	0.95	0.89
I	113.16	6.02	3.1	1.38	2.46	1.77	0.66	1.08	0.47	1.22
K	128.1792	9.74	-8.8	-1.5	-1.35	1.96	0.99	1.16	1.01	1.13
L	113.16	5.98	2.8	1.06	2.32	2.19	0.69	1.21	0.59	1.91
M	131.1977	5.47	3.4	0.64	1.68	2.27	0.59	1.45	0.6	1.25
N	114.104	5.41	-4.8	-0.78	-0.82	1.84	1.37	0.67	1.56	0.53
P	97.1171	6.48	-0.2	0.12	0.98	1.32	1.61	0.57	1.52	0
Q	128.131	5.65	-4.1	-0.85	-0.3	2.03	0.87	1.11	0.98	0.98
R	156.188	10.76	-12.3	-2.53	-1.37	1.94	1.07	0.98	0.95	0.67
S	87.0784	5.68	0.6	-0.18	-0.05	1.57	1.34	0.77	1.43	0.7
T	101.1054	5.87	1.2	-0.05	0.35	1.57	1.08	0.83	0.96	0.75
V	99.133	5.97	2.6	1.08	1.66	1.63	0.63	1.06	0.5	1.42
W	186.2139	5.89	1.9	0.81	3.07	1.9	0.76	1.08	0.96	1.33
Y	163.1756	5.67	-0.7	0.26	1.31	1.67	1.07	0.69	1.14	0.58

Table 2.5: The values of AA indices that include molecular weight (MolW), average isoelectric point (pI), Engelman's (EnH), Eisenberg's (EH), and Fauchere-Pliskas (FH) hydrophobicity indices, and average medium contact (Mc), coil tendency (P_c), helix tendency (P_h), turn tendency (P_t), and average power to be at helix N-terminal (Pn).

drophobic and hydrophilic tendencies. Table 2.6 demonstrates R groups.

Electronic groups: AAs can be classified based on their tendency to accept or donate electrons (Ganapathiraju *et al.*, 2004). Amino acid Cysteine has special properties which results in categorizing this amino acid separately. Cysteine is a sulfur-containing amino acid, which enables it to play a key role in stabilizing extracellular proteins. Cysteine can react with itself to form an oxidized dimer by formation of a disulfide bond. Electronic groups are shown

groups	subgroups	AAs	groups	subgroups	AAs
R groups	Nonpolar aliphatic	AVLIMG	Exchange	(A)	C
	Polar uncharged	SPTCNQ	groups	(C)	AGPST
	Positively charged	KHR		(D)	DENQ
	Negative	DE		(E)	KHR
	Aromatic	FYW		(F)	ILMV
				(G)	FYW
Hydrophobicity	Hydrophobic	VLIMAFPW	Other	Tiny	AG
		YCG			
groups	Hydrophilic basic	KHR	groups	Bulky	FHWYR
	Hydrophilic acidic	DE		Polar	NQ
	Hydrophilic polar	STNQ		uncharged	
Other	Charged	DEKHRVLI	Electrouic	Electron donor	DEPA
groups	Polar	DEKHRNT	groups	Weak Ed	VLI
		QSYW			
	Aromatic	FHWY		Electron acceptor	KNR
	Small	AGST		Weak Ea	FYMTQ
				Neutral	GHWS
				Special AA	C

Table 2.6: Amino acid property groups.

in Table 2.6.

Exchange groups: Unlike the other groupings that are based on amino acid physicochemical attributes, exchange groups are supported by statistical studies, see Table 2.6. In exchange groups AAs are clustered based on accepted point mutation to represent conservative replacements through evolution. Mutation replaces one amino acid with another one due to mostly natural selection (Wang *et al.*, 2000).

Other groups: Despite the overlap between groups, each group is considered as a separate attribute. Other groups are defined based on molecular weights,

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

Table 2.7: Chemical groups for amino acids

that is, tiny AAs with weight less than 80 Da, small AAs between 80 and 101 Da, and bulky AAs (more than 120 Da) (Hobohm and Sander, 1995), see Table 2.6.

Chemical groups: Chemical groups are defined based on the composition of a chemical group that constitute the side chains (Ganapathiraju *et al.*, 2004). Table 2.7 shows the chemical composition of AAs.

2.1.5 Secondary Structure Prediction

The knowledge of the protein structures helps to learn about their possible function and also their interactions in cellular processes. For example, knowledge of the structure of normally functioning proteins can help us to understand how defective protein structures can cause diseases. Knowing how an amino acid chain folds in three dimensional shape, we can conclude which amino acids may be involved in an interaction with other molecules.

A branch of proteomics aims to bridge the gap between the large number of determined sequences and the smaller number of determined structures using both protein structure determination methods and bioinformatics techniques. The reason for the gap lies in the experimental methods which have been used to solve the protein structure. Two main experimental methods are Nuclear Magnates Resonance (NMR) spectroscopy and X-ray crystallography. Both mentioned methods are relatively expensive, time consuming, and thus they can not provide a high throughput experimental structure determination.

Computational methods to predict the protein structures allow us to learn the structure of the protein directly from the protein sequence that has already been obtained. Community-wide experiments such as Critical Assessment of protein Structure Prediction (CASP) biennially evaluates the progress of the computational techniques in protein structure prediction.

The computationally intractable problem of predicting the 3D structure of proteins has motivated the development of knowledge based approaches that solve simpler intermediate problems such as the prediction of the protein secondary structure. It seems obvious that the secondary structure prediction should be easier than the tertiary structure prediction. One way to predict tertiary structure might be to predict helices and strands first and then to assemble them. Secondary structure prediction is a problem of moderate complexity. The problem of predicting secondary structure has been tackled using various machine learning algorithms, including neural networks (Rost and Sander, 1993; Jones, 1999; Pollastri *et al.*, 2002; Przybylski and Rost, 2002; Lin

et al., 2005), hidden Markov models (Bystroff *et al.*, 2000), support vector machines (SVMs) (Ward *et al.*, 2003; Guo *et al.*, 2004; Hua and Sun, 2001; Birzele and Kramer, 2006) and "jury-of-experts" based methods (Montgomerie *et al.*, 2006). While a significant body of research has been dedicated to secondary structure predictions, current state-of-the-art prediction techniques are able to guarantee accuracy of just above 80%.

In our work, we use two complementary secondary structure prediction methods, PROFsec (Rost and Sander, 1993) and PSIPRED (Jones, 1999), and thus this section focuses on these two methods.

A generic three-layer sequence-to-structure neural network design that is commonly used to predict secondary structure (Lesk, 2002) is shown in Figure 2.7.

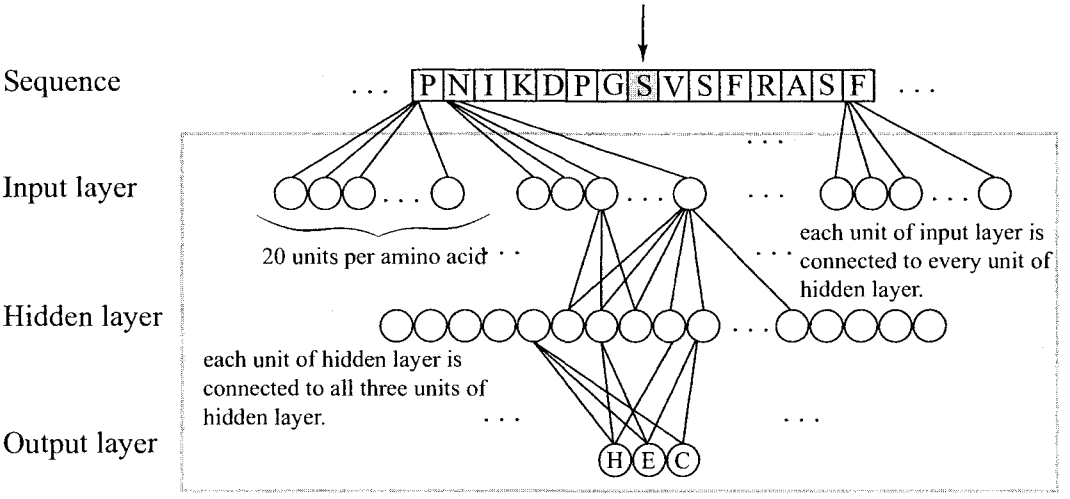

Figure 2.6: The generic model of a three layer sequence-to-structure neural network.

- The input layer takes care of a sliding n-residue window in the sequence to predict the structure of the central residue and then it moves one residue forward. Each of n residues in the sliding window corresponds to 20 nodes in the input layer.

- The hidden layer contains around 100 nodes and each node in this layer is fully connected to all nodes in the input and output layers.

- The output layer contains three nodes that predict the structure of the central residue as helix, strand or coil conformation, respectively.

Most of the neural network-based methods for secondary structure prediction feed the input layer not just with the amino acid characteristics, but with a profile generated from a multiple sequence alignment. Recent neural network based secondary structure prediction methods use a second level of prediction to identify correlations among neighboring predictions. In this level the predictions of several neighboring residues using another neural network are combined to generate the final prediction.

In both PROFsec (Rost and Sander, 1993) and PSIPRED (Jones, 1999), the evolutionary based information is used as input to the neural network and both of these methods use a second level of prediction, in which a second neural network is used to filter successive outputs from the main network.

PROFsec

- Input to the first level (sequence-to-structure) consists of two contributions: *local* information which is taken from a 13 residue long sliding window containing a $13 \cdot 21$ matrix extracted from PSSM (the extra input in each group is used to indicate that the window spans a chain terminus), and *global* information e.g., percentage of each amino acid in protein. This way PROFsec uses more than 1000 input nodes in first neural network. Output layer in first neural network contains three nodes.

- In the second neural network (structure-to-structure), Profsec just takes into account the predictions of the adjacent residues which leads to $3 \cdot 4$ units as inputs (the extra input in each group is used as spacer). There are three neurons in output layer that correspond to the predicted 3 structures.

PSIPRED

- This method uses the position-specific scoring matrix (PSSM) based information from PSI-BLAST (Altschul *et al.*, 1997) as input to the neural network.

- The neural network in the first level (sequence-to-structure), consists of a $15 \cdot 21$ input nodes, which receive the information which is taken from a 15 residue long sliding window over PSSM (the extra input in each group is used to indicate that the window spans a chain terminus). The hidden layer consists of 75 nodes and the output layer has three units.

- The network in the second level has an input layer comprising of 60 input units, divided into 15 groups of four (the extra input in each group is used to indicate that the window spans a chain terminus), which takes into account the correlations among all 15 predictions from the first level. The hidden layer has 60 nodes and the output layer consists of three neurons.

2.1.6 Secondary Structure Content

The knowledge about the amount of protein secondary structure elements or more specifically secondary structure content is yet another simplification in the characterization of the protein structure.

Considering 3-state secondary structure for each amino acid, i.e., the structure $S \in \{helix(H), strand(E), coil(C)\}$ for each protein sequence, we define n_S to be the number of AAs having the structure S. The protein secondary structure content of the structure S for a protein of length N is defined as:

$$l_S^t = \frac{n_S}{N}. \tag{2.2}$$

Fig. 2.7 demonstrates the sequence and secondary structure of 1n0w_B protein. The helix content (l_H^t) for this protein equals 0.27, while l_E^t and l_C^t are 0.09 and 0.64, respectively.

```
Sequence    PTLLGFHTASGKKVKIAKESLDKVKNLFDEKEQ
DSSP        CHHHCCEECCCCECCCCHHHHHHCCCCCCCCCC
```

Figure 2.7: The sequence (top line) and secondary structure (bottom line) of 1n0w_B protein.

In many applications, the crucial information about the involved proteins is represented by the protein structure content. The content prediction constitutes an important problem with a large number of applications in several areas of modern protein science. Some examples of these are structural class prediction (Kurgan *et al.*, 2006; Kurgan and Chen, 2007) and analysis of interactions between the CapZ protein and cell membranes (Smith *et al.*, 2006). The content computed from the predicted secondary structure or true (actual) secondary structure was also used in the prediction of coding and non-coding RNAs (Liu *et al.*, 2006), analysis of prion proteins (Concepcion *et al.*, 2005), prediction of folding rates (Ivankov and Finkelstein, 2004; Gong *et al.*, 2003; Gromiha and Selvaraj, 2008), distinguishing between enzyme and non-enzyme proteins (Dobson and Doig, 2008), prediction of enzyme classes (Dobson and Doig, 2005) and prediction of folding transition-state position (Huang and Cheng, 2007). Thus, improving accuracy of the protein structure content would have large impact on a variety of bioinformatics applications.

2.1.7 Evaluation Procedures for Content Prediction Methods

The structural content prediction methods have been evaluated using several approaches.

The *re-substitution* evaluation method assesses the prediction model on the same dataset used for training. In k fold *cross validation* assessment method, the training set is divided into k folds, training is performed using $k-1$ folds and the model is tested on the remaining fold. This procedure is repeated k times and the average prediction error is reported.

The quality of prediction is measured with the *average absolute error*. The absolute prediction error for a given structure $S \in \{H, E, C\}$ and protein sequence is defined as the absolute value of the difference between the actual content l_S^t and the predicted content l_S, i.e., $e = |l_S^t - l_S|$. The average absolute error is the average of the absolute errors for the individual sequences taken over the whole data set.

When comparing two methods with average absolute errors $\bar{e}_1$ and $\bar{e}_2$, where $\bar{e}_2 < \bar{e}_1$, we say that the second method delivers the improvement of $(\bar{e}_1 - \bar{e}_2)/\bar{e}_1 \cdot 100\%$ according to the *error rate reduction* formula.

A *paired t-test* is used to compare two population means, where there are two samples in which the observations in one sample can be compared to the observations in the other sample. Suppose the structural content of a dataset, composed of n proteins, is predicted by two content prediction methods (called x and y). We define l_S^x, l_S^y and l_S^t to be the predicted content by methods x and y, and the actual content for a given structure $S \in \{H, E, C\}$, respectively. By e_S^x and e_S^y, we denote the difference between the actual content l_S^t and the predicted content by methods x and y, respectively, i.e., $e_S^x = l_S^t - l_S^x$ and $e_S^y = l_S^t - l_S^y$. The procedure to compute the paired t-test on the predicted content by methods x and y is as follows:

- Calculate the difference of the prediction errors of two content prediction methods (x and y) for the ith protein $d_{S,i} = e_{S,i}^x - e_{S,i}^y$, $i = 1, 2, \ldots, n$.
- Calculate the mean difference, i.e., $\bar{d}_S = \frac{1}{n} \sum_{i=1}^{n} d_{S,i}$.
- Calculate the standard error:

$$SE_S = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (d_{S,i} - \bar{d}_S)^2}$$

- Calculate the t-value:

$$T = \frac{\bar{d}_S}{SE_S}$$

This statistics follows a distribution on $n - 1$ degrees of freedom.

- Use tables of the distribution to compare the calculated value for T to the t_{n-1} distribution. This will give the p-value for the paired t-test.
- If the p-value is smaller than 0.05, the difference between predictions of the two methods is statistically significant.

2.2 Methods

2.2.1 Support Vector Machine Regression

Support vector regression (SVR) which was introduced by Vapnik (Vapnik, 1998) is a statistical technique widely used for a larger variety of classification and prediction tasks. The basic problem setup is prediction of data from unknown distribution based on a number of observations and subject to a number of constrains. The observations can be associated with samples from the test set while the function which solves the regression problem can be used to predict the results on the test set.

General Regression Problem

Consider a set of l observations $(\vec{x}^i, y^i)$, $i = 1, 2, \ldots, l$ where $\vec{x}^i \in \mathbb{R}^n, y^i \in \mathbb{R}$ and assume that this set is generated from an unknown probability distribution $P(\vec{x}, y)$. Consider a class of functions

$$F = \{f | f : \mathbb{R}^n \to \mathbb{R}\} .$$

The basic regression problem is to minimize a risk functional

$$R(f) = \int L(y - f(\vec{x}), \vec{x}) dP(\vec{x}, y)$$

where $L(y - f(\vec{x}), \vec{x})$ is a cost function indicating the penalty for difference $y - f(\vec{x})$ at the point $\vec{x}$ (Smola, 1996). Since the distribution $P(\vec{x}, y)$ is unknown one can also calculate the empirical risk function

$$R_{\text{emp}} = \frac{1}{l} \sum_{i=1}^{l} L(y^i - f(\vec{x}^i), \vec{x}^i)$$

and the bound risk R as $R_{\text{emp}} + R_{\text{gen}}$ where R_{gen} is the upper bounds generalization error (Vapnik, 1998).

Depending on the application several risk functions are used. Some examples include

- $L(\eta) = \eta^2$ quadratic cost function resulting in the least square minimization of the empirical risk.

- ϵ-precision approximation

$$L(\eta)_\epsilon = \begin{cases} \infty, & \text{if } \eta < -\epsilon \\ 0, & \text{if } \eta \in [-\epsilon, \epsilon] \\ \infty, & \text{if } \eta > \epsilon \ . \end{cases}$$

  which indicates that deviation of no more than ϵ is not penalized, on the other hand deviation to more than ϵ is penalized with infinite cost.

Linear Regression

Consider a set of linear functions

$$F = \{f | f = (\vec{\omega}, \vec{x}) + b, \vec{\omega} \in \mathbb{R}^n, b \in \mathbb{R}\} \ .$$

and minimize (Vapnik, 1998)

$$G(\vec{\omega}, b) = \frac{1}{2}||\vec{\omega}||_2^2$$

subject to

$$\begin{aligned} \epsilon - (f(\vec{x}^i) - y^i) > 0 \quad &\text{for} \quad i = 1, 2, \ldots, l \\ \epsilon - (y^i - f(\vec{x}^i)) > 0 \quad &\text{for} \quad i = 1, 2, \ldots, l \end{aligned}$$

where $f(\vec{x}) = (\vec{\omega}, \vec{x}) + b$.

Considering a more general problem where ϵ_i are possible deviations for each point i we minimize

$$R(\vec{\omega}, b, \vec{\eta}, \vec{\eta^*}) = \frac{1}{2}||\vec{\omega}||_2^2 + C \sum_{i=1}^{l} (L(\eta_i + \epsilon_i)_\epsilon + L(\eta_i^* + \epsilon_i^*)_\epsilon)$$

subject to

$$
\begin{aligned}
f(\vec{x}^i) - y^i &\leq \eta_i + \epsilon_i \quad \text{for} \quad i = 1, 2, \ldots, l \\
y^i - f(\vec{x}^i) &\leq \eta_i^* + \epsilon_i^* \quad \text{for} \quad i = 1, 2, \ldots, l \\
\eta_i, \eta_i^* &\geq 0 \quad i = 1, 2, \ldots, l
\end{aligned}
$$

where $f(\vec{x}) = (\vec{\omega}, \vec{x}) + b$ and C is a complexity parameter.

WEKA software (Witten and Frank, 2005) is used to implement SVR algorithm (Smola, 1996) to solve regression tasks.

Assume that the set of functions has a basis $f_1, f_2, \ldots, f_n$ i.e. any function f can be represented as

$$f(\vec{x}) = \sum_{i=1}^{n} \alpha_i f_i(\vec{x}) \ .$$

Then solving optimization problem is equivalent to finding coefficients α_i which determine the optimal function f. The optimal solution of the above problem is often expressed as a linear combination of $(\vec{x}^i, \vec{x}^j)$, however this can be generalized to arbitrary functions $K(\vec{x}^i, \vec{x}^j)$, kernel functions.

Nonlinear regression, general kernels

Some of the popular kernels include (see (Vapnik, 1995))

- Polynomial

$$K(\vec{x}, \vec{x}') = ((\vec{x}, \vec{x}'))^p$$

- RBF

$$K(\vec{x}, \vec{x}') = \exp(-||\vec{x} - \vec{x}'||_2^2/(2\sigma^2))$$

- two-layer feed forward neural network

$$K(\vec{x}, \vec{x}') = \tanh(\kappa(\vec{x}, \vec{x}') - \theta)$$

2.2.2 Principal Component Analysis

Principal Component Analysis (PCA) is an orthogonal linear transformation of a set of data vectors (measurements) that brings the data to a new coordinate

system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

Consider an $M \times N$ matrix $\vec{X}$ which consist of N M-dimensional data vectors (columns). Assume that $\vec{X}$ has zero empirical mean. Then the PCA transformation is given by matrix $\vec{W}$ such that

$$\vec{Y}^{\mathrm{T}} = \vec{X}^{\mathrm{T}}\vec{W} = \vec{V}\vec{\Sigma}$$

where $\vec{V}\vec{\Sigma}\vec{W}^{\mathrm{T}}$ is a singular value decomposition (SVD) of $\vec{X}^{\mathrm{T}}$.

PCA can be used for dimensionality reduction of the data i.e. by keeping these characteristics of the data set which contribute most to the variance and ignoring higher order components which do not contribute much to the variance. This property of PCA can be important for feature selection properties where one would like to select a small set of features capturing the most important properties of the data.

2.2.3 Correlation Coefficient

Correlation between two random variables X and Y is defined as

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2}\sqrt{E(Y^2) - (E(Y))^2}}$$

where $E(\cdot)$ denotes expectation. Correlation $\rho_{X,Y} \in [-1, 1]$ is a measure of independence between the variables. For normal random variables $\rho_{X,Y} = 0$ is equivalent to the statement that X and Y are independent. Conversely, large absolute value of $\rho_{X,Y}$ shows strong dependence between X and Y.

If a series of n measurements from X and Y is available written as x_i and y_i for $i = 1, 2, \ldots, n$ then Pearson-product moment correlation coefficient statistics can be used to estimate the correlation $\rho_{X,Y}$. The Pearson coefficient can be written as

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

2.3 Datasets

EVA (http://cubic.bioc.columbia.edu/eva/) is a web-based server which continuously assesses the accuracy of protein structure prediction methods. Currently EVA evaluates several automated servers for the secondary structure, residue contact, protein structure modeling, and fold recognition prediction. Everyday, EVA downloads the protein sequences of the newest experimentally determined structures that are added to PDB. The sequences are sent to each prediction server and then the prediction results are collected. Every week, the assembled results are compared to the experimental structures, the evaluation is updated and the results are published on the EVA web pages.

We use the EVA server to construct the training, evaluation, and test sets required to design and empirically evaluate the proposed LAMICA method. The EVA server was selected because it provides unbiased predictions of several representative secondary structure prediction methods. In addition, EVA only accepts targets that show no significant sequence similarity to proteins known at the time when the structure was solved for the target (Rost and Eyrich, 2001; Birzele and Kramer, 2006).

Our training set, EVA977, is composed of 977 sequences that were released in EVA between April 2000 and April 2002. All our preliminary tests and parameter optimizations have been performed on the evaluation set of 149 proteins, named EVA149, containing EVA targets added to the server between May 2002 and January 2003. Our test set, EVA150, includes 150 proteins that were published on EVA between January 2004 and July 2005. The latter set was compiled by Birzele and Kramer (Birzele and Kramer, 2006). Prediction results of PSIPRED, PHD (Rost, 1996), PROFsec, and PG (Birzele and Kramer, 2006) are available for EVA150. A fair comparison between the content prediction methods should be based on a blind testing setup. Therefore, we constructed our training set from the proteins released in EVA before the proteins included in the evaluation and the test sets.

Chapter 3

Related Works

In Section 2.1.2, eight-state protein secondary structure was introduced. Several researchers investigated eight-state secondary structure content prediction (Chou, 1999; Cai *et al.*, 2003; Liu and Chou, 1999; Lee *et al.*, 2006). Eight-state secondary structure can be reduced to three-state secondary structure using DSSP defined assignment. In this research we focus on three-state protein secondary structure content prediction.

The prediction of secondary structure content is usually performed through an intermediate step, in which the protein sequences are converted into a fixed set of features. Using feature values and an established prediction model, the protein structural content is predicted.

As the first attempt to predict the three state (helix, strand, and coil) content, Krigbaum and Knutton introduced a multiple linear regression (MLR) method, which applied the amino acid (AA) composition of the protein sequence as the input (Krigbaum and Knutton, 1973). Subsequent attempts used either neural networks or MLR models and combined them with a variety of other features computed from the protein sequence. Examples include the molecular weight of a protein (Muskal and Kim, 1992), auto-correlation functions based on hydrophobicity (Zhang *et al.*, 1998, 2001; Lin and Pan, 2001), pair-coupled composition (Chou, 1999; Liu and Chou, 1999; Cai *et al.*, 2003), composition moment vector (Ruan *et al.*, 2005), evolutionary infor-

mation encoded using PSI-BLAST profiles (Lee *et al.*, 2006), and various physicochemical properties of amino acids combined with their composition (Homaeian *et al.*, 2007). Several researchers also investigated the impact of a priori knowledge of structural classes on the quality of the content prediction (Zhang *et al.*, 1998b, 2001). However, the scope of the latter two methods was limited to the proteins with a known structural class.

In this work, our results are compared with the predictions of three content prediction methods, which we discuss in detail. The first method is a primary sequence based technique (Zhang *et al.*, 1998) that uses MLR. The second method is also a MLR based technique, but it uses the knowledge of structural classes (Zhang *et al.*, 2001), and the third method is the most recent predictor that combines several amino acid physicochemical properties (Homaeian *et al.*, 2007).

The MLR based content prediction method of (Zhang *et al.*, 1998) uses the knowledge of protein primary sequence. In this method, the amino acid composition (defined as x_i, $i = 1, 2, \ldots, 20$) and the auto-correlation functions of the hydrophobicity indices of residues along the sequence, were used for secondary structure content prediction. It was previously shown that the amino acid composition is important to define the structural classes (Nakashima *et al.*, 1986; Chou, 1995), and for the secondary structure content prediction (Krigbaum and Knutton, 1973; Muskal and Kim, 1992; Eisenhaber *et al.*, 1996; Zhang *et al.*, 1996).

The autocorrelation function of each sequence denoted by ρ_n, using the hydrophobicity index proposed by (Fauchere and Pliska, 1983), is defined by:

$$\rho_n = \frac{\sum_{j=1}^{N-n} p_j p_{j+n}}{N-n} \qquad n = 1, 2, \ldots, 10 \; . \tag{3.1}$$

where p_j is the hydrophobicity index (Fauchere and Pliska, 1983) for the j^{th} residue (see Table 2.5) in the primary sequence and N is the protein sequence length. For convenience, the auto-correlation functions ρ_n ($n = 1, 2, \ldots, 10$) defined above are denoted by $x_{21}, x_{22}, \ldots, x_{30}$, respectively.

The central hypothesis of the technique proposed by (Zhang *et al.*, 1998)

is that helix and strand content denoted by f_α and f_β, respectively, are linear functions of $x_1, x_2, \ldots, x_{30}$, which are defined by:

$$f_\alpha = a_0 + \sum_{i=1}^{30} a_i x_i \qquad i = 1, 2, \ldots, 30, \tag{3.2}$$

$$f_\beta = b_0 + \sum_{i=1}^{30} b_i x_i \qquad i = 1, 2, \ldots, 30. \tag{3.3}$$

where a_i and b_i for $i = 0, 1, \ldots, 30$ are the regression coefficients, which are determined once the regression model is set up on the data in the training set.

The following equations show two regression formulae for f_α and f_β, proposed by (Zhang *et al.*, 1998):

$$\begin{aligned} f_\alpha = {} & 1.942x_1 + 0.008x_2 + 0.020x_3 + 0.231x_4 + 0.055x_5 - 0.792x_6 \\ & + 0.671x_7 - 0.212x_8 + 1.121x_9 + 1.070x_{10} + 1.029x_{11} + 0.661x_{12} \\ & - 0.934x_{13} + 1.038x_{14} + 0.637x_{15} + 0.042x_{16} - 0.711x_{17} - 1.455x_{18} \\ & - 1.773x_{19} + 0.796x_{20} + 0.105x_{21} - 0.377x_{22} + 0.156x_{23} + 0.301x_{24} \\ & - 0.073x_{25} - 0.076x_{26} + 0.215x_{27} + 0.093x_{28} - 0.025x_{29} - 0.139x_{30}, \end{aligned} \tag{3.4}$$

$$\begin{aligned} f_\beta = {} & 1.254x_1 - 0.202x_2 - 0.012x_3 + 0.480x_4 + 0.752x_5 + 0.763x_6 \\ & - 1.671x_7 + 0.665x_8 - 0.461x_9 + 0.402x_{10} + 0.025x_{11} - 0.514x_{12} \\ & - 0.279x_{13} - 0.422x_{14} + 0.219x_{15} + 0.660x_{16} + 1.397x_{17} + 2.401x_{18} \\ & + 2.453x_{19} + 0.172x_{20} - 0.184x_{21} + 0.340x_{22} - 0.171x_{23} - 0.203x_{24}. \end{aligned} \tag{3.5}$$

The MLR based content prediction method of (Zhang *et al.*, 2001) uses the prior knowledge of protein secondary structure classes, which is shown to improve the content prediction. Protein secondary structure class and content are strongly related to each other. We note that the protein class prediction is difficult and the state-of-the-art class prediction methods achieve just about 80% accuracy (Kurgan *et al.*, 2008b).

In this study similarly to the above mentioned method, sequences were represented by the amino acid composition and the auto-correlation functions

class	content
All α	f_α=0.515–1.758 x_2–3.289 x_{13}+1.038 x_1+1.456 x_{10}
	f_β=–0.002+0.214 x_{17}+0.156 x_2+0.053 x_1+0.012 x_{21}
All β	f_α=0.033+0.051 x_{29}+0.789 x_{11}–0.139 x_2–0.350 x_{13}
	f_β=0.458–0.678 x_{13}–0.109 x_{21}–0.171 x_2–0.140 x_{23}+0.949 x_{18}
	–1.026 x_7
$\alpha\beta$	f_α=0.297+0.685 x_{10}+0.143 x_{24}–0.239 x_2+0.462 x_1–0.428 x_{12}
	–0.364 x_{20}–0.594 x_{17}+0.655 x_8–0.287 x_{16}–1.599 x_{13}
	f_β=0.199+0.676 x17+0.782 x_{10}–0.036 x_{24}+0.555 x_9–0.224 x_1
	–0.246 x_3+0.773 x_{18}

Table 3.1: Regression formulae for helix and strand content prediction of proteins with known structural classes

of the hydrophobicity indices (Fauchere and Pliska, 1983) of residues along the sequence. The main difference between (Zhang *et al.*, 2001) and (Zhang *et al.*, 1998) is that separate MLRs were trained for α, β and $\alpha\beta$ structural classes (defined in SCOP). Considering the structural class of each protein sequence, this method introduces two regression formulae for helix and strand content denoted by f_α and f_β, which is shown in Table 3.1:

The other MLR based content prediction method, which is called PSSC-core (Homaeian *et al.*, 2007), uses a set of amino acid physicochemical properties for each of the helix and strand content predictions (see Section 2.1.4 for the common amino acid properties definition). This method uses a feature selection method, which first selects the feature with the highest correlation coefficient value with the content value and then the method proceeds by adding one feature at a time provided that the feature leads to a more accurate prediction model. PSSC-core represents additional features for the strand content prediction to cover long range interactions that are characteristic for β-sheets. These indices are based on probability distributions of amino acid pairs (A_i and A_j), which belong to two strands connected by hydrogen bonds

to form a sheet. In addition, a set of polypeptide composition based features are proposed by PSSC-core, which is defined to find a set of polypeptides of the same length that are frequently observed in each of the helix and strand structures.

Chapter 4

Proposed Prediction System

4.1 Overview

To construct novel features for our method, we investigated the quality of information retrieved from the predicted protein secondary structure with respect to the content prediction. Our study of several structure prediction methods reported on the EVA server (Rost and Eyrich, 2001) shows that the quality of the predicted helix and coil content varies with the length of the protein. While the majority of the secondary structure prediction methods have high helix and coil content error on short protein sequences, some methods, for example PROFsec, are characterized by the prediction quality that is less dependent on the protein length.

Fig 4.1 shows the helix and coil content prediction error rate of PSIPRED and PROFsec as a function of protein length. The experiment was performed on the EVA977 data set. The results indicate that, although PSIPRED has lower absolute average error of content, PROFsec showed better performance for short protein sequences. In other words, the PROFsec-based predictions for protein secondary structure are of higher quality for short protein sequences, while PSIPRED-based predictions are better for long sequences. Therefore, to improve the prediction of the helix and coil content, we classify the proteins into two categories: long and short. We also define a parameter N_0 to be the

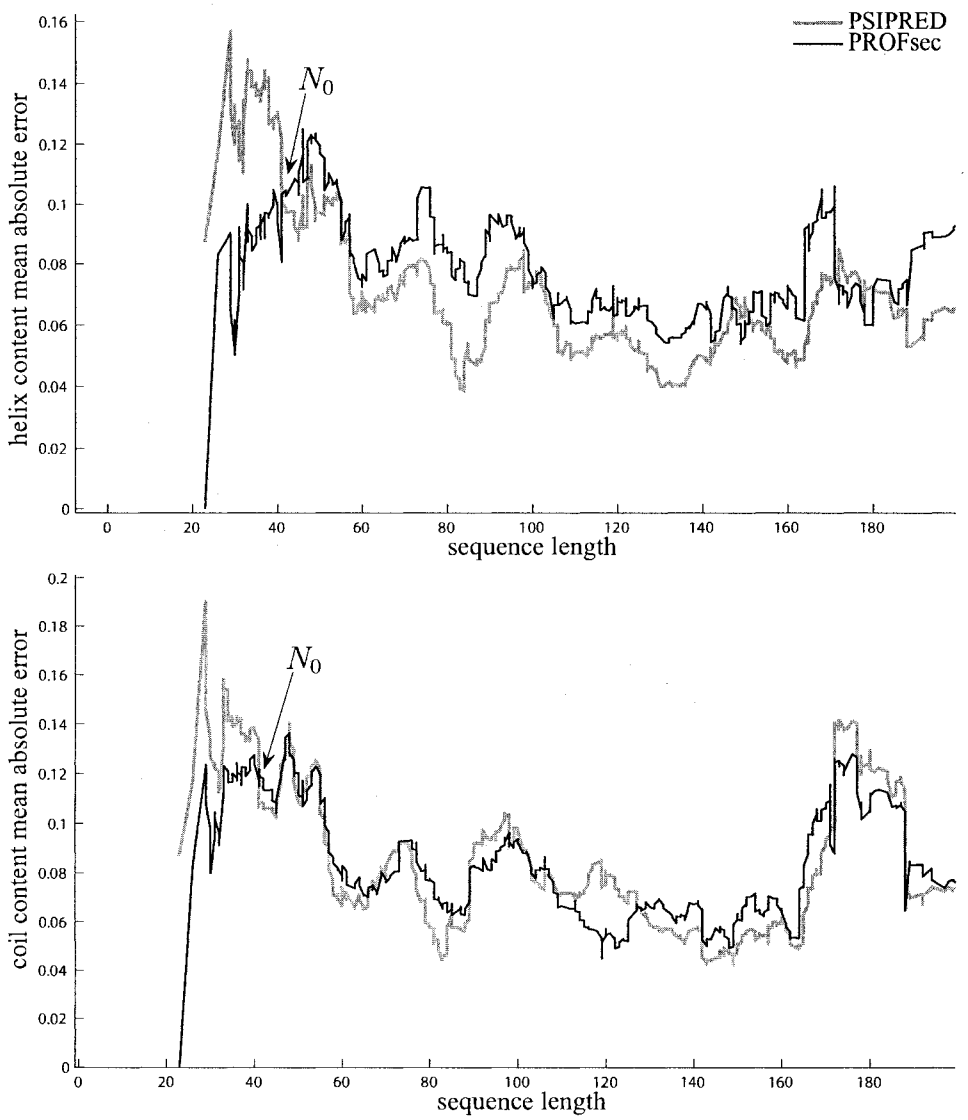

Figure 4.1: Comparison of the content error of different prediction methods as a function of protein sequence length.

length separation threshold. The error curves for PSIPRED and PROFsec shown in Fig 4.1 cross at 40 for both the helix and coil predictions, suggesting that N_0 should equal 40. These two curves might also cross for several length values larger than 40 (depending on the data set), but the difference between the two errors is not significant for larger lengths (see Fig 4.1).

Therefore, in the proposed algorithm, we use the PROFsec prediction to

calculate structural features for proteins of length smaller or equal to 40. For sequences longer than 40 we calculate features from PSIPRED prediction. We construct two separate learning models, one trained on short sequences, the other trained on long sequences. During the actual prediction we are using the short sequence model to predict the content of the short input sequences and the long sequence model to predict the content of the long ones.

4.2 Feature Generation

During the feature generation process, each input protein sequence is encoded into a vector of numbers (feature vector). Each feature is computed according to some property of the corresponding protein.

4.2.1 Structural features.

We compute a number of structural features using secondary structure prediction result of PROFsec (short sequences) or PSIPRED (long sequences). We consider 3-state prediction for each amino acid, i.e., the predicted structure denoted by S can be helix (H), strand (E) or coil (C). We construct structural features according to (Chen and Kurgan, 2007) and (Kurgan *et al.*, 2008), where these features were used to address protein fold prediction and structural class assignment. For each protein sequence, we define n_{S}^{j} to be the number of segments of length j having predicted structure S. For example, n_{H}^{2} denotes the number of HH pairs in a predicted secondary structure. By T_{S}, we indicate a total number of segments in a protein sequence with predicted structure S. The range for the index k was defined by (Chen and Kurgan, 2007) and (Kurgan *et al.*, 2008) based on the average and standard deviation of segment sizes in the training set. Let us define normalized segment counts

as

$$N_{\mathrm{C}}^{k} = \frac{\sum_{j=k}^{20} n_{\mathrm{C}}^{j}}{T_{\mathrm{H}} + T_{\mathrm{E}} + T_{\mathrm{C}}} \qquad k = 2, 3, \ldots, 20, \tag{4.1}$$

$$N_{\mathrm{E}}^{k} = \frac{\sum_{j=k}^{20} n_{\mathrm{E}}^{j}}{T_{\mathrm{H}} + T_{\mathrm{E}}} \qquad k = 2, 3, \ldots, 20, \tag{4.2}$$

$$N_{\mathrm{H}}^{k} = \frac{\sum_{j=k}^{20} n_{\mathrm{H}}^{j}}{T_{\mathrm{H}} + T_{\mathrm{E}}} \qquad k = 3, 4, \ldots, 20. \tag{4.3}$$

Here, the shortest helix segment is assumed to include at least three residues. The count of coil segments is normalized by the total number of all segments while the counts of strand and helix segments are normalized by the total number of strand and helix segments. These different normalizations aim to accommodate for the proteins that may not include any strand and helix segments, respectively. We define M_{S} to be the length of the longest segment having structure S in the protein sequence. By m_{S}, we denote the average length of segments with structure S. The normalized maximum and average segment lengths for a protein of length N are given by

$$\bar{M}_{\mathrm{S}} = \frac{M_{\mathrm{S}}}{N} \tag{4.4}$$

$$\bar{m}_{\mathrm{S}} = \frac{m_{\mathrm{S}}}{N}. \tag{4.5}$$

Another group of features called composition moment vectors are used to quantify position specific contents. Let n_{S} be the number of occurrences (frequency) of S in the predicted protein secondary structure and $n_{\mathrm{S},j}$ be the index of the jth occurrence of the structure S. The composition moment vector of order k, $\mathrm{CMV}_{\mathrm{S}}^{k}$, is defined as

$$\mathrm{CMV}_{\mathrm{S}}^{k} = \frac{\sum_{j=1}^{n_{\mathrm{S}}} n_{\mathrm{S},j}^{k}}{\prod_{i=0}^{k}(N-i)} \qquad k = 0, 1, \ldots, 5. \tag{4.6}$$

For $k = 0$, CMV reduces to a composition vector (CV), where CV_{S} is equivalent to the secondary structure content.

4.2.2 Sequence based features.

Several studies pointed out that similarities in physical, chemical, energetic and conformational properties enable AAs to conserve their ideal environments

and spatial positions in the folded conformation of proteins (Prabhakaran and Ponnuswamy, 1979). Therefore, careful construction of physicochemical features, which are able to capture the essence of conformational environments, may provide invaluable information for the protein content prediction. To generate the first subset of sequence based features, we considered 48 diverse AA properties analyzed by Tomii and Kenehisa (Tomii and Kanehisa, 1996). Also, some other AA characteristics, like hydrophobicity indices (Engelman *et al.*, 1989), (Eisenberg *et al.*, 1984) and (Fauchere and Pliska, 1983), molecular weight and isoelectric point indices, were examined, see Table 2.5. These features are known to improve content prediction (Homaeian *et al.*, 2007; Zhang *et al.*, 2001). Let $k = 1, 2, \ldots, 53$ index the mentioned 53 amino acid properties. We denote the value of the property k of the jth amino acid in the protein sequence by $p_j^{(k)}$. For each property k, we define the autocorrelation $\rho_n^{(k)}$ (with shift n), as

$$\rho_n^{(k)} = \frac{\sum_{j=1}^{N-n} p_j^{(k)} p_{j+n}^{(k)}}{N-n} \qquad n = 1, 2, \ldots, 6\ . \tag{4.7}$$

We denote the average (mean), and the standard deviation of the amino acid property k over the protein sequence by $\overline{p}^{(k)}$ and $\sigma^{(k)}$, respectively

$$\overline{p}^{(k)} = \frac{\sum_{j=1}^{N} p_j^{(k)}}{N} \tag{4.8}$$

$$\sigma^{(k)} = \sqrt{\frac{1}{N-1} \sum_{j=1}^{N} (p_j^{(k)} - \overline{p}^{(k)})^2}. \tag{4.9}$$

We use equations (4.7) and (4.9) to calculate sequence based features from the AA sequences.

The second subset of sequence based features is generated from the AA groups, i.e., hydrophobicity, exchange, electronic, chemical, side chain (or R) and the other groups. For each protein sequence, we calculate the number of AAs that belong to a particular group. Detailed discussion, definition, and motivation for these groups can be found in (Ganapathiraju *et al.*, 2004; Hobohm and Sander, 1995).

The third subset of sequence based features comprises the composition vector, CV, and the composition moment vectors of the first and the second degree, CMV^1 and CMV^2 (Ruan *et al.*, 2005), defined by

$$\mathrm{CMV}_i^k = \frac{\sum_{j=1}^{n_i} n_{ij}^k}{\prod_{d=0}^{k}(N-d)} \qquad k = 0, 1, 2, \tag{4.10}$$

where n_{ij} is the jth position of the ith AA, $i = 1, 2, \ldots, 20$ in the protein. The value n_i is the frequency of ith AA in the sequence, and k is the order of the CMV. For $k = 0$, CMV reduces to CV. $\mathrm{CV_S}$ is equivalent to the percentage of each AA in the primary sequence, while CMV takes into account the position of each AA in the sequence.

4.3 Feature Selection

The feature generation process described in Section 4.2 produces a feature vector composed of 552 feature values for each protein sequence. Two sets of feature vectors are generated, one for proteins of length smaller or equal to N_0 and the other for proteins of length larger than N_0. For each set, we perform an independent feature selection such that two distinct learning models are created: one for short proteins and the other for long proteins. We perform a two-stage feature selection aimed at increasing predictive performance of the model and decreasing the risk of overfitting. In the first step, the Pearson correlation coefficient between each feature and the class label (content value) in the training set is computed using the EVA977 data set to find the strength and direction of a linear relationship between each feature and the secondary structure content. We note that each feature is evaluated independently to minimize the risk of overfitting of this large set of features into the applied data sets. We define a feature to be *significant* if the absolute correlation coefficient between the feature and secondary structure content is higher than a selected threshold ρ. Fig. 4.2 illustrates the number of features selected in the first stage as a function of the threshold ρ. The numbers of selected features

calculated for five values of $\rho = 0.4,0.5,0.6,\ldots,0.8$ are given in Fig. 4.2(a) for short sequences and in Fig. 4.2(b) for long sequences.

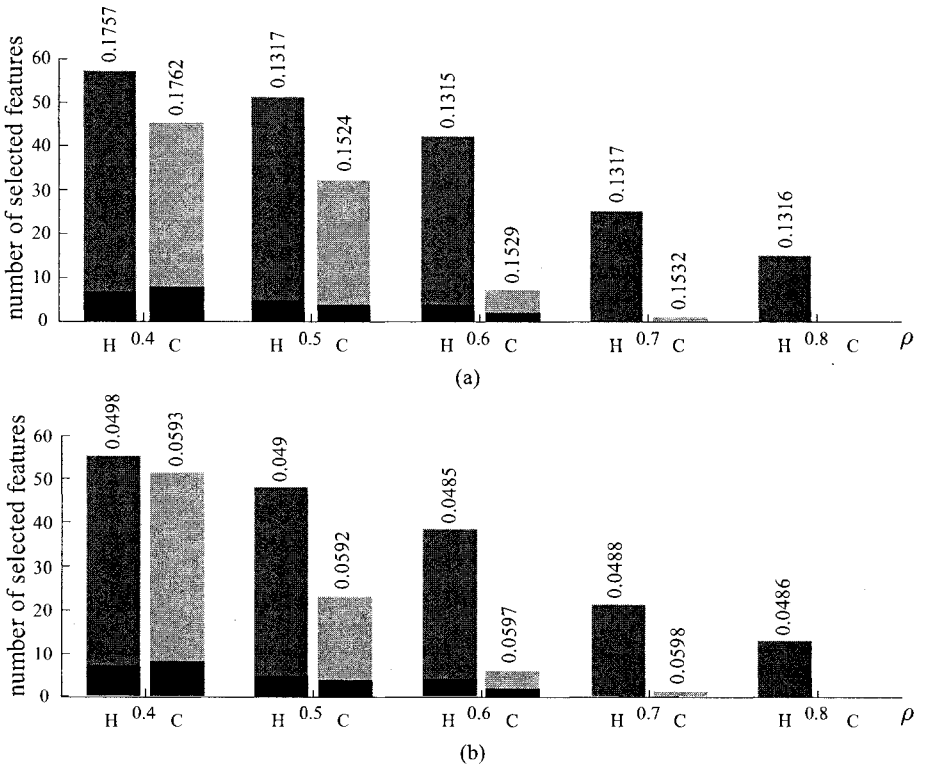

Figure 4.2: Comparison of number of selected features as a function of correlation coefficient in helix and coil content; The numbers on the top of the bars show the average absolute content prediction error for the models based on that corresponding feature set, which were trained on the EVA977 data set and tested on the EVA149 data set.(a) for short sequences; (b) for long sequences.

Bars located in the right side of the correlation coefficient threshold, correspond to features selected for the coil content prediction model, and the bars positioned on the left side correspond to the features selected for helix content prediction. The lighter shading in the upper part of each bar corresponds to the number of the selected structural features, and the darker shading in lower part of the bar corresponds to the number of the selected sequence based

features. The Figure shows that the majority of selected features are based on the predicted secondary structure, but the sequence based features are still shown to provide useful input. The numbers on the top of the bars show the average absolute content prediction error for the models based on that corresponding feature set that were trained on the EVA977 data set and tested on the EVA149 data set. We observe that the threshold $\rho = 0.6$ gives the lowest error for the prediction of the helix content, while, for the coil content prediction, the optimal value of the correlation threshold was found to be $\rho = 0.5$. We note that the feature groups corresponding to the optimal threshold values contain both the structural features as well as the sequence based features.

In the second stage of the feature selection, we further reduce the dimensionality using eigenvalue ranking, which is applied to the feature set selected in the first stage. The goal of this process is to select a relatively small group of features that are weakly correlated with each other and, at the same time, that would provide better prediction capability than the full set could. We compute a covariance matrix for the feature set selected at the first stage and find the corresponding eigenvectors and eigenvalues. We remove the least valuable features, i.e., the one which when removed would result in the lowest reduction of variance in the data set and we repeat this process. This continues until we obtain a set of features for which removal of extra features would worsen the content that gives the most accurate content prediction on the evaluation data set, EVA149.

Table 4.1 shows the final number of selected features after stage two for the helix and coil content prediction, and for short and long protein sequence models. The sequence based features include average coil tendency (P_c), average turn tendency (P_t), and average helix tendency (P_h) (Chou and Fasman, 1978), the average medium contact (Mc) and the power to be at helix N-terminal (Pn) in each protein sequence (Nakai *et al.*, 1988). These features belong to 48 diverse amino acid properties, which were shown to be correlated with the stability of the protein (Tomii and Kanehisa, 1996). The selected structural features include the helix content (CV_H), coil content (CV_C), com-

Pred. target	Seq. length	Selected features		
		Total #	Sequence based	Structural based
helix content	$N \leqslant 40$	3	average P_c	CMV_C^1, CV_H
	$N > 40$	5	average Mc	CMV_H^5, CMV_C^1 CV_H, N_H^{16}
coil content	$N \leqslant 40$	6	average P_h	N_H^{19}, CV_H
			average P_t	CMV_C^5, CV_C
	$N > 40$	2	average Pn	CV_C

Table 4.1: Features selected at the second stage of the feature selection process for the helix and coil content prediction, and for short and long protein sequence models.

position vector for coils (CMV_C^1, CMV_C^5) and helices (CMV_H^5), and count of long helical segments (N_H^{16} and N_H^{19}). This is consistent with the overall performance of secondary structure prediction methods, which provides greater accuracy for prediction of helices and coils than for the strands (Lin *et al.*, 2005; Birzele and Kramer, 2006). We can see that the selected features include both structural and sequence based features. This indicates that both structural and physicochemical components are required to create an accurate content prediction model.

4.4 Parametrization of Prediction Model

Support Vector Machine Regression (SVR) was selected to implement our content prediction method. For each helix and coil content prediction, two SVRs were computed, one for large and the other for small proteins. The classification algorithms used to develop and compare the proposed method were implemented in Weka (Witten and Frank, 2005).

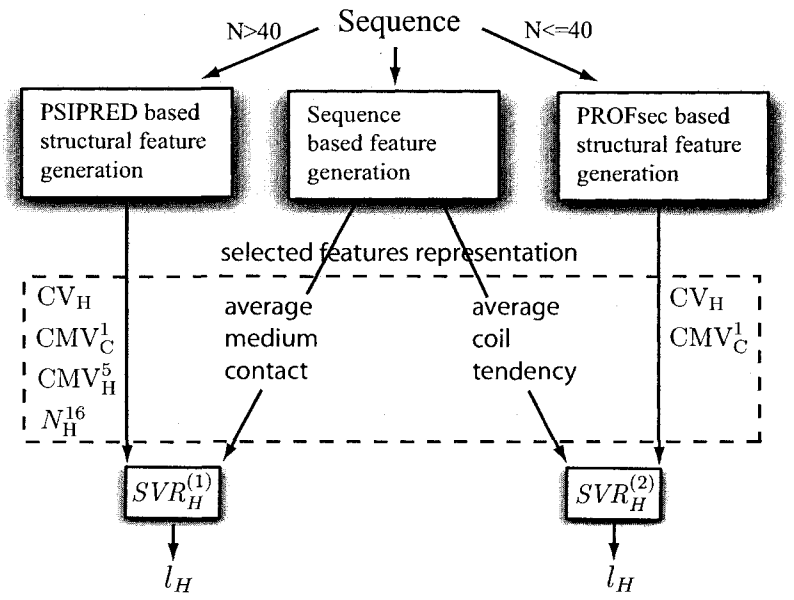

Figure 4.3: Architecture of the proposed helix content prediction method.

The diagram of the proposed prediction method for the helix content prediction is shown in Fig. 4.3. For sequences of length $N > 40$, five features are calculated, and the prediction is performed using classifier $SVR_H^{(1)}$. For $N \leqslant 40$, three features are computed, and the prediction is performed using classifier $SVR_H^{(2)}$. The content prediction result for an helix is denoted by l_{H}. Predicted coil content, l_{C}, is computed using the same procedure as depicted in Fig. 4.3, but with features selected for coil prediction (shown in Table 4.1). The strand content l_{E} is calculated as $l_{\mathrm{E}} = 1 - l_{\mathrm{H}} - l_{\mathrm{C}}$.

All four SVRs, employed for helix and coil content prediction, use the polynomial kernel. For helix content prediction, the complexity parameter $C = 1$ for $SVR_H^{(1)}$ and $C = 25$ for $SVR_H^{(2)}$. For coil content prediction, the complexity parameter $C = 21$ for $SVR_C^{(1)}$, which performs the prediction for sequences of length $N > 40$, and $C = 42$ for $SVR_C^{(2)}$. The parameters were selected by optimization performed on the evaluation set EVA149 and then applied with the same groups of features to the test set EVA150.

Chapter 5

Experiments and Results

In this section, the performance of the proposed method is compared with nine competing secondary structure and content prediction methods using two independent data sets.

5.1 Experimental Evaluations

The average absolute errors for the helix content prediction produced by the considered prediction methods on EVA149 and EVA150 are shown in Table 5.1. The errors are calculated separately for the subset composed of proteins of length $N \leqslant 40$ (first row), subset of proteins of length $N > 40$ (second row) and the whole set (third row). Each column of the table corresponds to the results of one prediction method, and the last column shows the results of the proposed LAMICA method.

Our prediction results were compared to content calculated from nine prediction methods. Six of these methods, PSIPRED, PROFsec, PHD, PHDpsi (Przybylski and Rost, 2002), SSPRO (Pollastri *et al.*, 2002), and PG (Birzele and Kramer, 2006) predict secondary structure, and the content is computed from these predictions. Three other methods, Zhang98 (Zhang *et al.*, 1998), Zhang01 (Zhang *et al.*, 2001), and PSSC-core (Homaeian *et al.*, 2007) predict the content directly from the sequence.

5.1.1 Helix Content

The bolded prediction results concerning whole datasets (see Table 5.1) show that the proposed prediction method delivers lower prediction errors for the *helix content* when compared with the considered competing methods. For helix content predictions on EVA149, the proposed method gives 11% improvement over the second best method, PSIPRED, and 14% improvement over the third best method, PROFsec. For EVA150, the improvement over the second best method, PG, is 5%, and the improvement over the third best method, PROFsec, equals 10%. These improvements were found to be significant using a paired t-test at 95% significance level. Table 5.2 gives the t-values and the degree of significance of the difference between the content predicted with LAMICA and content predicted with the second and the third best methods. More specifically, paired t-test compares pairs of corresponding content prediction errors (prediction of two methods on the same sequence) over all sequences in the EVA150 data set. Here, ++/- - means that LAMICA provides statistically significantly better/worse prediction than the method listed in the corresponding column; $\sim$ shows that the difference is not statistically significant. The results in the first row of Table 5.2 show that the proposed method significantly outperforms all other considered methods in the helix content prediction.

target	dataset	Seq. size	PSIPRED	PROFsec	PHD	PHDPSI	SSPRO	PG	Zhang98	Zhang01	PSSC-core	LAMICA
Helix	EVA149	N $\leqslant$ 40	0.2080	0.1192	0.1370	0.1447	0.2176	-	0.1698	0.2969	0.1814	0.1192
		N > 40	0.0471	0.0560	0.0671	0.0603	0.0580	-	0.1131	0.0928	0.1023	0.0467
		All	**0.0589**	**0.0607**	**0.0723**	**0.0666**	**0.0698**	-	**0.1173**	**0.0949**	**0.1081**	**0.0521**
	EVA150	N $\leqslant$ 40	0.1559	0.0931	0.1450	-	-	0.1037	0.2167	0.1698	0.1421	0.0933
		N > 40	0.0695	0.0783	0.0766	-	-	0.0703	0.1198	0.0929	0.1152	0.0680
		All	**0.0851**	**0.0810**	**0.0889**	-	-	**0.0763**	**0.1372**	**0.0944**	**0.1201**	**0.0725**
Coil	EVA149	N $\leqslant$ 40	0.1954	0.1475	0.1434	0.1411	0.1694	-	0.1369	0.1035	0.1250	0.1452
		N > 40	0.0641	0.0577	0.0740	0.0739	0.0729	-	0.1094	0.0960	0.1004	0.0577
		All	**0.0738**	**0.0644**	**0.0791**	**0.0788**	**0.0800**	-	**0.1114**	**0.0961**	**0.1022**	**0.0641**
	EVA150	N $\leqslant$ 40	0.1353	0.0786	0.1142	-	-	0.0982	0.1245	0.1686	0.0976	0.0832
		N > 40	0.0761	0.0814	0.0824	-	-	0.0842	0.1213	0.1112	0.0644	0.0642
		All	**0.0867**	**0.0809**	**0.0881**	-	-	**0.0867**	**0.1219**	**0.1123**	**0.0704**	**0.0676**
Strand	EVA149	N $\leqslant$ 40	0.1231	0.1429	0.1578	0.1679	0.1568	-	0.1873	0.1934	0.1396	0.1382
		N > 40	0.0489	0.0488	0.0674	0.0588	0.0597	-	0.1025	0.1054	0.0976	0.0471
		All	**0.0544**	**0.0558**	**0.0741**	**0.0669**	**0.0668**	-	**0.1088**	**0.1063**	**0.1007**	**0.0483**
	EVA150	N $\leqslant$ 40	0.0643	0.068	0.1163	-	-	0.0931	0.2086	0.0195	0.1555	0.0806
		N > 40	0.0508	0.0637	0.0632	-	-	0.0617	0.1092	0.0880	0.0939	0.0488
		All	**0.0532**	**0.0645**	**0.0728**	-	-	**0.0674**	**0.1271**	**0.0867**	**0.1050**	**0.0508**

Table 5.1: Comparison of the average absolute content prediction error for helix, coil and strand prediction between LAMICA and nine competing methods; "-" denotes results that were not originally reported and that cannot be duplicated.

Predicted targets	PSIPRED	PROFsec	PG	PSSC-core
Helix	++	++	++	++
	2.20	7.53	3.55	3.56
Coil	$\sim$	++	++	++
	0.53	2.28	2.47	4.63
Strand	++	$\sim$	$\sim$	++
	6.79	0.04	1.44	2.79

Table 5.2: Paired t-test based comparison between LAMICA and four best performing competing methods for helix, coil, and strand content predictions on EVA150 (at the 95% level); ++/- - means that LAMICA provides statistically significantly better/worse prediction than the method listed in the corresponding column; The numbers indicate the t-values.

Content prediction errors for short and long sequences are presented in Table 5.1 in rows denoted as $N \leqslant 40$ and $N > 40$, respectively. We observe that the performance of prediction methods, which were used to generate inputs to LAMICA, i.e., PSIPRED and PROFsec, are consistent with our observations in Section 4.1. For both data sets, PROFsec outperforms PSIPRED for the helix prediction on the set of short proteins, i.e., $N \leqslant 40$, while for long proteins PSIPRED shows better results.

5.1.2 Coil Content

The *coil content* prediction results shown in Table 5.1 demonstrate that the proposed method provides improvement in coil prediction as well. For EVA149, our method offers 0.5% improvement over the second best method (PROFsec) and 13% improvement over the third best method (PSIPRED). For EVA150, the improvement over the second best method (PSSC-core) equals 4%, and the improvement over the third best method (PROFsec) is 16%. The improvements were found to be significant using a t-test (see Table 5.2). Although for coil content prediction PSIPRED provides statistically comparable results,

positive t-value indicate that LAMICA provides more accurate predictions. At the same time, the proposed method provides statistically significantly better content predictions when compared with the remaining considered competitors.

5.1.3 Strand Content

The prediction results of the *strand content* (see Table 5.1) are calculated from predicted helix and coil contents. The prediction results, given in bold, show that strand prediction is also improved by the proposed method. For EVA149, our method provides an 11% improvement over the second best method, PSIPRED, and 13% improvement over the third best method, PROFsec. For EVA150, the improvement over the second best method, PSIPRED, equals 4% and the improvement over the third best method, PROFsec, is 21%. These improvements were found to be significant when compared against PSIPRED. For PROFsec and PG, the differences are not statistically significant; however, LAMICA is shown to provide lower errors (see Table 5.2).

5.1.4 Comparison with Composition Computed from Predictions of PSIPRED and PROFsec

We also compare predictions of the proposed method with the content computed from secondary structure predicted with PSIPRED and PROFsec methods, where PROFsec is used for proteins of length $N \leqslant 40$, and PSIPRED is used for proteins with length $N > 40$. This illustrates the performance of combination of two separate prediction models from short and long sequences, as used in the proposed method. The difference between the results based on PSIPRED/PROFsec content and results of the proposed method shows the value added by utilizing the prediction system. The average absolute error of the PSIPRED/PROFsec predicted helix content, for EVA149 is 0.0524 and for EVA150 equals 0.0738. In contrast, using LAMICA results in 0.5% improvement for EVA149 and 2% improvement for EVA150. For coil content pre-

diction the average absolute error based on predicted secondary structure for EVA149 equals 0.0702 and for EVA150 equals 0.0765. Application of our prediction system gives 8% improvement for the EVA149 and 11% improvement for the EVA150. The improvements for the strand content prediction equal 13% and 5% on the two data sets, respectively. The consistency and amount of improvement validate the quality of the proposed prediction method.

Fig 5.1 demonstrates the performance of the proposed method when compared to PROFsec, which is the second best method for α-helix and coil content on EVA150. A point with coordinates (x, y) corresponds to one protein sequence where x is the predicted content and y is the actual content. Hollow circles correspond to the PROFsec predictions and the black dots correspond to the proposed prediction method. We observe that the PROFsec prediction results are spread much wider from the diagonal line (perfect predictions) as compared to our predictions. This effect is even stronger for the coil prediction (Fig. 5.1 (b)). For each set of predictions, we draw two straight lines which indicate the confidence interval area, i.e., the predictions located within the standard deviation from the mean error. These lines can be expressed by points with coordinates $(z + \mu + \sigma, z)$ and $(z + \mu - \sigma, z)$, where $z \in (0, 1)$ is a parameter, μ is the mean of the prediction error, and σ is a standard deviation of the prediction error. For PROFsec prediction results, the lines are dashed, while for LAMICA the lines are solid. We observe that standard deviations of the proposed method are smaller than for PROFsec in both cases. We also note that PROFsec's predictions tend to underestimate α-helix content (lines are shifted downwards) and overestimate coil content, while LAMICA's predictions are centered closer to the diagonal.

Finally, we observe that, although, on average, the errors for the EVA150 data set are larger than for the EVA149 data set, the same relationships between the predictions of individual methods are observed for both sets. We emphasize that these differences are not a result of using EVA149 during the design of our method since we observe the same differences for other prediction methods, in which case no bias to a specific data set could be attributed.

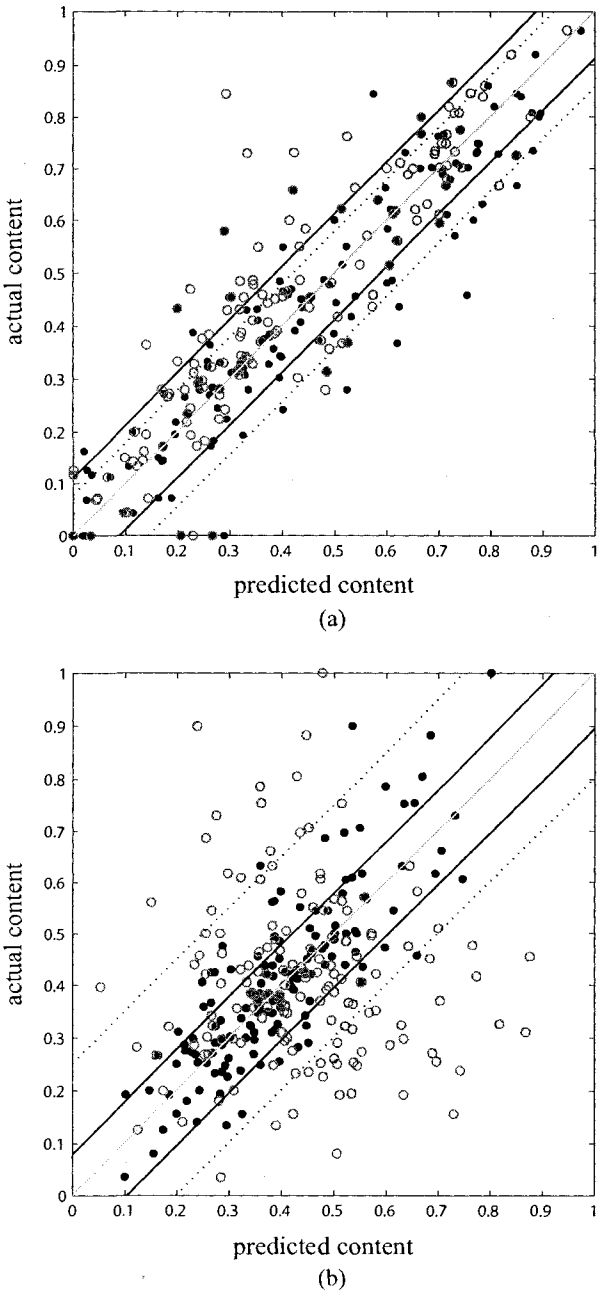

Figure 5.1: Comparison of helix and coil content in PROFsec and the proposed method on EVA150; (a) helix content; (b) coil content; Hollow circles correspond to the PROFsec predictions and the black dots correspond to the proposed prediction method.

5.2 Applications

We discuss two applications of the proposed content prediction method to improve structural class assignment and the secondary structure prediction.

5.2.1 Structural Class Assignment

Information about the structural classes is used in a variety of predictive tasks addressing protein structure and function. More specifically, the knowledge of structural classes was applied to improve the accuracy of secondary structure prediction (Gromiha and Selvaraj, 1998), to reduce the search space of possible conformations of the tertiary structure (Chou, 1995; Bahar *et al.*, 1997), to implement a heuristic approach to find tertiary structure (Carlacci *et al.*, 1991), to discriminate outer membrane proteins (Gromiha and Suwa, 2005), predict protein folding rates (Gromiha, 2005b) and unfolding rates (Gromiha *et al.*, 2006), and to predict DNA-binding sites (Kuznetsov *et al.*, 2006) and protein folds (He *et al.*, 2002). Prior studies have shown that secondary structure content can improve the quality of the protein structural class prediction (Kurgan *et al.*, 2006; Kurgan and Chen, 2007). In this work we use the predicted content to perform class assignment according to Eisenhaber's method (Eisenhaber *et al.*, 1996), which was recently shown to outperform other class assignment methods (Kurgan *et al.*, 2008), using EVA150 data set. Class assignment which uses LAMICA content is compared with the assignment based on content extracted from PSIPRED and PROFseq secondary structure prediction. The results indicate that use of LAMICA-predicted content as an input gives 89.32% class assignment accuracy and delivers 8.3% error rate reduction compared to use of PSIPRED content and a 26.6% reduction compared to use of PROFsec content. This experiment confirms that LAMICA is capable of supplying more accurate content estimates for the structural classes assignment.

5.2.2 Secondary Structure Prediction

The predicted content of LAMICA was also applied to improve the secondary structure predicted by PSIPRED on EVA150. We used each AA's secondary structure probability (generated by PSIPRED), structural content of the PSIPRED prediction, and the predicted content of the proposed method to improve the predicted secondary structure. Since the AA's secondary structure probabilities of PSIPRED were not available for EVA150 on the EVA server, we obtained PSIPRED structure prediction and the probabilities using PSIPRED server (McGuffin *et al.*, 2000) (using PSIPRED version 2.5). To improve the predicted secondary structure accuracy using the content predicted by LAMICA we adjust the number of AAs predicted as helical residues. This is motivated by recent study of the distribution of the predicted N-terminus positions of helical segments (Wilson *et al.*, 2004). We compute the number of helical residues that should be removed or added, assuming that LAMICA delivers more accurate content prediction, by comparing the helix content of PSIPRED with the content of LAMICA. Removal corresponds to changing the structure predicted as helix to coil. We consider helix probabilities of the AAs in helical N-terminal and C-terminal positions, and change the structure of AA with lowest helix probability to a coil. Adding corresponds to changing the structure of the AAs positioned at the interface of helical segments (immediately before or after), that have the highest helix probability to helix. Q_3 scores (the percentage of residues correctly predicted in all three states) of PSIPRED V2.5 and PROFsec for the whole EVA150 set equal 77.9% and 77.7% respectively. For short protein sequences in EVA150 the corresponding Q_3 scores are 71.3% and 75.1%. Applying the content predicted by LAMICA and the abovementioned rules we obtain 78.3% accuracy for the whole EVA150 set and 76.0% accuracy for the short sequences, showing substantial improvements. Fig. 5.2 demonstrates the above procedure for 1sse_A protein predicted by PSIPRED V2.5. The number of helical residues predicted by LAMICA for this protein equals 17, while PSIPRED V2.5 predicted 24. Therefore, we convert 7 helices

to coils. This experiment confirms that LAMICA is capable of supplying more accurate content estimates which can be used to improve secondary structure prediction.

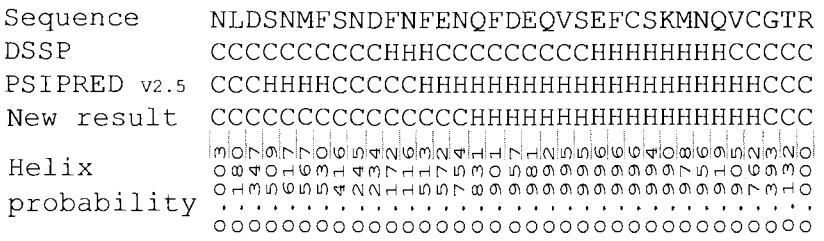

Figure 5.2: Example of the procedure used to improve the predicted secondary structure based on the predicted content.

Chapter 6

Summary and Conclusion

This chapter summarizes the contributions of this thesis, and draws the conclusions.

6.1 List of Contributions

The main contributions of the presented work are as follows:

- We design a novel content prediction method which outperforms state-of-the-art predictions in terms of the average absolute error.
- We show that combining structural and sequence based features is a suitable approach to tackle content prediction problems.
- We reveal that separate consideration of short and long proteins, separate training, feature selection and prediction improves the result.
- We designed and implemented an accurate feature selection algorithm based on a two-stage approach where the first stage selects the features that are the most correlated with the content value while the second stage is based on PCA analysis.
- We show that the established method provides useful input for a number of other practical prediction tasks such as structural class assignment and

secondary structure prediction.

6.2 Conclusion

A novel machine learning method called LAMICA for prediction of protein secondary structure content is proposed. Two sets of learning features are generated, the structural features and sequence based physicochemical features. To reduce the prediction error, two separate SVR based learning models, one for short and one for long sequences, are constructed. Experimental results obtained using two independent test sets demonstrate that the prediction error of LAMICA is smaller than the error of current prediction techniques reported in the literature, including content predictions performed directly from sequence and predictions computed from predicted secondary structure. LAMICA improves predictions for all three content values (helix, coil, and strand). It is also shown how content predicted by LAMICA can be used to improve results of other related prediction tasks, such as structural class prediction and prediction of the protein secondary structure. The major reasons for improvement of content prediction results are combining structural and sequence based features, classifying the protein sequences to short sequences and long ones, and separate training and feature selection. We note that performance improvement offered by the proposed method depends on the number of short protein sequences in data sets.

Bibliography

Altschul,S.F., Maden,T.L., Schaffer,A.A., Zhang,Z., Miller,W., Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucl. Acids. Res.*, **25**, 3389-3402.

Bahar,I., Atilgan,A. R., Jernigan,R. L., Erman,B. (1997) Understanding the Recognition of Protein Structural Classes by Amino Acid Composition, *Proteins*, **29**, 172-185.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne,P. (2000) The protein data bank, *Nucleic Acids Res.*, **28**, 235-242.

Birzele,F., Kramer,S. (2006) A new representation for protein secondary structure prediction based on frequent patterns, *Bioinformatics*, **22**, 2628-2634.

Bystroff,C., Thorsson,V., Baker,D. (2000) HMMSTR: a Hidden Markov Model for Local Sequence-Structure Correlations in Proteins, *J. Mol. Biol.*, **301**, 173-190.

Cai,Y.D., Liu,X.J., Chou,K.C. (2003) Prediction of protein secondary structure content by artificial neural network, *J Computational Chemistry*, **24**, 727-731.

Carlacci,L., Chou,K.C., Maggiora,G.M. (1991) A Heuristic Approach to Predicting the Tertiary Structure of Bovine Somatotropin, *Biochemistry*, **30**, 4389-4398.

Chen,K., Kurgan,L.A. (2007) PFRES: Protein Fold Classification by Using Evolutionary Information and Predicted Secondary Structure, *Bioinformatics*, **23**, 2843-2850.

Chou,K.C. (1999) Using pair-coupled amino-acid composition to predict protein secondary structure content, *J Protein Chemistry*, **18**, 473-480.

Chou,K.C., Zhang,C.T. (1995) Prediction of Protein Structural Classes, *Critical Review in Biochem. and Molecular Biology*, **30**, 275-349.

Chou,K.C. (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, *Proteins*, **21**, 319-344.

Chou,P.Y., Fasman,G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence, *Adv. Enzymol.*, **47**, 45-148.

Christianini,N., Shaw-Taylor,J. (2000) An Introduction to Support Vector Machines and other Kernel-Based Learning Methods, *Cambridge University Press, Cambridge.*

Concepcion,G.P., David,M.P., Padlan,E.A. (2005) Why don't humans get scrapie from eating sheep? A possible explanation based on secondary structure predictions, *Med Hypotheses*, **64**, 919-924.

Dobson,P.D., Doig,A.J. (2003) Distinguishing enzyme structures from non-enzymes without alignments, *J Mol Biol.*, **330**, 771-783.

Dobson,P.D., Doig,A.J. (2005) Predicting enzyme class from protein structure without alignments, *J Mol Biol.*, **345**, 187-199.

Eisenberg,D., Schwarz,E., Komaromy,M., Wall,R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot, *J Mol Biol.*, **179**, 125-142.

Eisenhaber,F., Frommel,C., Argos,P. (1996) Prediction of Secondary Structural Content of Proteins From Their Amino Acid Composition Alone. II. The Paradox With Secondary Structural Class, *Proteins*, **25**, 169-179.

Engelman,D.M., Steitz,T.A., Goldman,A. (1989) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu Rev Biophys Chem*, **15**, 321-353.

Fauchere,J.L., Pliska,V.E. (1983) Hydrophobic parameters of amino-acid sidechains from the partitioning of N-acetyl-amino-acid amide, *Eur J Med Chem*, **18**, 369-375.

Ganapathiraju,M.K., Klein-Seetharaman,J., Balakrishnan,N., Reddy,R. (2004) Characterization of protein secondary structure, *IEEE Sig. Pro. Mag*, **15**, 78-87.

Gong,H., Isom,D.G., Srinivasan,R., Rose,G.D. (2003) Local secondary structure content predicts folding rates for simple, two-state proteins, *J Mol Biol.*, **327**, 1149-1154.

Gromiha,M.M., Selvaraj,S. (1998) Protein Secondary Structure Prediction in Different Structural Classes, *Protein Engineering*, **11**, 249-251.

Gromiha,M.M., Suwa,M. (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy, *Bioinformatics*, **21**, 961-968.

Gromiha,M.M. (2005) A statistical model for predicting protein folding rates from amino acid sequence with structural class information, *J. Chem Inf Model.*, **45**, 494-501.

Gromiha,M.M., Selvaraj,S., Thangakani,A.M. (2006) A Statistical method for predicting protein unfolding rates from amino acid sequence, *J Chem Inf Model*, **46**, 1503-1508.

Gromiha,M.M., Selvaraj,S. (2008) Bioinformatics approaches for understanding and predicting protein folding rates, *Current Bioinformatics*, **3**, 1-9.

Guo,J., Chen,H., Sun,Z., Lin,Y. (2004) A Novel Method for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles, *Proteins*, **54**, 738-743.

He,H., McAllister,G., Smith,T.F. (2002) Triage protein fold prediction, *Proteins*, **48**, 654-663.

Hobohm,U., Sander,C. (1995) A sequence property approach to searching protein databases, *J. Mol. Biol.*, **251**, 390-399.

Holm,L., Sander,C. (1994) The FSSP database of structurally aligned protein fold families, *Nucleic Acids Res*, **22**, 3600-3609.

Holm,L., Sander,C. (1998) Dictionary of recurrent domains in protein structures, *Proteins*, **33(1)**, 88-96.

Homaeian,L., Kurgan,L.A., Ruan,J., Cios,K.J., Chen,K. (2007) Prediction of protein secondary structure content for the twilight zone sequences, *Proteins*, **69**, 486-498.

Hua,S., Sun,Z. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, *J. Mol. Biol.*, **308**, 397-407.

Huang,J.T., Cheng,J.P. (2007) Prediction of folding transition-state position (βT) of small, two-state proteins from local secondary structure content, *Proteins*, **68**, 218-222.

Ivankov,D.N., Finkelstein,A.V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure, *Proc Natl Acad Sci U S A*, **101**, 8942-8944.

Jones,D.T. (1999) Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices, *J. Mol. Biol.*, **292**, 195-202.

Krigbaum,W.R., Knutton,S.P. (1973) Prediction of the Amount of Secondary Structure in a Globular Protein from its Aminoacid Composition, *Proc Natl Acad Sci USA*, **70**, 2809-2813.

Kurgan,L.A., Rahbari,M., Homaeian,L. (2006) Impact of the Predicted Protein Structural Content on Prediction of Structural Classes for the Twilight Zone Proteins, *5th Intern. Conf. on Machine Learning and Applications*, **5**, 180-186.

Kurgan,L.A., Chen,K. (2007) Prediction of protein structural class for the twilight zone sequences, *Biochem Biophys Res Commun.*, **357**, 453-460.

Kurgan,L.A., Zhang,T., Zhang,H., Shen,S., Ruan,J. (2008) Secondary Structure Based Assignment of the Protein Structural Classes, *Amino Acids*, accepted.

Kurgan,L.A., Cios,K., Chen,K. (2008) SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences, *BMC Bioinformatics*, **9**, 198-226.

Kuznetsov,I.B., Gou,Z., Li,R., Hwang,S. (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins, *Proteins*, **64**, 19-27.

Lee,S., Lee,B., Kim,D. (2006) Prediction of Protein Secondary Structure Content Using Amino Acid Composition and Evolutionary Information, *Proteins*, **62**, 1107-1114.

Lesk,A.M. (2002) Introduction to Bioinformatics, Oxford University Press, New York.

Levitt,M., Chothia,C. (1976) Structural patterns in globular proteins, *Nature*, **261**, 552-558.

Lin,K., Simossis,V.A., Taylor,W.R., Heringa,J. (2005) A simple and fast secondary structure prodiction method using hidden neural networks, *Bioinformatics*, **21**, 152-159.

Lin,Z., Pan,X. (2001) Accurate prediction of protein secondary structural content, *J Protein Chemistry*, **20**, 217-220.

Liu,W., Chou,K.C. (1999) Protein secondary structural content prediction, *Protein Engineering*, **12**, 1041-1050

Liu,J., Gough,J., Rost,B. (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines, *PLoS Genet.*, **2**, 529-536.

Lodish,H., Berk,A., Zipursky,S.L., Matsudaria,P., Baltimore,D., Darnell,J.E. (2000) Distinguishing protein-coding from non-coding RNAs through support vector machines, *Molecular Cell Biology*, **4**, W.H. Freeman and Company, New York.

McGuffin,L.J., Bryson,K., Jones,D.T. (2000) The PSIPRED protein structure prediction server, *Bioinformatics*, **16**, 404-405.

Montgomerie,S., Sundararaj,S., Gallin,W.J., Wishart,D.S. (2006) Improving the accuracy of protein secondary structure prediction using structural alignment, *BMC Bioinformatics*, **7**, 301-314.

Murzin,S., Brenner,S., Hubbard,W.J., Chothia,D.S. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, **247**, 536-540.

Muskal,S.M., Kim,S.H. (1992) Predicting protein secondary structure content. A tandem neural network approach, *J Mol Biol.*, **225**, 713-727.

Nakai,K., Kidera,A., Kanehisa,M. (1988) Cluster analysis of amino acid indices for prediction of protein structure and function, *Protein Engng*, **2**, 93-100.

Nakashima,H., Nishikawa,K., Ooi,T. (1986) The folding type of a protein is relevant to the amino acid composition, *J. Biochem.*, **99**, 153-162.

Nelson,D., Cox,M. (2000) Lehninger Principles of Biochemistry Amino, *Worth Publishers.*

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B., Thornton,J.M. (1997) CATHa hierarchic classification of protein domain structures, *Structure*, **5(8)**, 1093-1108.

Pauling,L., Corey,R.B., Branson,H.R. (1951) The structure of proteins: two hydrogen bonded helical conformations of the polypeptide chain, *Proc. Natl. Acad. Sci. USA*, **37**, 205-211.

Pollastri,G., Przybylski,D., Rost,B., Baldi,P. (2002) Improving the Prediction of Protein Secondary Structure in three and egight Classes Using Recurrent Neural Networks and Profiles, *Proteins*, **47**, 228-235.

Prabhakaran,M., Ponnuswamy,P.K. (1979) The spatial distribution of physical, chemical, energetic and conformational properties of amino acid residues in globular proteins, *J. theoret. Biol.*, **80**, 485-504.

Przybylski,D., Rost,B. (2002) Alignments grow, secondary structure prediction improves, *Proteins*, **46**, 197-205.

Rost,B., Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy, *J Mol Biol*, **2**, 584-599.

Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks, *Methods Enzymol*, **266**, 525-539.

Rost,B., Eyrich,V.A. (2001) EVA: large-scale analysis of secondary structure prediction, *Proteins*, **5**, 192-199.

Ruan,J., Wang,K., Yang,J., Kurgan,L., Cios,K.J. (2005) Highly Accurate and Consistent Method for Prediction of Helix and Strand Content from Primary Protein Sequences, *Artificial Intelligence in Medicine*, **35**, 19-35.

Smith,J., Diez,G., Klemm,A.H., Schewkunow,V., Goldmann,W.H. (2006) CapZ-lipid membrane interactions: A computer analysis, *Theo. Bio. and Med. Model.*, **3**, 30-37.

Smola,A. (1996) Regression Estimation with Support Vector Learning Machines, *Technische University Mnchen.*

Tomii,K., Eyrich,M. (2001) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, *Protein Eng.*, **9**, 27-36.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory.* Springer, New York, ISBN 0-387-94559-8.

Vapnik,V. (1998) *Statistical Learning Theory.* John Wiley and Sons, New York.

Wang,J., Ma,Q., Shasha,D., Wu,C.H. (2000) 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.

Ward,J.J., McGuffin,L.J., Buxton,B.F., Jones,D.T. (2003) Secondary structure prediction with support vector machines, *Bioinformatics*, **19**, 1650-1655.

Wilson,C.L., Boardman,P.E., Doig,A.J., Hubbard,S.J. (2004) Improved Prediction for N-Termini of α-Helices Using Empirical Information, *Proteins*, **57**, 322-330.

Witten,X., Frank,B. (2005) Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, San Francisco.

Yang,X., Wang,B. (2003) Weave amino acid sequences for protein secondary structure prediction. In: Proceedings of the eight ACM SIGMOD workshop on research issues in data mining and knowledge discovery, **9**, 80-87.

Zhang,C.T., Zhang,Z., He,Z. (1996) Prediction of the secondary structure content of globular proteins based on structural classes, *J Protein Chemistry*, **15**, 775-786.

Zhang,C.T., Lin,Z.S., Zhang,Z., Yan,M. (1998) Prediction of helix/strand content of globular proteins based on their primary sequences, *Protein Engng.*, **11**, 971-979.

Zhang,C.T., Zhang,Z., He,Z. (1998) Prediction of the secondary structure contents of globular proteins based on three structural classes, *J Protein Chemistry*, **17**, 261-272.

Zhang,Z., Sun,Z., Zhang,C.T. (2001) A New Approach to Predict the Helix/Strand Content of Globular Proteins, *J. theor. Biol.*, **208**, 65-78.