PROTEIN STRUCTURAL CLASS PREDICTION USING PREDICTED SECONDARY STRUCTURE AND HYDROPATHY PROFILE

by

Syeda Nadia Firdaus

Bachelor of Science in Computer Science and Engineering

The University of Asia Pacific, Bangladesh, 2005

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Science

in the Program of

Computer Science

Toronto, Canada, 2013

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Protein Structural Class Prediction Using Predicted Secondary Structure and Hydropathy Profile

Syeda Nadia Firdaus

Master of Science in Computer Science

Ryerson University, 2013

Abstract

This thesis explores machine learning models based on various feature sets to solve the protein structural class prediction problem which is a significant classification problem in bioinformatics. Knowledge of protein structural classes contributes to an understanding of protein folding patterns, and this has made structural class prediction research a major topic of interest. In this thesis, features are extracted from predicted secondary structure and hydropathy sequence using new strategies to classify proteins into one of the four major structural classes: all-α, all-β, α/β, and $\alpha+\beta$. The prediction accuracy using these features compares favourably with some existing successful methods. We use Support Vector Machines (SVM), since this learning method has well-known efficiency in solving this classification problem. On a standard dataset (25PDB), the proposed system has an overall accuracy of 89% with as few as 22 features, whereas the previous best performing method had an accuracy of 88% using 2510 features.

Acknowledgements

I would like to express sincere gratitude to my supervisor Dr. Eric Harley for his continuous support, encouragement and time throughout the two years of graduate studies. Dr. Eric Harley has guided me with patience and care to improve my work in thesis. Without his kind cooperation and support, completion of this thesis and publication in a conference on this thesis work would not be possible. Working under the supervision of Dr. Eric Harley has been a great and memorable experience for me.

I would like to thank my thesis committee members: Dr. Isaac Woungang, Dr. Alireza Sadeghian, and Dr. Abdolreza Abhari for their time, patience and proficiency in judging my thesis. Their valuable opinions were very helpful to improve my work.

I would also like to convey my gratitude to the faculty members of the Department of Computer Science, Ryerson University. Attending courses under guidance of committed professors helped me to advance my knowledge in computer science.

My sincere appreciation goes to staff members of Computer Science department and fellow graduate students for their continuous support over the last two years.

Lastly, I am extremely grateful for the support and inspiration of my family. Without them I would not be able to attain my goal and fulfil my dream to work in my field of interest. No specific word of appreciation would be enough to convey my love and thankfulness towards them.

Dedication

To my family

Contents

List of Tables

List of Figures

List of Abbreviations

A	Amino acid with "ambivalent" state
Acc	Accuracy
Avg	Average
C	Amino acid in random coil
CATH	Class, Architecture, Topology and Homologous superfamily
CFS	Correlation based Feature Selection
CMV	Composition Moment Vector
CV	Cross Validation
E	Amino acid in beta-sheet (context: predicted secondary structure)
E	Amino acid with "external" state (context: hydrophathy profile)
H	Amino acid in alpha-helix
HS	Hydropathy Sequence
I	Amino acid with "internal" state
MCC	Matthews Correlation Coefficient
PDB id	Protein Data Bank Id
Prec	Precision
PredSSS	Predicted Secondary Structure Sequence
PSI-BLAST	Position Specific Iterative Basic Local Alignment and search Tool
PSIPRED	PSI-BLAST Predict Secondary Structure
PSSM	Position Specific Scoring Matrix

SCOP	Structural Classification of Protein
SCMV	State Change Moment Vector
Spec	Specificity
Sens	Sensitivity
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency

Chapter 1

Introduction

In the fast paced scientific world, the amount of biological data is already vast and continues to grow rapidly. These data are handled by applications developed in the field of bioinformatics. In this field of science, the discipline of biology, computer science, and information technology merge together to face the challenges of biological science [1]. The research done in bioinformatics is mainly focused on managing biological data and extracting useful information from them. Structural bioinformatics is a sub-section of bioinformatics which is concerned with the use of biological structures like proteins, DNA, RNA, etc., to advance the knowledge of biological systems [2]. Research is being done on biological macromolecule structure prediction and structural classification. Predicting the structural class of protein is a major area of research in structural bioinformatics due to its importance in understanding the nature and function of protein. Protein is the basic building block of every living cell and participates essentially in every biological process within cells. Understanding the function of each protein is very important to generate insight into biological systems. Predicting the structural class of a protein has become a major topic of interest due to its contribution towards understanding protein folding patterns and their impact on function.

1.1 Motivation

Protein structural class prediction is an important area of research within the field of overall protein structure prediction. Protein structural class focuses on one global aspect in our understanding of protein folding. Each protein has an unique 3D shape created by its folding

which determines its function. Structure prediction is an important area of research since it helps one to understand or discover the function of unknown proteins. Details of the 3D structures of proteins are very complicated, irregular, and expensive to determine. Researchers try to find out overall topological folding patterns of a protein which are simple, regular, and similar to other proteins. The goal of protein structural class prediction methods is to find out some simple or regular patterns from complicated or irregular 3D structures and then apply these patterns to predict the desired but still unknown information about proteins. Since folding can determine protein function and a wrongly folded protein causes disease, predicting structural class is of interest to the researchers from the drug industry as well.

The importance of determining protein structural class to obtain knowledge about the overall shape and function of protein, made us interested to do research in this area. Protein structural class prediction is a mature area of research, but problem has not been fully solved yet. The latest paper achieved 87% accuracy with very high dimensional feature vector. With our research, we tried to contribute some new measures to predict protein structural class more accurately and with fewer features.

1.2 Research Statement

The object of these research is to develop methods which can predict the structural class of a protein accurately. Generally if a method follows the SCOP classification scheme [3], then it classifies a protein into one of the following main four structural classes: all-α, all-β, α+β, and α/β. If a method follows CATH classification scheme [4] then it classifies a protein into one of the following main four classes: mainly α, mainly β, mixed α,β, and a few secondary structures. (Description of SCOP and CATH are in Section 2.6).

More specifically, the goal of our research is to develop a structural class prediction method which follows SCOP classification scheme and can predict proteins with twilight-zone similar sequences into structural classes more accurately with less number of features. Twilight-zone similar sequences are the sequences which are very less similar to each other. In protein structural class prediction research, use of low similar sequences as training and testing data is very important because classification models are developed based on machine learning techniques. The performance of these classification models is measured using the training and

testing sequences from the same dataset. If the sequences in the dataset are very similar, then the model is trained and tested with similar sequences, resulting in misleadingly high accuracy. When sequences in the dataset are less similar then the prediction accuracy will truly show its performance towards unknown, maybe less similar data.

We use 25PDB dataset which is very popular for work with low similar sequences. The Support Vector Machine (SVM) soft computing technique is used to build the classification model. We have generated several feature sets and checked the performance of the resulting models. Our objective is to extract effective features from protein sequences, in order to obtain more accurate prediction using less features.

1.3 Objective

The objective of this research is to explore some new ideas for extracting features from protein amino acid sequence and predicted secondary structure sequence in order to predict structural classes more accurately using fewer features than other published methods. To achieve this objective, the plan of work is as follows:

- Construct feature sets including new features from predicted protein secondary structure sequence and hydropathy sequence corresponding to amino acid sequence of protein and evaluate their performances.
- Evaluate the effectiveness of using the Term Frequency-Inverse Document Frequency Technique to extract useful patterns from protein secondary structure sequences to determine protein structural class.
- Construct a feature set using patterns extracted from the sequence constructed using hydropathy profile of amino acids in protein amino acid sequence and check its performance.
- Check the performance of combinations of feature sets for structural class prediction.

1.4 Assumption and Scope

The objective of this thesis is to predict structural classes using fewer features than other published methods and obtaining more accurate results. We assume that using less features will reduce computational time and resource usage. We use the Support Vector Machine classification algorithm, assuming that it will give good performance for our classification model as it has for previous published methods. We did not check other classification algorithms like Neural Network and Fuzzy Logic as want to compare our results with other published results which used the SVM classification algorithm.

We restrict the scope of this research to the 25PDB dataset, since it is a benchmark dataset for low similarity sequences.

1.5 Organization of Chapters

The thesis is organized as follows:

- Chapter 2 describes some introductory information about proteins, structures and structural classes of proteins. The information presented here is related to later discussion in the thesis.
- Chapter 3 presents some recent significant published research in the area of protein structural class prediction.
- Chapter 4 is presents the materials and method used in the thesis. The dataset, feature sets and classification method are described in this chapter.
- Chapter 5 presents the results of the thesis work. This chapter includes the comparison of performance of various feature sets developed for this thesis. The comparison of performance of the best performing feature sets with some major published work is also presented in this chapter.
- Chapter 6 concludes the thesis work along with some proposals for future research.

Chapter 2

Preliminaries

In this chapter, some preliminary information regarding proteins, protein structures and structural classes are provided. The concept presented here are relevant for later description of thesis work.

2.1 Protein

Protein is an essential component to the structure and function of all living cells. It is a complex molecular compound consisting of amino acids joined by peptide bonds. The peptide bonds link the carboxyl group (-COOH) of one amino acid to the amino group ($-NH_2$) of another amino acid [5]. There are 20 different amino acids known as residues (see Table 2.1) [6]. The chain of amino acids comprising a protein is folded into a unique three dimensional shape. The unique sequence and shape of a protein determines its function. The shape of a protein is described using four levels of structure: primary, secondary, tertiary, and quaternary.

2.2 Protein Primary Structure

The linear sequence of amino acids in a protein is referred to as the primary structure of the protein. For example, the primary structure of protein, "Angiotensin I" (PDB id of "Angiotensin I" is 1N9U) is: "asp - arg - val - tyr - ile - his - pro - phe - his - leu" [7]. Each amino acid in this example is represented by its three-letter code. The sequence is also written as "D-R-V-Y-I-H-P-F-H-L", in one letter codes. 20 amino acids and their corresponding one-letter and three-letter codes are given in Table 2.1 [6].

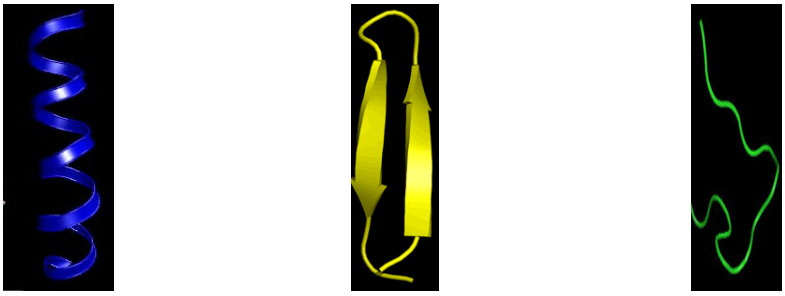
Table 2.1: 20 amino acids and their one-letter and three-letter code [6].

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

2.3 Protein Secondary Structure

In a protein, amino acids adjacent to one another interact to form segments with defined structure called secondary structure. The most common secondary structure elements are α-helix and β-sheet, introduced by Linus Pauling and coworkers in 1951 [8].

An α-helix segment is a single, spiral chain of amino acids stabilized by hydrogen bonds [9]. A β-sheet segment consists of two or more polypeptide chains, called β-strand, where hydrogen bonds between the chains form a twisted and pleated structure [10-11]. A segment with neither α-helix nor β-sheet structure is referred to as a random coil segment [12]. The β-sheets are said to be parallel or antiparallel, depending on whether the β-strands run in the same or opposite directions, respectively, where direction is by the amino-carboxyl orientation of the amino acids in the chain. Visualizations of α-helix (H) [13], β-sheet segment with two anti-parallel β-strands (E) [14], and random coil (C) segments are shown in Figure 2.1 [15].



(a) (b) (c)

Figure 2.1: Visualizations of (a) α-helix (H) segment [13], (b) β- sheet segment with two β-strands (E) [14], and (c) random coil (C) segment [15].

A protein's secondary structure is sometimes represented as a linear sequence of the letters H, E and C, according to whether the corresponding amino acid is in an α-helix, β- strand or random coil. This corresponding secondary structure sequence can be predicted with an accuracy of about 77% by some excellent methods including PSIPRED [16] and YASPIN [17]. An example of amino acid sequence of a protein along with its corresponding predicted secondary structure sequence (generated using PSIPRED) is given below:

Amino acid sequence of protein, "Probable translation initiation factor 2 beta subunit"

EILIEGNRTIIRNFRELAKAVNRDEEFFAKYLLKETGSAGNLEGGRLILQRR

Predicted secondary structure sequence:

CEEECCCHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHCCCCCCCCCCEEEEEEC

Here, each letter in the amino acid sequence represents the identity of the amino acid by its one-letter code, and each letter in the secondary structure represents the secondary structural state that the amino acid participates in.

2.4 Protein Tertiary Structure

The relative orientation of secondary structure elements with respect to each other determines the protein's unique three-dimensional shape known as the tertiary structure of the protein. The tertiary structure is very difficult and expensive to determine. Visualisation of a protein's 3D structure is shown in Figure 2.2 (created using Polyview visualization software [18] at http://polyview.cchmc.org/polyview3d.html) where the protein is colored according to its secondary structure segments (α-helix segments are colored as red, β-sheet segments are colored as green, and coil segments are colored as blue).

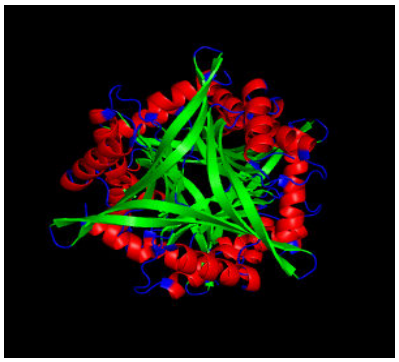


Figure 2.2: Visualization of tertiary structure of protein "Electron Transport" (PDB id of "Electron Transport" is 1naq) .

2.5 Protein Quaternary Structure

Many large proteins are composed of several individual protein chains. For these proteins, multiple, disconnected amino acid chains interact to form a larger structure referred to as the quaternary structure of the protein. An example visualization of a protein's quaternary structure is shown in Figure 2.3 (Created using Polyview visualization software [18]) where the two different chains in a protein are rendered by two different colors.

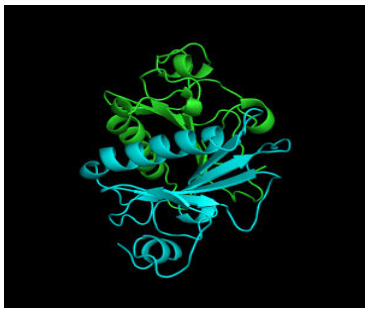


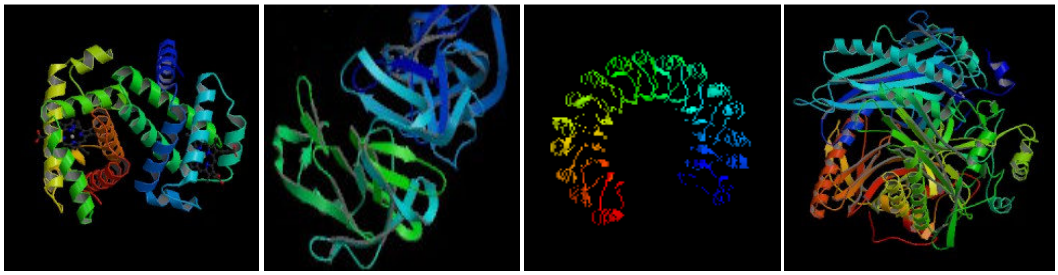
Figure 2.3: Visualization of quaternary structure of protein "Protein kinase inhibitor " (PDB id of " Protein kinase inhibitor " is 1AV5).

2.6 Protein Structural Classification

Protein structural classification is the clustering of proteins into groups based primarily on shape but also on other features. Structural classification is primarily based on simple and local folding patterns which reflect evolutionary relationships and structural similarities. The two most popular hierarchical protein structure classification schemes are SCOP (Structural Classification of Proteins) and CATH (Class, Architecture, Topology and Homologous superfamily). In SCOP proteins are assigned to the following hierarchical levels [3]:

- Family: Proteins having the same evolutionary origin are clustered into a family which is determined by either their sequence similarity or their structural and functional similarity.
- Superfamily: Families whose proteins share common structural and functional features are clustered into a superfamily.
- Common fold: Superfamilies whose proteins have same secondary structural elements in the same arrangements and common architecture are clustered into a common fold.

- Class: Common folds are clustered into a class based on their secondary structural content and some other features. The current version of the SCOP database, v. 1.75 (release in June,2009) classifies proteins into eleven structural classes: i) all-α, ii) all-β, iii) α/β, iv) $\alpha+\beta$, v) multi-domain protein, vi) membrane and cell-surface proteins, vii) small proteins, viii) coiled-coil proteins, ix) low resolution proteins, x) peptide, and xi) designed proteins.



(a) (b) (c) (d)

Figure 2.4: Visualisations of representative proteins belonging to the four structural classes: a) all-α (Name: Hemoglobin a, PDB id: 2hbc) [19], b) all-β (Name: jacalin alpha chain, PDB id: 1ku8) [20], c) α/β (Name: Ribonuclease inhibitor, PDB id: 1bnh) [21], and d) $\alpha + \beta$ (Name : Pyruvoyl-dependent histidine decarboxylase, PDB id: 1pya) [22]

Among these 11 classes, the four major structural classes are all-α, all-β, α/β and $\alpha+\beta$. Most researchers deal with these four structural classes, since they contain the majority of protein sequences. The main features of proteins belonging to these four structural classes are described below [3].

- all-α : The all-α class contains proteins that are basically composed of α-helix folding.
- all-β : The all- β class contains proteins that are basically composed of β-sheet folding.
- α/β : The α/β class includes proteins having alternating α-helix and β-strand.
- $\alpha+\beta$: The $\alpha+\beta$ class includes proteins where folds are formed by scattered α-helices and β-strands.

Visualisations of example proteins belonging to all these four classes are given in Figure 2.4.

The CATH structural classification scheme assigns proteins to the following hierarchal levels [4]:

- Homologous superfamily : Proteins having similar sequence, structures and functions are clustered into a homologous superfamily.
- Topology: Homologous superfamilies whose proteins share common arrangement and order of secondary structures are clustered into a topology.
- Architecture: This level of grouping is based on gross orientation or arrangement (example: barrel, roll or sandwich) of secondary structures in proteins.
- Class: This level of grouping is based on secondary structural content. In this level proteins are grouped into the following four categories: i) mainly α, ii) mainly β, iii) mixed α - β, and iv) few secondary structures.

One of the main differences between SCOP and CATH scheme lies in the class level. In the CATH scheme, there is only one class to represent mixed α and β. In the SCOP scheme, there are two separate classes (the α+β and the α/β class) to represent protein with both α and β secondary structure.

2.7 Support Vector Machine

Since its development by Vapnik and his group in former AT&T Bell Laboratories [23], Support Vector Machine (SVM) has proved to be an efficient technique for data classification and regression. The basic idea behind SVM technique is to create a linear separating hyperplane which maximizes the distance between two classes. Figure 2.5 [24] illustrates, with triangles and ovals, data belonging to two different classes. The classes can be fully separated by a hyperplane $w^T v + b = 0$, where v is a variable vector (x,y), w is a weight vector (w1,w2), and b is essentially another weight. The weights represent the model which the SVM binary machine infers from the training data. A binary SVM classifies data point v_i as belonging to the +1 class if $w^T v_i + b > 0$, and data point v_i is classified as -1 if $w^T v_i + b < 0$.

The decision boundary is the hyperplane from which the distance of nearest data point of each class is maximum. The distance between two classes of data should be as large as possible. Two

more hyperplanes H1 and H2 are considered that can also separate data and there is no data point between them. The distance between H1 and H2 is called the margin. The objective of SVM is to maximize the margin to reduce the probability of misclassification. It sets the value of w and b to maximize the distance between the planes $w^T v + b = \pm 1$.

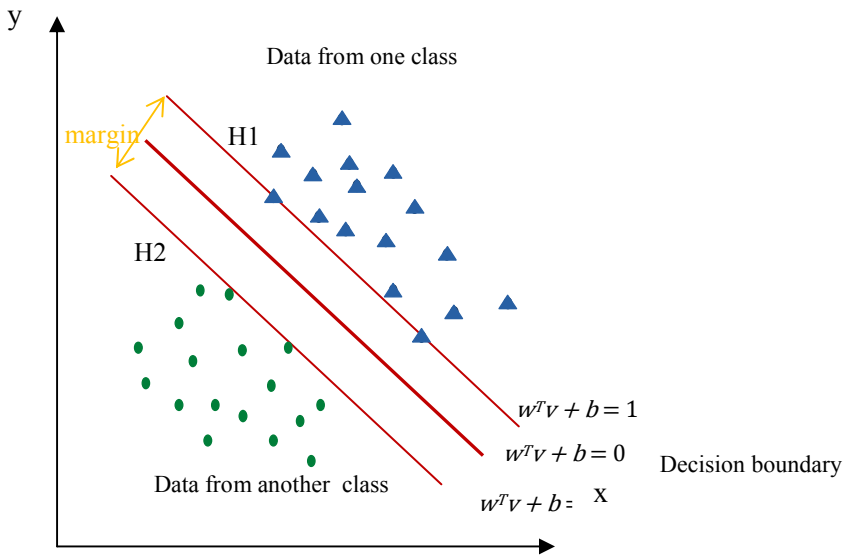
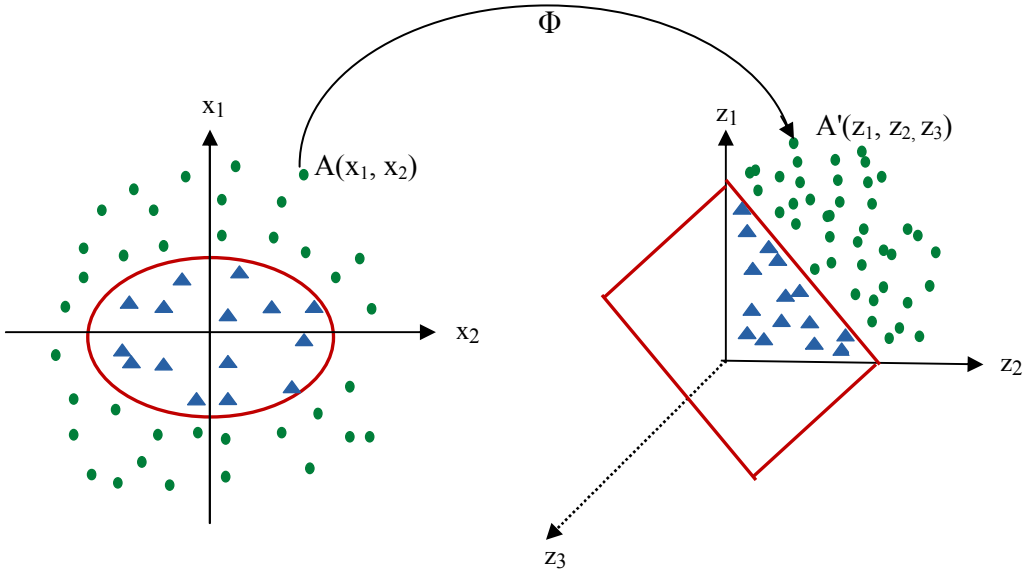


Figure 2.5: Linearly separable data [24]

In many cases, real world datasets are not perfectly linearly separable. SVM solves this problem by introducing a "soft margin" design. When there is no hyperplane that can separate the full data clearly, then the soft margin method selects a hyperplane that allows some data points of one class to be classified as a different class while separating data as well as possible.



(a) (b)

Figure 2.6: (a) Non separable data in 2D space, (b)Separated data in 3D space by hyperplane where transformation Φ is done by a kernel function [25].

When data are not linearly separable at all and the soft margin option alone does not help, then SVM handles the problem by mapping the input space into a higher dimensional feature space where there is more possibility to find a separating hyperplane. This mapping is done by a kernel function. SVM maps every data point of the input space into high-dimensional space via some transformation Φ: $x \rightarrow \varphi(x)$. For example, Figure 2.6(a) is showing non linearly separable data in 2D feature space [25]. SVM maps the data into 3D feature space where they are separable as shown in Figure 2.6(b) [25]. In Figure 4.3, data point 'A' in 2D space is (x_1, x_2) which is mapped to $A'(z_1, z_2, z_3)$ in 3D space.

Some basic kernel functions used by SVM techniques are as follows where x_i and x_j are feature vectors in input the space "." is a dot product, and K determines the mapping Φ [23][26]:

i) Polynomial (homogeneous) : $K(x_i, x_j) = (x_i . x_j)^d$

ii) Polynomial (inhomogeneous) : $K(x_i, x_j) = ((x_i . x_j) + 1)^d$

iii) Radial basis function $K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$

iv)Hyperbolic tangent: $K(x_i, x_j) = tanh(\rho(x_i, x_j) + c))$ for some $\rho > 0$ and $c > 0$

The main concept of SVM is based on binary classification. It is extended to do multi-class classification where a multi-class problem is considered as multiple binary class problem. The most commonly used multi-class classification approaches are as follows:

i) One-against-all: In this approach, for an n class problem, n binary classifiers are created where each classifier distinguishes data between one class and the remaining (n-1) classes. For example, for a 3 class (A,B,C) problem, 3 binary classifiers will classify test data as A / ~A, B / ~B and C / ~C, where "X / ~X" means "belong to class X or not belong to class X". Every classifier will calculate a decision function value regarding whether the test data belong to that class. Finally a test data is classified as belonging to the class for which the decision function value is highest.

ii) One-against-one: In this approach for the n class problem, $\frac{n(n-1)}{2}$ binary classifiers are designed. For each pair of classes, a binary classifier will classify data between that pair of classes. For example, for 3 class (A, B, C) problem, 3 binary classifier will classify data as A / B, A / C and B / C. Finally the classification is done using maximum win voting strategy. In this strategy when a binary classier assigns a test instance to one of the two classes then that class will get a vote. Lastly the class with the highest vote is considered as the true class for that test data.

Chapter 3

Related Work

The concept of protein structural class was proposed by Levitt and Chothia in 1976 [27]. They used a diagrammatic two dimensional representation to illustrate the known structure of 31 proteins. They classified proteins into 4 major structural classes (all-α, all-β, α/β, and $\alpha+\beta$) by visually inspecting the representation. The CATH structural classification scheme classifies a protein following the methods proposed by Levitt and Chothia [27] except for mixing the α/β and $\alpha+\beta$ class to create Mixed α-β class. In SCOP classification scheme, classification of proteins is done by visual inspection and comparison of sequence [28]. Both SCOP and CATH use visual inspection and automatic tools, but the papers are not clear on which tools are used.

In contrast to SCOP and CATH classification which rely on 3D structural analysis, there have been attempts to predict structural classes based on sequence and properties of amino acids. Protein structural class prediction is a significant problem studied by bioinformatics researchers for a long time. During the past 3 decades, many methods have been developed to address this problem. Success in this field is very slow due to the large number of possible protein structures and the lack of knowledge concerning factors influencing protein physical structure stability. Protein structural class prediction methods typically have two main steps. First, class discriminating features are extracted from amino acid sequences of proteins. Each protein can then be represented by a feature vector whose dimension is fixed, regardless of the length of the protein. In the second step those features are fed into suitable classification mechanisms to classify proteins into one of the main four structural classes. A classification algorithm indentifies the class of a new unseen instance based on the knowledge obtained during the

training phase. In the training phase, the data along with their known class information is fed into the classification model to prepare the model for testing phase. The main differences among the published methods are their strategies to extract class discriminating features from sequences and the choice of classification algorithm. Some of the features used in published methods are described in Section 3.1, and classification algorithms used by some successful methods are described in Section 3.2.

3.1 Feature Extraction Strategies

3.1.1 Features Extracted From Amino Acid Sequence of Proteins

The earliest methods of classification of proteins used only features extracted directly from the amino acid sequences. In those methods, the researchers established a correlation between amino acid sequences and the corresponding structural classes. Zerin et al. [29] used the occurrence frequency of each of 20 amino acid and all possible combination of three consecutive amino acids known as triplets in the protein sequence. Using support vector machine classification algorithm their method achieved 71.4% prediction accuracy. Subsequent research showed that information related to only amino acid composition, such as the frequency of each amino acid or peptide, may have limited prediction ability, since the folding pattern of a proteins is the result of collective interaction among the residues in protein sequence [30]. To improve the accuracy of predictions, features representing amino acid position and order were introduced in the later research. Wu et al. [31] combined amino acid word frequency, word position and physiochemical properties of amino acid to represent proteins, where a word is a short sequence of amino acids of length *n* also referred to as "n-gram" pattern. They calculated the position information of amino acids based on the concept of measuring inter-nucleotide distances as described in [32-33]. They transformed the amino acid sequence into a numerical sequence which contains position information of each element. For each of the 20 amino acids they used the interval distance between the two nearest positions of that amino acid and calculated the probability of occurrence of that amino acid at that interval. They also calculated the 1-word frequency (frequency of word with length "1") of hydropathy states in the sequences after

transforming amino acid sequence of protein to hydropathy sequence, based on the hydropathy profile of amino acids.

Zhang et al. [34] constructed a 46 dimensional feature vector, where 20 values represent amino acid frequency, 20 values represent amino acid correlation at various distances, and 6 values represent frequency of hydrophobic amino acid couples. The calculations of the amino acid distance correlation are described by Equation (1) - (5) of [34]. Hydrophobic amino acids are those that avoid interaction with water. The distance correlation is relevant because amino acids which are far apart in the sequence may be close neighbours after folding. They used the support vector machine algorithm based on a binary tree as described in [35]. Ding et al. [36] used the concept of pseudo amino acid composition (PseAA) introduced by Chou [37] to incorporate information about the order of amino acid residues in proteins as features. They used eight physiochemical properties like volume, polarity, and hydrophobic value to construct eight PseAA vectors to represent each protein. Each of these eight vectors was a 40 dimensional vector, where 20 values were the frequency of the 20 amino acids and 20 values were the correlation values between k-tier contiguous residues. Using each of these physiochemical property, they used Equations (2)-(6) of [37] to generate correlation values between k-tier (k=1,...,20) contiguous residues in the protein chain. The difference between Ding et al. [36] and chou's [37] method is that Ding et al. used eight different physiochemical property to generate eight PseAA vectors, whereas Chou [37] constructed only one PseAA vector using three physiochemical property values. For multiclass classification, Ding et al. [36] used dual layer fuzzy support vector machine (FSVM) as establish by Abe [38]. For each protein sample, eight PseAA vectors were fed into eight FSVM in the first layer. Outputs of the first layer generated by eight FSVM classifiers were again reclassified in the second layer. Their dual layer FSVM network showed 92.6% overall accuracy on a dataset taken from [39]. This accuracy is higher than accuracies reported in this thesis, however, as mentioned earlier, accuracy is influenced by choice of dataset. The high accuracy was achieved on dataset constructed by Chou [39]. There are a few more reports [40-42] based on extracting features from amino acid sequence of proteins.

3.1.2 Features Extracted from PSI-BLAST Profiles of Sequence

PSI-BLAST [43] (Position-Specific Iterative Basic Local Alignment Search Tool) profiles of sequences have also been used in structural class prediction methods as they reflect the evolutionary relationship among sequences [44-45]. PSI-BLAST [43] generates a position-specific scoring matrix (PSSM) or profile from multiple sequence alignment which reflects how closely a query sequence is to the database of collected sequences. Taigang et al. [44] transformed the PSSM generated by PSI-BLAST into a fixed length feature vector by auto covariance (AC) transformation. They used AC transformation as it is a powerful statistical tool for analyzing sequence vectors in other areas of bioinformatics [46-49]. Their model, using a combination of PSSM and the AC method, showed good performance (74.1% accuracy for a dataset with low similar sequences) while reflecting evolutionary information and sequence order information at the same time.

3.1.3 Features Based on Functional Domains of Sequence

Functional domains are the regions in an amino acid sequence of protein that carry out a specific function. Proteins typically have several functional domains. Using these functional domains as features in the structural class prediction problem, some researchers tried to capture the relationship among distant amino acids which is crucial for protein folding [50-51]. Chou et al. [50] used an integrated domain database [52] (InterPro database) which contains many sequences along with functional domain information. InterPro release 6.2 documents 7785 different functional domains (http://www.ebi.ac.uk/interpro). Chou et al. [50] represented each protein as a 7785 dimensional vector, where each feature is Boolean. A "1" represents the presence of a particular functional domain, and a "0" represents the absence of that functional domain in a protein. They suggest that functional domains of a protein correlate well with its structural class.

Amin et al. [51] followed Chou et al. [50] and used functional domains as class discriminating features. They used InterPro Release 30.0 which contains 21,178 functional domain entries. Of the 21,178 functional domains they only considered the domains which appear in the proteins of their dataset. Thus, their method used 2,400 functional domains as features. They also extracted features from predicted protein secondary structure. To reduce the dimension of the feature vector and select the most effective features they used the correlation based feature selection

(CFS) method [53]. CFS is a filtering method to select from the original feature set a smaller set of non-redundant features which have powerful class discriminating ability. They also checked their method of class prediction on intrinsically disordered proteins (biologically active proteins with no specific full 3D structure) and achieved reasonable prediction accuracy (76.20%).

3.1.4 Features Extracted from Predicted Protein Secondary Structure Sequence

Recently, many good methods have been developed using only features extracted from predicted protein secondary structure sequence [54-56]. The structural class of a protein mainly depends on its secondary structural content. Some researchers extracted features from predicted secondary structure sequence instead of amino acid sequence of protein. In these methods the researchers used secondary structure sequences predicted by methods like PSIPRED [16] and YASPIN [17]. Liu and Jia [54] constructed three novel features to differentiate between the $\alpha+\beta$ class and the α/β class more accurately. They used some previously used effective features from research [57-58] like content of α-helix (H) and β-strand (E) in the sequence, maximum and average length of H and E segments, and composition moment vector of H and E in the sequence. One would expect the predicted states H and E to alternate more frequently in a protein belonging to the α/β class than in a protein belonging to the $\alpha+\beta$ class where α-helix and β-strands are isolated. Therefore, one of their newly developed features was the normalized alternating frequency of H and E. They also included two newly developed features based on count of anti-parallel β-strands in the sequence, considering the fact that β-sheets in the α/β class proteins are usually composed of parallel β-strands whereas in the $\alpha+\beta$ class proteins, β-sheets are normally composed of anti-parallel β-strands. They also showed that their newly constructed features had good impact in identifying the α/β and the $\alpha+\beta$ class proteins with accuracies of 81.5% and 76% respectively.

Along with some previously used features from [54,58], Zhang et al. [55] introduced some new features to capture the distribution of α-helix (H) and β-strand (E) in the sequence. They made a reduced representation of sequences using only H and E, while ignoring Coil (C). Then, using a transition probability matrix, they computed features based on probability of transition from H to E and E to H. They showed that their newly developed features based on transition probability matrix made good contribution in the overall accuracy of 83.9%. Ding et al. [56] also

constructed several new features to extract information from predicted secondary structure sequences, such as the following: the variance of the length of H and E segments; variance of the positions of H and E in the secondary structure sequence; average length of H and E segments in the sequences, while ignoring coil segments. They also showed that their newly designed features are good for predicting the $\alpha+\beta$ and α/β class proteins compared with some established methods.

3.1.5 Features from both Amino Acid Sequence and Predicted Secondary Structure Sequence

Some successful methods used both amino acid sequences and predicted secondary structure sequences to extract features [45,51,58,59]. These methods try to incorporate useful class discriminating information from both amino acid sequence and predicted secondary structure sequence of protein. In 2008, Kurgan et al. [58] proposed a structural class prediction method popularly known as SCPRED. For this research, initially they extracted 2146 features from the amino acid sequence. These 2146 features included physiochemical values of amino acid based features, amino acid component based features like 1st and 2nd order composition moment vector, and property groups based features. The 20 amino acids can be subdivided into groups based on any one of several physiochemical properties. For example, according to electronic property, 20 amino acids are classified into following five groups: electron donor, weak electron donor, electron acceptor, weak electron accepter and neutral [58]. Features like composition percentage of electronic groups of amino acids, composition percentage of hydrophobic groups of amino acids were calculated. They also extracted 176 features from predicted protein secondary structure sequences which include maximum and average length of secondary structure segment, composition moment vector of secondary structural state. They reduced the dimension of their initial feature set from 2322 to 9 by using Hall's [53] correlation based feature selection method. The algorithm chose 8 features from secondary structure sequence and 1 from amino acid sequence, confirming the class discriminating quality of features extracted from secondary structure sequences.

Mohammad and Hampapathalu [59] extracted features from secondary structure sequences, but also considered the solvent accessibility information of amino acid residues and residue pairs

in the amino acid sequence. They used features like frequency of each amino acid to occur in a particular secondary structural states (H, E or C). Frequencies of amino acids pairs predicted as secondary structural state H, E and C were also measured for their model. They calculated the solvent accessibility state information for each amino acid in the protein from ACCpro [60]. Solvent accessibility of an amino acid residue is described as a binary value, either buried or exposed in terms of the degree of its interaction with the water molecules. They calculated features like frequency of 'buried' or 'exposed' residues. They also calculated the frequencies of amino acid pairs having solvent accessibility state 'buried', 'exposed' or 'partially buried'. Here they considered an amino acid pair to be in 'buried' or 'exposed' state only if both the residues were predicted in 'buried' or 'exposed' state, respectively, otherwise the pair was considered as in 'partially buried' state. Finally they checked their prediction model with different individual feature set and with the combination of these feature sets. They successfully showed that using information from both protein sequence and predicted secondary structure sequence could give better prediction accuracy (80.9%) than some other contemporary methods like SCPRED [58] with79.7% accuracy and Kurgan and Chen [61] with 62.7% accuracy .

Amin et al. [51] followed Kurgan et al. [58] and Yang et al. [62] to extract features from predicted secondary structure sequence of protein. Along with functional domain features they checked the contribution of features from secondary structure sequence in predicting structural classes. They used the CFS method to select the effective class discriminating features from initial feature set, resulting in only 77 functional domain features from the initial 2400 features, and 34 secondary structural features from the initial 110 features. This study too confirmed the effectiveness of secondary structural features in solving this problem.

Mizianty and Kurgan [45] used features based on the PSI-BLAST profile of proteins along with features from amino acid sequence and predicted secondary structure sequence of protein. After checking a combination of feature sets, they found that a combination of features from PSI-BLAST and predicted secondary structure sequence gave the best class discriminating performance (83.5% prediction accuracy) for a twilight zone dataset.

3.2 Classification Algorithm

Soft computing techniques like Support Vector Machine (SVM), Artificial Neural Network (ANN), Fuzzy Logic (FL) are machine learning techniques often used to construct classification models for protein structural class prediction. The strength of soft computing techniques to give human-like expert decisions and to handle ambiguous and uncertain situations as well as to process approximate and flexible information with low solution cost make them an effective choice to be used by bioinformatics researchers who need to deal with a large amount of faulty uncertain data.

Support Vector Machine seems to be the most popular among all soft computing techniques to solve the protein structural classification problem as it has been used in many projects [29,31,44-45,51,54-56,58-59]. The basic idea behind this supervised machine learning method is to create a hyperplane which not only separate but also maximizes the distance between two classes. It then assigns the prediction label according to on which side of this hyperplane a test case falls. (described in Chapter 2, Section 2.7). SVM solves the multiclass problem by creating either one-against-one binary classifiers or one-against-all binary classifiers. Amin et al. [51], Kurgan et al. [58] and Mohammad and Hampapathalu [59] developed one-against-one binary classifiers. Amin et al. used the initial predictions by binary classifiers to predict final class labels by pair-wise coupling technique as presented by [63]. Mohammad and Hampapathalu [59] used a voting scheme to assign the most probable class to the query protein sequence. Wu et al. [31] constructed one-against-all binary classifiers. Some methods combined SVM with other techniques to get a more efficient result [34][36]. Zhang et al. [34] developed SVM based on binary tree methods whereas Ding et al. [36] combined fuzzy logic with SVM.

Some methods used Artificial Neural Networks (ANN) to solve the prediction problem [64-65]. They chose ANN due to its self-organizing and self-adaptability properties. After learning and training on representative proteins, it refers the relevant features of proteins, and then it can assign a query protein to a specific structural class.

Chou et al. [50] used an intricate sorting method to do class prediction, where a similarity score was calculated between the query protein and all proteins in the dataset. The query protein is assigned to the class of a protein in the dataset which is most similar to it as in nearest neighbour approaches.

Chapter 4

Materials and Methods

This chapter provides the description of materials and methods which are used to generate and test protein structural class prediction models for this research.

4.1 Dataset

In order to be able to compare our results with other published results, we test our method on the frequently used 25PDB protein sequence dataset. 25PDB is a benchmark dataset for research on fairly dissimilar sequences, since pair-wise sequence similarity is not more than 25% in this dataset. The dataset was created using 25% PDBSELECT list [66] and was published in [67]. It contains high quality and low similar proteins, that is, proteins with high resolution structures and pair-wise similarity not more than 25%. The proteins were selected from PDB release February, 2005 to generate the 25PDB dataset. 25PDB contains 1673 protein sequences, where 443 sequences belong to the all-α class, 443 sequences belong to the all-β class, 441 sequences belong to the α+β class and 346 sequences belong to the α/β class. 25PDB was used in many standard methods [45,51,54-59,62]. We obtained the 25PDB dataset along with the corresponding secondary structure sequence set from http://biomine.ece.ualberta.ca/SCPRED/SCPRED.htm. Mohammad et al. [59] also used the dataset from the above mentioned web address. The predicted secondary structure sequences corresponding to protein amino acid sequences were predicted using the PSIPRED method [16] .

4.2 Generation of Feature Sets

The generation of a feature set is the most important step in the process of developing a protein structural class prediction model, once the machine learning software has been selected. Proteins are represented by their amino acid sequences which are alphabetical sequences over an alphabet of 20 letters representing amino acids. To feed these amino acid sequences into a classification algorithm like a support vector machine, different length sequences should be represented by a fixed length numeric feature vector. Various techniques are used to calculate numeric feature values from these sequences. The performance of a prediction model greatly depends on the techniques of extracting effective, class-discriminating features from sequences. In order to develop a more accurate protein structural class prediction model, we experimented with four different feature sets and their combinations. Since secondary structural content and spatial arrangement plays an important role in structural class allocation, two feature sets were developed based on predicted secondary structural class information. Another feature set was developed based on both secondary structural state and hydropathy profile. The fourth feature set was based on only hydropathy profile of amino acids. The hydropathy profile was chosen due to its importance in protein folding. This extends the work of Wu et al. [31] where they used hydropathy profile to create a reduced representation of protein sequence over alphabet {I,E,A} corresponding to three hydropathy states internal, external and ambivalent, and then extracted six features based on 1-word frequencies and position information.

4.2.1 Feature Set Constructed from Predicted Secondary Structural State Profile

An example of amino acid sequence and corresponding predicted secondary structure sequence of protein is given below.

Example 1

Amino Acid Sequence (AAS):

PVITLPDGSQRHYDHAVSPMDVALDIGPGLAKACIAGRVNGELVDACDLIEN

Predicted Secondary Structure Sequence (PredSSS):

CEEECCCCCEEECCCCCCHHHHHHHCCHHHHCCEEEEEECCEECCCCCCCCC

For each protein we extract 22 features based on the corresponding predicted secondary structure sequence. Ten of these features are re-used from [55] and [58], and 12 features are newly constructed. The details of these features calculated from the predicted secondary structure sequence along with their values using the PredSSS of Example 1 are described in the following paragraphs.

Features also used in previous research:

1. Probabilities of secondary structural states α-helix, $p(H)$, and β-strand, $p(E)$ in a secondary structural sequence are used due to their proven ability to discriminate among structural classes. The probabilities are calculated using the following formula:

$$p(i)=\frac{n_i}{N} \quad (1)$$

where n_i= total number of occurrences of secondary structural state i in the sequence, for $i \in\{H, E\}$, and N= length of sequence. Using PredSSS of Example 1, values for 2 features $p(H)$ and $p(E)$ are 0.2115 and 0.2692 respectively where n_H = 11, n_E = 14, and N = 52.

2. Since the lengths of the α-helix and β-strand structural components can reflect their spatial arrangement and have influence in forming shapes, the following features are also extracted from secondary structural sequence. Again N is the length of the sequence of the whole protein.
    - Two features based on normalized length of the longest segment are calculated by

$$NMaxSeg_i=\frac{MaxSeg_i}{N} \quad (2)$$

where $MaxSeg_i$ = length of longest i- segment in the sequence for $i \in \{H, E\}$. Using PredSSS of Example 1, values for $NMaxSeg_H$ and $NMaxSeg_E$ are 0.1346 and 0.1153 respectively where $MaxSeg_H = 7$, $MaxSeg_E = 6$ and $N = 52$.

- Two features based on normalized average length of the segment are calculated by

$$NAvgSeg_i = \frac{AvgSeg_i}{N} \tag{3}$$

where $AvgSeg_i$ = average length of i-segment in the sequence for $i \in \{H, E\}$. Using PredSSS of Example 1, value for $NAvgSeg_H$ and $NAvgSeg_E$ are 0.1058 and 0.0673 respectively where $AvgSeg_H = 5.5$, $AvgSeg_E = 3.5$ and $N = 52$.

3. The composition moment vector, CMV, encodes both the secondary structural state composition and position in the predicted secondary structure sequence. The 1st order composition moment vectors for secondary structural state component α-helix (H) and β-strand (E) are calculated using the following formula:

$$CMV_i = \frac{1}{N(N-1)} \sum_{j=1}^{n_i} x_{ij} \tag{4}$$

where $i \in \{H, E\}$ and N is the length of the secondary structure sequence for the whole protein, x_{ij} is the index of the j^{th} position of the i-structural state, and n_i is the total number of residues in the i-structural state in the sequence. In PredSSS of length $N = 52$ from Example 1, there are two segments of α-helix starting at indices 19 and 28. The composition moment vector for H is , $CMV_H = \frac{1}{(52)(51)}(19 + 20 + 21 + 22 + 23 + 24 + 25 + 28 + 29 + 30 + 31) = 0.1026$. There are 4 β-strand segments in PredSSS of Example 1 starting at indices 2, 10, 34 and 42. Then $CMV_E = \frac{1}{(52)(51)}(2 + 3 + 4 + 10 + 11 + 12 + 34 + 35 + 36 + 37 + 38 + 39 + 42 + 43) = 0.1305$.

4. Let the word "segment" refer to a maximal subsequence of the secondary structure where the sequence is all one state. Probabilities of α-helix (H) and β-strand (E) segments are calculated using the following formula:

$$p_{seg}(i) = \frac{TotalSeg_i}{TotalSeg} \tag{5}$$

where $TotalSeg$ = total number of segments (H, E or C) in the sequence and $TotalSeg_i$ = total number of *i*-segment in the sequence for $i \in \{H, E\}$. In PredSSS of Example 1, total number of H segment, $TotalSeg_H$ = 2, total number of E-segment, $TotalSeg_E$ = 4, and total number of C-segment. $TotalSeg_C$ = 7. Total number of segments, $TotalSeg$ = 2+4+7 = 13. The values of $p_{seg}(H)$ and $p_{seg}(E)$ are 0.1538 and 0.3077, respectively.

Newly constructed features in this thesis:

5. State change probabilities in predicted secondary structure sequence are calculated and used as features. We define a state change as a change of secondary structural state from one state to different state like a change from *α*-helix (H) to *β*-strand (E) or from *α*-helix (H) to coil (C), in the predicted secondary structure sequences. These probabilities are calculated using formula (1), with n_i= total number of occurrences of predicted secondary structural state change *i* in the sequence, for $i \in \{HE, EH, HC, CH, EC, CE\}$, and N = total number of state changes in the sequence. In PredSSS of Example 1, there are a total of 12 state changes including 4 from C to E, 4 from E to C, 2 from C to H, and 2 from H to C. So, $p(CE) = p(EC) = \frac{4}{12}$, and $p(CH) = p(HC) = \frac{2}{12}$. Other state change probabilities (*HE, EH*) are 0 for this example. In a sequence belonging to the all-*α* class, probabilities of state change from *β*-strand (E) to coil is low whereas for sequences belonging to the all-*β* class the situation is reverse. In sequences belonging to *α*/*β* and *α*+*β* class, the probability of state change from *α*-helix (H) to *β*-strand (E) is greater than in the other two classes as they are composed of both *α*-helix (H) and *β*-strand (E) states. In Zhang et al. [55], probabilities of state transition from *α*-helix (H) to *β*-strand (E) and from *β*-strand (E) to *α*-helix (H) were calculated in a different manner, since they converted secondary structure sequence to a reduced segment sequence composed of only *α*-helix (H) and *β*-strand (E) segments before calculating probabilities. We use state change probabilities from *α*-helix (H) to *β*-strand (E) , from *α*-helix (H) to coil (C), from coil (C) to *α*-helix (H), from coil (C) to *β*-strand (E), from *β*-strand (E) to coil (C) and from *β*-strand (E) to *α*-helix (H) as features.

6. State change moment vectors (SCMV) are calculated to reflect the position at which the state changes in secondary structure sequences. These features are displayed to differentiate between *α*+*β* and *α*/*β* class sequences as sequences in these classes possess all types of state

change, but their arrangement and positions may have distinguishing characteristics. Six State Change Moment Vectors (SCMV) are calculated using the following formula:

$$SCMV_i = \frac{1}{N(N-1)} \sum_{j=1}^{n_i} x_{ij} \quad (6)$$

where $i \in \{HC, HE, CH, CE, EC, EH\}$, and N is the length of the protein's secondary structure sequence, x_{ij} is the position in the sequence of the j^{th} occurrence of a state change i, and n_i is the total number of times the state changes i. In PredSSS of length N=52 from Example 1, there are 4 state changes from C to E at indices 2, 10, 34, and 42. Therefore, state change moment vector for CE is, $SCMV_{CE} = \frac{1}{(52)(51)}(2 + 10 + 34 + 42) = 0.0332$. There are 4 state changes from E to C at indices 5, 13, 40 and 44, making $SCMV_{EC} = \frac{1}{(52)(51)}(5 + 13 + 40 + 44) = 0.0346$. There are 2 state changes from C to H at indices 19 and 28, making $SCMV_{CH} = \frac{1}{(52)(51)}(19 + 28) = 0.0177$. There are 2 state changes from H to C at indices 26 and 32, making $SCMV_{HC} = \frac{1}{(52)(51)}(26 + 32) = 0.0219$. In this example there are no state changes from H to E or E to H, making $SCMV_{HE} = SCMV_{EH} = 0$. While our work was in progress, Ding et al. [53] published work using features based on two 2nd order composition moment vectors CMV_{HE} and CMV_{EH}, which may represent the same concept as our $SCMV_{HE}$ and $SCMV_{EH}$ (Ding et al. [53] did not give the equation they use for calculating CMV_{HE} and CMV_{EH}).

4.2.2 Feature Set Constructed from Predicted Secondary Structure and Hydropathy Profile

Physiochemical properties of amino acids are generally believed to have significant impact in forming protein structures, since these properties affect the tendency for certain amino acid side chains to be exposed to water. Therefore it is not surprising that various physiochemical properties of amino acids such as hydropathy, polarity, isoelectric point and flexibility have been used to predict structural classes [31,36,58,68-69]. For this study, the hydropathy profile of amino acids has been chosen, assuming it to have major impact on protein folding. The

hydropathy profile describes the hydrophobic and hydrophilic nature of segments of a protein based on amino acid sequence of protein. Liu and Wang [70] categorized the 20 amino acids into three groups according to hydropathy: Internal (I), External (E) and Ambivalent (A). Amino acids belonging to the Internal group are likely to be found in the interior of protein's structure, whereas amino acids from the External group are likely to appear at the surface. We use the following rule from Liu and Wang [70] to categorize amino acids according to hydropathy:

$$F(S(i)) = \begin{cases} I \ \ if \ \ S(i) = F, I, L, M, V \\ E \ \ if \ S(i) = D, E, H, K, N, Q, R \\ A \ \ if \ S(i) = S, T, Y, C, W, G, P, A \end{cases} \quad (7)$$

Here *S(i)* represent the i^{th} amino acid in amino acid sequence of protein, and *F(S(i))* represents its corresponding replacement according to its hydropathic nature. Below is an example of amino acid sequence of protein and its corresponding hydropathy sequence generated using formula (7).

Example 2.1

Amino Acid Sequence (AAS):

PVITLPDGSQRHYDHAVSPMDVALDIGPGLAKACIAGRVNGELVDACDLIEN

Hydropathy sequence (HS):

AIIAIAEAAEEEAEEAIAAIEIAIEIAAAIAEAAIAAEIEAEIIEAAEIIEE

For each protein in the dataset, the corresponding hydropathy profile sequence (HS) is used to construct 72 features. Details of these features along with their values calculated using the HS of Example 2.1 are given below:

1. Probabilities of hydropathy states Internal, *p(I),* External, *p(E), and* Ambivalent, *p(A)* in the hydropathy profile sequence are calculated as features using the formula (1) where n_i= total number of occurrences of hydropathy state *i* in the sequence for $i \in \{I, E, A\}$, and *N*= length of sequence. For HS in Example 2, the values of 3 features *p(I), p(E),* and *p(A)* are 0.2885, 0.3077, and 0.4038 respectively where n_I=15, n_E = 16, n_A = 21, and N = 52.

2. Normalized longest length and normalized average length of the three different hydropathy blocks, *I* block, *E* block and *A* block, in the sequence are calculated using formulas (2) and

(3) for $i \in \{I, E, A\}$. For HS of Example 2.1, the values for $NMaxSeg_I$, $NMaxSeg_E$, and $NMaxSeg_A$ are 0.0385, 0.0577, and 0.0577 respectively where $MaxSeg_I = 2$, $MaxSeg_E = MaxSeg_A = 3$, and $N = 52$. The values for $NAvgSeg_I$, $NAvgSeg_E$, and $NAvgSeg_A$ are 0.0243, 0.0403, and 0.0288 respectively where $AvgSeg_I = 1.2727$, $AvgSeg_E = 1.25$, and $AvgSeg_A = 1.5$.

3. The conditional probabilities of a hydropathy state occurring at position *i* given a certain hydropathy state at the previous position (*i-1*) are calculated to reflect the spatial arrangement of different hydropathy states in the sequences. Nine conditional probabilities are calculated using the following formula:

$$p(h_i \mid h_{i-1}) = \frac{p(h_{i-1},\ h_i)}{p(h_{i-1})} \tag{8}$$

where $p(h_{i-1}, h_i) =$ the probability of state h_{i-1} being followed by state h_i in the sequence and $p(h_{i-1}) =$ probability of state *i-1* in the sequence. Here, $p(h_{i-1}, h_i)$ and $p(h_{i-1})$ are calculated using formulas (9) and (10), respectively.

$$p(h_{i-1}, hi) = \frac{Total_{h_{i-1}h_i}}{N-1} \tag{9}$$

$$p(h_{i-1}) = \frac{Total_{h_{i-1}}}{N} \tag{10}$$

where $Total_{h_{i-1}h_i}$= total number of occurrences of state pair $[h_{i-1}, h_i]$ in the sequence, *N*= length of the sequence, and $Total_{h_{i-1}}$= total number of occurrences of state h_{i-1} in the sequence. State h_i belongs to hydropathy state space *{I, E, A}*. Thus, $p(h_{i-1}, h_i)$ produces 9 combinations. For HS of Example 2.1, the probability that E occurs at a position given that A occurs at the previous position, $p(E \mid A)$ is $\frac{p(A,\ E)}{p(A)} = \frac{0.1346}{0.4038}$ where $p(A, E) = \frac{7}{51}$ and $p(A) = \frac{21}{52}$, as the total number of occurrence of A followed by E, $Total_{AE} = 7$, total number of occurrence of A, $Total_A = 21$, and length of sequence, $N = 52$.

4. To reflect the impact of two consecutive hydropathy state elements on two consecutive secondary structural state elements at the same position, another conditional probability is used. Here we calculate 54 conditional probability values which reflect the impact of

hydropathy state pair h which belongs to set *{EE, EI, EA ,II, IE, IA, AA, AI, AE}* on secondary structural state pair s which belongs to the set *{HC, CH, EC, CE, HE, EH}* using following formula:

$$p(s|h) = \frac{p(h,\ s)}{p(h)} \tag{11}$$

where $p(s \mid h) =$ probability that secondary structure pair s occurs given that hydropathic pair h is at the same position in the sequence, $p(h) =$ probability that hydropathic state pair h occurs in the sequence, and $p(h,s)$ = probability of h and s co-occurs at a given location. Here, $p(h,s)$ and $p(h_j)$ are calculated using formulas (12) and (13), respectively:

$$p(h,s) = \frac{Total_{s,h}}{N-1} \tag{12}$$

$$p(h) = \frac{Total_h}{N-1} \tag{13}$$

where $Total_{s,h}$= total number of times structure pair s occurs where hydropathic pair h occurs, N=length of the sequence, and $Total_h$= total number of occurrences of state pair h in the sequence.

Example 2.2

Hydropathy sequence (HS):

AIIAIAEAAEEEAEEAIAAIEIAIEIAAAIAEAAIAAEIEAEIIEAAEIIEE

Predicted Secondary Structure Sequence (PredSSS):

CEEECCCCCEEECCCCCCHHHHHHHCCHHHHCCEEEEEECCEECCCCCCCCC

Using Example 2.2, the conditional probability of secondary structural state pair EC in PredSSS if EA occurs at the same positions in HS, $p(EC/EA)$ is $\frac{p(EA,\ EC)}{p(EA)} = 0.1176$, since the probability that secondary structure pair EC co-occurs with the hydropathic pair *EA* is $p(EA,EC) = \frac{1}{51}$, and the probability of hydropathy state pair EA occurring in the sequence is $p(EA) = \frac{6}{51}$. There is only one occurrence of secondary structural state pair EC in PredSSS and

hydropathy state pair EA in HS at same position (index 12), so $Total_{EC,EA}$= 1. The total number of occurrence of hydropathy state pair EA in HS is $Total_{EA}$= 6, and the length of the sequence is N = 52.

4.2.3 Feature Set formed by Extracting n-gram Patterns from Predicted Secondary Structure Sequence

A protein secondary structure sequence is a string of any length over the set Σ= {H, E, C}. An n-gram pattern is defined to be a block of length n consisting of n characters over set Σ. To reflect the information related to predicted secondary structural content and its arrangement in a secondary structure sequence, n-gram patterns are extracted from the predicted secondary structure sequence. Frequencies of n-gram patterns are used as features to describe the sequence where n = 2, 3, 4, 5. Two feature sets are developed using frequencies of n-gram patterns. The initial feature set is developed using the frequencies of all n-gram patterns. To extract n-gram patterns, a sliding window of length n is moved from left to right one character at a time over the sequence as shown in Figure 4.1, and then frequencies of all n-gram patterns are calculated. The dimension of the initial feature set is $3^2 + 3^3 + 3^4 + 3^5 = 360$, since for a window of length n there are 3^n different possible patterns that can be counted.

The second feature set is a reduced version of the initial one. In this case, the Term Frequency - Inverse Document Frequency (TF-IDF) technique is used to select "important" patterns from the sequences. Yang et al. [71] used TF-IDF technique on amino acid sequences of proteins to extract features in order to classify proteins into different functional groups. TF-IDF is a numeric statistic to find important terms from a set of documents [72]. It is based on the fact that if a term is frequent in many documents, then it has low quality for distinguishing documents, and if it is present in few documents frequently then it has the ability to differentiate those documents from others. The TF-IDF technique assigns weights to terms present in documents. The terms with higher weight are assumed to be important terms, having the capability to separate a document from others.

For this thesis, we construct a feature set which includes important class discriminating patterns from protein secondary structure sequences. Here each protein sequence is treated as a document. Weights of different length patterns from sequences are calculated using the TF-IDF

technique. Patterns with weight higher than a chosen threshold are selected as important patterns. The frequency of selected patterns in a sequence are calculated to form the feature vector for that sequence. According to the TF-IDF technique, the weight $w_{p,s}$ of a pattern p in sequence s is calculated using the following formula:

$$w_{p,s} = f_{p,s} \; ln \frac{N}{n_p} \tag{14}$$

where $f_{p,s}$ is the frequency of pattern p in sequence s, N is the total number of sequences in the dataset, and n_p is the number of sequences in which pattern p occurs. The final weight w_p of pattern p is calculated using the following formula:

$$w_p = \max_{s \in S} w_{p,s} \tag{15}$$

where S denote the whole dataset.

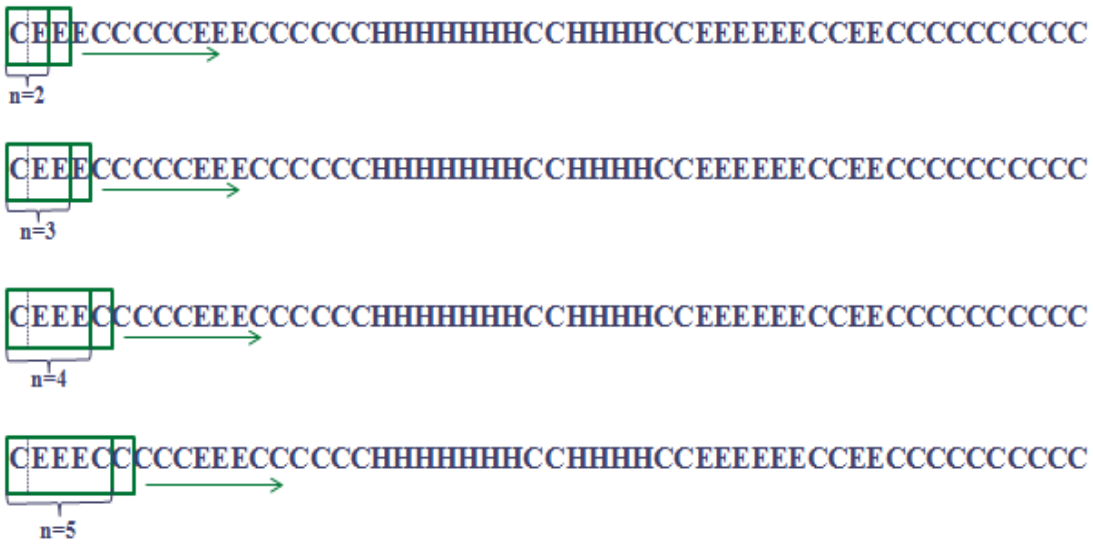


Figure 4.1: Sliding window technique to extract n-gram patterns from predicted secondary structure sequence of protein. Window sizes of 2,3,4, and 5 are each shown at two positions.

For this dataset (25PDB), the weight of patterns usually lies between 0 and 10, with very few exceptions where the weight is more than 10. In the next steps we set the threshold at 5 so that the patterns having weight more than 5 are selected for inclusion as features. This way we select approximately the best 50% of the patterns. Frequencies of these selected patterns are used to form a feature vector. Thus, the TF-IDF technique is used to select patterns having good

distinguishing quality. In this case 140 patterns are selected, which reduces the dimension of feature vector from the initial one by 60%.

4.2.4 Feature Set formed by Extracting n-gram Patterns from Hydropathy Profile

In order to assess the impact of hydropathy blocks in structural class allocation, we extracted n-gram patterns from the hydropathy sequence. Here, n-gram patterns constructed over the alphabet {I, E, A}, with n = 2, 3, 4, 5 are extracted using the same process as in the previous case with the predicted secondary structural patterns. We construct and test the complete set of hydropathic n-gram patterns, a 360-dimensional feature vector. Since the classification ability of this feature set turned out to be less than satisfactory (as described in Chapter 5, Section 5.1), the pattern filtration using the TF-IDF process is not applied.

4.3 Classification Algorithm

Protein structural class prediction is a typical classification problem in the area of bioinformatics which can be solved by supervised machine learning technique. Supervised machine learning methods can classify unknown test data based on the knowledge gained during a training process. Several machine learning methods like Artificial Neural Network (ANN), Support Vector Machine (SVM) , Bayesian Network, and Markov models are used in bioinformatics research [73]. Among all these machine learning techniques, SVM is gaining popularity in bioinformatics due to its good performance in real word problems, ability to handle high dimensional noisy data and efficiency in handling variable length sequences and graphs [74]. Many successful protein structural class prediction models use SVM [29,31,44-45,51,54-56,58-59] due to its proven competence. After reviewing the research on predicting protein structural classes based on SVM, we decided to use SVM as the classification algorithm.

To implement SVM, we use the LIBSVM [75-76] tool in MATLAB. We downloaded LIBSVM software version 3.12 from http://www.csie.ntu.edu.tw/~cjlin/libsvm. The radial basis function is used here because of its proven superiority in solving classification problems [77-78]. Parameters C=4 and *gamma* = 0.5 are chosen by the grid search strategy available in LIBSVM

software [75-76]. Protein structural class prediction is a multi-class classification problem, since a new protein is to be predicted as belonging to one of the four structural classes (all-α, all-β, α+β, and α/β). We implement the one-against-all multi-class approach where four binary classifiers are generated to classify data as all-α / ~all-α, all-β / ~all-β, α+β / ~α+β, and α/β / ~α/β. The accuracy of these classifiers was calculated by their ability to classify test data correctly. (Details of methods used to measure the accuracy of models is discussed in the next section).

4.4 Performance Measure

To measure the performance of the SVM prediction model, we use the statistical method of k-fold cross validation. In supervised learning, a certain amount of labeled data is available for training the prediction model. The performance of a prediction model depends on its efficiency on detecting the labels of unlabeled data. To estimate performance one can set aside some of the labeled data for testing, making sure that the test data is not also used for training. Where the available data is limited, then the process of training on part of the labeled data and testing on the remaining part can be repeated to improve the estimate of accuracy. In k-fold cross validation, the total dataset is divided into k parts where k-1 parts are used for training the model and the remaining part is used for testing using the model trained by training data. The process is then repeated k times, so that each instance in the dataset is used once as a test instance. Figure 4.2 shows an example of k = 3 fold cross validation. In Turn 1, the left $\frac{2}{3}$ of the data are used to train the model, and the right third of the data is used to test the inferred model. The result is the model's accuracy on that test. Turns 2 and 3 use different thirds for testing and the rest for training. The results are averaged to give the estimated accuracy of the model trained on the whole dataset. The accuracy of each model, for example Model 1 in Figure 4.2, is defined by the following formula:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{16}$$

where *TP, TN, FP,* and *FN* correspond to the number of true positive, true negative, false positive and false negative classifications, respectively. True positive refers to actual positive data classified as positive, true negative refers to actual negative data classified as negative, false positive refers to actual negative data classified as positive, and false negative refers to actual positive data classified as negative.

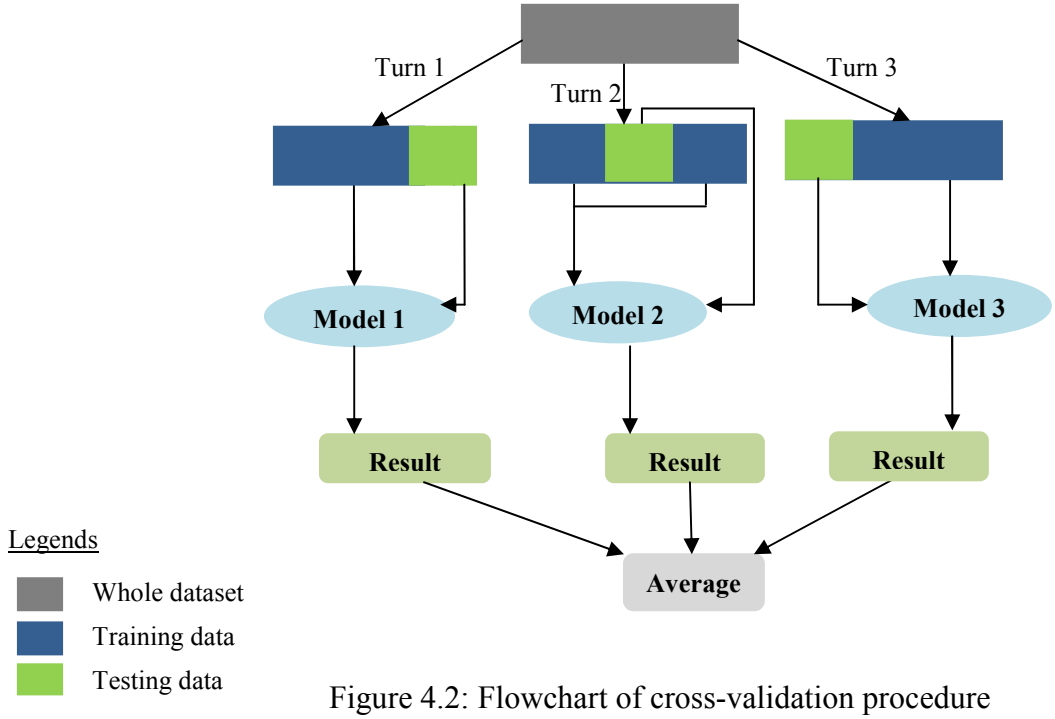


Figure 4.2: Flowchart of cross-validation procedure

The k-fold cross validation prediction accuracy is defined as the average accuracy over the *k* models developed in this way. The measure is used as the estimate of accuracy of a model built on all the labeled data. We used 10-fold, 15-fold, and 20-fold cross validation to measure the performance of each of the four binary classifiers developed with each of the nine different feature sets. The overall accuracy of these models for a given feature set was calculated using following formula:

$$Overall\ accuracy = \frac{\sum_{i=1}^{4} accuracy_i * total_i}{\sum_{i=1}^{4} total_i} \quad (17)$$

where the summation is over the four classes, $total_i$ is total number of sequences in class *i*, and $accuracy_i$ is the 10-fold / 15-fold / 20-fold cross validation prediction accuracy for class *i*. The final overall accuracy of each model (shown in the rightmost column of Table 5.2 in Section 5.1) is the average of accuracies measured using 10-fold, 15-fold and 20-fold cross validation techniques.

The Jackknife test is also used to measure the accuracy of the SVM model, after choosing the best performing feature set. The Jackknife test is k-fold cross validation, where k is equal to number of proteins in the dataset. In this case, each protein is used as the test dataset in turn while the model is trained with the rest of the sequences. Again, the overall prediction accuracy of the model was calculated using formula (17).

To evaluate the effectiveness of every classification model, four additional measures were used: Matthews correlation coefficient (MCC), Sensitivity, Specificity and Precision [79-80]. Sensitivity, specificity, precision and MCC score for each type of class prediction are given by formulas (18) - (21).

$$Sensitivity = \frac{TP}{FN + TP} \tag{18}$$

$$Specificity = \frac{TN}{FP+TN} \tag{19}$$

$$Precision = \frac{TP}{FP + TP} \tag{20}$$

$$MCC = \frac{((TP)(TN) - (FP)(FN))}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{21}$$

Sensitivity refers to the fraction of actual positives correctly identified for a given class. Specificity refers to the fraction of actual negatives correctly identified for a given class. Precision refers to the fraction of positives that are true positives. Matthews correlation coefficient ($-1 \leq MCC \leq +1$) corresponds to prediction quality where +1 represents perfect prediction, and -1 represents total disagreement between prediction and learning.

4.5 Overall Approach

The steps to implement our classification models are shown in Figure 4.3. Our input file (25PDB dataset) contains amino acid sequences of proteins, corresponding predicted secondary structure sequences and their true class labels. We used both predicted secondary structure sequences and amino acid sequences to create feature sets F1-F8. Programming language C was used to extract features from sequences as well as to create input file for LIBSVM. When we used predicted secondary structure sequences then features were directly extracted from them. When we used amino acid sequences then firstly those amino acid sequences were transformed to hydropathy sequences and then features were extracted from hydropathy sequences. To create input file for LIBSVM, class information are added to features. In the input file for LIBSVM, every protein sequence was represented by a fixed dimensional feature vector along with its true class label. Nine classification models were developed using nine different feature sets. The input files were fed into the LIBSVM version 3.12. The cross validation, training, and testing tasks were done in MATLAB interface using LIBSVM software to give final results.

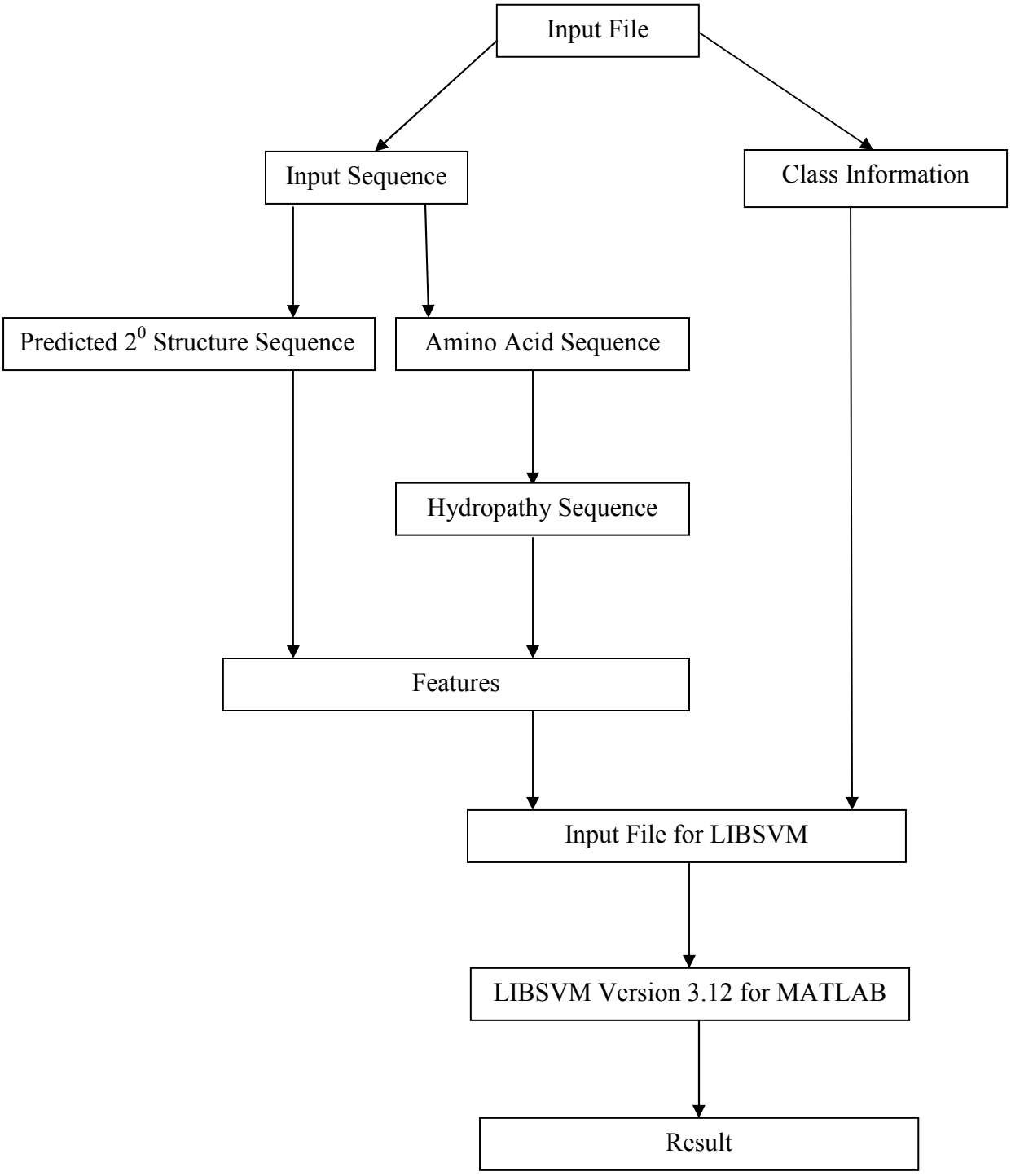


Figure 4.3: Flowchart of implementation steps

Chapter 5

Results

We develop several different feature sets for structural class prediction. These feature sets and combinations thereof are used to generate SVM classifiers which in turn are tested. A list of all of these feature sets is given in Table 5.1. The performance of the SVM models trained on the different feature sets is given in Table 5.2. The final overall accuracy of each model is measured as the average of its accuracies calculated using 10-fold, 15-fold, and 20-fold cross validation methods. 10-fold cross validation was also used in [58] and [59] to measure the performance of protein structural class prediction model. Comparisons of these results with those of published methods are described below.

Table 5.1: Feature sets used for class discrimination

Index	# of features	Feature Set Definition
F1	22	Predicted secondary structure state profile
F2	72	Predicted secondary structure and hydropathy state profile
F3	360	All n-gram patterns from predicted secondary structure sequence
F4	140	Filtered n-gram patterns from predicted secondary structure seq.
F5	360	All n-gram patterns from hydropathy sequence
F6	94	F1 + F2
F7	212	F2 + F4
F8	162	F1 + F4
F9	234	F1 + F2 + F4

Table 5.2: Performance of feature sets measured by Cross Validation (CV)

Feature Set	Fold & Avg. class-wise acc(%)	Class-wise accuracy(%)				Overall Acc(%)	Average Acc(%) (Standard Deviation)
		All α	All β	α+ β	α/ β		
F1	10F CV	94.44	91.33	79.68	89.20	88.64	**88.71** (0.08)
	15F CV	94.32	91.51	79.86	89.66	88.80	
	20F CV	94.32	91.39	79.68	89.60	88.70	
	Average	94.36	91.78	79.74	89.49		
F2	10F CV	93.78	88.52	73.69	83.56	84.97	85.26 (0.08)
	15F CV	93.78	88.64	73.70	84.10	85.12	
	20F CV	93.60	88.46	73.70	84.40	85.09	
	Average	93.72	88.54	73.70	84.02		
F3	10F CV	94.68	90.91	76.27	79.49	85.68	85.77 (0.08)
	15F CV	94.68	91.09	76.63	79.50	85.83	
	20F CV	94.63	91.03	76.69	79.44	85.80	
	Average	94.66	91.01	76.53	79.48		
F4	10F CV	94.62	90.85	75.85	79.32	85.51	85.56 (0.04)
	15F CV	94.62	91.03	75.91	79.32	85.57	
	20F CV	94.62	90.97	76.03	79.32	85.59	
	Average	94.62	90.95	75.93	79.32		
F5	10F CV	76.47	73.87	70.98	79.74	75.01	75.50 (0.42)
	15F CV	77.23	73.58	73.64	79.32	75.75	
	20F CV	77.17	73.58	73.64	79.32	75.73	
	Average	76.96	73.68	72.75	79.46		
F6	10F CV	94.26	91.33	80.81	89.54	88.96	88.91 (0.06)
	15F CV	94.20	91.27	80.75	89.24	88.85	
	20F CV	94.32	91.45	80.51	89.54	88.93	
	Average	94.26	91.35	80.69	89.44		

Table 5.2 Continued

Feature Set	Fold & Avg. class-wise acc(%)	Class-wise accuracy(%)				Overall Acc(%)	Average Acc(%)
		All α	All β	α+ β	α/ β		
F7	10F CV	93.84	90.97	79.43	88.40	88.15	88.27 (0.11)
	15F CV	94.32	91.09	79.80	88.11	88.35	
	20F CV	94.14	91.03	79.92	88.11	88.32	
	Average	94.10	91.03	79.72	88.21		
F8	10F CV	94.80	91.69	80.87	89.96	89.30	**89.25** (0.04)
	15F CV	94.80	91.57	80.87	89.78	89.23	
	20F CV	94.68	91.63	80.81	89.90	89.23	
	Average	94.76	91.63	80.85	89.88		
F9	10F CV	94.20	91.15	80.63	89.54	88.86	89.00 (0.14)
	15F CV	94.26	91.27	81.35	89.71	89.13	
	20F CV	94.14	91.33	80.99	89.78	89.03	
	Average	94.20	91.25	80.99	89.43		

5.1 Performance Comparison Among the Various Feature Sets

Table 5.2 shows that for the all-α class, every feature set except F5 performs very well (accuracy 93%-94%), and for the all-β class, every feature set except F2 and F5 shows 90%-92% accuracy. F2 shows ~89% accuracy for the all-β class, which is better than the accuracy of ~74% given by F5. For the α+β and α/β classes, performance of feature sets F1 and F6-F9 are considerably better than the others. The poor performance of feature set F5 for all classes, where F5 is constructed based on frequencies of n-gram hydropathy patterns, suggests that this information alone does not possess very good structural class distinguishing ability. Feature set F2, which does better than F5, is based on both hydropathy information and secondary structural information. Feature set F1, which has only 22 features and which is based on predicted secondary structural content, gives very good overall prediction accuracy. This corroborates the expected impact of predicted secondary structural content in predicting structural classes. The feature sets F6-F9 are

combinations of feature sets, and these show slightly better results than F1. The improved accuracy is at the cost of higher dimension.

Feature sets F3 and F4 both are developed based on n-gram secondary structural patterns in predicted secondary structure sequences. Feature set F3 is developed using all n-gram secondary structural patterns, whereas F4 is constructed based on filtered n-gram patterns. The predicted accuracy of the model based on feature set F3 is less than one percentage point better than F4, whereas in F4 the dimension of the feature set is reduced by 60% using TF-IDF. This suggests that the TF-IDF method is a good choice for filtering this type of pattern. Between F3 and F4, low dimensional feature set F4 is combined with other feature sets to check the performance of combinations. Feature set F5, based on n-gram hydropathy patterns, is not filtered using the TF-IDF method, because of its relatively poor performance compared to other feature sets. Feature set F8 with 162 features from sets F1 and F4 shows the highest average overall accuracy (89.25%). Feature set F9 with 234 features from F1, F2, and F4 shows an average overall accuracy of 89% which is slightly less than F8. Note that F9 is a superset of F8, and it does a little worse than F8. Although the difference between the accuracies with F8 and F9 is not statistically significant according to t-test they are consistent with the well known concept of data mining that increasing the number of features does not necessarily increase the performance of a model, and may even decrease the performance. The performance of a model depends less on the number of features, and more on the class discriminating quality of features.

The average of the 10k, 15k, and 20k cross-validation accuracies of each model based on feature sets F1-F9 are shown in Figure 5.1 (a-d). This summarizes average class-wise accuracies presented in Table 5.2. Figure 5.1 shows that models based on any feature set other than F5 are similar in their prediction accuracy of the all-α and the all-β classes. The overall prediction quality of the models mainly differs according to the α+β and α/β class data.

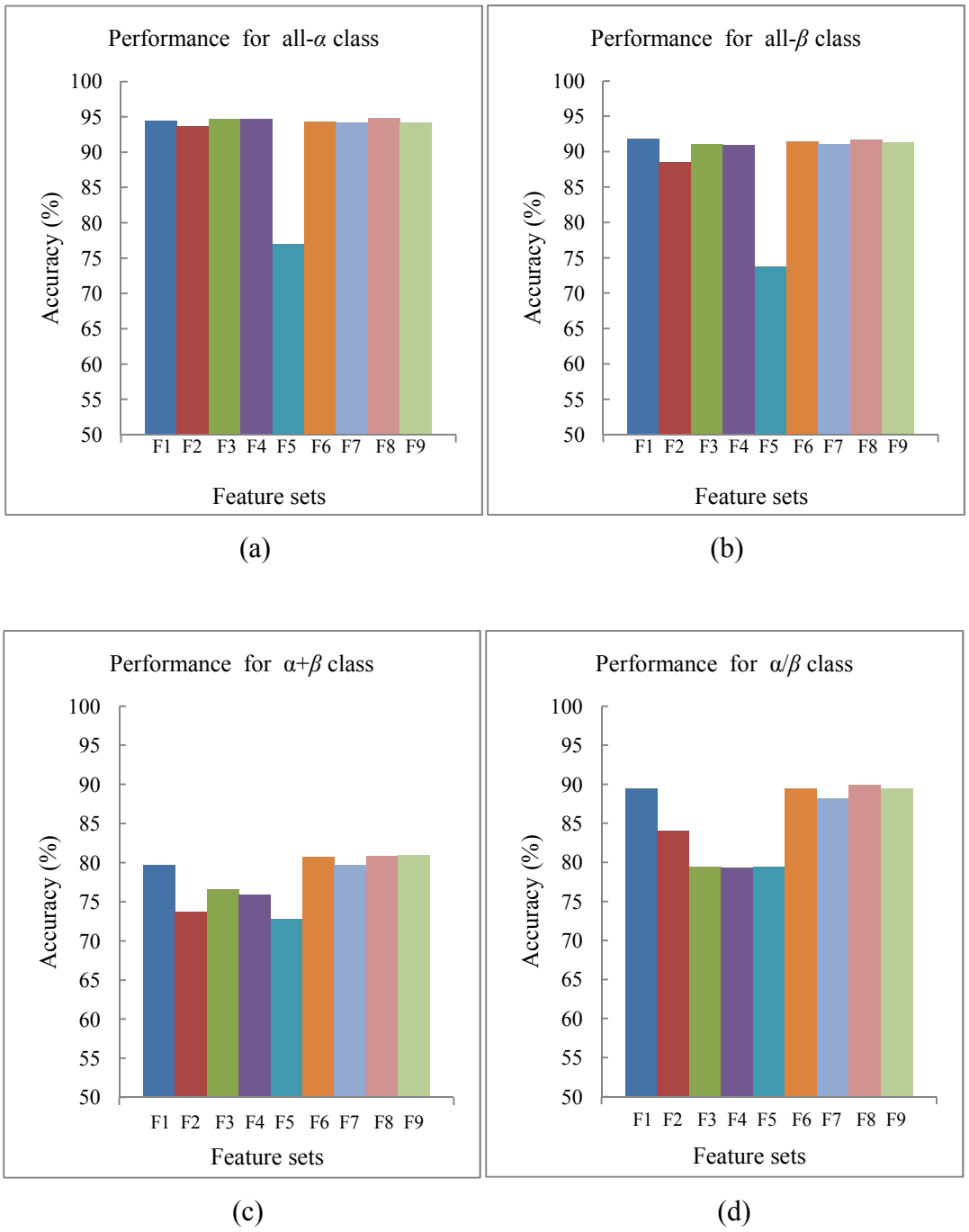


Figure 5.1: Performance of feature sets F1 - F9 for (a) all-α class, (b) all-β class, (c) α+β class, and (d) α/β class

5.2 Prediction Quality of Different Models

Most research papers use not only cross validation accuracy for the multiclass classification but also calculate the specificity, sensitivity, precision, and MCC score to evaluate the prediction quality of classification model. We measure these statistics for prediction quality twice using different training and testing sets. For both observations, the whole dataset is divided into training and testing sets where 20% of the data are selected randomly for testing and the rest 80% data are used for training the models. The prediction qualities in terms of average specificity, average sensitivity, average precision, and average MCC score for each class of nine models for the two experiments are shown in Table 5.3.

Table 5.3 shows that the specificity score (average 91.5%) is greater than the sensitivity score (average 72.6%) (These are averages over the 36 classes representing nine feature sets and four structural classes). This indicates that all these model are better in predicting the negative class (-1) data than positive class (+1) class data. The average MCC score for predicting the all-α and the all-β class is more than 0.80 and 0.70, respectively, for all these models except the one based on F5. This indicates fairly good prediction ability in classifying the all-α and all- β class data. The MCC score for classification of the α/β class data is between 0.60 and 0.76 for all models except F5. These scores are not as good as the MCC score for the all-α and all-β class, but they are better than the MCC scores for predicting the $\alpha+\beta$ class data (which is less than 0.50). This indicates that these feature sets are not as good in supporting prediction of the $\alpha+\beta$ class data as they are in supporting prediction of the other three classes.

Table 5.3: Performance of feature sets F1 - F9 using average of Specificity, Sensitivity, Precision, and MCC scores. The average of two experiments is shown, plus or minus the range.

Feature Set	Structural Class	Avg Spec (%)	Avg Sens (%)	Avg Prec (%)	Avg MCC
F1	all α	96.5 ±1.5	90.0 ±1.0	91.5 ±2.5	0.87 ±3.0
	all β	97.5 ±1.5	78.5 ±1.5	93.5 ±4.5	0.82 ±4.0
	$\alpha+\beta$	85.5 ±0.5	70.0 ±11	48.5 ±6.5	0.48 ±4.0
	α/β	95.0 ±1.0	80.5 ±6.5	81.5 ±5.5	0.76 ±1.0
F2	all α	91.0 ±2.0	92.5 ±0.5	79.5 ±4.5	0.83 ±1.5
	all β	86.0 ±0.0	85.5 ±1.5	74.0 ±5.0	0.70 ±2.5
	$\alpha+\beta$	95.5 ±1.5	28.5 ±1.5	63.5 ±11.5	0.33 ±3.0
	α/β	89.5 ±1.5	73.5 ±5.5	63.5 ±0.5	0.60 ±2.5
F3	all α	95.0 ±1.0	91.5 ±0.5	89.0 ±3.0	0.86 ±1.0
	all β	94.5 ±2.5	85.5 ±1.5	87.0 ±8.0	0.81 ±5.5
	$\alpha+\beta$	88.5 ±1.5	58.0 ±2.0	57.5 ±1.5	0.46 ±3.0
	α/β	92.5 ±1.5	63.5 ±9.5	62.5 ±0.5	0.56 7.5

Table 5.3 Continued

Feature Set	Structural Class	Avg Spec(%)	Avg Sens (%)	Avg Prec (%)	Avg MCC
F4	all α	95.0 ±0.0	93.0 ±0.0	89.5 ±0.5	0.87 ±0.0
	all β	95.0 ±2.0	84.0 ±3.0	89.0 ±6.0	0.80 ±5.5
	$\alpha+\beta$	85.5 ±2.5	62.0 ±3.0	50.5 ±3.5	0.44 ±5.5
	α/β	95 ±0.0	50.5 ±7.5	64.0 ±1.0	0.56 ±12.0
F5	all α	67.5 ±1.5	78.5 ±0.5	51.0 ±5.0	0.42 ±2.0
	all β	75.5 ±0.5	68.0 ±0.0	51.0 ±1.0	0.41 ±0.5
	$\alpha+\beta$	95.0 ±2.0	8.5 ±1.5	34.0 ±12.0	0.10 ±0.0
	α/β	92.0 ±2.0	26.0 ±3.0	46.5 ±3.5	0.22 ±0.0
F6	all α	96.5 ±0.5	86.0 ±3.0	93.0 ±1.0	0.84 ±1.0
	all β	97.0 ±1.0	75.5 ±0.5	92.5 ±2.5	0.77 ±1.5
	$\alpha+\beta$	87.5 ±1.5	60.5 ±9.5	50.5 ±0.5	0.45 ±4.5
	α/β	91.5 ±1.5	89.0 ±4.0	66.0 ±7.0	0.72 ±2.5

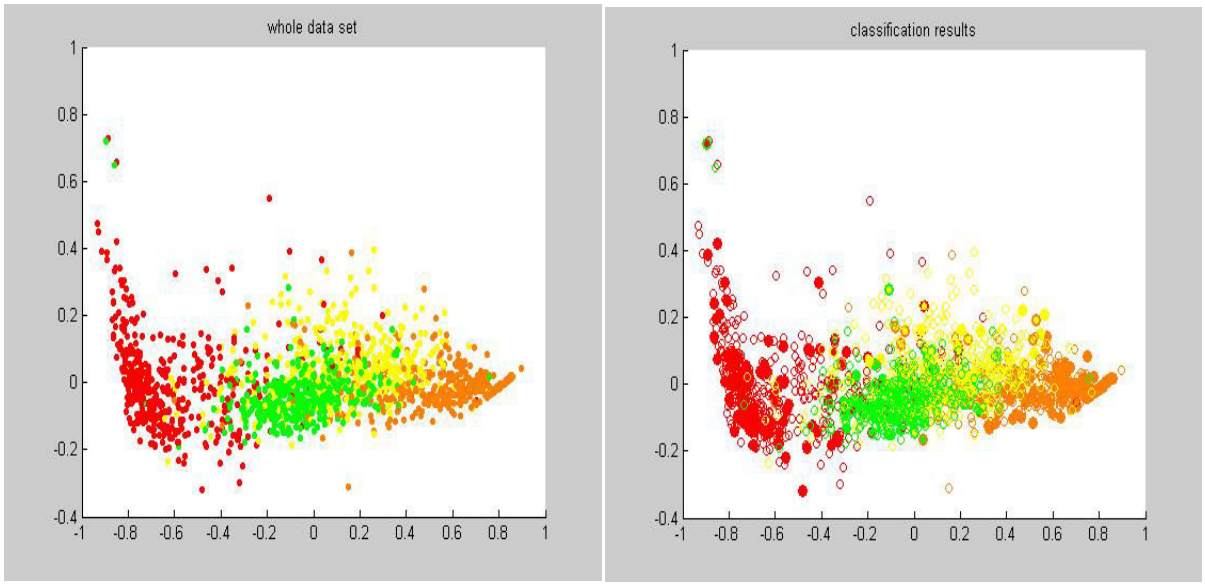
Table 5.3 Continued

Feature Set	Structural Class	Avg Spec (%)	Avg Sens (%)	Avg Prec (%)	Avg MCC
F7	all α	96 ±1.0	89.0 ±1.0	92.5 ±1.5	0.86 ±1.0
	all β	96.5 ±0.5	77.0 ±1.0	92.0 ±2.0	0.77 ±1.0
	α+ β	88.5 ±1.5	56.0 ±7.0	50.0 ±3.0	0.42 ±1.0
	α/β	91.5 ±1.5	85.0 ±8.0	65.5 ±5.5	68.5 ±1.5
F8	all α	95.6 ±0.5	91.2 ±0.5	87.5 ±0.5	0.86 ±0.0
	all β	96.0 ±1.0	81.5 ±2.5	88.5 ±3.5	0.79 ±4.0
	α+ β	87.5 ±0.5	60.0 ±3.0	63.5 ±0.5	0.49 ±2.0
	α/β	92.0 ±2.0	80.5 ±0.5	73.0 ±4.0	0.71 ±3.5
F9	all α	96.5 ±0.5	89.5 ±1.5	90.5 ±2.5	0.87 ±0.5
	all β	96.0 ±0.0	80.5 ±0.5	89.0 ±1.0	0.79 ±0.0
	α+ β	88.0 ±1.0	62.0 ±7.0	59.5 ±3.5	0.49 ±3.0
	α/β	91.5 ±2.5	81.0 ±0.0	73.0 ±7.0	0.71 ±4.5

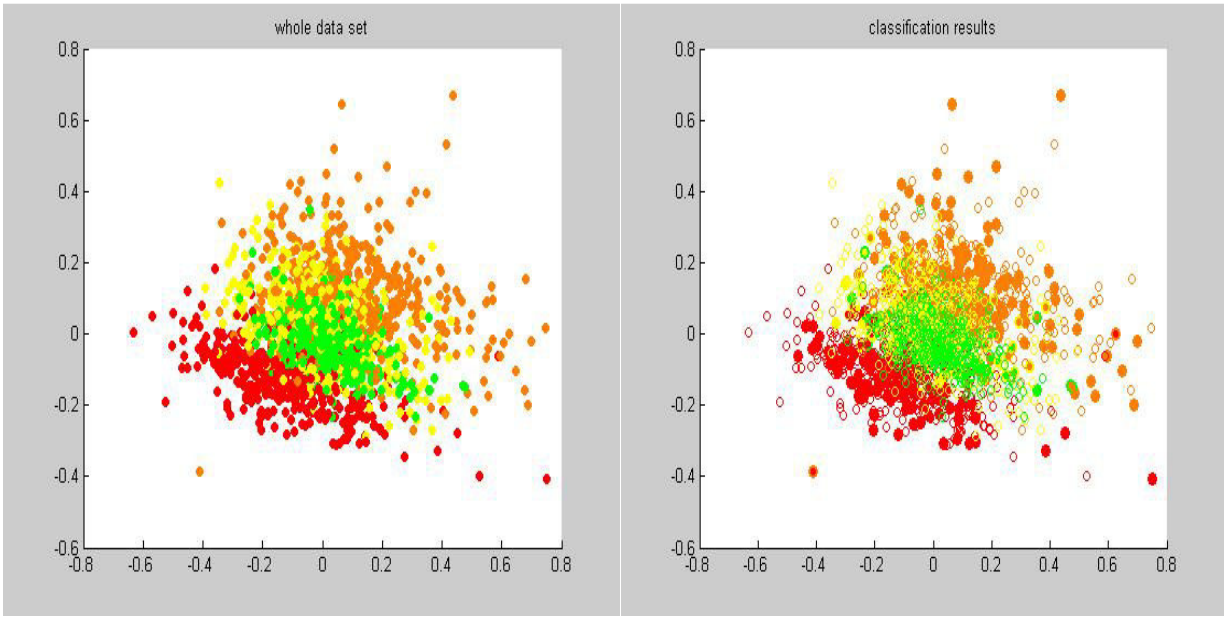
5.3 Visualization of Clustering and Prediction Quality of Different Models

Figure 5.2 presents visualizations obtained using MATLAB LIBSVM to show the clustering as well as the prediction quality of nine models. These visualizations are based on the training and testing sets used in the first experiment of Section 5.2 to measure the prediction quality of different models. The X and Y axis of these visualizations are arbitrarily chosen by MATLAB LIBSVM code. In these visualizations, four different colors are used to represent four different classes as following: red for the all-α, orange for the all-β class, yellow for the α+β class, and green for the α/β class data. The visualizations of clustering with the best performing feature set F8 (Figure 5.2(h)) and worst performing feature set F5 (Figure 5.2(e)) show the difference in their clustering quality. In Figure 5.2(e) the clusters/class data represented by 4 different colored circles are jumbled together. No class is clearly separable from the others, which implies the poor clustering ability of model with feature set F5. In Figure 5.2(h) the class data represented by red (all-α) and orange (all-β) circles are clearly separable. The class data represented by yellow (α+β) and green (α/β) circles are not as plainly separable as red (all-α) and orange (all-β) class data, but the clusters are clearly visible. The poor classification using F5 can be seen in Figure 5.2(e) where the misclassified data (the edge color and fill color of circles are not the same) are more visible than they are in Figure 5.2(h) using feature set F8.

From all these visualizations it is evident that all these feature sets except F5 are good in clustering and predicting the all-α and all-β class data, since in every case the clusters with red and orange colored circles are fairly separable. Using feature set F5, the clusters with red (all-α) and orange (all-β) colored circles are not separable which confirms their poor prediction ability (comparing to models with other feature sets) for the all-α and all-β class data.

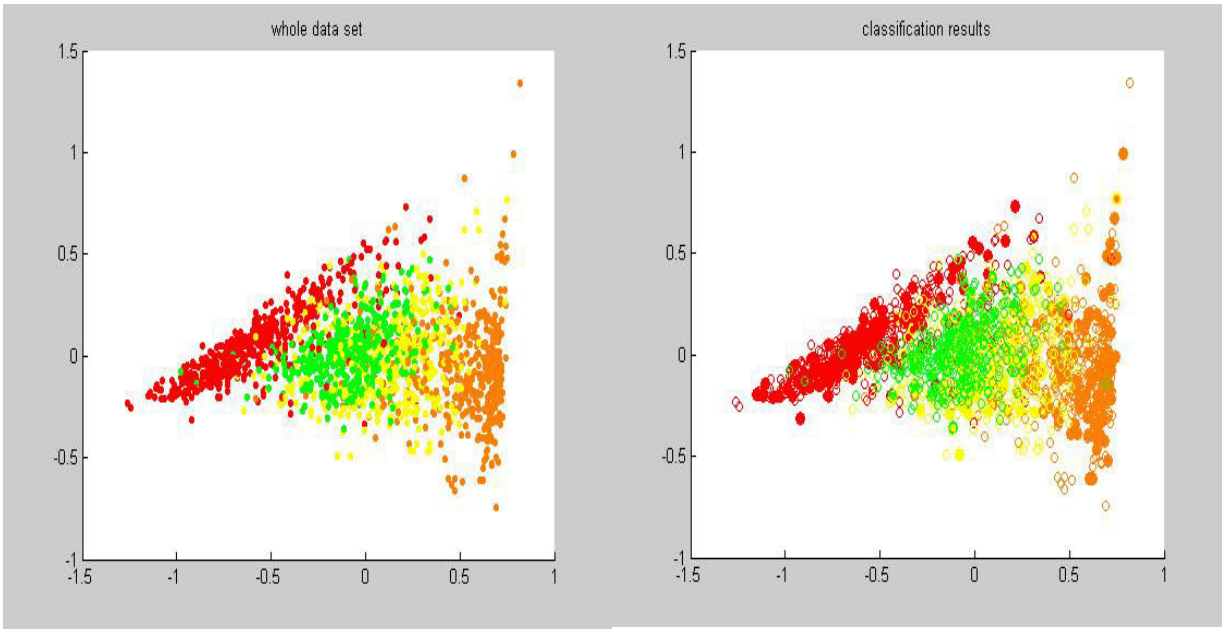


(a) Clustering using F1 Classification using F1

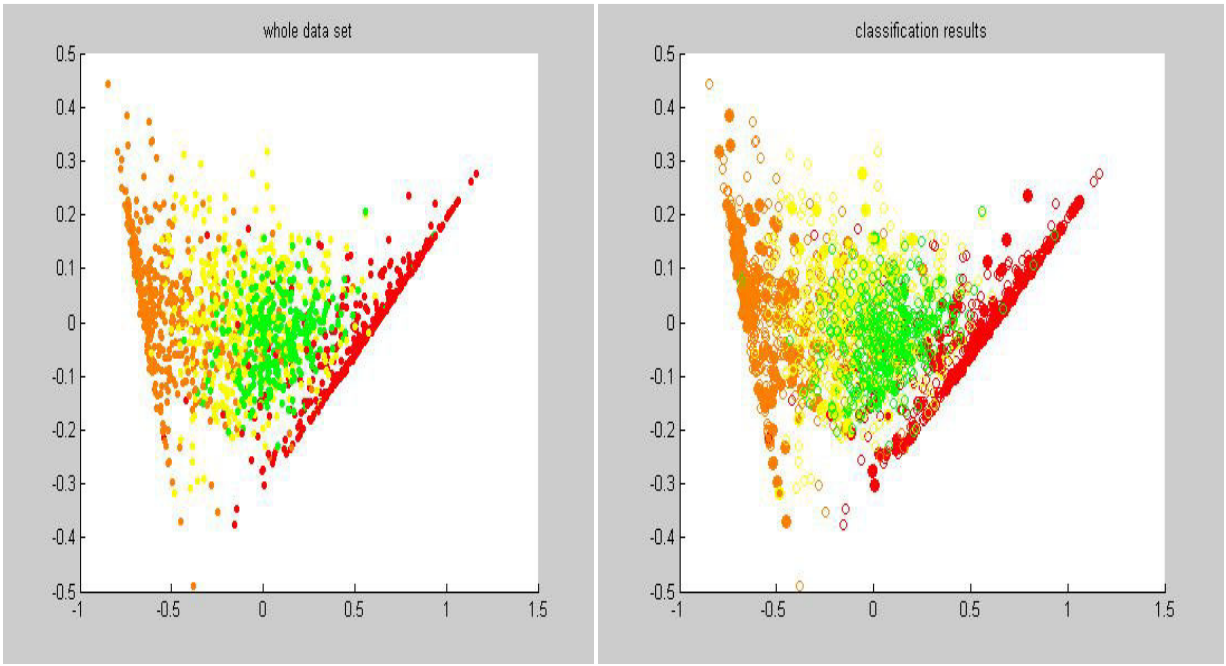


(b) Clustering using F2 Classification using F2

Figure 5.2: Visualizations of example clustering and classification using feature sets F1-F9. In each case, the left panel shows the clustering where four different colored filled circles represent four different classes of data (red for the all-α, orange for the all-β, yellow for the α+β, and green for the α/β class data). The right panel shows the classification, where unfilled circles represent training data, and filled circles represent test data. The fill color represents the class label determined by the SVM model and the edge color represents the true class label. (Circle colors are same for class data as in clustering in left panel).

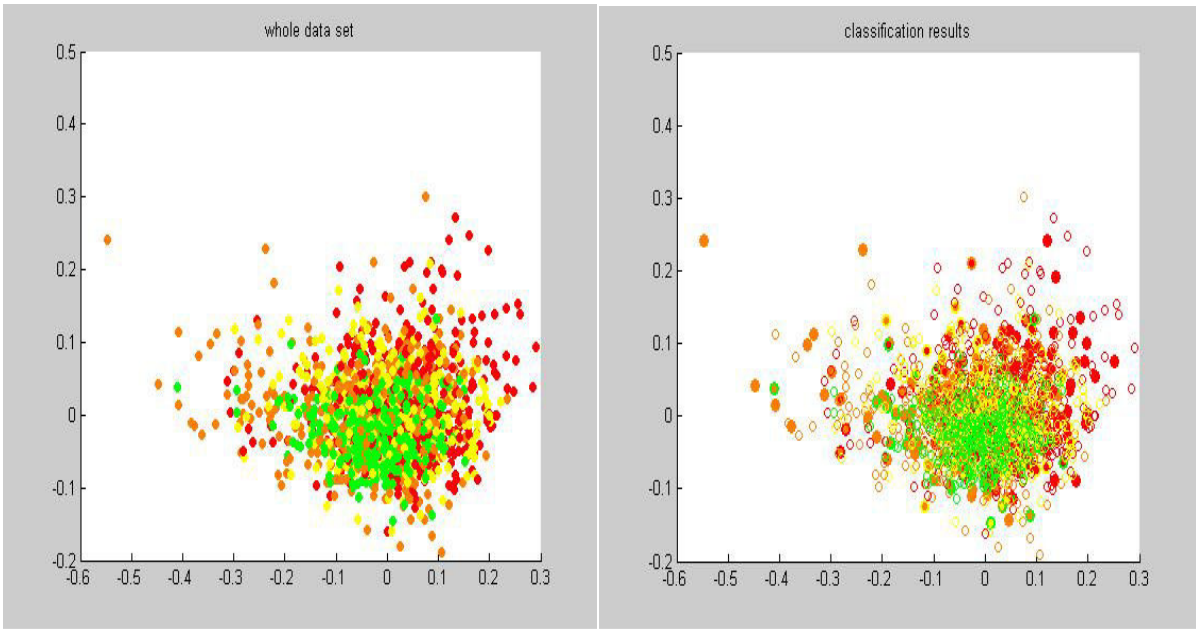


(c) Clustering using F3 Classification using F3

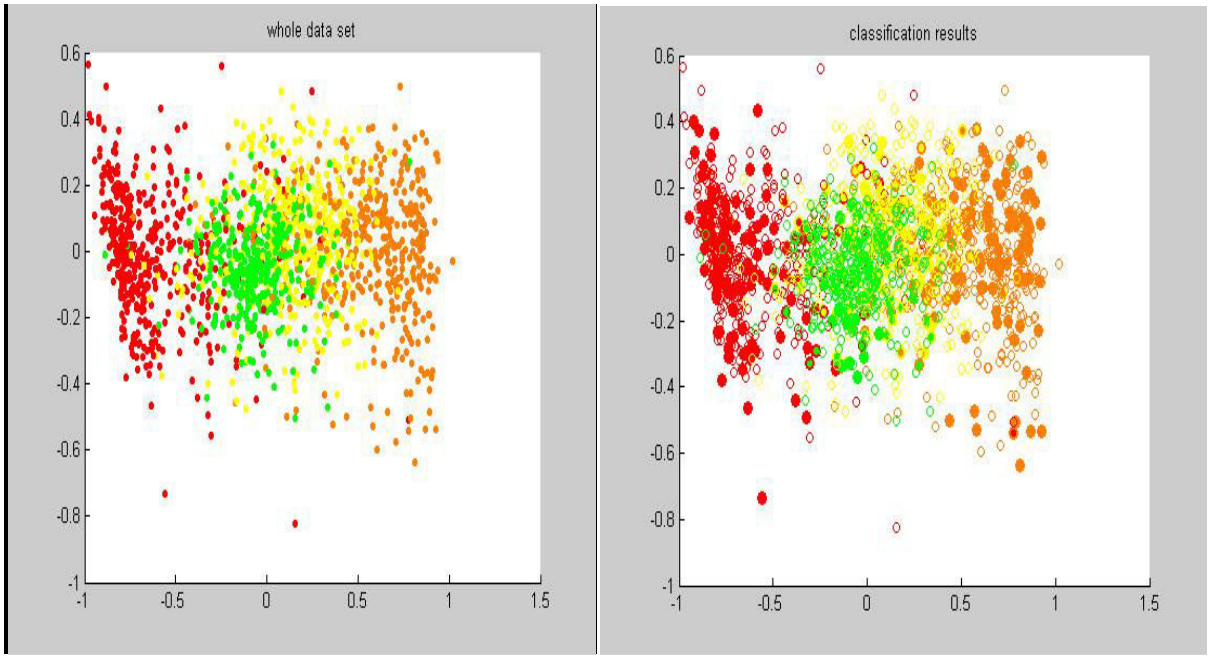


(d) Clustering using F4 Classification using F4

Figure 5.2 Continued

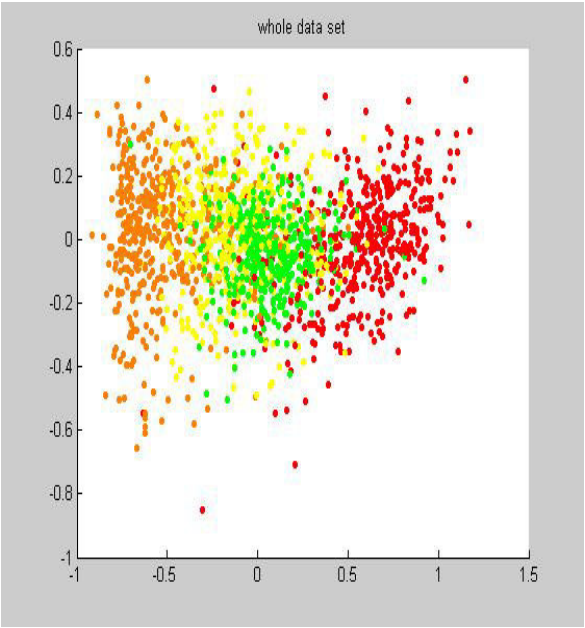


(e) Clustering using F5 Classification using F5

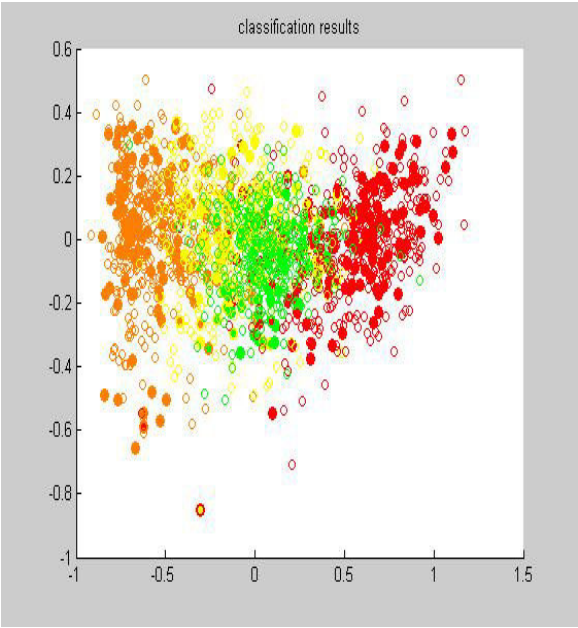


(f) Clustering using F6 Classification using F6

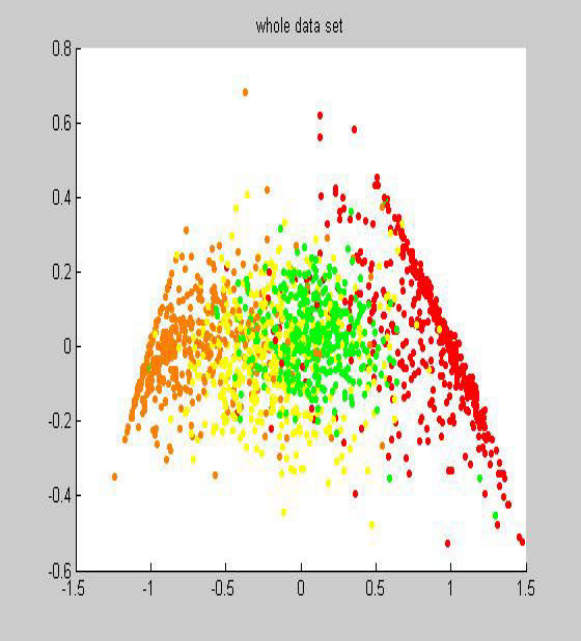
Figure 5.2 Continued



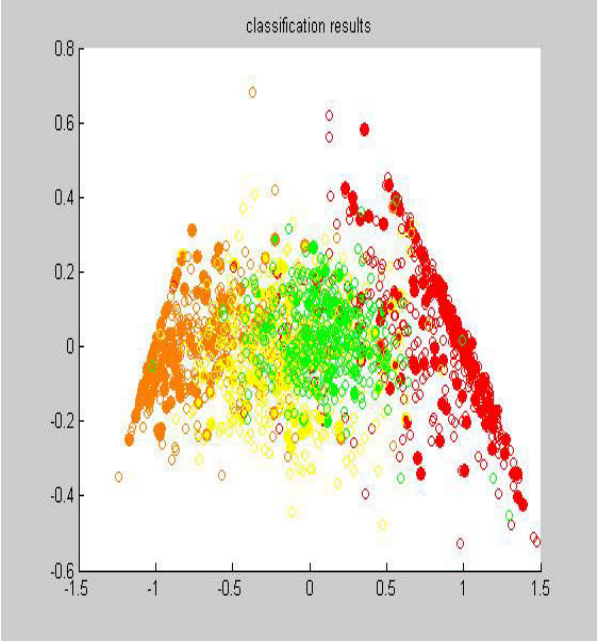
(g) Clustering using F7



Classification using F7

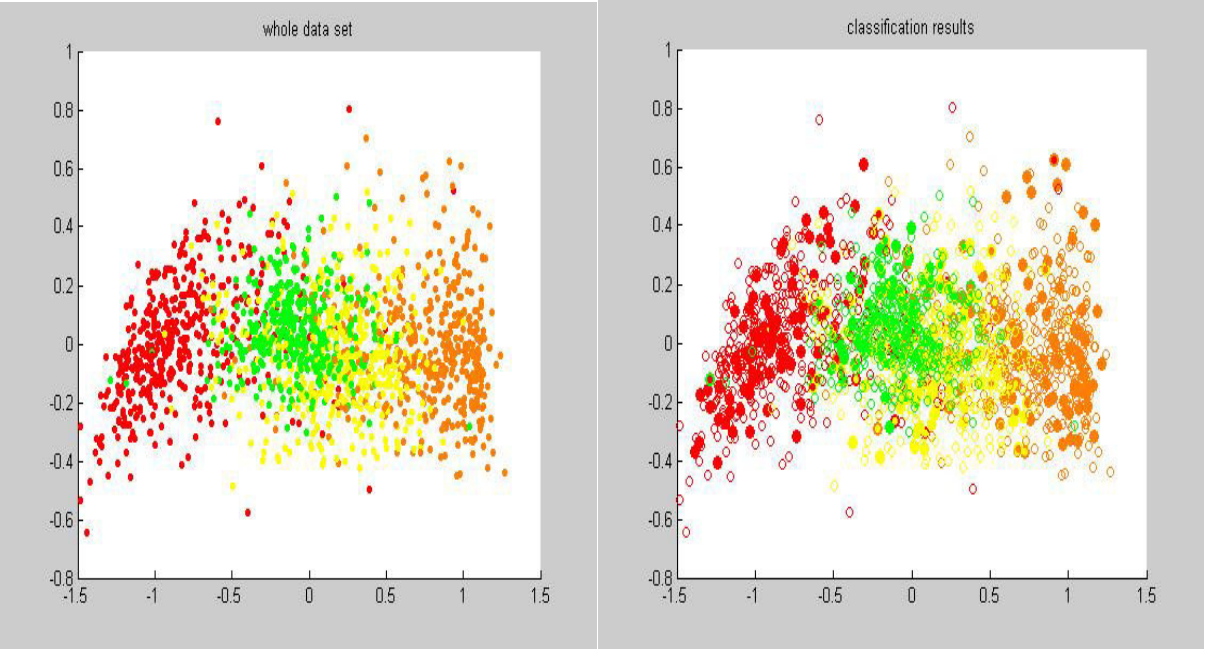


(h) Clustering using F8



Classification using F8

Figure 5.2 Continued



(i) Clustering using F9 Classification using F9

Figure 5.2 Continued

5.4 Performance Comparison with Published Methods

Our best performing models (according to cross validation accuracy) are based on feature sets F8 (combination of F1 and F4) and F1(feature set constructed from predicted secondary structure state profile). These models are compared with some published results in Table 5.4 (All these published results are based on 25PDB dataset). Set F1, although not quite as accurate as F8, is included in the comparison because of its low dimension. Since most of these published results are based on the jackknife test (see Chapter 4, Section 4.4), our results are also measured with jackknife test in this table. Table 5.4 shows that our classifiers based on F1 or F8 give higher prediction accuracies for all-β, $\alpha+\beta$ and α/β class than the results reported in the literature for the 25PDB dataset. Our classifier's (F8) best overall prediction accuracy (89.25%) is marginally better than the previous best performing method (87.8%) [51]. The latter result is based on 2510 features whereas our model uses only 234 features. Amin et al. [51] reduced the number of features using correlation based feature selection method, but after filtering down to 57 features,

their overall prediction accuracy is 86.67% which is not as good as our model which has only 22 features (88.69%). The model proposed in [56] has slightly higher prediction accuracy than our method for the all-α class (95% vs. 94.7%), but does considerably worse than our method for the other classes (81.3% vs. 91.6% for all-β, 77.6% vs. 80.9% for α+β, and 83.2% vs. 89.7% for α/β).

Ding et al. [56] used only 11 features, all based on predicted secondary structure, and obtained accuracies of 95.0% for all-alpha and 83.4% overall, whereas our F1 set has 22 such features and does a little worse on all-alpha (94.3%) but better overall (88.7%). Thus, it appears that adding more features based on secondary structural content and arrangement improves the performance on the three classes other than all-α.

Our best performing model with feature set F8 has accuracy 91.6% on the all-β class, which is 6% better than the highest published accuracy (85.6%) obtained by Zhang et al [55] and 11.53% better accuracy than the lowest published accuracy (80.1%) given by SCPRED [58] as shown in Table 5.4. This shows that the features used in our design are more effective in classifying the all-β class data and give 6%-11% more accurate results than the published methods mentioned in Table 5.4.

For the α+β class data, Liu et al. [44] obtained 55.3% accuracy whereas our model with F8 shows 80.93% accuracy, which is 25.6% better. This suggests that PSI-Blast profiles of protein sequences which represent the evolutionary relationship information of sequences used by Liu et al. [44] are not very effective in classifying the α+β class data. For classifying the α+β class data our model with F8 shows (80.9%) accuracy which is 3.4% better accuracy than the accuracy obtained by Ding et al. (77.6%) [56] and 4.9% better accuracy than Liu et al. (76%) [54]. All these three methods (our model with F8, [56], and [54]) extracted information only from predicted protein secondary structure sequences, which shows the effectiveness of information obtained from secondary structure sequence in predicting the α+β class data.

For the α/β class data, our model with F8 obtained 89.8% accuracy which is 6% better than the accuracy given by Ding et al. (83%) [56]. Our model with F8 (89.8%) shows 16.1% better accuracy than accuracy obtained by Liu et al. (73.7%) [44]. This shows that the proposed method with feature set F8 in this thesis is 6% to 16% more accurate in classifying α/β class data than the published methods mentioned in Table 5.4.

Table 5.4: Performance comparison of different methods using Jackknife test for 25PDB dataset

Method	Refe-rence	Class-wise accuracy(%)				Overall acc(%)
		all-α	all-β	$\alpha + \beta$	α/β	
SCPRED, with 9 features	[58]	92.6	80.1	71.0	74.0	79.7
Method using secondary structural information with 11 features	[54]	92.6	81.3	76.0	81.5	82.9
Method for low-similarity sequence based on secondary structure with 11features	[55]	**95.0**	85.6	73.2	81.5	83.9
SVM based method with 63 features	[59]	93.7	82.4	65.8	75.5	80.4
Method using auto-covariance transformation of PSI-BLAST profile with 140 features	[44]	85.3	81.7	55.3	73.7	74.1
Method based on predicted secondary structure with 11 features	[56]	**95.03**	81.26	77.55	83.24	84.34
Method based on functional domain and predicted secondary structure with 57 features	[51]	-	-	-	-	**86.67**
Method based on functional domain and predicted secondary structure with 2510 features	[51]					**87.80**
Method based on predicted secondary structure with 22 features (F1)	This Paper	94.26	**91.33**	**79.91**	**89.36**	**88.69**
Method based on predicted secondary structure with 162 features (F8)	This Paper	94.74	**91.63**	**80.93**	**89.78**	**89.25**

Table 5.4 shows that on the all-α class data, our proposed two models (F1 and F8) have similar accuracy to the published models (above 90% accuracy). For all-β class data our proposed two models with F1 (91.33%) and F8 (91.63%) show more than 90% accuracy, whereas all the published methods in Table 5.4 obtained accuracy in between 80%-90%. Table 5.4 also shows that for the $\alpha+\beta$ and α/β class data, the methods using features extracted from only predicted secondary structure sequences [54-56] achieved more than 70% and 80% accuracy, respectively. Our proposed two models F1 and F8, based on only features extracted from predicted secondary structure sequence also achieved more than 70% and 80% accuracy in predicting the $\alpha+\beta$ and α/β class data, respectively. Mohammad and Hampapathalu [59] used some features from predicted secondary structure sequence along with features extracted from amino acid sequence of protein and achieved 65.8% and 75.5% accuracy for the $\alpha+\beta$ and α/β class data, respectively. This suggests that use of information from predicted secondary structure sequence does not ensure higher prediction accuracy for the $\alpha+\beta$ and α/β class data. The higher prediction accuracy depends on extraction of effective information from the sequences.

Chapter 6

Conclusion and Future Work

Protein structural class prediction is a mature area of research, but there is still some room for improvement, especially when training and testing on proteins with low sequence similarity. We built SVM models to predict the structural class of these proteins based on information gathered from predicted secondary structure and the hydropathy profile. We did our experiments on the 25PDB dataset (20-25% pairwise sequence identity), which is a popular benchmark for research with twilight-zone similar sequences. The Contributions of this thesis are as follows:

- We constructed a feature set including new features from predicted protein secondary structure sequence and evaluated its performance.
- We constructed a new feature set using the hydropathy profile of amino acids and checked its performance.
- We evaluated the effectiveness of using Term Frequency-Inverse Document Frequency technique to extract useful patterns from protein secondary structure sequences to determine protein structural class. To our knowledge nobody else used the TF-IDF approach for extracting important patterns from secondary structure sequence.
- We constructed a feature set using patterns extracted from sequence constructed using hydropathy profile of amino acids in protein amino acid sequence and checked its performance. We found that features based solely on hydropathy profile do not provide high prediction accuracy.
- We checked the performance of combinations of feature sets for structural class prediction.

- We showed that the performance of our method using various feature sets compares favourably with published state-of-the art systems. The slightly superior prediction ability of our method mainly depends on extraction of extra spatial information from predicted secondary structure sequence.
- We showed that the performance of a classifier using one of our newly created feature sets is at least as accurate as current models, and uses 10 times fewer features.
- Our feature set F9 (with 216 features) is a superset of best performing feature set F8 (with 162 features) but gives lower prediction accuracy than F8. This shows that increasing the number of features does not increase prediction ability and that adding irrelevant features can decrease the prediction quality.

The limitations of this research work are as follows:

- We compared the performance of developed models with some published methods. The ideal way is to replicate the previously published models before comparing them with newly developed model. We could not replicate the published methods due to the lack of implementation information provided in respective research papers. We just compared our results with the results shown in published literatures.
- We assumed that use of less features would require less computational time. We did not check the actual computational time of newly developed models for this thesis.
- We used only one dataset to test the performance of newly developed method due to time constraint.
- We found the feature sets which give better accuracy than other published methods, but did not find the exact features which are responsible for this improvement in accuracy.

In future work, we would like to include the following tasks:

- According to the test results, the specificity scores of binary classifiers are always greater than their sensitivity scores. This might be caused by the design of one-against-all binary classifiers to solve the multiclass problem. In our proposed method one-against-all binary classifiers are trained by a larger amount of -1 class data and little +1 class data. For example, only 443 positive labeled (+1) instances and 1230 negative labeled (-1) instances are available to train and test the all-α/~all-α binary classifier. For this

unbalanced proportion, the classifiers can be biased in classifying the -1 class data over classifying +1 class data. In future work we would like to develop SVM models with the same feature sets using the one-against-one multiclass classification method to see if there is any change in the specificity and sensitivity scores.

- Each of the SVM models using one of the nine different feature sets shows better performance in predicting the all-α, all-β, and α/β class data than in predicting the $\alpha+\beta$ class data. We would like to search for effective features to increase the efficiency in predicting the $\alpha+\beta$ class data. The all-α, all-β, and α/β class sequences follow some regular trends. For example, the all-α and all-β class sequences are mainly composed of α-helix and β-strand patterns, respectively. The α/β class sequences follow $\beta\alpha\beta$ motifs, where β-strands alternate with α-helices. In the $\alpha+\beta$ class sequences, α-helix and β-strands appear separately with no specific trends. This lack of regularity may make it difficult to extract specific features to predict the $\alpha+\beta$ class data. In the future, we would try to search for some features which would be helpful for predicting the $\alpha+\beta$ class data more accurately.

Bibliography

[1] Luscombe, Nicholas M., Dov Greenbaum, and Mark Gerstein. "What is bioinformatics? A proposed definition and overview of the field." *Methods of information in medicine* 40, no. 4 (2001): 346-358.

[2] Bourne, Philip E., and Gu, Jenny (Editors), *Structural Bioinformatics* (2nd edition), John Wiley & Sons, New York, 2009.

[3] Murzin, Alexey G., Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. "SCOP: a structural classification of proteins database for the investigation of sequences and structures." *Journal of molecular biology* 247, no. 4 (1995): 536-540.

[4] Orengo, Christine A., A. D. Michie, S. Jones, David T. Jones, M. B. Swindells, and Janet M. Thornton. "CATH–a hierarchic classification of protein domain structures." *Structure* 5, no. 8 (1997): 1093-1109.

[5] "Using molecular marker technology in studies on plant genetic diversity", IPGRI and Cornell University, 2003, http://www2.bioversityinternational.org/Publications/Molecular_Markers_Volume_1_en/molmarkers/PDF/Protein%20basics.pdf , Last visited: July 3, 2013.

[6] "The 20 Amino Acids", http://www.cryst.bbk.ac.uk/education/AminoAcid/the_twenty.html, Last visited: July 23, 2013.

[7] "Protein primary structure", Ophardt, Charles E., Virtual Chembook, Elmhurst College (2003), http://www.elmhurst.edu/~chm/vchembook/565proteins.html, Last visited: July 3, 2013.

[8] Pauling, Linus, Robert B. Corey, and Herman R. Branson. "The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain." *Proceedings of the National Academy of Sciences* 37, no. 4 (1951): 205-211.

[9] "alpha helix", http://www.thefreedictionary.com/alpha+helix, Last visited: July 3, 2013.

[10] "Beta sheet", https://en.wikipedia.org/wiki/Beta_sheet, Last visited: July 3, 2013.

[11] "beta sheet", http://www.thefreedictionary.com/beta+sheet, Last visited: July 3, 2013.

[12] "random coil", http://www.thefreedictionary.com/random+coil, Last visited: July 3, 2013.

[13] "Chapter 2-cell structure and organization", http://eglobalmed.com/core/VirtualMicrobiology/www.bact.wisc.edu/Microtextbook/indexd811.html?module=Book , Last visited: July 3, 2013.

[14] "*β*-sheet", Jeff D Cronk, Gonzaga University, http://guweb2.gonzaga.edu/faculty/cronk/biochem/B-index.cfm?definition=beta, Last visited: July 3, 2013.

[15] "Folding@Home – Distributed Computing", http://www.yespleasestudio.co.za/foldinghome-distributed-computing/, Last visited: July 3, 2013.

[16] Jones, David T. "Protein secondary structure prediction based on position-specific scoring matrices." *Journal of molecular biology* 292, no. 2 (1999): 195-202.

[17] Lin, Kuang, Victor A. Simossis, Willam R. Taylor, and Jaap Heringa. "A simple and fast secondary structure prediction method using hidden neural networks." *Bioinformatics* 21, no. 2 (2005): 152-159.

[18] Porollo, Aleksey, and Jaroslaw Meller. "Versatile annotation and publication quality visualization of protein complexes using POLYVIEW-3D." *BMC bioinformatics* 8, no. 1 (2007): 316.

[19] "Information on 2hbc", ArchDB, http://sbi.imim.es/cgi-bin/archdb//loops.pl?pdb=2hbc ,

Last visited: July 3, 2013.

[20] "Information on 1ku8", ArchDB, http://sbi.imim.es/cgi-bin/archdb//loops.pl?pdb=1ku8, Last visited: July 3, 2013.

[21] "Alpha/ Beta Topologies", Birkbeck College, 1995

http://www.cryst.bbk.ac.uk/PPS95/course/8_folds/alph_bet_wnd.html, Last visited: July 3, 2013.

[22] "Information on 1pya", ArchDB, http://sbi.imim.es/cgi-bin/archdb//loops.pl?pdb=1pya, Last visited: July 3, 2013.

[23] Vapnik, Vladimir. *The nature of statistical learning theory*. springer, 2000.

[24] "SVM Tutorial", Zoya Gavrilov, http://web.mit.edu/zoya/www/SVM.pdf, Last visited: July 3, 2013.

[25]"Support Vector Machine Tutorial", Jason Weston, NEC Labs America, http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf, Last visited: July 3, 2013.

[26] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003).

[27] Levitt, Michael, and Cyrus Chothia. "Structural patterns in globular proteins." *Nature* 261, no. 5561 (1976): 552.

[28]Brenner, Steven E., Cyrus Chothia, Tim JP Hubbard, and Alexey G. Murzin. "[37] Understanding protein structure: Using scop for fold interpretation." *Methods in Enzymology* 266 (1996): 635-643.

[29] Zerrin Isik, Berrin Yanikoglu, and Ugur Sezerman. "Protein structural class determination using support vector machines." In *Computer and Information Sciences-ISCIS 2004*, pp. 82-89. Springer Berlin Heidelberg, 2004.

[30] Zhou, Guo-Ping. "An intriguing controversy over protein structural class prediction." *Journal of Protein Chemistry* 17, no. 8 (1998): 729-738.

[31] Wu, Li, Qi Dai, Bin Han, Lei Zhu, and Lihua Li. "Prediction of protein structural class using a combined representation of protein-sequence information and support vector machine." In *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on*, pp. 101-106. IEEE, 2010.

[32] Nair, Achuth Sankar S., and T. Mahalakshmi. "Visualization of genomic data using inter-nucleotidedistance signals." *Proceedings of IEEE Genomic Signal Processing* (2005): 11-13.

[33] Afreixo, Vera, Carlos AC Bastos, Armando J. Pinho, Sara P. Garcia, and Paulo JSG Ferreira. "Genome analysis with inter-nucleotide distances." *Bioinformatics* 25, no. 23 (2009): 3064-3070.

[34] Zhang, T-L., and Y-S. Ding. "Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes." *Amino Acids* 33, no. 4 (2007): 623-629.

[35] Tang, Fa-ming, Zhong-dong Wang, and Mian-yun Chen. "On multiclass classification methods for support vector machines." *Control and Decision* 20, no. 7 (2005): 746.

[36] Ding, Yong-Sheng, Tong-Liang Zhang, and Kuo-Chen Chou. "Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network." *Protein and peptide letters* 14, no. 8 (2007): 811-815.

[37] Chou, Kuo-Chen. "Prediction of protein cellular attributes using pseudo-amino acid composition." *Proteins: Structure, Function, and Bioinformatics* 43, no. 3 (2001): 246-255.

[38] Abe, Shigeo. "Fuzzy LP-SVMs for multiclass problems." In *Proceedings of European Symposium on Artificial Neural Networks (ESANN'2004) Bruges, Belgium*, pp. 429 - 434. 2004.

[39] Chou, Kuo-Chen. "A key driving force in determination of protein structural classes." *Biochemical and biophysical research communications* 264, no. 1 (1999): 216-224.

[40] Klein, Petr, and Charles Delisi. "Prediction of protein structural class from the amino acid sequence." *Biopolymers* 25, no. 9 (1986): 1659-1672.

[41] Bu, Wei-Shu, Zhi-Ping Feng, Ziding Zhang, and Chun-Ting Zhang. "Prediction of protein (domain) structural classes based on amino-acid index." *European Journal of Biochemistry* 266, no. 3 (1999): 1043-1049.

[42] Chou, Kuo-Chen. "A key driving force in determination of protein structural classes." *Biochemical and biophysical research communications* 264, no. 1 (1999): 216-224.

[43] Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research* 25, no. 17 (1997): 3389-3402.

[44] Liu, Taigang, Xingbo Geng, Xiaoqi Zheng, Rensuo Li, and Jun Wang. "Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles." *Amino acids* 42, no. 6 (2012): 2243-2249.

[45] Mizianty, Marcin, and Lukasz Kurgan. "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences" *BMC bioinformatics* 10.1 (2009): 414.

[46] Dong, Qiwen, Shuigeng Zhou, and Jihong Guan. "A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation." *Bioinformatics* 25, no. 20 (2009): 2655-2662.

[47] Guo, Yanzhi, Lezheng Yu, Zhining Wen, and Menglong Li. "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences." *Nucleic acids research* 36, no. 9 (2008): 3025-3030.

[48] Guo, Yanzhi, Menglong Li, Minchun Lu, Zhining Wen, and Zhongtian Huang. "Predicting G-protein coupled receptors–G-protein coupling specificity based on autocross-covariance transform." *Proteins: structure, function, and bioinformatics* 65, no. 1 (2006): 55-60.

[49] Wu, Jiang, Meng-Long Li, Le-Zheng Yu, and Chao Wang. "An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition." *The Protein Journal* 29, no. 1 (2010): 62-67.

[50] Chou, Kuo-Chen, and Yu-Dong Cai. "Predicting protein structural class by functional domain composition." *Biochemical and biophysical research communications* 321, no. 4 (2004): 1007-1009.

[51] Ahmadi Adl, Amin, Abbas Nowzari-Dalini, Bin Xue, Vladimir N. Uversky, and Xiaoning Qian. "Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences." *Journal of Biomolecular Structure and Dynamics* 29, no. 6 (2012): 1127-1137.

[52] Apweiler, Rolf, Terri K. Attwood, Amos Bairoch, E. Birney, M. Biswas, P. Bucher, L. Cerutti et al. "The InterPro database, an integrated documentation resource for protein families, domains and functional sites." *Nucleic acids research* 29, no. 1 (2001): 37-40.

[53] Hall, Mark A. "Correlation-based feature selection for machine learning." PhD diss., The University of Waikato, 1999.

[54] Liu, Tian, and Cangzhi Jia. "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information." *Journal of theoretical biology* 267, no.3 (2010): 272-275.

[55] Zhang, Shengli, Shuyan Ding, and Tianming Wang. "High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure." *Biochimie* 93, no. 4 (2011): 710-714.

[56] Ding, Shuyan, Shengli Zhang, Yang Li, and Tianming Wang. "A novel protein structural classes prediction method based on predicted secondary structure." *Biochimie* 94, no. 5 (2012): 1166-1171.

[57] Kurgan, Lukasz A., Tuo Zhang, Hua Zhang, Shiyi Shen, and Jishou Ruan. "Secondary structure-based assignment of the protein structural classes." *Amino Acids* 35, no. 3 (2008): 551-564.

[58] Kurgan, Lukasz, Krzysztof Cios, and Ke Chen. "SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences." *BMC bioinformatics* 9, no. 1 (2008): 226.

[59] Mohammad, Tabrez Anwar Shamim, and Hampapathalu Adimurthy Nagarajaram. "Svm-based method for protein structural class prediction using secondary structural content and structural information of amino acids." *Journal of Bioinformatics and Computational biology* 9, no. 4 (2011): 489-502.

[60] Cheng, Jianlin, Arlo Z. Randall, Michael J. Sweredoski, and Pierre Baldi. "SCRATCH: a protein structure and structural feature prediction server." *Nucleic acids research* 33, no. suppl 2 (2005): W72-W76.

[61] Kurgan, Lukasz, and Ke Chen. "Prediction of protein structural class for the twilight zone sequences." *Biochemical and biophysical research communications* 357, no. 2 (2007): 453-460.

[62] Yang, Jian-Yi, Zhen-Ling Peng, and Xin Chen. "Prediction of protein structural classes for low-homology sequences based on predicted secondary structure." *BMC bioinformatics* 11, no. Suppl 1 (2010): S9.

[63] Hastie, Trevor, and Robert Tibshirani. "Classification by pairwise coupling." *The annals of statistics* 26, no. 2 (1998): 451-471.

[64] Cai, Yu-Dong, and Guo-Ping Zhou. "Prediction of protein structural classes by neural network." *Biochimie* 82, no. 8 (2000): 783-785.

[65] Chandonia, John-Marc, and Martin Karplus. "Neural networks for secondary structure and structural class predictions." *Protein Science* 4, no. 2 (1995): 275-285.

[66] Hobohm, Uwe, and Chris Sander. "Enlarged representative set of protein structures." *Protein Science* 3, no. 3 (1994): 522-524.

[67] Kurgan, Lukasz A., and Leila Homaeian. "Prediction of structural classes for protein sequences and domains—impact of prediction algorithms, sequence representation and

homology, and test procedures on accuracy." *Pattern Recognition* 39, no. 12 (2006): 2323-2343.

[68] Jiang, Kai, Shuming Ye, Hang Chen, and Fei Gu. "Protein Structural Class Prediction Using Physiochemical Property Based Grouped Weighted Encoding Index." In *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*, pp. 275-278. IEEE, 2008.

[69] Li, Zhan-Chao, Xi-Bin Zhou, Zong Dai, and Xiao-Yong Zou. "Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis." *Amino Acids* 37, no. 2 (2009): 415-425.

[70] Liu, Na, and Tianming Wang. "Protein-based phylogenetic analysis by using hydropathy profile of amino acids." *FEBS letters* 580, no.22 (2006): 5321-5327.

[71] Yang, Yang, Bao-liang Lu, and Wen-yun Yang. "Classification of protein sequences based on word segmentation methods." *Proceedings of the 6th Asia-Pacific Bioinformatics Conference*. Vol. 6. 2008.

[72] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." *Information processing & management* 24, no. 5 (1988): 513-523.

[73] "Machine Learning Techniques in Bioinformatics", Winter School on "Data Mining Techniques and Knowledge Discovery in Agricultural Datasets": 335-344, http://iasri.res.in/ebook/win_school_aa/notes/supervised_machine.pdf, Last visited: July 3, 2013.

[74] Noble, William Stafford. "Support vector machine applications in computational biology." In Schoekkopf,B., Tsuda,K. and Vert,J.-P. (eds), *Kernel Methods in Computational Biology* (2004), MIT Press,Cambridge, MA, pp. 71–92.

[75] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, no.3 (2011): 27.

[76] "LIBSVM-A Library for Support Vector Machine", Chang, Chih-Chung, and Chih-Jen Lin, http://www.csie.ntu.edu.tw/~cjlin/libsvm, Last visited: July 3, 2013.

[77] Yuan, Zheng, and Bixing Huang. "Prediction of protein accessible surface areas by support vector regression." *Proteins: Structure, Function, and Bioinformatics* 57, no. 3 (2004): 558-564.

[78] Yuan, Zheng, Timothy L. Bailey, and Rohan D. Teasdale. "Prediction of protein B-factor profiles." *Proteins: Structure, Function, and Bioinformatics* 58, no.4 (2005): 905-912.

[79] Fawcett, Tom. "ROC graphs: Notes and practical considerations for researchers." *Machine Learning* 31 (2004): 1-38.

[80] Matthews, Brian W. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405, no. 2 (1975): 442-451.