Two-stage support vector machines for protein structure and solvent prediction

Nguyen, Ngoc Minh

2005

Nguyen, N. M. (2005). Two-stage support vector machines for protein structure and solvent prediction. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/2407>

<https://doi.org/10.32657/10356/2407>

Nanyang Technological University

Downloaded on 26 Apr 2023 21:30:50 SGT

Two-Stage Support Vector Machines for Protein Structure and Solvent Prediction

Nguyen Ngoc Minh

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2005

Table of Contents

Table of Contents	ii
List of Tables	v
List of Figures	ix
Abstract	xii
Acknowledgements	xiv
Publications	xv
Abbreviations and Symbols	xviii
1 Introduction	1
1.1 Background	1
1.2 Motivation	4
1.3 Major Contributions	8
1.4 Organization of the Thesis	10
2 Bioinformatics Approaches to Protein Secondary Structure Prediction: A Review	12
2.1 First Generation: single amino acid statistics	14
2.1.1 Chou-Fasman Method	14
2.1.2 GOR I Method	15
2.2 Second Generation: segment statistics	17
2.2.1 Statistical Methods	17
2.2.2 Neural Networks	18
2.2.3 Nearest-Neighbor Methods	21
2.3 Third Generation: evolutionary information	23

2.3.1	Multiple Sequence Alignments	23
2.3.2	PSI-BLAST Profiles	23
2.4	Recent Developments	25
2.4.1	Combining Predictions	25
2.4.2	Bayesian Method	26
2.4.3	Recurrent Neural Networks	29
2.4.4	Support Vector Machines	29
2.4.5	Combining Evolutionary Information	32
2.4.6	Hidden Markov Models	32
2.5	Comparison	34
2.6	Summary	37
3	Two-Stage Binary SVMs to Protein Relative Solvent Accessibility Prediction	40
3.1	Introduction	41
3.2	Two-Stage SVM Approach	43
3.2.1	First Stage	44
3.2.2	Second Stage	45
3.3	Generalization in Two-Stage SVMs	48
3.4	Experimental Results	61
3.4.1	Dataset 1 (RS126)	61
3.4.2	Dataset 2 (Manesh)	62
3.4.3	RSA and Prediction Accuracy Assessment	63
3.4.4	Results	64
3.4.5	Time and Space Complexities	78
3.5	Summary	79
4	Two-Stage Multi-Class SVMs to Protein Secondary Structure Prediction	82
4.1	Introduction	83
4.2	Two-Stage MSVM Approach	85
4.2.1	First Stage	86
4.2.2	Second Stage	87
4.3	Generalization in Two-Stage Multi-Class SVMs	89
4.4	Vapnik Chervonenkis (VC) Dimension	96
4.5	Comparison of Two-Stage MSVM and Two-Stage MLP Methods for PSS Prediction	97
4.6	Experimental Results	99
4.6.1	Dataset 1 (RS126)	99

4.6.2	Dataset 2 (CB396)	100
4.6.3	Dataset 3 (EVAsec)	101
4.6.4	Protein Secondary Structure Definition	101
4.6.5	Prediction Accuracy Assessment	102
4.6.6	Results	104
4.6.7	Time and Space Complexities	117
4.7	Summary	119
5	Discussion and Conclusion	121
5.1	Two-Stage Approaches to Protein Secondary Structure Prediction . .	122
5.2	Two-Stage SVM Approaches for RSA and PSS Predictions	126
5.3	Directions of Future Work	128
5.3.1	Classification of Protein Domains	129
5.3.2	3-D Structure Prediction from Predicted Secondary Structure and Solvent Accessibility	130
5.3.3	Prediction of Protein-Protein Interactions	131
5.3.4	Prediction of Function of an Unknown Protein	132
5.3.5	Summary	133
	Bibliography	142

List of Tables

2.1	Six programs used in the Jpred method and three-state per-residue accuracy (Q_3) [149] of each method on the RS126 dataset of 126 non-homologous globular proteins.	25
2.2	Performances and web servers of recent methods for protein secondary structure prediction. The notation (ext) indicates that the corresponding method uses PSI-BLAST profiles instead of multiple sequence alignments.	35
2.3	Summary of various generations and timeline of techniques among NNs, HMMs, and SVMs.	37
3.1	The list of 30 proteins used for training the single-stage and two-stage SVM approaches with Manesh dataset.	62
3.2	Performance of RSA prediction by the single-stage and two-stage SVM approaches with different window lengths w_1 and the selected window $w_2 = 21$ on the RS126 and Manesh datasets at a 25% threshold. . . .	65
3.3	Performance of RSA prediction by the single-stage and two-stage SVM approaches with different type of kernel functions on RS126 and Manesh datasets at a 25% threshold with $\gamma_1 = \gamma_2 = 1.0$. The neighborhood windows of size 13 and 21 were used in the single-stage and two-stage SVM approach, respectively.	67

3.4	Performance of RSA prediction by the single-stage and two-stage SVM approaches with different parameters of Gaussian kernel on the Manesh dataset at a 25% threshold.	67
3.5	Comparison of performances of two-stage SVM approach with other methods in RSA prediction on the RS126 dataset with position specific scoring matrices generated by PSI-BLAST. The notation - indicates that the corresponding result was not available from the literature. . .	69
3.6	Comparison of performances of two-stage SVM approach in RSA prediction based on position specific scoring matrices generated by PSI-BLAST, with other methods on the Manesh dataset.	70
3.7	Comparison of performances of two-stage SVM approach on 10 proteins (1lts, 1nba, 1afw, 3cox, 2wsy, 7rsa, 1amm, 1mai, 1knb, 1kte) based on PSI-BLAST profiles with two different training sets of 30 and 205 proteins at a 5% threshold.	75
3.8	The properties of 20 amino acids: their average occurrences, probabilities of exposures, and the error in RSA prediction on the Manesh dataset at a 25% threshold.	76
3.9	Comparison of performances of two-stage SVM approach on the Manesh dataset based on position specific scoring matrices generated by PSI-BLAST with two different NR databases.	77
3.10	The training time and memory space of the first stage and the second stage SVM for the Manesh and RS126 datasets on a supercomputer, Compaq AlphaServer SC45.	77
4.1	The dataset of 126 nonhomologous globular protein chains from Rost and Sandar (1993) used for the seven-fold cross validation.	99
4.2	The list of 90 proteins extracted from PSIPRED dataset, in 1999, used for training the two-stage MSVM approach.	101
4.3	The list of 64 proteins of the EVAsec dataset used for testing the two-stage MSVM approach.	102

4.4	Accuracies of PSS prediction by the single-stage MSVM approach on the RS126 and CB396 datasets. The window length indicates the size of neighborhood taken as the input for the single-stage approach. . .	105
4.5	Accuracies of PSS prediction by the two-stage MSVM approach on the RS126 and CB396 datasets at different neighborhood window sizes of the second stage. The size of the first stage neighborhood was taken as 15.	106
4.6	Accuracies of PSS prediction by the single-stage and two-stage MSVM approaches with different window lengths w_1 and the selected window $w_2 = 21$ on the RS126 and CB396 datasets.	107
4.7	Comparison of Q_3 accuracy of the single-stage and two-stage MSVM approaches with different type of kernel functions on RS126 and CB396 datasets with $\gamma_1 = \gamma_2 = 0.5$. The neighborhood windows of size 15 and 21 were used in the single-stage and two-stage MSVM approach, respectively.	108
4.8	Comparison of Q_3 accuracy of the single-stage and two-stage MSVM approaches with different parameters of Gaussian kernel on the CB396 dataset.	108
4.9	Comparison of performances of single-stage and two-stage MSVM approaches with other methods in protein secondary structure prediction on RS126 dataset with position specific scoring matrices generated by PSI-BLAST and the following reduction: H, G to (H); E, B to (E); the remainder to (C). The notation - indicates that the corresponding results were not available from the literature.	112
4.10	Comparison of performances of single-stage and two-stage MSVM approaches in PSS prediction on the CB396 dataset with PSI-BLAST profiles.	113

4.11	Comparison of performances of two-stage MSVM approach based on position specific scoring matrices generated by PSI-BLAST, with other methods on the EVAsec dataset.	114
4.12	Comparison of performances of two-stage MSVM approach with PSIPRED method on PSIPRED dataset.	114
4.13	The properties of 20 amino acids: their average occurrences, the error in PSS prediction, and probabilities of α -helix, β -strand, and coil on the CB396 dataset.	115
4.14	The training time and memory space of the first stage and the second stage MSVM for the RS126, CB396, and PSIPRED datasets on a supercomputer, Compaq AlphaServer SC45.	116
5.1	Comparison of performances of single-stage and two-stage approaches with MSVM at the second stage in PSS prediction on the RS126 dataset.	124
5.2	Comparison of performances of single-stage and two-stage approaches with MSVM at the second stage in PSS prediction on the CB396 dataset.	125

List of Figures

1.1	The names, abbreviations and R-groups of the twenty amino acids. Atoms are colored white for Carbon, light gray for Nitrogen, dark for Oxygen, and black for Sulphur. Small atoms are Hydrogen. Bonds connecting R-groups to main-chain atoms are drawn in bold. Note that the proline side chain joins the main chain at both C ^α and N atoms.	2
1.2	The amino acid sequence and its corresponding three-dimensional structure, from the light chain of the anti-prion FAB 3F4 protein complex (PDB ID 1CR9).	3
2.1	Illustration of folding a linear chain of amino acids into a three-dimensional (3-D) protein structure.	13
2.2	Architecture of a neural network model for protein secondary structure prediction.	19
2.3	Architecture of binary support vector machine (SVM) approach to protein secondary structure prediction.	30
2.4	Comparison of three-state per-residue accuracy (Q_3) [149] of different predictors of protein secondary structure on the RS126 dataset of 126 nonhomologous globular proteins. The notation (ext) indicates that the corresponding method uses position specific scoring matrices generated by PSI-BLAST instead of multiple sequence alignments.	33

2.5	Results of different methods for predicting secondary structure of the redox switch domain of the E. Coli Hsp33 protein (PDB ID 1XJH:A) from EVA, http://cubic.bioc.columbia.edu/eva/sec/prot/1xjh_A.html	36
3.1	The two-stage binary SVMs for relative solvent accessibility prediction from amino acid sequences.	46
3.2	A set of lines shattering three points.	52
3.3	Performance of relative solvent accessibility prediction by the single-stage and two-stage SVM approaches on the RS126 dataset at a 25% threshold. The window length indicates the size of neighborhood taken as the input for the single-stage and two-stage SVM methods. For two-stage SVM, the window length indicates the size of neighborhood taken as the input for the second stage, with the first stage operating with its optimal window size 13.	64
3.4	Performance of relative solvent accessibility prediction by the single-stage and two-stage SVM approaches on the Manesh dataset with different window lengths at a 25% threshold. For two-stage SVM, the window length indicates the size of neighborhood taken as the input for the second stage, with the first stage operating with its optimal window size 13.	65
3.5	The distribution of prediction scores obtained by two-stage SVMs for the benchmark 185 proteins of the Manesh dataset at a 5% threshold based on PSI-BLAST profiles.	71
3.6	The distribution of prediction scores obtained by two-stage SVMs for the benchmark RS126 dataset at a 5% threshold based on PSI-BLAST profiles.	72
3.7	The ROC curves on the Manesh dataset for single-stage and two-stage SVM approaches for RSA prediction.	73
3.8	The ROC curves on the RS126 dataset for single-stage and two-stage SVM approaches for RSA prediction.	74

4.1	The two-stage Multi-class Support Vector Machine (MSVM) approach for protein secondary structure prediction	88
4.2	A decision directed acyclic graph with discriminant functions $f_2^{k/l} \in \mathcal{F}$ at decision nodes k/l where $k, l \in \Omega_T$ and the leaves labeled by the classes H , E , and C	92
4.3	Comparison of Q_3 accuracy of different predictors in protein secondary structure prediction on RS126 dataset of 126 nonhomologous globular proteins. The notation (ext) indicates that the corresponding method uses position specific scoring matrices generated by PSI-BLAST instead of multiple sequence alignments.	110
4.4	Bar graph showing the distribution of Q_3 scores obtained by two-stage MSVMs for the benchmark RS126 set based on PSI-BLAST profiles.	111
4.5	Comparison of Q_3 accuracy of different predictors in protein secondary structure prediction on CB396 dataset of 396 nonhomologous proteins.	111
5.1	Illustration of two-stage approaches with Multi-class Support Vector Machine (MSVM) at the second stage for protein secondary structure prediction.	123
5.2	Illustration of the two-stage SVM approach for (a) relative solvent accessibility prediction and (b) protein secondary structure prediction, from amino acid sequences.	136

Abstract

Information of a protein's 3-D structure provides valuable clues to the protein's function that is vital to many aspects of living organism, such as those of enzymes, hormones, and structural material, etc., as well as to the design of new drugs for combating disease. Unfortunately, the protein structure prediction problem is a combinatorial optimization problem, which so far has an eluded solution because of the exponential number of potential solutions. One of the current approaches is to predict the Relative Solvent Accessibility (RSA) and/or Protein Secondary Structure (PSS), which are intermediate representations of the full knowledge of the 3-D structure to aid the prediction of ultimate 3-D structure.

Most RSA and PSS prediction approaches are single-stage methods in the sense that the types of solvent accessibility and secondary structure elements are considered as a complex function of the neighboring elements in their amino acid sequences. This complex mapping between the input and output is insufficient to extract the structure from the input sequence because (1) the solvent accessibility or secondary structure at a particular residue of a sequence depends not only on the amino acid residues in the neighborhood but also on the information of structural formations, such as solvent accessibilities or secondary structures, of the residues in the neighborhood, (2) the problem of a neighborhood window associated with the amino acid sequence considers only a local neighborhood, and (3) the generalization of the mapping realized by one stage is insufficient. In this research, we introduce another prediction layer to incorporate the interactions or contextual information among the elements of the solvent accessibility or secondary structure sequence as one way to overcome these

limitations. We have proposed Support Vector Machines (SVMs) as the second stage because of their capacity to minimize the risk of an already prediction.

This thesis investigates the use of two-stage SVMs for RSA and PSS predictions by using a second SVM to enhance the output of the classical SVM approaches. The inputs to the two-stage SVMs are based on the position specific scoring matrices generated by PSI-BLAST (Position Specific Iterative - Basic Local Alignment Search Tool) of the input sequence. We show that the SVM at the second stage minimizes the generalization error made by the first stage with the incorporation of the contextual relationships among RSA or PSS elements. Two-stage approach was first introduced in PHD (Profile network from HeiDelberg) approach which uses two Multi-Layer Perceptrons (MLPs) in cascade for PSS prediction. MLPs are not optimal classifiers in terms of generalization capabilities over unseen input patterns. Further, a cascade of two MLPs merely increases the input window size of sequences for prediction. On the other hand, SVMs provide classifiers with optimal margins and hence have the best generalization capabilities among the classifiers. As shown analytically, SVMs are optimal classifiers for the second stage because they minimize not only the empirical risk of known sequences but also the actual risk of unknown sequences. Additionally, two stages are proven to be sufficient to find an optimal classifier for RSA or PSS prediction as the second stage SVM minimizes the generalization error at the output of the first stage by solving an optimization problem at the second stage SVM.

By incorporating the state-of-the-art methods based on the position specific scoring matrices generated by PSI-BLAST and SVMs in two-stage approaches, we are able to report the best accuracies to date for RSA and PSS predictions on several benchmark amino acid sequences.

Acknowledgements

First of all, I would like to express my gratitude and sincere appreciation to my supervisor, Assoc. Prof. Jagath C. Rajapakse, who encouraged me to pursue an advanced degree and gave continuous advice and excellent comments that enabled me to complete the whole research successfully.

I wish to thank to my colleagues in the research group for interesting and helpful discussions on subject details. I would like to express my thanks to Dr. Jean-Philippe Vert for his explanations of SVM theory, Dr. Burkhard Rost, Dr. James A. Cuff, and Dr. Shandar Ahmad for providing the relevant datasets and useful suggestions, Dr. John Platt and Dr. Chih-Jen Lin for providing SVM learning software.

I am greatly indebted to Nanyang Technological University (NTU) and the School of Computer Engineering (SCE) for granting the necessary financial resources to pursue my PhD degree. I am also grateful to the staffs in the BioInformatics Research Centre (BIRC), Dr. Tan Choong Leong, Ms Wendy Tham Wei Ping, Ms Wong Lee Chin, Ms Norhana Binte Ahmad, and Mr. Toh Ban Kuan, and my friends Mr. Ho Sy Loi and Mr. Nguyen Viet Thang, who assisted me in many ways throughout the duration of the research.

Finally, without my parents and sister, I would not even have studied anything. Many thanks to them.

Publications

Journal Papers

1. M. N. Nguyen and J. C. Rajapakse. Prediction of protein relative solvent accessibility with two-stage SVM approach. *Proteins: Structure, Function, and Bioinformatics*, vol 59, pp. 30-37, 2005.
2. J. C. Rajapakse and M. N. Nguyen. Two-stage support vector machines for protein structure learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, (First revision submitted).
3. M. N. Nguyen and J. C. Rajapakse. Two-stage support vector machines for protein secondary structure prediction. *Neural, Parallel and Scientific Computations*, vol. 11, pp. 1-18, 2003.
4. J. C. Rajapakse and M. N. Nguyen. General framework for two-stage approaches to protein secondary structure prediction. *Asia Pacific Biotech News*, vol. 7, no 3, 2003.
5. J. Ma, M. N. Nguyen, and J. C. Rajapakse. Classification of HLA Sequences through Codon Usage Patterns with SVM. *Bioinformatics*, (submitted).
6. M. N. Nguyen and J. C. Rajapakse. Two-stage SVR approach for predicting accessible surface areas of amino acids. *Proteins: Structure, Function, and Bioinformatics*, (First revision submitted).

Book Chapters

1. M. N. Nguyen and J. C. Rajapakse. Multi-class support vector machines for protein secondary structure prediction. *Genome Informatics* 14, M. Gribskov, et al. (Eds.), Universal Academic Press Inc., pp. 218-227, 2003.

Conference Papers

1. J. Ma, M. N. Nguyen, G. W. L. Pang, and J. C. Rajapakse. Gene classification using codon usage patterns and SVMs. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, San Diego, USA, 2005.
2. K. S. Toh, M. N. Nguyen, and J. C. Rajapakse. LVQ approach using AA indices for protein subcellular localisation prediction. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, San Diego, USA, 2005.
3. M. N. Nguyen and J. C. Rajapakse. Two-stage multi-class support vector machines for protein relative solvent accessibility prediction. *International Conference on Neural Information Processing (ICONIP)*, Taiwan, 2005.
4. M. N. Nguyen and J. C. Rajapakse. Two-stage multi-class SVMs for protein secondary structure prediction. *Pacific Symposium on Biocomputing (PSB)*, Hawaii, USA, 2005.
5. M. N. Nguyen and J. C. Rajapakse. Two-stage support vector machines to protein relative solvent accessibility prediction. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, San Diego, USA, 2004.
6. M. N. Nguyen and J. C. Rajapakse. Prediction of protein secondary structure using bayesian method and support vector machines. *International Conference*

- on Neural Information Processing (ICONIP)*, Singapore, vol. 2, pp. 616-620, 2002.
7. M. N. Nguyen and J. C. Rajapakse. Optimal approach to sequence-to-sequence prediction: Applications in bioinformatics. *International Conference on Neural Information Processing (ICONIP)*, Singapore, vol. 5, pp. 2254-2258, 2002.
 8. M. N. Nguyen, J. C. Rajapakse, and S. L. Ho. Combining GOR techniques with support vector machines for protein secondary structure prediction. *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Singapore, vol. 3, pp. 1528-1533, 2002.
 9. S. L. Ho, J. C. Rajapakse, and M. N. Nguyen. Augmenting HMM with neural network for finding gene structure. *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Singapore, vol. 3, pp. 1522-1527, 2002.

Awards

1. Award of Japanese Society for Bioinformatics (JSBi) for the recognition as a young scientist based on the work to be presented at the Fourteenth International Conference on Genome Informatics in Yokohama, Japan, 2003.

Abbreviations and Symbols

Abbreviations

3-D	three-dimensional
ASA	accessible surface area
CV	cross-validation
DNA	deoxyribonucleic acid
HLA	human leukocyte antigen
HMM	hidden Markov model
MB	megabyte
ML	maximum likelihood
MLP	multilayer perceptron
MSVM	multi-class support vector machine
NMR	nuclear magnetic resonance
NN	neural network
NR	non-redundant
OSH	optimal separating hyperplane
PAC	probably approximately correct
PSI-BLAST	position specific iterative - basic local alignment search tool
PSS	protein secondary structure
PSSM	position-specific scoring matrix
QP	quadratic programming
RBF	radial basis function
RSA	relative solvent accessibility
SVM	support vector machine
SMO	sequential minimal optimization
SOV	segment overlap measure
SRM	structural risk minimization
VC	Vapnik-Chervononkis (dimension)

Important Symbols

$\mathbf{r}$	an amino acid sequence
Ω_R	the set of 20 amino acids
$\mathbf{a}$	a solvent accessibility sequence
Ω_A	the set of solvent accessibilities
$\mathbf{t}$	a secondary structure sequence
Ω_T	the set of secondary structures
n	the length of the sequence
$\mathcal{F}$	a set of discriminant functions
I	information function
$\mathbb{R}$	the set of reals
N	number of training exemplars
M	number of testing exemplars
$\mathbf{v}_i$	vectors representing a 21-dimensional coding from PSSM profiles
$\mathbf{r}_i$	input patterns at the first stage
q_i	desired classification
$\mathcal{K}$	kernel function
ϕ	map into feature space
$\mathbf{w}$	weight vector
b	bias
α_i	Lagrange multiplier at the first stage
γ	positive constant
Q	quadratic function
f	discriminant function of RSA
$\mathbf{d}_i$	input patterns at the second stage
β_i	Lagrange multiplier at the second stage
Γ	the set of input patterns
$\mathcal{P}$	probability distribution of RSA
$\text{err}_{\mathcal{P}}$	generalization error
L	loss function
Γ_{train}	the training set
$\text{err}_{\Gamma_{\text{train}}}$	empirical risk

Γ_{test}	the testing set
δ	small positive value
$\epsilon(\delta)$	pac bound
$\text{fat}_{\mathcal{F}}$	fat-shattering dimension
mar	margin
w	neighborhood window
ξ_i	slack variables
f^k	discriminant functions of PSS
$\mathcal{Q}$	probability distribution of PSS
z	VC dimension
m	dimension of a feature vector
O	time complexity

Chapter 1

Introduction

1.1 Background

Proteins play an important role in molecular biology. They are large macromolecules with complex structures and constitute to the bulk of living organisms: enzymes, hormones, and structural material, etc. The genetic make-up of an organism is composed of thousands of genes, most of which are *protein-coding*: they specify instructions for building proteins. To understand the life processes of an organism, it is necessary to first know the functions of the proteins produced by the organism. Proteins serve great many functions in an organism, with some proteins serving multiple functions, such as transporting small molecules (e.g., hemoglobin transports O₂ in the blood-stream), catalyzing biological functions, providing structure to collagen and skin, regulating hormones [110]. The function of a protein molecule in a given environment is determined by its structure - for instance, antibodies in the human immune system recognize antigens by having a complementary surface to that of antigen [29]. In order to fully understand the function and biological activity of proteins, one needs to have the knowledge of their structures.

Figure 1.1: The names, abbreviations and R-groups of the twenty amino acids. Atoms are colored white for Carbon, light gray for Nitrogen, dark for Oxygen, and black for Sulphur. Small atoms are Hydrogen. Bonds connecting R-groups to main-chain atoms are drawn in bold. Note that the proline side chain joins the main chain at both C^α and N atoms.

The information of life is stored in the genes with a four-letter alphabet (that is, nucleotides in DNA). Proteins are, among others, the macromolecules that perform all important tasks in organisms. Therefore, genes are the blueprints or library, and proteins are the machinery of life. Proteins are formed by joining amino acids by peptide bonds into a stretched chain. A protein sequence comprises of a translation of the four-letter DNA alphabet into a twenty-letter alphabet of native amino acids. Figure 1.1 lists the twenty amino acids and their abbreviations, and illustrates each

Figure 1.2: The amino acid sequence and its corresponding three-dimensional structure, from the light chain of the anti-prion FAB 3F4 protein complex (PDB ID 1CR9).

R-group. Proteins differ in the length (from 30 to over 30,000 amino acids), and in the arrangement of the amino acids (dubbed residues, when joined in proteins). In a solvent environment, a protein folds according to an energy cost that favors some conformations over the others. This energy cost depends on the number and type of weak interactions (hydrogen bonds, ionic bonds, Van der Waals forces, dipole and hydrophobic) formed between pairs of atoms in the protein and between atoms of the protein and atoms of the solvent [29]. Proteins are typically designed by evolution to exist in a particular solvent environment which we call their *native environment*. In its native environment, a protein quickly folds into a single and stable conformation that corresponds to the global minimum Gibbs free energy over the space of the conformation. This conformation is called the *native fold* of the protein. The tertiary structure of a protein refers to the three-dimensional (3-D) shape of its native fold

(as shown in Figure 1.2).

The protein structure prediction problem, also called the *protein folding problem*, is one of the major problems in computational biology. This problem is that of predicting the tertiary structure of a protein, given its amino acid sequence. When a newly sequenced protein is determined from genomic sequencing and translation, the most important task for characterizing is to determine its function. Knowing the tertiary structure of a protein can help determine its function. Regrettably, ascertaining the structure of a protein is a difficult and expensive task, which explains why few proteins have been categorized in this regard. Virtual protein models created on computers may provide a cost-effective solution to accurate prediction of protein structure.

1.2 Motivation

The ultimate objective of this research is to propose an approach to predict a protein structure from its amino acid sequence. In fact, the protein structure prediction problem is one of the most important unsolved problems of molecular biology and biophysics. Not only would a successful prediction algorithm be a tremendous advance in the understanding of the biochemical mechanisms of proteins, but, since such an algorithm could conceivably be used to design proteins to carry out specific functions, it would have profound, far-reaching effects on biotechnology and the treatment of disease [29].

The most direct method of determining the 3-D structure of a protein molecule is to observe it experimentally. Two technologies are often used: X-ray diffraction and nuclear magnetic resonance (NMR) [110]. Both are complex and expensive, and each

has unique drawbacks. X-ray diffraction is very time consuming because of creating a crystal of the target protein. This is not always possible, especially for proteins whose native environment is not an aqueous solution e.g., trans-membrane proteins. NMR techniques are limited to small proteins.

The difficulty of direct observation of protein structure has created an interest in protein structure prediction and led to the development of various methods [38]. The most direct prediction methods are theoretical and attempt to find a structure having the global minimum potential energy. One way of finding this minimum energy state is by molecular force-field simulation of the physical process of folding [96]. This approach involves modeling the interaction of each protein and solvent atom in a system of many thousands of atoms, with time steps measured in thousandths of picoseconds, for a period of milliseconds to seconds. Current supercomputers are not powerful enough to simulate protein folding more than a small fraction of time required. It is estimated that twenty years of steady increase in computer power remain before the smallest proteins can be folded in a molecular dynamics simulation. A second theoretical approach tries to find the global minimum of the potential energy function analytically. The complexity of the energy function and the number of parameters make this approach unfeasible [113].

As protein structure prediction using physical/chemical energy functions appears currently intractable, one instead computes statistical pseudo-energy functions from frequencies of certain amino acids known to lie in proximity to others in conformations of a representative sampling of the protein database. The bioinformatics approach is an attempt to align in parallel both the sequence and the structure. Up to now, several bioinformatics techniques have been proposed to predict the 3-D structure

[57, 154]. However, three major factors hinder the success of these methods [38].

1. *Lack of sufficient data.* The main source of data is the Protein Data Bank (PDB) [14, 154], which contains all known protein 3-D structures. As of 05 October 2004, the PDB holds 27,570 experimentally observed proteins. This is a small fraction of the possible set of known protein sequences [110].
2. *Bias in the data.* The proteins in the PDB are not representative of the set of all proteins. Most protein families do not have a single member in the PDB, and many other families are over-represented.
3. *Natural selection against highly stable structures.* Biological organisms require their proteins to be easily disassembled so that their components may be recycled once they have served their function. As such, protein structures must be stable enough to function, but unstable enough to be disassembled when necessary. Nature therefore hides from us the prime examples of well-folding proteins.

Since the protein 3-D structure prediction directly from amino acid sequences still remains an open problem, the general and reliable way is to first predict Relative Solvent Accessibility (RSA) and/or Protein Secondary Structure (PSS), which projects the very complicated 3-D structure onto one dimension [29, 110]. The successful prediction of RSA is helpful in elucidating the relationship between protein sequence and structure [23, 151]. Knowledge of the secondary structure, even from the predicted secondary structure, can improve the accuracy of 3-D structure prediction methods by 25% [57]. Another study reports a near 50% reduction of root-mean-squared-deviation errors in 3-D structure prediction when constraints derived from secondary structure

prediction are incorporated [26]. The problem focused in the course of the present study is therefore the solvent accessibility and secondary structure predictions.

Most existing solvent accessibility and secondary structure prediction methods can be considered as single-stage approaches, except those combine two neural networks [149, 84]. In our opinion, the current success rates of RSA and PSS prediction methods are not high because of the following reasons:

1. The single-stage approaches are insufficient to find complex relations (correlations) between different elements because in RSA and PSS predictions both the amino acid and solvent accessibility or secondary structure sequences are presumed to be contextual in the sense that the type of one element depends on the types of its neighbors. This could be improved by incorporating the interactions or contextual information among the elements of the solvent accessibility or secondary structure sequence. Recent analysis by information theory indicates that correlations between neighboring secondary structures are much stronger than those of neighboring amino acids [34].
2. A problem with the methods combining two neural networks for the solvent accessibility and secondary structure predictions is that the feedforward neural networks are not optimal classifiers in terms of generalization capabilities over unseen input patterns [174].

Therefore, the aim of this thesis is to provide an approach to capture the contextual information among the secondary structural elements and the relative solvent accessibilities, and minimize the generalization error over unseen input patterns. For this purpose, we propose two-stage Support Vector Machines (SVMs) for RSA and PSS predictions as a solution.

1.3 Major Contributions

The major contributions of this thesis are listed as entries below:

1. Two-Stage Binary SVMs to RSA Prediction

Bioinformatics techniques to Relative Solvent Accessibility (RSA) prediction are mostly single-stage approaches; they predict solvent accessibility of proteins by taking into account only the information available in amino acid sequences. We, in this thesis, propose two-stage binary Support Vector Machines (SVMs) by using a binary SVM predictor as a second stage following the existing SVM classifier for RSA prediction problem to improve the accuracy. The purpose of the second stage is to capture the contextual relationship of solvent accessibility elements in a neighborhood in determining the solvent accessibility at a particular site. SVM is introduced as the second stage at the output of the single-stage technique, because SVM is an optimal margin classifier, which has capacity to minimize the risk of an already prediction. By using the position specific scoring matrices generated by PSI-BLAST, the two-stage binary SVM approach achieves substantial improvements of accuracies over the highest scores published on the Manesh dataset of 215 protein structures and the RS126 dataset of 126 nonhomologous globular proteins [119, 120].

2. Two-Stage Multi-Class SVMs to PSS Prediction

The characteristics and limitations of bioinformatics techniques to Protein Secondary Structure (PSS) prediction are discussed in detail. In this thesis, we propose two-stage Multi-class Support Vector Machines (MSVMs) approach where a MSVM predictor is introduced to the output of the first stage MSVM

to capture the sequential relationship among secondary structure elements for the prediction. We argue that it is feasible in enhancing the present single-stage MSVM approach farther by augmenting with another prediction scheme at their outputs and propose to use MSVM as the second stage to improve the accuracy for PSS prediction. The results show that it is possible to obtain higher accuracies with combined hierarchical classifiers than single-stage classifiers alone, in the secondary structure prediction [116, 115, 117, 122, 143].

3. Generalization Capabilities of Two-Stage SVM Approaches

The disadvantages of the classical SVM approach for RSA prediction are discussed to indicate that the errors are introduced at the output of the first stage SVM for not taking into account the contextual information of RSA types or for not selecting the optimal values of parameters in practice. We prove that the generalization error made in the first stage is further minimized by the second stage of the two-stage SVM approach [119]. Additionally, we show that the SVM is an optimal classifier for the second stage by minimizing the generalization error of the output of single-stage based on solving the optimization problem at the second stage. Furthermore, two-stage SVMs are sufficient for RSA prediction because they minimize both the generalization error based on interactions among amino acids and the generalization error of the output of the first stage SVM by capturing the contextual information of solvent accessibilities [120].

We have proposed the two-stage MSVM approach for PSS prediction to improve the prediction accuracy of the single-stage. MSVM is more suitable for

prediction of PSS values than classical methods by combining binary SVM classifications because MSVM has the capacity to solve the optimization problem in one step. We prove that two-stage MSVMs take into account the sequential relationships among secondary structure elements, which are predicted by the first stage, to minimize the generalization error further in the classification [121, 142]. MSVM is an optimal classifier for the second stage in terms of the margin of separation; it attempts to minimize not only the empirical risk of known sequences but also the actual risk for unknown sequences. Two stages of MSVM are sufficient to find an optimal classifier for PSS prediction as the second stage MSVM attempts to minimize the generalization error of the first stage by solving the optimization problem at the second stage. Further, we demonstrate that two-stage MSVMs have better performance than the PHD method using two Multi-Layer Perceptrons (MLPs) in cascade for PSS prediction [118, 142].

The various contributions achieved so far in the present research have been accepted to three journals, one book chapter, nine international conferences, and are presently under review to three journal (see Author's Publications); each publication emphasized a different aspect of the thesis.

1.4 Organization of the Thesis

The thesis begins with Chapter 2 where a comprehensive introduction to PSS prediction techniques is given. We discuss the characteristics and limitations of the approaches to secondary structure prediction. The most well-known algorithms for

predicting secondary structure are reviewed in detail.

In Chapter 3, we give an introduction to RSA prediction problem and propose a two-stage binary SVM approach for RSA prediction by using a second SVM to enhance the output of the classical SVM. We take an in-depth study and extend the theory of SVM to prove that the two-stage approach using SVM to optimize the predictions from the first stage are better than the single-stage classifier alone. We demonstrate the performance of two-stage binary SVMs by using the Manesh dataset of 215 protein structures and the RS126 dataset of 126 nonhomologous globular proteins.

Chapter 4 investigates the use of two-stage MSVMs to PSS prediction by adding MSVM as the second stage following the existing MSVM classifier. We prove that the generalization error made in the first stage is further minimized by the second stage of the two-stage MSVM approach. We show that the prediction accuracy of two-stage MSVMs outperforms the result of two-stage MLPs for PSS prediction. The difference between the VC (Vapnik Chervonenkis) dimension and the margin analysis based on the theory of Bartlett and Shawe-Taylor is discussed in detail. The performance of two-stage MSVMs is evaluated on the CB396 dataset of 396 nonhomologous proteins, RS126, and EVAsec datasets.

Finally, chapter 5 concludes the thesis with a discussion of promising avenues of research to continue development. We demonstrate that it is possible to improve the PSS prediction accuracy by the second MSVM classifier at the output of the existing secondary structure prediction schemes.

Chapter 2

Bioinformatics Approaches to Protein Secondary Structure Prediction: A Review

Proteins are large biological molecules with complex structures and constitute to the bulk of living organisms: enzymes, hormones, and structural material. The function of a protein molecule in a given environment is determined by its 3-dimensional (3-D) structure [29]. Since the prediction of protein 3-D structure directly from amino acid sequences is difficult, bioinformatics continuously improve methods by predicting simplified aspects of the structure. Protein Secondary Structure (PSS) is an intermediate step to simplify the protein structure prediction problem by projecting the very complicated 3-D structure onto one dimension. The goal of the secondary structure prediction is to classify a pattern of residues in amino acid sequences as protein secondary structure elements: an α -helix (H), β -strand(E), or coil (C , the remaining type). Figure 2.1 illustrates the pathway for folding a sequence of amino acids into 3-D structure.

In order to understand the process of PSS predictions and explain our algorithm in the next chapters, we define the following:

Figure 2.1: Illustration of folding a linear chain of amino acids into a three-dimensional (3-D) protein structure.

Definition 2.0.1. An amino acid sequence is a tuple $\mathbf{r} = (r_1, r_2, \dots, r_n)$ where $r_i \in \Omega_R$ denotes the i th element of the amino acid sequence and Ω_R denotes the set of 20 amino acids, $\Omega_R = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$; n is the length of the sequence.

Definition 2.0.2. A secondary structure sequence is a tuple $\mathbf{t} = (t_1, t_2, \dots, t_n)$ where $t_i \in \Omega_T$ denotes the i th element of the secondary structure sequence and $\Omega_T = \{H, E, C\}$; n is the length of the sequence; H , E , and C denote the secondary structure elements, α -helix, β -strand, and coil, respectively.

Definition 2.0.3. PSS prediction problem is to predict the secondary structure sequence, $\mathbf{t} \in \Omega_T^n$; from the given amino acid sequence, $\mathbf{r} \in \Omega_R^n$. That is, to find the

mapping F such that

$$F : \Omega_R^n \longrightarrow \Omega_T^n,$$

where n is the length of the sequence.

The previously used methods of predicting PSS are discussed below as first generation, second generation, third generation, and recent advanced methods. This classification of approaches for PSS prediction is based on the paper of Rost [148].

2.1 First Generation: single amino acid statistics

The suggestion of Pauling and Corey for the existence of α -helices and β -strands opened the field of secondary structure prediction [130, 131]. The first PSS prediction methods analyze propensities for amino acids in a sequence to form α -helices or β -strands or coils [16, 17, 161, 46, 128, 56, 145, 28, 146, 64]. Typical examples of these approaches are Chou-Fasman and GOR I methods [28, 64].

2.1.1 Chou-Fasman Method

This method is based on calculating the frequency of each of the twenty amino acids in an α -helix, β -strand, and coil. The frequency of amino acid $r_i \in \Omega_R$ in structure $t_i \in \Omega_T$ is then divided by the frequency of all residues in structure t_i to produce probability values for each type of secondary structure.

To predict secondary structure of a new amino acid sequence, the Chou-Fasman method scans the sequence to find contiguous regions of residues that have a high probability of forming one type of the secondary structure. If four of six amino acids

have a high probability > 1.03 of being an α -helix, they are assigned to α -helices. β -strands are predicted when three of five amino acids with a probability > 1.00 of being β -strand. If both α -helical and β -strand regions are predicted, the higher probability of prediction is used. A turn (coil) is predicted when the average probability value for each of the four amino acids being in a turn is greater than the probabilities for an α -helix and β -strand in the region and the probability value in the turn tetrapeptide is greater than 7.5×10^{-5} . The regions are then extended along the sequence by using these rules.

2.1.2 GOR I Method

The GOR I method is based on the principle of maximization of the mutual information between input and output, which suggests that the sequence of secondary structure is obtained by maximizing the information transferred from the input sequence of residues. The information function, I , measures the mutual information present between the amino acid sequence and the sequence of secondary structure elements, which can be written as

$$I(t_i, r_i) = \ln \left(\frac{P(t_i|r_i)}{P(t_i)} \right), \quad (2.1.1)$$

where $P(t_i)$ is the prior probability of having a secondary structural type t_i and $P(t_i|r_i)$ is the conditional probability for the presence of a secondary structural type t_i at the location of a residue r_i .

In theory, the secondary structural type of any residue should depend on the whole sequence. However, in practice, it is only possible to use a window-based approach, considering the local neighborhood sequence around the residue of interest. A window

of 17 residues is used to compute information difference for each of three secondary structural types of a residue i as follows:

$$I(\Delta t_i, \mathbf{r}_i) = I(t_i, \mathbf{r}_i) - I(\bar{t}_i, \mathbf{r}_i), \quad (2.1.2)$$

where $\bar{t}_i$ is the complement of a secondary structural type t_i , and $\mathbf{r}_i = (r_{i-8}, \dots, r_{i+8})$ represents residues within the window around a residue r_i .

In GOR I [64], the following approximation is used for the computation of the information difference:

$$\begin{aligned} I(\Delta t_i, \mathbf{r}_i) &\approx \sum_{j=-8}^8 I(\Delta t_i, r_{i+j}), \\ &\approx \sum_{j=-8}^8 \ln \left(\frac{f(t_i, r_{i+j})}{f(\bar{t}_i, r_{i+j})} \right) + 17 \ln \left(\frac{f(\bar{t}_i)}{f(t_i)} \right), \end{aligned} \quad (2.1.3)$$

where $f(t_i, r_{i+j})$ is the frequency of a secondary structural type t_i given the type of a residue r_{i+j} . This approximation renders the information that a residue carries about another residue's secondary structure, that does not depend on the other residues, i.e., the independent influence that each residue in the window has on the secondary structural type of the central residue. The secondary structural type of each residue is then specified from the highest information difference.

The assumption made by these early methods is that each amino acid in the sequence influences the secondary structure independently, which is a major drawback. Another assumption made is that the scoring rule does not account for misses or gaps. As a consequence, the Chou-Fasman and GOR I methods are only about 50%-60% accurate in predicting secondary structure.

2.2 Second Generation: segment statistics

The second generation methods improve the disadvantages of the first generation by taking into account local interactions between amino acids in the scoring function; this is achieved by establishing a neighborhood window around a particular amino acid of interest.

2.2.1 Statistical Methods

In statistical methods, the likelihood of each amino acid being one of the three types of secondary structures is estimated [86, 70, 63, 15, 65, 103, 175, 85, 48, 99]. Typical examples are GOR III and GOR IV methods [70, 63], which treat the amino acid sequence and the sequence of secondary structure as two messages related by a translation process.

Like the GOR I method, Gibrat *et al.* [70] use the following approximation to compute the information difference in GOR III:

$$I(\Delta t_i, \mathbf{r}_i) \approx \sum_{j=-8}^8 \ln \left(\frac{f(t_i, r_{i+j}, r_i)}{f(\bar{t}_i, r_{i+j}, r_i)} \right) + 17 \ln \left(\frac{f(\bar{t}_i, r_i)}{f(t_i, r_i)} \right) \quad (2.2.1)$$

where $f(t_i, r_{i+j}, r_i)$ is the frequency that t_i , r_{i+j} and r_i jointly occur. In this approximation, the assumption is that the pairwise combination of the central residue and a residue in the window influences the secondary structural type of the central residue.

In GOR IV [63], another approximation is used to take into account all possible pairs formed by each residue in the window. The assumption used in this approximation is that the pairwise combinations of residues in the window influence the

secondary structural type of the central residue:

$$I(\Delta t_i, \mathbf{r}_i) \approx \frac{2}{17} \sum_{j=-8}^7 \sum_{k=j+1}^8 \ln \left(\frac{f(t_i, r_{i+j}, r_{i+k})}{f(\bar{t}_i, r_{i+j}, r_{i+k})} \right) - \frac{15}{17} \sum_{j=-8}^{j=8} \ln \left(\frac{f(t_i, r_{i+j})}{f(\bar{t}_i, r_{i+j})} \right) + \ln \left(\frac{f(\bar{t}_i)}{f(t_i)} \right). \quad (2.2.2)$$

Using the amount of information, the probability that the residue r_i presents the secondary structure t_i is computed as

$$P(t_i | \mathbf{r}_i) = \frac{1}{1 + \frac{f(\bar{t}_i)}{f(t_i)} e^{-I(\Delta t_i, \mathbf{r}_i)}}. \quad (2.2.3)$$

The above conditional probability can be used to determine the secondary structure at a given location of the amino acid sequence. The secondary structure prediction of each residue is given by

$$t_i = \arg \max_{t \in \Omega_T} P(t | \mathbf{r}_i). \quad (2.2.4)$$

2.2.2 Neural Networks

Neural networks, namely feedforward networks, use residues in a local neighborhood, as inputs, to predict the secondary structure at a particular location of an amino acid sequence by finding an arbitrary non-linear mapping [18, 140, 76, 93, 166, 182, 105, 23]. The neural network approaches are theoretically able to extract more information and complex features from sequences than the statistical methods [140]. The architecture of a typical neural network model used for PSS prediction is illustrated in Figure 2.2.

The input to neural networks is based on a conventional orthogonal encoding of the single sequence. The first neural network, called *sequence-to-structure* network, attempts to predict secondary structure at a particular element of the input sequence by using a symmetric window of size 13-17 around that element. Let the input to

Figure 2.2: Architecture of a neural network model for protein secondary structure prediction.

the network to predict the secondary structure at the location i of the sequence, $\mathbf{v}_i = (v_{i-w_1}, v_{i-w_1+1}, \dots, v_i, \dots, v_{i+w_1})$ where v_i denotes the input profile at i and the size of the input window is $2w_1 + 1$. The network has three output neurons, each neuron representing one secondary structure type of protein. The output activation of the neuron representing the secondary structure $k \in \Omega_T$, o_k^1 , can be written as

$$o_k^1(\mathbf{v}_i) = g^k \left(\sum_j w_{kj}^1 g_j(\mathbf{w}_j^1 \mathbf{v}_i) \right), \quad (2.2.5)$$

where $\mathbf{w}_j^1$ denotes the weight matrix connecting input to the hidden layer neuron j and w_{kj}^1 is the weight connecting the j th hidden neuron to the k th output neuron. The activation functions $g^k(\cdot)$ and $g_j(\cdot)$ are sigmoidal functions. The back-propagation training adjusts the values of weights used to modify the signals from the input layer

to the hidden layer and from the hidden layer to the output layer [140, 149]. The goal is to have the weights balance the input signals so that the model output correctly recognizes the known secondary structure of the central amino acid in a window of a protein sequence.

Since the sequence-to-structure neural network is trained to classify mutually independent segments of residues in terms of the state of a single residue, there is no explicit representation of the fact that consecutive patterns are correlated, like for an α -helix consisting at least three consecutive patterns. The correlation can be taken into account in part, by using a second stage, *structure-to-structure* network. In the work of Qian *et al.* [140] and Rost *et al.* [149], a window of 17 secondary structure predictions is used as input to fully connected structure-to-structure network.

For the second stage, the input at location i , $\mathbf{o}_i = (o_k^1(\mathbf{v}_{i-w_2}), \dots, o_k^1(\mathbf{v}_i), \dots, o_k^1(\mathbf{v}_{i+w_2}) : k \in \Omega_T)$. The network has three outputs each corresponding to one secondary structural type. The output activation o_k^2 of the $k \in \Omega_T$ is given by

$$o_k^2(\mathbf{o}_i) = f^k \left(\sum_j w_{kj}^2 f_j(\mathbf{w}_j^2 \mathbf{o}_i) \right) \quad (2.2.6)$$

where the activation functions $f^k(\cdot)$ and $f_j(\cdot)$ are sigmoidal functions; w_{kj}^2 is the weight connecting j th hidden node to k th output neuron and $\mathbf{w}_j^2$ denotes the weight matrix connecting j the hidden neuron to the input neurons.

The predicted secondary structure is chosen as the largest of the three outputs [140, 149, 150]. The second network significantly improves the accuracy and makes the prediction more realistic in terms of predicted mean lengths of secondary structure segments [149].

2.2.3 Nearest-Neighbor Methods

Like neural networks, nearest-neighbor methods also use a neighborhood window surrounding a residue. The basic idea of nearest-neighbor approaches is that the secondary structural type of an amino acid in the query sequence is predicted by identifying sequences of known structures that are similar to the query sequence [94, 158, 179, 165, 61]. Therefore, a choice of scoring table for evaluation of segment similarity is a key element in any nearest-neighbor prediction algorithm.

Yi and Lander [179] developed a scoring system by combining a sequence similarity matrix with the local structural environment scoring method of Bowie *et al.* [19]. The method proposed by Bowie *et al.* [19] assigns every residue of a protein with known 3-D structure to an environment class based on the local structural features of the residue position, such as the solvent accessibility, polarity and secondary structure. In the method of Yi and Lander [179], 15 environmental classes, e_j , $j = 1, \dots, 15$, involving three types of secondary structures and five types of accessibility/polarity categories, c_j , $j = 1, \dots, 5$, are firstly defined. Then, 110 amino acid sequences from the Brookhaven Protein Data Bank are converted into 3-D structure profiles (a 20-dimensional amino acid alphabet into an 15-dimensional environment class alphabet). Finally, segments from a test protein are matched against all the other environmental sequences in the training set by the scoring table, which assigns a score for the alignment of each local environment class with each amino acid type. The score for matching a residue, r_i , with a local structural environment, e_j , is given by:

$$\text{Score}(r_i, e_j) = \log_{10} \left(\frac{P(r_i|e_j)}{P(r_i)} \right), \quad (2.2.7)$$

where $P(r_i|e_j)$ is the probability of finding residue r_i in environment e_j , and $P(r_i)$ is the probability of finding residue r_i in any environment [179].

The test residue is considered as the center of w consecutive amino acid residues of a sequence window. n nearest neighbors are selected by comparing the test window sequence with all w residue windows from the training set using the similarity score measure (see Eq. (2.2.7)) averaged over all window residues. The secondary structural type of a nearest neighbor is taken as the type of the center residue in the corresponding window. The secondary structure of a test residue is therefore predicted as the type of the majority of its nearest neighbors

$$t_i = \arg \max_{t \in \Omega_T} f_i^t, \quad (2.2.8)$$

where f_i^H , f_i^E , and f_i^C are the numbers of nearest neighbors of the residue r_i with α -helix, β -strand, and coil types, respectively.

The method of Yi and Lander [179] achieved a prediction accuracy of 68.0% with the window size of $w = 19$ amino acid residues and the number of nearest neighbors of $n = 50$ on the dataset of 110 proteins.

The assumption made by the second generation methods is that the interactions among amino acid residues are local. The main drawback of the above approaches is that they are unable to incorporate useful information from evolutionary profiles. Further, the methods are unable to take into account the interactions between distant amino acids in the sequence and the dependencies of sliding the window. The second generation methods achieve accuracies about 60%-70% in predicting secondary structure.

2.3 Third Generation: evolutionary information

2.3.1 Multiple Sequence Alignments

The third generation methods build on the second generation by using multiple sequence alignments instead of single sequences [68, 149, 101, 150, 165, 80, 107, 59, 67, 92, 144, 147, 62, 156, 157, 10, 78, 118]. It is reasonable to believe that there is a correlation between hidden patterns of substitutions of amino acids in similar proteins and similar structures, as the same mutations due to evolution are responsible for both. A typical example of the third generation is PHD (Profile network from HeiDelberg) method [149]. Rost and Sander have gained significant success by using sequence profiles from multiple alignments. A BLAST (Basic Local Alignment Search Tool) search of the input sequence is conducted to identify similar but not closely identical sequences, and a multiple alignment of the sequences is transformed into a sequence profile. A profile thus consists of the frequencies of the 20 amino acids in each column of the multiple alignment. In the work of Rost and Sander [149], the frequencies are used as inputs to the neural network instead of single sequences, i.e., the usual representation of an amino acid by orthogonal coding is replaced by 20 real numbers. Note that the sequence profiles differ in the number of sequences in the family and in the similarity of the aligned sequences, to the input sequence.

2.3.2 PSI-BLAST Profiles

Recently, many computational techniques solving the PSS prediction problem use the position specific scoring matrices generated by PSI-BLAST (Position Specific Iterative - Basic Local Alignment Search Tool) as inputs [84, 127, 36, 109, 137, 139,

176, 138, 89, 117, 75, 121]. Jones pioneered using PSI-BLAST profiles to achieve impressive improvement of accuracies. Firstly, the values of raw matrices of PSI-BLAST [3] are obtained from NR (Non-Redundant) or SWISS-PROT databases. The low-complexity regions, transmembrane regions, and coil-coil segments are then filtered from these databases by PFILT program [84]. Finally, the E-value cut-off, iterations, a similarity matrix, and gap penalties are used for searching the non-redundant sequence database to generate position specific scoring matrix (PSSM) profiles. The elements of PSSM profiles in the range $[-7, 7]$ are then scaled to the $[0, 1]$ range by using the standard logistic function [84]

$$f(x) = \frac{1}{1 + e^{-x}}$$

or the following function [89]

$$f(x) = \begin{cases} 0.0 & \text{if } x \leq -5 \\ 0.5 + 0.1x & \text{if } -5 < x < 5 \\ 1.0 & \text{if } x \geq 5 \end{cases},$$

where x is the value from PSSM profiles. The input vector represents a 21-dimensional coding of each residue, where 20 elements take the values from PSSM profiles ranging from $[0, 1]$ and the last element is used as the padding space to indicate the end of the sequence. The PSI-BLAST profiles contain more useful information than the multiple sequence alignments: the probability of each residue at a specific position is properly computed; the amount of significant information of each sequence is weighted; more distant homologues are found [84]. Therefore, the accuracies of methods using PSI-BLAST profiles are significantly higher than the results obtained by using multiple sequence alignments [84, 36, 117, 75].

Program	Method	Accuracy $Q_3(\%)$
NNSSP	Nearest Neighbor	72.7
DSC	Linear Discrimination	71.1
PREDATOR	Hydrogen Bonding Propensities (Neural Network)	70.3
MULPRED	Single Sequence Method Combination	67.2
Zpred	Conservation Number Weighted Prediction	66.7
PHD	Neural Network	70.8
Jpred	Consensus	74.8

Table 2.1: Six programs used in the Jpred method and three-state per-residue accuracy (Q_3) [149] of each method on the RS126 dataset of 126 nonhomologous globular proteins.

In the next section, we discuss in detail recent enhancements of the above approaches to PSS prediction.

2.4 Recent Developments

2.4.1 Combining Predictions

The basic probability theory suggests that if a novel prediction method provides new relevant information, then it should be combined with existing methods to obtain improved predictions [81]. This idea is the basis of combining different classifiers, parallely, into a single superior predictor for PSS prediction [37, 164, 73, 35, 91, 74]. One prediction program using a hybrid of these methods is Jpred [37] which is accessible on the Internet [30]. This program sends a protein sequence through six different

prediction techniques and returns a prediction based on a majority voting scheme calculated from these results. The programs are chosen so as to apply different methods while being the best in their category (see Table 2.1). More complex approaches for combining different methods based on neural networks, linear discrimination [127], and multi-category SVM [74] have been presented. However, these methods depend on the performances of individual single models and also do not overcome the limitation of single-stage methods.

Recently, Liu *et al.* have applied several graphical chain models to solve the combination problem and showed that they are consistently more effective than the traditional window-based methods [100].

2.4.2 Bayesian Method

Bayesian segmentation of protein secondary structure is proposed by Schmidler *et al.* [162] to provide a framework that takes into account non-local interactions between amino acids. The Bayesian approach incorporates the prior and likelihood of secondary structural segments in the relationships of amino acid sequences and the sequences of secondary structure. Further, it takes into account interactions between distant elements in the sequence, which do not depend on a window of neighborhood.

The Bayesian technique attempts to segment the secondary structure sequence: say the segmented sequence, $\mathbf{T} = (T_1, T_2, \dots, T_m)$, where T_i indicates the i subsequence in which the elements belong to the same secondary structure type. When all the segments are concatenated, the secondary structure sequence representing the given amino acid sequence is obtained. The sequence of sites, $\mathbf{d} = (d_1, d_2, \dots, d_m)$,

characterizes the segmentation of the amino acid sequence with the m positions denoting the end of each individual secondary structural segment such that $d_{i-1} + 1$ and d_i denotes the beginning and the end of the i th segment.

Let us denote the subsequence of amino acids between sites i and j , by $\mathbf{r}_{[i:j]}$. In order to find the optimal segmentation of the secondary structures, let us define the concepts of forward (α) and backward (β) probabilities:

- $\alpha_i(d, t) = P(d_i = d, T_i = t, \mathbf{r}_{[1:d]} | \theta)$ to be the probability of the observation sequence of d residues, $\mathbf{r}_{[1:d]}$, with the secondary structural type of this segment t , given the model $\theta = (m, \mathbf{d}, \mathbf{T})$.
- $\beta_i(d, t) = P(\mathbf{r}_{[d+1:n]} | d_i = d, T_i = t, \theta)$ to be the probability of the observation sequence, $\mathbf{r}_{[d+1:n]}$, given the model θ , the segment i ending at d with the secondary structural type of this segment t .

The residue at a particular site in a segment is assumed to be independent of the residues at sites of the other segments. The algorithm attempts to find a model $\theta = (m, \mathbf{d}, \mathbf{T})$ that segments the given sequence of residues optimally. By using the chain rule of probability:

$$\begin{aligned} \alpha_i(d, t) &= \sum_{j=0}^{d-1} \sum_{v \in \Omega_T} \alpha_{i-1}(j, v) P(\mathbf{r}_{[j+1:d]} | d_{i-1} = j, d_i = d, T_i = t) \\ &\quad \times P(d_i = d | d_{i-1} = j, T_i = t, \theta) P(T_i = t | T_{i-1} = v, \theta). \end{aligned} \quad (2.4.1)$$

A similar estimation holds for the probability $\beta_i(d, t)$:

$$\begin{aligned} \beta_i(d, t) &= \sum_{j=d+1}^n \sum_{v \in \Omega_T} \beta_{i+1}(j, v) P(\mathbf{r}_{[d+1:j]} | d_i = d, d_{i+1} = j, T_{i+1} = v) \\ &\quad \times P(d_{i+1} = j | d_i = d, T_{i+1} = v, \theta) P(T_{i+1} = v | T_i = t, \theta). \end{aligned} \quad (2.4.2)$$

Then, the probability of the amino acid sequence $\mathbf{r}$ with a secondary structural type t given at the residue r_i is given by

$$\begin{aligned}
 P(t_i = t, \mathbf{r}|\theta) &= \sum_{j=0}^{i-1} \sum_{d=i}^n \sum_{v \in \Omega_T} \sum_{i'=1}^{j+1} P(\mathbf{r}, d_{i'-1} = j, T_{i'-1} = v, d_{i'} = d, T_{i'} = t|\theta) \\
 &= \sum_{j=0}^{i-1} \sum_{d=i}^n \sum_{v \in \Omega_T} \sum_{i'=1}^{j+1} \alpha_{i'-1}(j, v) \beta_{i'}(d, t) P(\mathbf{r}_{[j+1:d]} | d_{i'-1} = j, d_{i'} = d, T_{i'} = t) \\
 &\quad \times P(d_{i'} = d | d_{i'-1} = j, T_{i'} = t, \theta) P(T_{i'} = t | T_{i'-1} = v, \theta).
 \end{aligned} \tag{2.4.3}$$

Here, the probability of the subsequence $\mathbf{r}_{[j+1:d]}$ under the assumption that a secondary structural type t begins at $j + 1$ and ends at d can be achieved by using the N-cap and C-cap model [7, 162].

The probability that the secondary structural type of the residue r_i is t , given the amino acid sequence $\mathbf{r}$ and the model θ can be computed from Eq. (2.4.3) and by using Bayesian formula:

$$P(t_i = t | \mathbf{r}, \theta) = \frac{P(t_i = t, \mathbf{r}|\theta)}{P(\mathbf{r}|\theta)}, \tag{2.4.4}$$

where $P(\mathbf{r}|\theta)$ is considered as a normalizing constant. The secondary structure of each residue is given by

$$t_i = \arg \max_{t \in \Omega_T} P(t_i = t | \mathbf{r}, \theta). \tag{2.4.5}$$

An assumption made by the model is that segments are independent, so that the probability of the sequence can be expressed in the product of the segment probabilities. The technique achieved an accuracy of 68.8% on a non-redundant set of 452 globular protein structures, which is an impressive result for single sequence methods. Nevertheless, this technique is unable to incorporate the useful information from multiple sequence alignments or PSI-BLAST profiles into the prediction scheme.

2.4.3 Recurrent Neural Networks

Baldi *et al.* [10] proposed an approach for predicting secondary structure of a protein from its amino acid sequence by including information contained within long range interaction. A recurrent neural network is trained to extract features from multiple sequence alignments. The major idea of the method aims at incorporating the interactions among secondary structure elements by using an ensemble of bidirectional recurrent neural networks instead of adding the second level structure-to-structure network in other methods. Recently, Pollastri *et al.* [137] extended the method to extract features from PSI-BLAST profiles instead of multiple sequence alignments. By using the position specific scoring matrices generated by PSI-BLAST, the method yields sustained performances of 76%-78% correct prediction. However, recurrent neural networks are not optimal classifiers in terms of generalization capabilities over unseen input patterns.

2.4.4 Support Vector Machines

Support Vector Machines (SVMs) have been applied very broadly within the field of computational biology [33, 123], including PSS prediction [78, 176, 89, 75], which have strong foundations in statistical learning theory as shown by Vapnik [173, 174]. A typical SVM model using single sequences for PSS prediction is illustrated in Figure 2.3. Let v_i be the orthogonal binary vector representing 21-dimensional coding of the residue r_i and the input pattern of the residue at site i be $\mathbf{v}_i = (v_{i-w}, v_{i-w+1}, \dots, v_{i+w})$ where $2w + 1$ is the size of the neighborhood window around the element i . Three binary SVM classifiers, $\mathbf{C}^k$, $k \in \Omega_T$, are constructed, each predicting if the secondary structure at the local site i belongs to the secondary structure type k or not. The

Figure 2.3: Architecture of binary support vector machine (SVM) approach to protein secondary structure prediction.

input vectors are transformed to higher dimensions via kernel functions, $\mathcal{K}^k$, for each classifier and linearly combined to derive the outputs via parameter vectors, $\mathbf{w}^k$.

During training, knowing the primary and secondary sequences, the pairs $(\mathbf{v}_i, q_i^k)$ are derived from the sequences as training samples, where $q_i^k \in \{+1, -1\}$ corresponds to the desired output for the classifier $\mathbf{C}^k$ at site i . The output of the feature space is combined using parameters, $\mathbf{w}^k$, determined by α_i^k , $i = 1, 2, \dots, n; k \in \Omega_T$, that are found by maximizing the following quadratic function Q^k for the classifier $\mathbf{C}^k$:

$$Q^k = \sum_{j=1}^n \alpha_j^k - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n q_i^k q_j^k \alpha_i^k \alpha_j^k \mathcal{K}^k(\mathbf{v}_i, \mathbf{v}_j) \quad (2.4.6)$$

subject to $0 \leq \alpha_j^k \leq \gamma_k$ and $\sum_{j=1}^n \alpha_j^k q_j^k = 0$. The corresponding parameter vector is then given by $\mathbf{w}^k = \sum_{j=1}^n q_j^k \alpha_j^k \phi^k(\mathbf{v}_j)$. The summations of the maximizing function

Q^k run over all training patterns; $\mathcal{K}^k(\mathbf{v}_i, \mathbf{v}_j) = \phi^k(\mathbf{v}_i)\phi^k(\mathbf{v}_j)$ denotes the kernel function; q_i^k encodes the secondary structure such that a binary value +1 if the residue r_i corresponding to the secondary structure k or -1, otherwise. γ_k is positive constant used to decide the trade-off between training error and the margin. The above optimization procedure produces an optimal margin of separation of the two classes.

Once the parameters α_j^k are obtained from the above optimization, the resulting discriminant function for classifier $\mathbf{C}^k$, d^k , is given by

$$d^k(\mathbf{v}_i) = \sum_{j=1}^n q_j^k \alpha_j^k \mathcal{K}^k(\mathbf{v}_j, \mathbf{v}_i) + b_k, \quad (2.4.7)$$

where the bias b_k is chosen so that $q_j^k d^k(\mathbf{v}_j) = 1$ for any j with $0 < \alpha_j^k < \gamma_k$. The estimate of the secondary structural type, t_i , of the residue r_i is determined by the highest value of three discriminant function values:

$$t_i = \arg \max_{k \in \Omega_T} d^k(\mathbf{v}_i). \quad (2.4.8)$$

Like neural networks, SVMs have the capacity to incorporate the evolutionary information from multiple sequence alignments or PSI-BLAST profiles into the prediction scheme [78, 118, 176, 89, 75]. SVMs have shown better PSS prediction than other type of neural networks [78, 89] because, in bioinformatics, the sequences encountered for the prediction are usually uncharacterized before and SVMs generalize well in this situation. One of the drawbacks in these approaches is that the methods do not take into account the sequential relationship among the protein secondary structure elements. Additionally, SVM methods only construct a multi-class classifier by combining several binary classifiers.

2.4.5 Combining Evolutionary Information

Combining evolutionary information in optimization techniques with other methods has provided effective and efficient approaches for predicting PSS [58]. Recently, Meiler and Baker proposed a new method using the information of 3-D structure and PSI-BLAST profiles as inputs to a neural network [108]. By combining 3-D structure information, the new approach reported 4%-5% higher accuracies than the methods using only the local information of amino acid sequences. A drawback of this approach is its dependence on the structural models generated by ROSETTA [108] and, therefore, limited to a small number of proteins.

2.4.6 Hidden Markov Models

Hidden Markov models (HMMs) have been successfully applied to secondary structure prediction [20, 88]. HMMs are generative models, which assume that the data are generated by a particular model. Like the Bayesian method, these approaches predict PSS by calculating the joint distribution of observations $\mathbf{r}$ and states $\mathbf{t}$, $P(\mathbf{r}, \mathbf{t})$, and make predictions by using Bayes rules to compute $P(\mathbf{t}|\mathbf{r})$.

Zheng [183] introduced simple Markov models to solve the PSS prediction problem. The method focuses on inclusion of short range correlations of residues and conformation states in the models, and ignores the duration effect of length distribution because the length of protein conformation segments varies in a narrow range. Conformation-independent and conformation-dependent amino acid coarse-graining schemes are designed for the models by means of proper mutual information. Based on non-comparable datasets, the author estimated prediction accuracy to be around

Figure 2.4: Comparison of three-state per-residue accuracy (Q_3) [149] of different predictors of protein secondary structure on the RS126 dataset of 126 nonhomologous globular proteins. The notation (ext) indicates that the corresponding method uses position specific scoring matrices generated by PSI-BLAST instead of multiple sequence alignments.

69% using single sequences. However, the method is unable to combine the evolutionary information from multiple sequence alignments or PSI-BLAST profiles.

Based on the Bayesian method of Schmidler *et al.* [162], Crooks and Brenner [34] developed a relatively simple HMM that embodies three key approximations: protein sequences are statistically homogeneous; direct secondary structure to primary structure interactions are local along the chain; amino acids at neighboring sites are independent. Crooks and Brenner extended their HMM method to handle the evolutionary information of multiple sequence alignments and reported a prediction

accuracy of 72.2% on the CB513 dataset of 513 nonhomologous proteins [45].

Lin *et al.* [98] proposed the secondary structure prediction method YASPIN that uses a single neural network to predict the secondary structure elements in a 7-state local structure scheme and then optimizes the output using a HMM, which results in providing more information for the prediction.

2.5 Comparison

Figure 2.4 shows the performance of different secondary structure methods with the RS126 dataset of 126 nonhomologous globular proteins based on single sequences, multiple sequence alignments, and PSI-BLAST profiles. The methods are GOR I (1st generation) [64], GOR III (2nd) [70], GOR IV (2nd) [63], Zpred (2nd) [184], PHD (3rd) [149], NNSSP (3rd) [156], PREDATOR (3rd) [60], DSC (3rd) [92], Riis and Krogh (3rd) [144], BRNN (3rd) [10], Jpred (3rd) [35], SVMfreq (3rd) [78], and SVMpsi (3rd ext) [89]. The results of Zpred, NNSSP, PREDATOR, DSC and Jpred methods on the RS126 were obtained from Cuff and Barton [35], and the results of the refined neural network proposed by Riis and Krogh, SVMfreq, SVMpsi, BRNN, and PHD methods were published in their original publications [144, 78, 89, 10, 150]. We implemented and tested GOR I, GOR III, and GORIV techniques on the RS126 dataset. The best algorithm was found on the RS126 dataset to be the SVM method of Kim and Park with the PSI-BLAST profiles [89], which achieved 76.1% of Q_3 accuracy.

Table 2.2 lists performances and web servers of recent approaches for secondary structure prediction. In these methods: PHD, PSIPRED, Jnet, and PHDpsi predict PSS of a residue based on neural networks; Jpred and PROF combine the outputs of individual methods in parallel into a single superior predictor; SSpro uses ensembles

Method and Web address	Generation	$Q_3(\%)$
Rost and Sander, 1993 (PHD) [149, 150] http://cubic.bioc.columbia.edu/predictprotein	3 rd	70.8-71.6
Cuff and Barton, 1999 (Jpred) [35] http://www.compbio.dundee.ac.uk/~www-jpred/submit.html	3 rd	72.9-74.8
Jones, 1999 (PSIPRED) [84] http://bioinf.cs.ucl.ac.uk/psipred/psiform.html	3 rd (ext)	76.5-78.3
Ouali and King, 1999 (PROF) [127] http://www.aber.ac.uk/~phiwww/prof/	3 rd (ext)	76.7
Cuff and Barton, 2000 (Jnet) [36] http://www.compbio.dundee.ac.uk/~www-jpred/jnet/	3 rd (ext)	76.4
Przybylski and Rost, 2002 (PHDpsi) [139] http://cubic.bioc.columbia.edu/predictprotein	3 rd (ext)	75.0
Pollastri <i>et al.</i> , 2002 (SSpro) [137] http://www.igb.uci.edu/tools/scratch/	3 rd (ext)	76.0-78.0
Porollo <i>et al.</i> , 2003 (SABLE) [138] http://sable.cchmc.org/	3 rd (ext)	77.5-78.8
Meiler <i>et al.</i> , 2003 (JUFO3D) [108] http://www.jens-meiler.de/jufo3D.html	3 rd (ext)	76.0-82.0
Guo <i>et al.</i> , 2004 (PMSVM) [75] http://www.bioinfo.tsinghua.edu.cn/pmsvm/pmsvm.html	3 rd (ext)	74.0-75.2
Lin <i>et al.</i> , 2004 (YASPIN) [98] http://ibivu.cs.vu.nl/programs/yaspinwww/	3 rd (ext)	77.0

Table 2.2: Performances and web servers of recent methods for protein secondary structure prediction. The notation (ext) indicates that the corresponding method uses PSI-BLAST profiles instead of multiple sequence alignments.

of methods: take the N newest experimental structures added to PDB, send the sequences to all PSS prediction servers, collect the results, and accumulate a continuous evaluation of prediction accuracy every week. Figure 2.5 shows results of different PSS prediction servers from EVA for 1XJH:A protein.

2.6 Summary

Method	Generation II single sequence	Generation III multiple alignments	Generation III (ext) PSI-BLAST profiles
HMMs	Karplus 1998 [88] Bystroff 2000 [20] Zheng 2004 [183]	Crooks 2004 [34]	
NNs	Bohr 1988 [18] Qian 1988 [140] Holley 1989 [76] Kneller 1990 [93] Stolorz 1992 [166] Chandonia 1995 [23]	PHD 1993 [149] Livingstone 1994 [101] PREDATOR 1995 [60] Riis 1996 [144] BRNN 1999 [10]	PROF 1999 [127] PSIPRED 1999 [84] Jnet 2000 [36] PHDpsi 2002 [139] SSpro 2002 [137] JUFO3D 2003 [108]
SVMs		SVMpreq 2001 [78]	Ward 2003 [176] SVMpsi 2003 [89] PMSVM 2004 [75]

Table 2.3: Summary of various generations and timeline of techniques among NNs, HMMs, and SVMs.

Many computational methods have been proposed in the literature to predict secondary structures of proteins from their amino-acid sequences. However, the current

success rates of existing approaches to PSS prediction are far from most biologist's expectations; further improvement of the accuracy is necessary [148]. Most bioinformatics techniques for secondary structure prediction are single-stage approaches that suffer from the limited size of the local neighborhood; the sequential relationships among the secondary structures of residues are not taken into account in the prediction scheme. Crooks and Brenner use the analysis of information theory to demonstrate that correlations between neighboring amino acids are essentially uninformative and only one-fourth of the total information needed to determine the secondary structure is available from local inter-sequence correlations [34]. Since the information of local correlations among amino acid residues is insufficient to determine secondary structure accurately, there are two ways to overcome this lack. One is to utilize evolutionary information and use a second stage to capture the contextual information among the secondary structural elements. The alternative approach is to handle non-local interactions explicitly by methods such as Bayesian and Hidden Markov models. Table 2.3 lists the various generations and timeline of techniques among NNs, HMMs, and SVMs.

Recently, secondary structure prediction methods have achieved substantial improvements of accuracies by incorporating evolutionary information from multiple sequence alignments and position specific scoring matrices generated by PSI-BLAST. Improvements of a few percentage points of PSS prediction are significant, especially in the context of genome sequencing and structural proteomics projects [137]. Accuracies of PSS prediction methods will be further improved from the result of database growth and improvements of algorithms and search techniques.

A further improvement in PSS prediction accuracy could be obtained by training

on larger datasets. Most researchers have reported 1%-2% increases in accuracy by using larger training sets [149, 150, 78, 89, 75]. This is because the overall performance of any method based on evolutionary profiles suffers when very remote or no homologues are included [150]. Therefore, improvements of accuracies are observed when larger sequences and more homologous profiles are used in training.

Meiler and Baker [108] improve the performance of existing secondary structure prediction schemes by using the information of 3-D structure and PSI-BLAST profiles as inputs to a neural network. As an avenue for future research, the combination of the information of 3-D structure and PSI-BLAST profiles with other methods may achieve higher prediction accuracies.

Protein secondary structure prediction from amino acid sequences has been now recognized as one of the biggest challenges of computational biology and a popular first step toward predicting 3-D structure [148]. Secondary structure predictions are successfully used as one component of 3-D structure prediction methods [126, 102]. Eyrich *et al.* [54, 55] propose 3-D protein structure prediction method based on minimizing the energy of arranging predicted rigid secondary structure segments. Chen *et al.* [26] develop an approach to reduce the complexity of molecular dynamics simulations based on useful information of secondary structure predictions. Samudrala *et al.* [160, 159] suggest combining secondary structure prediction methods with a particular lattice simulation to enumerate all possible folds.

Chapter 3

Two-Stage Binary SVMs to Protein Relative Solvent Accessibility Prediction

Information on Relative Solvent Accessibility (RSA) of amino acid residues in proteins provides valuable clues to protein structure prediction. A two-stage approach with Support Vector Machines (SVMs) is proposed, where an SVM predictor is introduced to the output of the single-stage SVM approach to take into account the contextual relationships among solvent accessibilities. By using the position specific scoring matrices, generated by PSI-BLAST, the two-stage SVM approach achieves substantial improvements of accuracies up to 7.6% and 4% over the highest scores published on the Manesh dataset of 215 protein structures and the RS126 dataset of 126 nonhomologous globular proteins, respectively. A web server for protein relative solvent accessibility prediction using two-stage SVM method has been developed and is available at [177].

3.1 Introduction

The knowledge of protein structures is valuable for understanding mechanisms of diseases of living organisms and for facilitating discovery of new drugs. Protein structure can be experimentally determined by NMR spectroscopy and X-Ray crystallography techniques or by molecular dynamics simulations. However, the experimental approaches are marred by long experimental time, prone to difficulties, and expensive and, therefore, limited to small proteins [110]. Bioinformatics approaches are recently being sought to predict Relative Solvent Accessibility (RSA) to help elucidate the relationship between the protein's sequence and structure, and, thereby, predict the three-dimensional (3-D) structure of proteins [23, 151]. The studies of solvent accessibility have shown that the hydrophobic free energies of proteins are directly related to the accessible surface area of both polar and non-polar groups of amino acid in proteins [124]. Chan and Dill [22] have discovered the burial of core residues is a strong driving force in protein folding. Furthermore, the RSA prediction gives insight into the organization of 3-D structure: the position of protein hydration sites playing an important part in protein function could be predicted based on solvent accessibility [52]; the information of solvent accessibility has improved the prediction of protein subcellular location as the distribution of solvent accessibilities is correlated with its subcellular environments [5]. The main goal of RSA prediction is to classify a pattern of residues in amino acid sequences to a pattern of RSA types: buried (B) and exposed (E) residues.

Many different techniques have been proposed for RSA prediction, which broadly fall into the following categories: (1) Bayesian, (2) neural networks, and (3) information theoretical approaches. The Bayesian methods provide a framework to take into

account local interactions among amino acid residues, by extracting the information from single sequences or multiple sequence alignments to obtain posterior probabilities for RSA prediction [168]. Neural networks use residues in a local neighborhood, as inputs, to predict RSA of a residue at a particular location by finding an arbitrary non-linear mapping [129, 97, 136, 2]. The information theoretical approaches use mutual information between the sequences of amino acids and of solvent accessibility values, derived from a single amino acid residues or pairs of residues in a neighborhood for RSA prediction [111]. Recently, variants of these approaches with increased prediction accuracies have been proposed: Gianese *et al.* predicted RSA of a residue based on probability profiles computed on amino acid residues in the neighborhood [69]; Adamczak *et al.* proposed using neural networks based regression to find continuous approximations to RSA values [1].

Despite the existence of many approaches, the current success rates of existing techniques to RSA prediction are insufficient; further improvement of the accuracy is necessary. Most existing techniques for RSA prediction are single-stage approaches in the sense that the solvent accessibility type is directly predicted from amino acid sequences or profiles derived thereof. They suffer from the limited size of the local neighborhood used in the prediction; the sequential relationships among the solvent accessibilities of residues are not taken into account. In this chapter, we propose a two-stage approach to RSA prediction by using a second predictor, Support Vector Machine (SVM) classifier, introduced at the end of a single-stage RSA prediction scheme. The aim of the second stage is to take into account the influence on the RSA of a residue by those of its neighbors.

SVMs have been earlier shown to perform well in multiple areas of biological analysis [33, 123] including RSA prediction [180, 89], which have strong foundations in statistical learning theory as shown by Vapnik [173, 174]. SVMs implement a classifier that is capable of minimizing structural risk. Furthermore, SVMs offer several associated computational advantages such as the lack of local minima and a solution completely encompassed by the set of support vectors. In addition, SVM with its generalization capability is known to behave well compared to other statistical or machine learning methods for many biological problems [33]. Also, the generalization capability of SVMs is well suited for the prediction of RSA of novel amino acid sequences. All previous SVM approaches to RSA prediction are single-stage approaches.

We demonstrate the performance of two-stage SVM approach by using the Manesh [111] and the RS126 [151] datasets, and report prediction accuracies upto 90.4% and 90.2% on the Manesh and the RS126 datasets respectively.

3.2 Two-Stage SVM Approach

In the two-stage SVM approach, we use two SVMs in cascade to predict relative solvent accessibilities of residues in amino acid sequences.

Let us denote the amino acid sequence by $\mathbf{r} = (r_1, r_2, \dots, r_n)$ where $r_i \in \Omega_R$ and Ω_R is the set of 20 amino acid residues and the corresponding solvent accessibility sequence by $\mathbf{a} = (a_1, a_2, \dots, a_n)$ where $a_i \in \Omega_A = \{B, E\}$; n is the length of the sequence. The prediction of the sequence of RSA types, $\mathbf{a}$, from an amino acid sequence, $\mathbf{r}$, is the problem of finding the optimal mapping from the space of Ω_R^n to the space of Ω_A^n .

Firstly, the values of raw matrices of PSI-BLAST [4], used as inputs to the first stage SVM, are obtained from NR database that is available at [39]. The low-complexity regions, transmembrane regions, and coil-coil segments are then filtered from the NR database by PFILT program [84]. Finally, the E-value threshold of 0.001, three iterations, BLOSUM62 matrix, a gap open penalty of 11, a gap extended penalty of 1 are used for searching the non-redundant sequence database to generate position specific scoring matrix (PSSM) profiles. These arguments are consistent with those used in other methods [84, 89]. Let $\mathbf{v}_i$ be a vector representing a 21-dimensional coding of the residue r_i where 20 elements take the values from PSSM profiles ranging from $[0, 1]$ and the last element is used as the padding space to indicate the end of the sequence [89]; the padding element is set to 1 to indicate the end of the sequence or 0, otherwise. Let the input pattern to SVM at site i be $\mathbf{r}_i = (\mathbf{v}_{i-h_1^1}, \mathbf{v}_{i-h_1^1+1}, \dots, \mathbf{v}_i, \dots, \mathbf{v}_{i+h_2^1})$ where $\mathbf{v}_i$ denotes the center element, h_1^1 and h_2^1 denote the width of input segment on the two sides; $w_1 = h_1^1 + h_2^1 + 1$ is the size of the neighborhood window around the center element i .

3.2.1 First Stage

The first stage for RSA prediction consists of a binary SVM classifier, B/E , that maps the input patterns of class B to -1 and the patterns of class E to +1. The SVM transforms the input vectors to a higher dimension via a kernel function, $\mathcal{K}^1$, and linearly combines to derive the outputs with a weight vector, $\mathbf{w}_1$. The function $\mathcal{K}^1$ and vector $\mathbf{w}_1$ are determined to minimize the error in the prediction during the training phase. Let $\{(\mathbf{r}_j, q_j) : j = 1, 2, \dots, N\}$ denote the set of all training exemplars where q_j denotes the desired classification, B or E , for the input pattern $\mathbf{r}_j$ such that

the output of SVM is -1 if the correct RSA type is B or +1 if the type is E . When N is the number of training exemplars, the weight vector, $\mathbf{w}_1$, is determined by scalars α_j , $j = 1, 2, \dots, N$, that are found by maximizing the following quadratic function Q_1 :

$$Q_1 = \sum_{j=1}^N \alpha_j - \frac{1}{2} \sum_{j=1}^N \sum_{i=1}^N \alpha_j \alpha_i q_j q_i \mathcal{K}^1(\mathbf{r}_j, \mathbf{r}_i), \quad (3.2.1)$$

subject to $0 \leq \alpha_j \leq \gamma_1$ and $\sum_{j=1}^N \alpha_j q_j = 0$. $\mathcal{K}^1(\mathbf{r}_i, \mathbf{r}_j) = \phi^1(\mathbf{r}_i) \phi^1(\mathbf{r}_j)$ denotes the kernel function and ϕ^1 represents the mapping function to higher dimension; γ_1 is a positive constant used to decide the trade-off between the training error and the margin of the classifier [173, 174].

The weight vector is then given by $\mathbf{w}_1 = \sum_{j=1}^N q_j \alpha_j \phi^1(\mathbf{r}_j)$. Once the parameters α_j are obtained from the above algorithm, the resulting discriminant function, say f_1 , is given by

$$\begin{aligned} f_1(\mathbf{r}_i) &= \sum_{j=1}^N q_j \alpha_j \mathcal{K}^1(\mathbf{r}_j, \mathbf{r}_i) + b_1, \\ &= \mathbf{w}_1 \phi^1(\mathbf{r}_i) + b_1, \end{aligned} \quad (3.2.2)$$

where the bias b_1 is chosen so that $q_j f_1(\mathbf{r}_j) = 1$ for any j with $0 < \alpha_j < \gamma_1$.

In the single-stage SVM method, the solvent accessibility type a_i corresponding to the residue at site i , r_i , is determined by

$$a_i = \begin{cases} E & \text{if } f_1(\mathbf{r}_i) \geq 0 \\ B & \text{otherwise} \end{cases}. \quad (3.2.3)$$

3.2.2 Second Stage

The single-stage approach takes only the interactions among amino acid residues in the neighborhood into the prediction scheme. The RSA type of a residue is also

Figure 3.1: The two-stage binary SVMs for relative solvent accessibility prediction from amino acid sequences.

influenced by those in its neighborhood. A second SVM predictor is used in the two-stage approach to predict the RSA type of a residue by using the predictions from the first-stage, capturing the sequential relationships among the RSA values in the neighborhood. The architecture of the two-stage SVM prediction approach is illustrated in Figure 3.1.

The second SVM classifier improves the accuracy of the single-stage RSA prediction schemes by taking into account the sequential relationships among the RSA values of residues into the prediction. The second-stage SVM processes the estimated

RSA values at the first stage and minimizes the generalization error by incorporating the contextual information among RSA values. Rost and Sander [151] proposed a simple method to incorporate the sequential relationships of the estimated RSA types, in which an averaging filter is employed to take the average of neighboring outputs of the first neural network at each amino acid residue; then, the solvent accessibility is predicted as the type with the largest average. Two-stage SVM approaches were previously proposed for protein secondary structure prediction [118, 75].

The second stage SVM processes the output of the discriminant functions of the first stage to enhance the prediction. At the site i , the input to the second SVM is given by a vector $\mathbf{d}_i = (d_{i-h_1^2}, d_{i-h_1^2+1}, \dots, d_i, \dots, d_{i+h_2^2})$ where h_1^2 and h_2^2 are the length of the neighborhood on two sides, and $d_i = 1/(1 + e^{-f_1(\mathbf{r}_i)})$. The logistic sigmoid function restricts the input units of the second stage to the (0,1) interval. The SVM converts the input patterns, usually linearly inseparable, to a higher dimensional space by using the mapping ϕ^2 with a kernel function $\mathcal{K}^2(\mathbf{d}_i, \mathbf{d}_j) = \phi^2(\mathbf{d}_i)\phi^2(\mathbf{d}_j)$.

As in the first stage, the hidden outputs in the higher dimensional space are linearly combined by a weight vector, $\mathbf{w}_2$, to obtain the prediction output. Let the training set of exemplars for the second stage SVM be $\{(\mathbf{d}_j, q_j), j = 1, \dots, N\}$. The kernel function $\mathcal{K}^2$ and vector $\mathbf{w}_2$ are obtained by solving the following convex QP (quadratic programming) problem, over all the patterns seen in the training phase:

$$\begin{aligned} \max_{\beta} \quad & \sum_{j=1}^N \beta_j - \frac{1}{2} \mathbf{w}_2^T \mathbf{w}_2, \\ \text{such that } & 0 \leq \beta_j \leq \gamma_2 \text{ and } \sum_{j=1}^N \beta_j q_j = 0, \\ \text{where } \mathbf{w}_2 = & \sum_{j=1}^N q_j \beta_j \phi^2(\mathbf{d}_j). \end{aligned} \tag{3.2.4}$$

The discriminant function, f_2 , at the second stage is given by

$$\begin{aligned} f_2(\mathbf{d}_i) &= \sum_{j=1}^N q_j \beta_j \mathcal{K}^2(\mathbf{d}_j, \mathbf{d}_i) + b_2, \\ &= \mathbf{w}_2 \phi^2(\mathbf{d}_i) + b_2, \end{aligned} \quad (3.2.5)$$

where the bias b_2 is chosen so that $q_j f_2(\mathbf{d}_j) = 1$ for any j with $0 < \beta_j < \gamma_2$. The solvent accessibility type a_i corresponding to the residue r_i is given by

$$a_i = \begin{cases} E & \text{if } f_2(\mathbf{d}_i) \geq 0 \\ B & \text{otherwise} \end{cases}. \quad (3.2.6)$$

3.3 Generalization in Two-Stage SVMs

In this section, we prove that the second stage of two-stage SVMs minimizes the generalization error made by the first stage SVM for RSA prediction.

In the classical SVM approach, the function, f_1 , discriminates the type of RSA, based on the features or interactions among the residues in the input pattern. With optimal parameters, the SVM attempts to minimize the generalization error in the prediction. Let Γ^1 denote the set of input patterns seen by the SVM during both the training and testing phases and suppose that the training and testing patterns are drawn independently and identically according to a probability distribution $\mathcal{P}_1$.

Definition 3.3.1. [174] The *generalization error* made by a discriminant function f_1 is the error that random patterns, $(\mathbf{r}, q) \in \Gamma^1 \times \{-1, +1\}$ generated according to a

probability distribution $\mathcal{P}_1$, are misclassified by f_1 :

$$\text{err}_{\mathcal{P}_1}(f_1) = \int L(\text{sign}(f_1(\mathbf{r})), q) d\mathcal{P}_1(\mathbf{r}, q), \quad (3.3.1)$$

where the loss function $L(\cdot)$:

$$L(\text{sign}(f_1(\mathbf{r})), q) = \begin{cases} 0 & \text{if } \text{sign}(f_1(\mathbf{r})) = q \\ 1 & \text{if } \text{sign}(f_1(\mathbf{r})) \neq q \end{cases},$$

where q is the desired output for the input pattern $\mathbf{r}$.

Since the generalization error, $\text{err}_{\mathcal{P}_1}(f_1)$, is a measure of the quality of the chosen function f_1 , it is also referred as a *risk functional* [174].

Definition 3.3.2. The *fractional error* of a discriminant function f_1 on a set $S = \{(\mathbf{r}_i, q_i) \in \Gamma^1 \times \{-1, +1\} : i = 1, \dots, N\}$ is a fraction of patterns misclassified by f_1 in S :

$$\text{err}_S(f_1) = \frac{1}{N} \sum_{i=1}^N L(\text{sign}(f_1(\mathbf{r}_i)), q_i) \quad (3.3.2)$$

Definition 3.3.3. The *empirical risk* of a discriminant function f_1 , $\text{err}_{\Gamma_{\text{train}}^1}(f_1)$, is the fractional error of f_1 on the training set Γ_{train}^1 .

If the input pattern $\mathbf{r}$ corresponds to a site i , then $\mathbf{r} = \mathbf{r}_i$ and $\mathbf{r}_i = (\mathbf{v}_{i-h_1^1}, \mathbf{v}_{i-h_1^1+1}, \dots, \mathbf{v}_i, \dots, \mathbf{v}_{i+h_2^1})$. The first stage SVM only considers the sequences of amino acid residues and their interaction in the neighborhood into the prediction schemes. SVM has the capacity of finding the discriminant function f_1 to minimize the generalization error $\text{err}_{\mathcal{P}_1}(f_1)$ based on interactions among the residues in the input pattern $\mathbf{r}$. However, the RSA type of an amino acid residue is also influenced by RSA types of the residue in its neighborhood, for example, it accounts for the fact that the buried or exposed type consists of at least two consecutive residues. Hence, the errors are

introduced at the output of the classical SVM approach for not taking into account the contextual information of RSA types or for not selecting the optimal values of parameters in practice.

The aim of two-stage SVMs is to find an optimal function, f_2 , taking into account the sequential relationships among solvent accessibility elements, that are predicted by the first stage, to minimize the generalization error further in the classification. Let Γ^2 denote the set of input patterns seen by the second stage SVM in both training and testing phases. If the training and testing patterns of the second stage, $(\mathbf{d}, q) \in \Gamma^2 \times \{-1, +1\}$, are drawn independently and identically according to a probability distribution $\mathcal{P}_2$, the generalization error, $\text{err}_{\mathcal{P}_2}(f_2)$, is given by

$$\text{err}_{\mathcal{P}_2}(f_2) = \int L(\text{sign}(f_2(\mathbf{d})), q) d\mathcal{P}_2(\mathbf{d}, q). \quad (3.3.3)$$

If the input pattern $\mathbf{d}$ corresponds to a site i , then $\mathbf{d} = \mathbf{d}_i$ and $\mathbf{d}_i =$

$$\left(\left(1 + e^{-f_1(\mathbf{r}_{i-h_1^2})}\right)^{-1}, \left(1 + e^{-f_1(\mathbf{r}_{i-h_1^2+1})}\right)^{-1}, \dots, \left(1 + e^{-f_1(\mathbf{r}_i)}\right)^{-1}, \dots, \left(1 + e^{-f_1(\mathbf{r}_{i+h_2^2})}\right)^{-1} \right).$$

That is, the second stage takes into account the influences of the RSA values of residues in the neighborhood into the prediction. The application of the logistic sigmoid function to the outputs of the first stage has the advantage of constraining the input units of the second stage to the (0,1) interval that is similar to the range of the input units of the first stage. We consider the case where $h_1^2 = h_2^2 = 0$ and $f_2^*(\mathbf{d}) = \ln \frac{\mathbf{d}}{1-\mathbf{d}}$. It follows that $\mathbf{d} = \mathbf{d}_i = (1 + e^{-f_1(\mathbf{r}_i)})^{-1}$ and $f_2^*(\mathbf{d}) = \ln(e^{-f_1(\mathbf{r}_i)})^{-1} = f_1(\mathbf{r})$. The generalization error, $\text{err}_{\mathcal{P}_2}(f_2^*)$, can now be written as

$$\text{err}_{\mathcal{P}_2}(f_2^*) = \int_{L=0} L(\text{sign}(f_2^*(\mathbf{d})), q) d\mathcal{P}_2(\mathbf{d}, q) + \int_{L=1} L(\text{sign}(f_2^*(\mathbf{d})), q) d\mathcal{P}_2(\mathbf{d}, q)$$

$$\begin{aligned}
&= \int_{L=1} L(\text{sign}(f_2^*(\mathbf{d})), q) d\mathcal{P}_2(\mathbf{d}, q), \\
&= \mathcal{P}_2 \{ (\mathbf{d}, q) : \text{sign}(f_2^*(\mathbf{d})) \neq q; (\mathbf{d}, q) \in \Gamma^2 \times \{-1, +1\} \}, \\
&= \mathcal{P}_1 \{ (\mathbf{r}, q) : \text{sign}(f_1(\mathbf{r})) \neq q; (\mathbf{r}, q) \in \Gamma^1 \times \{-1, +1\} \}, \\
&= \int_{L=1} L(\text{sign}(f_1(\mathbf{r})), q) d\mathcal{P}_1(\mathbf{r}, q), \\
&= \int L(\text{sign}(f_1(\mathbf{r})), q) d\mathcal{P}_1(\mathbf{r}, q), \\
&= \text{err}_{\mathcal{P}_1}(f_1).
\end{aligned}$$

Note that the integral $\int_Z d\mathcal{P}_2(z)$ is written as $\mathcal{P}_2\{Z\}$ to simplify notations in next sections, where $z = (\mathbf{d}, q)$ and $Z \in \Gamma^2 \times \{-1, +1\}$.

Since there exists at least a function f_2^* such that $\text{err}_{\mathcal{P}_2}(f_2^*) = \text{err}_{\mathcal{P}_1}(f_1)$, the optimal function f_2 providing the smallest $\text{err}_{\mathcal{P}_2}(f_2)$ ensures $\text{err}_{\mathcal{P}_2}(f_2) \leq \text{err}_{\mathcal{P}_2}(f_2^*) = \text{err}_{\mathcal{P}_1}(f_1)$. However, finding the global minimum of generalization error $\text{err}_{\mathcal{P}_2}(f_2)$ is not a trivial problem because the form of the probability distribution $\mathcal{P}_2$ is unknown. We can instead consider the *probably approximately correct* (pac) bound $\epsilon(\delta)$ [173, 174, 33], of the generalization error satisfying

$$\mathcal{P}_2 \{ \Gamma_{\text{train}}^2 : \exists f_2 \in \mathcal{F} \text{ such that } \text{err}_{\mathcal{P}_2}(f_2) > \epsilon(\delta) \} < \delta, \quad (3.3.4)$$

where parameter δ is a small positive value specified in the training. This is equivalent to asserting that with probability greater than $1 - \delta$ over the training set $\Gamma_{\text{train}}^2 \subset \Gamma^2$, the generalization error of f_2 is bounded by the pac bound:

$$\text{err}_{\mathcal{P}_2}(f_2) \leq \epsilon(\delta).$$

In the following proofs, we assume that the training set $\Gamma_{\text{train}}^2 \subset \Gamma^2$ and the testing set $\Gamma_{\text{test}}^2 \subset \Gamma^2$ for the second stage contained N and M patterns, respectively. In order

Figure 3.2: A set of lines shattering three points.

to simplify the notation, we imply Γ_{train}^2 or Γ_{test}^2 as a set of input pattern $\mathbf{d}$ and its desired output q when Γ_{train}^2 or Γ_{test}^2 is used following the probability distribution $\mathcal{P}_2$.

Definition 3.3.4. [11] Let $\mathcal{F} = \{f_2 : \mathbf{d} \rightarrow \mathbf{w}_2 \phi^2(\mathbf{d}) + b_2; \|\mathbf{w}_2\| = 1; \mathbf{d} \in \Gamma^2\}$ be a set of discriminant functions and $S = \{(\mathbf{d}_i, q_i) \in \Gamma^2 \times \{-1, +1\} : i = 1, \dots, N\}$. We say that the set of points S is η_2 -shattered by $\mathcal{F}$ if there exist real numbers $\zeta_i, i = 1, \dots, N$, such that, for every binary classification, $q_i \in \{-1, +1\}$ on set S , there exists $f_2 \in \mathcal{F}$ satisfying $f_2(\mathbf{d}_i) \begin{cases} \geq \zeta_i + \eta_2 & \text{if } q_i = +1 \\ \leq \zeta_i - \eta_2 & \text{if } q_i = -1 \end{cases}$.

Definition 3.3.5. [11] The *fat-shattering dimension* $\text{fat}_{\mathcal{F}}(\eta_2)$ at scale η_2 is the size of the largest η_2 -shattered subset of Γ^2 .

We can view the real numbers ζ_i as individual thresholds of the classification

for each point while η_2 -shattering implies that every classification with margin η_2 , corresponding to the chosen thresholds is realized on the set S . For example, if $\mathcal{F}$ is the set of oriented straight line in 2-D space, so that for a given line, all points on one side are assigned the class $+1$, and all points in the other side, class -1 . The orientation shown in Figure 3.2 by an arrow specifies the slide of the points that are assigned the value $+1$. If we take three points in the 2-D space, we can orient the lines to classify the points correctly, no matter how the points are labels (see Figure 3.2).

Lemma 3.3.1. *Let $\mathcal{F} = \{f_2 : \mathbf{d} \rightarrow \mathbf{w}_2 \phi^2(\mathbf{d}) + b_2; \|\mathbf{w}_2\| = 1; \mathbf{d} \in \Gamma^2\}$. If $S = \{\mathbf{d}_i \in \Gamma^2 : i = 1, \dots, N\}$ is the largest η_2 -shattered set by $\mathcal{F}$, then every subset $S_0 \subseteq S$ satisfies*

$$\left\| \sum_{\mathbf{d}_i \in S_0} \phi^2(\mathbf{d}_i) - \sum_{\mathbf{d}_i \in S - S_0} \phi^2(\mathbf{d}_i) \right\| \geq N\eta_2. \quad (3.3.5)$$

Proof. Suppose that $S = \{\mathbf{d}_i \in \Gamma^2 : i = 1, \dots, N\}$ is the largest η_2 -shattered set by $\mathcal{F}$, witnessed by $\zeta_i \in \mathbb{R}$, $i = 1, \dots, N$. It follows that the fat-shattering dimension $\text{fat}_{\mathcal{F}}(\eta_2) = |S| = N$; for every binary classification, $q_i \in \{-1, +1\}$ on set S , there exists $\mathbf{w}_2$ and b_2 with $\|\mathbf{w}_2\| = 1$ satisfying $q_i (\mathbf{w}_2 \phi^2(\mathbf{d}_i) + b_2 - \zeta_i) \geq \eta_2$, $i = 1, \dots, N$. For a set $S_0 \subseteq S$, there are two cases.

In the first case, if

$$\sum_{\mathbf{d}_i \in S_0} (\zeta_i - b_2) \geq \sum_{\mathbf{d}_i \in S - S_0} (\zeta_i - b_2) \quad (3.3.6)$$

then we fix $q_i = 1$ if and only if $\mathbf{d}_i \in S_0$. It follows that

$$\begin{aligned} \mathbf{w}_2 \phi^2(\mathbf{d}_i) + b_2 - \zeta_i &\geq \eta_2 \quad \text{if } \mathbf{d}_i \in S_0, \\ \mathbf{w}_2 \phi^2(\mathbf{d}_i) + b_2 - \zeta_i &\leq -\eta_2 \quad \text{if } \mathbf{d}_i \in S - S_0. \end{aligned}$$

Hence,

$$\sum_{\mathbf{d}_i \in S_0} \mathbf{w}_2 \phi^2(\mathbf{d}_i) \geq \sum_{\mathbf{d}_i \in S_0} (\zeta_i - b_2) + |S_0| \eta_2, \quad (3.3.7)$$

$$\sum_{\mathbf{d}_i \in S - S_0} \mathbf{w}_2 \phi^2(\mathbf{d}_i) \leq \sum_{\mathbf{d}_i \in S - S_0} (\zeta_i - b_2) - |S - S_0| \eta_2. \quad (3.3.8)$$

The inequality (3.3.7), together with inequality (3.3.8), shows that

$$\begin{aligned} \mathbf{w}_2 \left(\sum_{\mathbf{d}_i \in S_0} \phi^2(\mathbf{d}_i) - \sum_{\mathbf{d}_i \in S - S_0} \phi^2(\mathbf{d}_i) \right) &\geq \sum_{\mathbf{d}_i \in S_0} (\zeta_i - b_2) - \sum_{\mathbf{d}_i \in S - S_0} (\zeta_i - b_2) + |S| \eta_2, \\ &\text{from the inequality (3.3.6) and } |S| = N, \\ &\geq N \eta_2. \end{aligned}$$

From $\|\mathbf{w}_2\| = 1$ and the Cauchy-Schwarz inequality, the following inequality is given

$$\begin{aligned} \left\| \sum_{\mathbf{d}_i \in S_0} \phi^2(\mathbf{d}_i) - \sum_{\mathbf{d}_i \in S - S_0} \phi^2(\mathbf{d}_i) \right\| &= \|\mathbf{w}_2\| \left\| \sum_{\mathbf{d}_i \in S_0} \phi^2(\mathbf{d}_i) - \sum_{\mathbf{d}_i \in S - S_0} \phi^2(\mathbf{d}_i) \right\|, \\ &\geq \left| \mathbf{w}_2 \left(\sum_{\mathbf{d}_i \in S_0} \phi^2(\mathbf{d}_i) - \sum_{\mathbf{d}_i \in S - S_0} \phi^2(\mathbf{d}_i) \right) \right|, \\ &\geq N \eta_2. \end{aligned}$$

In the second case, if the inequality (3.3.6) is not satisfied, that is

$$\sum_{\mathbf{d}_i \in S_0} (\zeta_i - b_2) < \sum_{\mathbf{d}_i \in S - S_0} (\zeta_i - b_2) \quad (3.3.9)$$

then we fix $q_i = 1$ if and only if $\mathbf{d}_i \in S - S_0$ and use an identical argument. $\square$

Lemma 3.3.2. For $S = \{\mathbf{d}_i \in \Gamma^2 : i = 1, \dots, N\}$ with $\phi^2(\mathbf{d}_i) \in \mathbb{R}^m$ and $\|\phi^2(\mathbf{d}_i)\| \leq R$, some subset $S_0 \subseteq S$ satisfies

$$\left\| \sum_{\mathbf{d}_i \in S_0} \phi^2(\mathbf{d}_i) - \sum_{\mathbf{d}_i \in S - S_0} \phi^2(\mathbf{d}_i) \right\| \leq R \sqrt{N}. \quad (3.3.10)$$

Proof. We choose S_0 randomly, by defining $S_0 = \{\mathbf{d}_i \in S : q_i = 1\}$, where $q_i \in \{-1, +1\}$, $i = 1, \dots, N$ are independent and uniform random variables. Then

$$\begin{aligned}
E \left\| \sum_{\mathbf{d}_i \in S_0} \phi^2(\mathbf{d}_i) - \sum_{\mathbf{d}_i \in S - S_0} \phi^2(\mathbf{d}_i) \right\|^2 &= E \left\| \sum_{i=1}^N q_i \phi^2(\mathbf{d}_i) \right\|^2, \\
&= \sum_{i=1}^N E(q_i \phi^2(\mathbf{d}_i)) \left(\sum_{j=1}^N q_j \phi^2(\mathbf{d}_j) \right), \\
&= \sum_{i=1}^N \left(\sum_{j \neq i} E(q_i \phi^2(\mathbf{d}_i)) (q_j \phi^2(\mathbf{d}_j)) + E \|q_i \phi^2(\mathbf{d}_i)\|^2 \right), \\
&= \sum_{i=1}^N E \|q_i \phi^2(\mathbf{d}_i)\|^2, \\
&= \sum_{i=1}^N \|\phi^2(\mathbf{d}_i)\|^2, \\
&\leq NR^2.
\end{aligned}$$

where the last equality follows from the fact that the q_i 's have zero mean and are independent.

Since there exists at least a set S_0 with the value less than or equal to the expectation, there must be a set S_0 for which

$$\left\| \sum_{\mathbf{d}_i \in S_0} \phi^2(\mathbf{d}_i) - \sum_{\mathbf{d}_i \in S - S_0} \phi^2(\mathbf{d}_i) \right\| \leq R\sqrt{N}.$$

□

Theorem 3.3.3. *Let $\mathcal{F}$ be restricted to points in a ball of m dimensions of radius R about the origin, that is $\phi^2(\mathbf{d}) \in \mathbb{R}^m$, $\|\phi^2(\mathbf{d})\| \leq R$, $\mathbf{d} \in \Gamma^2$, and $S = \{\mathbf{d}_i \in \Gamma^2 : i = 1, \dots, N\}$ denote the largest η_2 -shattered set by $\mathcal{F}$. Then, the fat-shattering dimension*

$$\text{fat}_{\mathcal{F}}(\eta_2) \leq \left(\frac{R}{\eta_2} \right)^2.$$

Proof. Theorem 3.3.3 extends the result of Bartlett and Shawe-Taylor [11] when the bias b_2 and the mapping ϕ^2 are used in the discriminant function $f_2 \in \mathcal{F}$. The proof of this theorem follows from two above lemmas, 3.3.1 and 3.3.2. Lemma 3.3.1 states that the sum of the vectors contained in any subset of the largest shattered set is far from the sum of the vectors contained in the remainder of that set. Lemma 3.3.2 states that if the norms of the input vectors are small, these sum cannot be too far apart. Using inequalities (3.3.5) and (3.3.10), the fat-shattering dimension

$$\text{fat}_{\mathcal{F}}(\eta_2) = |S| = N \leq \left(\frac{R}{\eta_2}\right)^2.$$

□

Definition 3.3.6. Let $\mathcal{F}$ be a set of discriminant functions on Γ^2 . A finite set of functions $\mathcal{B}$ is said to be a η_2 -cover of $\mathcal{F}$ on a set $S = \{\mathbf{d}_i \in \Gamma^2 : i = 1, \dots, N\}$ if, for all $f_2 \in \mathcal{F}$, there exists $g_2 \in \mathcal{B}$ such that $\max_{1 \leq i \leq N} (|f_2(\mathbf{d}_i) - g_2(\mathbf{d}_i)|) < \eta_2$.

Definition 3.3.7. Let $\mathcal{N}(\mathcal{F}, S, \eta_2)$ be the size of the smallest η_2 -cover of $\mathcal{F}$ on a set $S \subset \Gamma^2$. The η_2 -covering numbers of $\mathcal{F}$ is given by

$$\mathcal{N}(\mathcal{F}, N, \eta_2) = \max_{S \subset \Gamma^2, |S|=N} \mathcal{N}(\mathcal{F}, S, \eta_2).$$

Theorem 3.3.4. [33] Let $\mathcal{F}$ be a set of discriminant functions from $\Gamma^2 \rightarrow [a, b]$ and $\mathcal{P}_2$ denote a distribution over Γ^2 . If $0 < \eta_2 < 1$ and $h = \text{fat}_{\mathcal{F}}(\frac{\eta_2}{4})$, then

$$\mathcal{N}(\mathcal{F}, N, \eta_2) \leq 2 \left(\frac{4N(b-a)^2}{(\eta_2)^2} \right)^{h \log(2eN(b-a)/(h\eta_2))}.$$

This theorem shows how the covering numbers of a set of discriminant functions can be bounded in terms of the fat-shattering dimension. By using the theorem, the probability over an infinite set of discriminant functions is reduced to a finite set to find the pac bound of the generalization error $\text{err}_{\mathcal{P}_2}(f_2)$.

Definition 3.3.8. Let $\mathcal{F} = \{f_2 : \mathbf{d} \rightarrow \mathbf{w}_2\phi^2(\mathbf{d}) + b_2; \|\mathbf{w}_2\| = 1; \mathbf{d} \in \Gamma^2\}$ be a set of discriminant functions on Γ^2 . The *margin* of a function $f_2 \in \mathcal{F}$ with respect to the training set Γ_{train}^2 , $\text{mar}_{\Gamma_{\text{train}}^2}(f_2)$, is the minimum value of the Euclidean distances from the points in $\Gamma_{\text{train}}^2 \subset \Gamma^2$ to a separating hyperplane $\mathbf{w}_2\phi^2(\mathbf{d}) + b_2 = 0$ where $\mathbf{w}_2$ and b_2 are the weight vector and the bias of the function f_2 , respectively.

The margin at the second stage SVM with a discriminant function f_2 for the training set Γ_{train}^2 is given by [33]

$$\begin{aligned} \text{mar}_{\Gamma_{\text{train}}^2}(f_2) &= \min_{\mathbf{d} \in \Gamma_{\text{train}}^2} \frac{|\mathbf{w}_2\phi^2(\mathbf{d}) + b_2|}{\|\mathbf{w}_2\|}, \\ &= \min_{\mathbf{d} \in \Gamma_{\text{train}}^2} \frac{|f_2(\mathbf{d})|}{\|\mathbf{w}_2\|}. \end{aligned}$$

We imply that the value of $\text{mar}_{\Gamma_{\text{train}}^2}(f_2)$ is positive if f_2 correctly classifies Γ_{train}^2 .

Theorem 3.3.5. Let $f_2 \in \mathcal{F}$ be a discriminant function on the two classes, B and E , with a margin equal to η_2 on the training set Γ_{train}^2 . Then, the following probability is bounded:

$$\begin{aligned} \mathcal{P}_2\{\Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \quad \exists \quad f_2 \in \mathcal{F} \text{ such that } \text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0, \text{mar}_{\Gamma_{\text{train}}^2}(f_2) = \eta_2, \\ \text{and } \text{err}_{\Gamma_{\text{test}}^2}(f_2) > \epsilon(\delta)\} < \delta \end{aligned} \quad (3.3.11)$$

where $\epsilon(\delta) = \frac{1}{M} \left(h \log \frac{8e(N+M)}{h\eta_2} \log \frac{64(N+M)}{(\eta_2)^2} + \log \frac{2}{\delta} \right)$, $h = \text{fat}_{\mathcal{F}}\left(\frac{\eta_2}{8}\right)$, $\text{err}_{\Gamma_{\text{train}}^2}(f_2)$ and $\text{err}_{\Gamma_{\text{test}}^2}(f_2)$ are the fractional errors of f_2 on the training set Γ_{train}^2 and a random testing set Γ_{test}^2 , respectively.

Proof. Since $\text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0$, the margin of f_2 with respect to the training set Γ_{train}^2 satisfies $\eta_2 > 0$. We consider the smallest $(\eta_2/2)$ -cover $\mathcal{B}$ of $\mathcal{F}$ on the set $\Gamma_{\text{train}}^2 \cup \Gamma_{\text{test}}^2$ with respect to $\mathbf{w}_2$. From definition (3.3.6), we can therefore find $g_2 \in \mathcal{B}$ such that

$(|f_2(\mathbf{d}) - g_2(\mathbf{d})|) < \eta_2/2$ where $\mathbf{d} \in \Gamma_{\text{train}}^2 \cup \Gamma_{\text{test}}^2$. It follows that g_2 has $\text{mar}_{\Gamma_{\text{train}}^2}(g_2) > \text{mar}_{\Gamma_{\text{train}}^2}(f_2) - \eta_2/(2\|\mathbf{w}_2\|) = \eta_2/2 > 0$, and, therefore, $\text{err}_{\Gamma_{\text{train}}^2}(g_2) = 0$. Hence,

$$\begin{aligned} & \mathcal{P}_2 \left\{ \Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists f_2 \in \mathcal{F}; \text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(f_2) = \eta_2 \right\} \\ & \leq \mathcal{P}_2 \left\{ \Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists g_2 \in \mathcal{B}; \text{err}_{\Gamma_{\text{train}}^2}(g_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(g_2) > \eta_2/2 \right\}. \end{aligned}$$

Besides, if g_2 has a margin less than $\eta_2/2$ on a point $\mathbf{d} \in \Gamma_{\text{test}}^2$, that is $\text{mar}_{\mathbf{d}}(g_2) = |g_2(\mathbf{d})|/\|\mathbf{w}_2\| = |g_2(\mathbf{d})| < \eta_2/2$, then f_2 has a margin on $\mathbf{d}$, $\text{mar}_{\mathbf{d}}(f_2) = |f_2(\mathbf{d})| \leq \eta_2/2 + |g_2(\mathbf{d})| < \eta_2$, and, thus, $\mathbf{d}$ is misclassified by f_2 . Let $\text{nerr}_{\Gamma_{\text{test}}^2}(f_2)$ be the number of points misclassified of f_2 on Γ_{test}^2 , that is $\text{nerr}_{\Gamma_{\text{test}}^2}(f_2) = \text{err}_{\Gamma_{\text{test}}^2}(f_2)M$, and $\text{nerr}_{\Gamma_{\text{test}}^2}(g_2)$ denotes the number of points in Γ_{test}^2 for which g_2 has margin less than $\eta_2/2$. Then, the left hand side of inequality (3.3.11) is bounded:

$$\begin{aligned} & \mathcal{P}_2 \left\{ \Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists f_2 \in \mathcal{F}; \text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(f_2) = \eta_2; \text{err}_{\Gamma_{\text{test}}^2}(f_2) > \epsilon(\delta) \right\} \\ & = \mathcal{P}_2 \left\{ \Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists f_2 \in \mathcal{F}; \text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(f_2) = \eta_2; \text{nerr}_{\Gamma_{\text{test}}^2}(f_2) > \epsilon(\delta)M \right\}, \\ & \leq \mathcal{P}_2 \left\{ \Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists g_2 \in \mathcal{B}; \text{err}_{\Gamma_{\text{train}}^2}(g_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(g_2) > \eta_2/2; \text{nerr}_{\Gamma_{\text{test}}^2}(g_2) > \epsilon(\delta)M \right\}. \end{aligned}$$

By the permutation argument, for fixed g_2 at most $2^{-\epsilon(\delta)M}$ of the sequences obtained by swapping corresponding points in Γ_{test}^2 that g_2 has margin less than $\eta_2/2$.

Hence by the union bound

$$\begin{aligned} & \mathcal{P}_2 \left\{ \Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists g_2 \in \mathcal{B}; \text{err}_{\Gamma_{\text{train}}^2}(g_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(g_2) > \eta_2/2; \text{nerr}_{\Gamma_{\text{test}}^2}(g_2) > \epsilon(\delta)M \right\} \\ & \leq |\mathcal{B}| 2^{-\epsilon(\delta)M}, \end{aligned}$$

from the Definition (3.3.7),

$$\leq \mathcal{N}(\mathcal{F}, N + M, \eta_2/2) 2^{-\epsilon(\delta)M},$$

from the Theorem (3.3.4) (setting $[a, b]$ to $[-1, +1]$, η_2 to $\eta_2/2$, and N to $N + M$),

$$\leq 2 \left(\frac{64(N + M)}{(\eta_2)^2} \right)^{h \log(8e(N+M)/(h\eta_2))} 2^{-\epsilon(\delta)M},$$

where $h = \text{fat}_{\mathcal{F}}\left(\frac{\eta_2}{8}\right)$.

Let $\delta = 2\left(\frac{64(N+M)}{(\eta_2)^2}\right)^{h \log(8e(N+M)/(h\eta_2))} 2^{-\epsilon(\delta)M}$. It follows that

$$\epsilon(\delta) = \frac{1}{M} \left(h \log \frac{8e(N+M)}{h\eta_2} \log \frac{64(N+M)}{(\eta_2)^2} + \log \frac{2}{\delta} \right).$$

This value of $\epsilon(\delta)$ therefore ensures that the left hand side of inequality (3.3.11) is bound to be less than δ . $\square$

Lemma 3.3.6. *If the number of patterns of Γ_{train}^2 and Γ_{test}^2 are equal, that is $N = M$, then*

$$\begin{aligned} \mathcal{P}_2\{\Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \quad \exists \quad f_2 \in \mathcal{F} \text{ such that } \text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0, \text{mar}_{\Gamma_{\text{train}}^2}(f_2) = \eta_2, \\ \text{and } \text{err}_{\Gamma_{\text{test}}^2}(f_2) > \epsilon(\delta)\} < \delta \end{aligned} \quad (3.3.12)$$

where $\epsilon(\delta) = \frac{1}{N} \left(h \log \frac{16eN}{h\eta_2} \log \frac{128N}{(\eta_2)^2} + \log \frac{2}{\delta} \right)$ and $h = \text{fat}_{\mathcal{F}}\left(\frac{\eta_2}{8}\right)$.

Theorem 3.3.7. [172] *Let $\text{err}_{\mathcal{P}_2}(f_2)$ be the generalization error of a discriminant function f_2 at the output of the first stage. The margin of f_2 with respect to the training set Γ_{train}^2 is η_2 . If the number of patterns of Γ_{train}^2 and Γ_{test}^2 are equal, then*

$$\begin{aligned} \mathcal{P}_2 \left\{ \Gamma_{\text{train}}^2 : \exists f_2 \in \mathcal{F}; \text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(f_2) = \eta_2; \text{err}_{\mathcal{P}_2}(f_2) > 2\epsilon(\delta) \right\} \leq \\ 2\mathcal{P}_2 \left\{ \Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists f_2 \in \mathcal{F}; \text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(f_2) = \eta_2; \text{err}_{\Gamma_{\text{test}}^2}(f_2) > \epsilon(\delta) \right\}. \end{aligned} \quad (3.3.13)$$

This theorem is due to Vapnik [172] and is the key to bound the probability of inequality (3.3.4) by twice the probability of having zero fractional error on the training set Γ_{train}^2 but high error on a random testing set Γ_{test}^2 . Since the number of patterns of $\Gamma_{\text{train}}^2 \cup \Gamma_{\text{test}}^2$ is $2N$, there cannot be more than 2^{2N} classification functions on the right hand side of inequality (3.3.13). Therefore, the advantage of considering

errors over a finite set of $2N$ patterns is that the hypothesis space effectively become finite.

Theorem 3.3.8. *Suppose N patterns in the training set Γ_{train}^2 are classified by using a discriminant function $f_2 \in \mathcal{F}$ at the second stage on the two classes, B and E , with a margin equal to η_2 . Then, the pac bound of the generalization error $\text{err}_{\mathcal{P}_2}(f_2)$ is equal to*

$$\epsilon(\delta) = \frac{1}{N} \left(\frac{130R^2}{(\eta_2)^2} \log \frac{16eN}{\eta_2} \log \frac{128N}{(\eta_2)^2} + 2 \log \frac{4N}{\delta} \right). \quad (3.3.14)$$

Proof. Let us apply the result of Lemma 3.3.6 for a particular discriminant function $f_2 \in \mathcal{F}$ with a specified margin η_2 to bound the generalization error. Since the number of all possible patterns of $h = \text{fat}_{\mathcal{F}}\left(\frac{\eta_2}{8}\right)$ over the discriminant function is bounded by N , there are N of applications of inequality (3.3.12) for the two classes, B and E .

We let $\delta_i = \delta/N$ so that the sum $\sum_{i=1}^N \delta_i = \delta$. By choosing

$$\begin{aligned} \epsilon\left(\frac{\delta_i}{2}\right) &= \frac{1}{N} \left(\frac{65R^2}{(\eta_2)^2} \log \frac{16eN}{\eta_2} \log \frac{128N}{(\eta_2)^2} + \log \frac{4N}{\delta} \right), \\ &> \frac{1}{N} \left(\frac{R^2}{(\eta_2/8)^2} \log \frac{16eN}{h\eta_2} \log \frac{128N}{(\eta_2)^2} + \log \frac{2}{\delta_i/2} \right), \end{aligned}$$

from Theorem 3.3.3,

$$> \frac{1}{N} \left(h \log \frac{16eN}{h\eta_2} \log \frac{128N}{(\eta_2)^2} + \log \frac{2}{\delta_i/2} \right).$$

Lemma 3.3.6 ensures that the probability of any of the statements failing to hold is less than $\delta/2$

$$\mathcal{P}_2 \left\{ \Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists f_2 \in \mathcal{F}; \text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(f_2) = \eta_2; \text{err}_{\Gamma_{\text{test}}^2}(f_2) > \epsilon\left(\frac{\delta_i}{2}\right) \right\} < \frac{\delta_i}{2} < \frac{\delta}{2}.$$

By using the result of the Theorem 3.3.7, the probability $\mathcal{P}_2\{\Gamma_{\text{train}}^2 : \exists f_2 \in \mathcal{F}; \text{err}_{\Gamma_{\text{train}}^2}(f_2) = 0; \text{mar}_{\Gamma_{\text{train}}^2}(f_2) = \eta_2; \text{err}_{\mathcal{P}_2}(f_2) > 2\epsilon(\delta_i/2)\}$ is bound to be less

than δ . The pac bound of the generalization error $\text{err}_{\mathcal{P}_2}(f_2)$ is equal to $2\epsilon(\frac{\delta_i}{2}) = \frac{1}{N} \left(\frac{130R^2}{(\eta_2)^2} \log \frac{16eN}{\eta_2} \log \frac{128N}{(\eta_2)^2} + 2 \log \frac{4N}{\delta} \right)$. $\square$

From Eq. (3.3.14), maximizing the value of margin η_2 by the second stage SVM results in minimization of the generalization error $\text{err}_{\mathcal{P}_2}(f_2)$ at the output of the first stage SVM. Maximization of η_2 is done by solving the convex quadratic programming problem. Since the SVM at the second stage minimizes the generalization error of the output of the first stage by solving the optimization problem, two stages are sufficient to find an optimal classifier for RSA prediction with minimal generalization error, which takes into account the contextual information of solvent accessibility values of amino acid sequences.

3.4 Experimental Results

3.4.1 Dataset 1 (RS126)

The set of 126 nonhomologous globular protein chains, used in the experiment of Rost and Sander [151] and referred to as the RS126 set, was used to evaluate the accuracy of the prediction. Many current generation RSA prediction methods have been developed and tested on this dataset which is available at [42]. Since outputs of the RS126 dataset were sequences of two solvent accessibility elements: buried (B) and exposed (E), a two-stage binary SVM approach was proposed to address the binary prediction of RSA. The two-stage SVM approach was implemented, with the position specific scoring matrices generated by PSI-BLAST, and tested on the dataset, using a seven-fold cross validation to estimate the prediction accuracy. With seven-fold cross validation, approximately one-seventh of the dataset was left out while training and,

after training, the one-seventh of the dataset was used for testing. In order to avoid the selection of extremely biased partitions, the RS126 set was divided into subsets of same size and composition of each type of RSA. Recent methods for RSA predictions have used the seven-fold cross-validation for evaluation [151, 36, 89]. Therefore, we adopted the seven-fold cross-validation in order to provide an objective comparison of the prediction accuracy of two-stage SVM approach with the other methods.

3.4.2 Dataset 2 (Manesh)

1aba	1abr	1bdo	1beo	1bib	1bmf	1bnc	1btm	1btn	1cem
1ceo	1cew	1cfy	1chd	1chk	1cyx	1dea	1del	1dkz	1dos
1fua	1gai	1gpl	1gsa	1gtm	1hav	2ilb	2sns	3grs	3mdd

Table 3.1: The list of 30 proteins used for training the single-stage and two-stage SVM approaches with Manesh dataset.

The second dataset, generated by Manesh [111] and referred to as the Manesh dataset, consisted of 215 nonhomologous protein chains. The dataset contained proteins with less than 25% homology and is available at [43]. The NETASA prediction method [2] was developed and tested on this dataset. A set of 30 proteins containing 7545 residues was selected for training (see Table 3.1). The remaining 185 proteins with 43,137 residues were used for testing. We adopted these training and testing sets in order to provide an objective comparison of the prediction accuracy of the two-stage SVM approach with the results of the NETASA method [2] and the probability profile approach of Gianese *et al.* [69]. The two-stage SVM predicted the RSA types based on the position specific scoring matrices generated by PSI-BLAST.

The PSI-BLAST profiles contained probabilities of residues, taking into account the significance of each sequence and distant homologues [84].

3.4.3 RSA and Prediction Accuracy Assessment

RSA percentage (%) of an amino acid residue is defined as the ratio of the solvent accessible surface area of the residue observed in the 3-D structure to that observed in an extended tripeptide (Gly-X-Gly or Ala-X-Ala) conformation [171]. The value of RSA lies between 0% and 100% with 0% corresponding to a fully buried type and 100% to the fully exposed type. The solvent accessibility of an amino acid residue is considered as a buried (B) if the RSA value of the residue is smaller than a threshold, $c\%$, or an exposed (E) element, otherwise. We experimented with $c \in \{0, 5, 9, 10, 16, 20, 25, 50\}$ to fairly compare the two-stage SVM approach to previous methods. The residue solvent accessible surface areas of the RS126 set were computed with the DSSP program [87]. The ASC program [53] with the van der Waals radii of the atoms [124] was used to compute the residue solvent accessible surface areas for the Manesh dataset. The Ala-X-Ala oligopeptide in an extended conformation instead of Gly-X-Gly is used to calculate RSA in the Manesh dataset. The definition of RSA and programs used to compute it are consistent with those used by other authors whose methods are compared against the proposed approach.

The prediction accuracy is measured by the percentage of correctly predicted types of solvent accessibility of residues [151]; the sensitivity score indicates the proportion of exposed (E) residues that are correctly predicted as E ; the specificity measures the proportion of buried (B) residues that are correctly predicted as B . By changing the thresholds of RSA definition of the prediction, we get a range of sensitivities

Figure 3.3: Performance of relative solvent accessibility prediction by the single-stage and two-stage SVM approaches on the RS126 dataset at a 25% threshold. The window length indicates the size of neighborhood taken as the input for the single-stage and two-stage SVM methods. For two-stage SVM, the window length indicates the size of neighborhood taken as the input for the second stage, with the first stage operating with its optimal window size 13.

and specificities, which leads to Receiver Operation Characteristics (ROC) that plots sensitivity versus one minus specificity. The ROC curves offer for comparisons among different prediction methods irrespective of the threshold for determination of solvent accessibility type.

3.4.4 Results

We performed extensive experiments to find the optimal window sizes, kernel, and kernel and SVM parameters by first determining the window size of the single-stage SVM classifier for RSA prediction. The second stage optimal window size was then

Figure 3.4: Performance of relative solvent accessibility prediction by the single-stage and two-stage SVM approaches on the Manesh dataset with different window lengths at a 25% threshold. For two-stage SVM, the window length indicates the size of neighborhood taken as the input for the second stage, with the first stage operating with its optimal window size 13.

Dataset	Method	Window w_1									
		9	11	13	15	17	19	21	23	25	27
RS126	Single-stage	75.9	76.0	76.2	76.1	76.0	75.7	75.6	75.5	75.4	75.1
	Two-stage	77.0	77.1	77.2	77.2	77.0	76.9	76.8	76.7	76.6	76.6
Manesh	Single-stage	76.8	77.0	77.1	77.1	77.0	76.7	76.6	76.4	76.3	76.1
	Two-stage	78.0	78.1	78.1	78.1	78.0	77.8	77.8	77.7	76.6	77.5

Table 3.2: Performance of RSA prediction by the single-stage and two-stage SVM approaches with different window lengths w_1 and the selected window $w_2 = 21$ on the RS126 and Manesh datasets at a 25% threshold.

experimentally determined by fixing the window size of the first stage to its optimal value. Once, the optimal window size of the second stage is found, we further investigated the influence of input window size on the performance of RSA prediction.

The selected window sizes of 13 and 15 of previous methods for RSA prediction [151, 180, 90] were used as the initial selection of the two-stage SVM approach. Then, the window size was tried in the extended range of the initial selection, i.e. [9, 27], to find the optimal value.

Figures 3.3 and 3.4 illustrate the performance of the single-stage and two-stage SVM approaches against different neighborhood sizes (window length) on the RS126 and Manesh datasets at a 25% threshold. As seen, the neighborhood windows, w_1 of size 13 ($h_1^1 = h_2^1 = 6$), and w_2 of size 21 ($h_1^2 = h_2^2 = 10$), gave the optimal accuracy of the first stage and the second stage, respectively. Using window lengths in the interval [9, 17] at the first stage and [17, 25] at the second stage, the variation of the prediction accuracy was small, less than 0.3%. Table 3.2 indicates that the different results of two-stage SVM method with window lengths w_1 in the [9, 27] range and the selected window $w_2 = 21$ on the RS126 and Manesh datasets were not significant. These results show that the selected optimal window size and parameters in both learning stages were not biased by the test data chosen. Further, the accuracies of two-stage SVM method with $w_1 = 13$ and $w_2 = 1$ ($h_1^2 = h_2^2 = 0$) are 76.3% and 77.1% on the RS126 and Manesh datasets at a 25% threshold. The result indicates that the improvement of accuracy of two-stage SVM without the contextual information of solvent accessibility was very small.

For our knowledge, there does not exist any simple method to find the optimal

Dataset	Method	Kernel Function				
		Gaussian $\sigma_1 = 0.10$ $\sigma_2 = 0.15$	Linear	Polynomial $d = 2$	Polynomial $d = 3$	Polynomial $d = 4$
RS126	Single-stage	76.2	75.0	75.3	75.6	75.7
	Two-stage	77.2	76.4	76.6	76.8	76.8
Manesh	Single-stage	77.1	76.1	76.5	76.7	76.8
	Two-stage	78.1	77.2	77.5	77.8	77.9

Table 3.3: Performance of RSA prediction by the single-stage and two-stage SVM approaches with different type of kernel functions on RS126 and Manesh datasets at a 25% threshold with $\gamma_1 = \gamma_2 = 1.0$. The neighborhood windows of size 13 and 21 were used in the single-stage and two-stage SVM approach, respectively.

Parameter γ Gaussian σ	1.0				0.5	1.5	2.0
	0.01	0.05	0.1	0.15	0.1	0.1	0.1
Single-stage	76.6	76.8	77.1	76.9	76.7	77.0	76.8
Two-stage	77.5	77.8	78.0	78.1	77.8	78.1	77.9

Table 3.4: Performance of RSA prediction by the single-stage and two-stage SVM approaches with different parameters of Gaussian kernel on the Manesh dataset at a 25% threshold.

values of the kernel and other parameters of SVMs, and, therefore, they were empirically determined. The selected values of the kernel and SVM parameters of the previous SVM methods [180, 90] were used as the initial selection of the proposed method. The ranges $[0.01, 0.5]$ of SVM parameter and $[0.1, 2]$ of Gaussian kernel parameter are consistent with those used in previous methods [180, 90].

Table 3.3 shows RSA prediction accuracies of the SVM method with the Gaussian RBF (Radial Basis Function) $\mathcal{K}(\mathbf{x}, \mathbf{y}) = e^{-\sigma \|\mathbf{x} - \mathbf{y}\|^2}$, linear kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{y}$, and polynomial kernels $\mathcal{K}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}\mathbf{y} + 1)^d$ with different $d = 2, 3, 4$, on the RS126 and Manesh datasets at a 25% threshold with $\gamma_1 = \gamma_2 = 1.0$. The Gaussian kernel achieved the better results over the linear and polynomial kernels for RSA prediction. The main reason is that the Gaussian kernel can result in complex (but smooth) decision function, and therefore has the ability to better fit the data where simple discrimination by using a hyperplane or a low-dimensional polynomial surface is not possible. The use of Gaussian kernel showed the best performance when the dimension of feature space is infinite [163]. The performance of RSA prediction by the single-stage and two-stage SVM approaches with different parameters of Gaussian kernel on the Manesh dataset at a 25% threshold is shown in Table 3.4. The kernel selected here was the Gaussian with the parameters: $\sigma_1 = 0.1$ for the first stage and $\sigma_2 = 0.15$ for the second stage.

The SVM method was implemented using LIBSVM library [25] and SMO (sequential minimal optimization) algorithm [134] which are simple to implement without needing storage for matrices or to invoke an iterative numerical routine for each sub-problem.

Table 3.5 shows the performances of different solvent accessibility predictors and

Method / Threshold	0%	5%	9%	16%	25%
Rost and Sander, 1994 (PHDacc)	86.0	-	74.6	75.0	-
Gianese <i>et al.</i> , 2003 (PP)	-	-	76.8	75.1	-
Kim and Park, 2004 (Single-stage SVM)	86.2	79.8	-	77.8	-
Two-stage SVMs	90.2	83.5	81.3	79.4	77.2

Table 3.5: Comparison of performances of two-stage SVM approach with other methods in RSA prediction on the RS126 dataset with position specific scoring matrices generated by PSI-BLAST. The notation - indicates that the corresponding result was not available from the literature.

two-stage SVM approach on the RS126 set. Two-stage SVMs with PSI-BLAST profiles achieved accuracies of 90.2%, 83.5%, 81.3%, and 79.4% at thresholds of 0%, 5%, 9%, and 16%, respectively, which are the highest scores on the RS126 set to date. Compared to the newest method of Kim and Park, using single-stage SVM [89], the two-stage SVM method significantly obtained 4%, 3.7%, and 1.6% higher prediction accuracies at 0%, 5%, and 16% thresholds, respectively. On the RS126 dataset, the accuracies were improved by 4.5% and 4.3% at thresholds of 9% and 16% compared to the results of the probability profiles approach of Gianese *et al.* [69]. The prediction accuracy of two-stage SVMs outperformed the results by the multi-layer perceptron networks of PHDacc method proposed by Rost and Sander [151] at all thresholds.

Table 3.6 shows the performance of two-stage SVM approach on the Manesh dataset based on PSI-BLAST profiles and comparisons with other solvent accessibility predictors. The best algorithm was found to be the cascade of two SVMs, which achieved accuracies of 90.4%, 82.9%, 81.0%, 78.6%, 78.1%, and 79.1% at thresholds

Method / Threshold	0%	5%	10%	20%	25%	50%
Ahmad and Gromiha, 2002 (NETASA)	87.9	74.6	71.2	-	70.3	75.9
Gianese <i>et al.</i> , 2003 (PP)	89.5	75.7	73.4	-	71.6	76.2
Two-stage SVMs	90.4	82.9	81.0	78.6	78.1	79.1
Giorgi <i>et al.</i> , 1999 (PredAcc)	85.0	-	-	-	70.7	-
Cuff and Barton, 2000 (Jnet)	86.6	79.0	-	-	75.0	-
Li and Pan, 2001	-	-	-	71.5	-	-
Pollastri <i>et al.</i> , 2002 (BRNN)	86.5	81.2	-	-	77.2	-
Adamczak <i>et al.</i> , 2004 (SABLE)	-	76.8	77.5	77.9	77.6	-

Table 3.6: Comparison of performances of two-stage SVM approach in RSA prediction based on position specific scoring matrices generated by PSI-BLAST, with other methods on the Manesh dataset.

of 0%, 5%, 10%, 20%, 25%, and 50%, respectively. On the Manesh dataset, the accuracies were significantly improved by 2.5%, 8.3%, 9.8%, 7.8% and 3.2% for 0%, 5%, 10%, 25%, and 50% thresholds, respectively, compared to the results of NETASA method [2]. Comparing two-stage SVMs to the probability profiles method [69], substantial gains of 0.9% to 7.6% of prediction accuracy were observed for different thresholds. Manesh *et al.* [111] reported an accuracy value of 70% when the RSA threshold between buried and exposed residues was 9% on this dataset by using information theory. PredAcc approach [71] achieved accuracies of 85.0% and 70.7% at thresholds of 0% and 25%, respectively. Carugo [21] proposed an independent method for RSA prediction and reported 68.7% accuracy for a two-stage definition

Figure 3.5: The distribution of prediction scores obtained by two-stage SVMs for the benchmark 185 proteins of the Manesh dataset at a 5% threshold based on PSI-BLAST profiles.

on 338 proteins. Jnet method of Cuff and Barton [36] using PSI-BLAST profiles obtained accuracy values of 86.6%, 79.0%, and 75.0% at thresholds of 0%, 5% and 25%, respectively, on a dataset of 480 proteins. Li and Pan [97] developed a multiple linear regression method and reported 71.5% prediction accuracy at 20% threshold on 704 proteins using single sequences. Pollastri *et al.*'s BRNN [136] achieved accuracies of 86.5%, 81.2%, and 77.2% at thresholds of 0%, 5% and 25%, respectively, based on PSI-BLAST profiles. Recently, SABLE method proposed by Adamczak *et al.* [1] using neural networks based regression reported 76.8%, 77.5%, 77.9%, and 77.6%

Figure 3.6: The distribution of prediction scores obtained by two-stage SVMs for the benchmark RS126 dataset at a 5% threshold based on PSI-BLAST profiles.

prediction accuracies at 5%, 10%, 20%, and 25% thresholds on a training set of 890 proteins and a testing set of 174 proteins. Compared to previous SVM approaches for RSA prediction, the prediction accuracies of two-stage SVMs were higher than those of Yuan *et al.* [180].

Figures 3.5 and 3.6 present the distributions of prediction scores obtained by two-stage SVMs for the benchmark Manesh and RS126 datasets with a 5% threshold based on PSI-BLAST profiles. The ROC curves on the Manesh and RS126 datasets for single-stage and two-stage SVM approaches at different thresholds are illustrated

Figure 3.7: The ROC curves on the Manesh dataset for single-stage and two-stage SVM approaches for RSA prediction.

in Figures 3.7 and 3.8. As shown, the prediction accuracy of two-stage SVMs outperformed the single-stage SVM methods for RSA prediction at all thresholds.

For RSA prediction, the accuracy of two-stage SVMs using PSI-BLAST profiles was significantly higher than results based on multiple sequence alignments. For example, the accuracy of two-stage SVM method on RS126 dataset was only 78.6% at a threshold of 5% based on multiple sequence alignments. As mentioned [84], PSI-BLAST profiles contain more information of homologous protein structures than multiple sequence alignments. Additionally, improvements of accuracies are observed when larger sequences and more homologous profiles are used in training. As shown in Table 3.7, by using a set of 205 proteins instead of 30 proteins for training, the prediction accuracies of 10 sequences, obtained from the tails of the histogram in Figure 3.5 (1lts, 1nba, 1afw, 3cox, 2wsy, 7rsa, 1amm, 1mai, 1knb, 1kte), were improved at a threshold of 5%. These observations suggest that the performance of two-stage

Figure 3.8: The ROC curves on the RS126 dataset for single-stage and two-stage SVM approaches for RSA prediction.

SVM method based on PSI-BLAST profiles for a novel amino acid sequence suffers if it lacks in the homologous structures in the training set. For a completely new protein whose homologous proteins are not used in training, two-stage SVM method predicts its solvent accessibilities with a low accuracy. For our knowledge, Rost and Adamczak [151, 1] concluded that the overall performance of any method based on evolutionary profiles suffers when very remote or no homologues are included.

Table 3.8 lists the properties of 20 amino acids and their average occurrence and probabilities for exposure and error in RSA prediction on the Manesh dataset at a 25% threshold. Nelson and Cox [112] based on the polarity or tendency to interact with water of R group at biological pH to group 20 amino acids into five main classes. According to the statistical data, amino acids, Ala, Val, Leu, Ile, Phe, and Cys were easy to predict while Gly, Pro, Trp, Thr, Arg, and His were difficult to predict by two-stage SVMs. As shown, two-stage SVM method frequently predicted A, V, L, I, M, F, W, Y, C to be buried, and G, P, S, T, N, Q, K, R, H, D, E to be exposed.

Training	1lts	1nba	1afw	3cox	2wsy	7rsa	1amm	1mai	1knb	1kte
30 proteins	70.9	71.4	71.8	73.6	75.2	91.1	92.0	93.3	93.3	93.3
205 proteins	72.6	71.8	71.8	73.9	76.9	91.0	93.1	94.1	93.3	93.3

Table 3.7: Comparison of performances of two-stage SVM approach on 10 proteins (1lts, 1nba, 1afw, 3cox, 2wsy, 7rsa, 1amm, 1mai, 1knb, 1kte) based on PSI-BLAST profiles with two different training sets of 30 and 205 proteins at a 5% threshold.

The statistical data confirms that the non-polar R groups (hydrophobic) tend to be buried, i.e., in the interior of a protein, and the polar R groups (hydrophilic) tend to be on the surface (exposed), except for Gly, Pro, and Cys [27]. This is because two Cys are readily oxidized to form a disulfide bond and disulfide-linked residues are hydrophobic. Gly tends to be exposed as it contributes little in general to the stability of folded proteins [112]. Pro commonly appears at exposed sites in proteins, such as loops, turns, and N-terminal first turn of helix [112]. The results from Table 3.8 suggest that the amino acid residues that tend to be buried (A, V, L, I, M, F, W, Y, C) are predicted with higher accuracies than exposed ones (G, P, S, T, N, Q, K, R, H, D, E).

As shown in Tables 3.5 and 3.6, predictions were best for buried residues, e.g., 90.2% and 90.4% of the completely buried sites were correctly predicted at a threshold of 0% on the RS126 and Manesh datasets, respectively. The two-stage SVM method achieved the highest prediction accuracy for the extreme case of fully buried type

Amino acid	Occurrence (%)	Exposure (%)	Error in RSA prediction (%)
<i>Non-polar R group (hydrophobic)</i>			
Gly G	7.5	55.8	27.1
Ala A	7.7	39.7	19.0
Val V	6.8	16.7	16.2
Leu L	8.8	14.1	15.7
Ile I	5.7	12.1	14.8
Met M	2.2	20.8	21.7
Pro P	4.5	64.8	27.5
<i>Aromatic R group (hydrophobic)</i>			
Phe F	4.1	10.5	16.5
Trp W	1.4	12.3	25.1
Tyr Y	3.8	18.8	24.8
<i>Polar, uncharged R group (hydrophilic)</i>			
Ser S	5.9	63.7	24.7
Thr T	5.6	53.2	25.6
Cys C	1.6	12.5	15.1
Asn N	4.6	74.4	25.0
Gln Q	3.9	79.4	24.6
<i>Positively R charged (hydrophilic)</i>			
Lys K	6.1	84.5	19.8
Arg R	4.8	72.5	28.3
His H	2.2	51.2	30.5
<i>Negatively R charged (hydrophilic)</i>			
Asp D	6.3	80.9	23.2
Glu E	6.4	84.7	21.4

Table 3.8: The properties of 20 amino acids: their average occurrences, probabilities of exposures, and the error in RSA prediction on the Manesh dataset at a 25% threshold.

Database / Threshold	0%	5%	10%	20%	25%	50%
1,581,064 NR	90.2	82.8	80.9	78.6	78.1	79.0
2,745,128 NR	90.4	82.9	81.0	78.6	78.1	79.1

Table 3.9: Comparison of performances of two-stage SVM approach on the Manesh dataset based on position specific scoring matrices generated by PSI-BLAST with two different NR databases.

Dataset	Number of proteins	Memory space (MB)	Training time (minute)
<i>First stage</i>			
Manesh	30	14	1
RS126	126	43	12
<i>Second stage</i>			
Manesh	30	2	0.3
RS126	126	6	2.5

Table 3.10: The training time and memory space of the first stage and the second stage SVM for the Manesh and RS126 datasets on a supercomputer, Compaq AlphaServer SC45.

because the accessibility of completely buried residues is best conserved in 3-D homologous structures [151]. Residues in α -helix and β -strand structure segments were predicted better than ones in coil segments, e.g., 80.7%, 82.2%, and 77.5% residues were correctly predicted in α -helix, β -strand, and coil segments, respectively, on the Manesh dataset at a 25% threshold.

We also estimated the effect of the growing size of NR databases used to generate position scoring matrices by PSI-BLAST on the accuracy of two-stage SVM method. Two NR databases were used: one as of December 22, 2003 with 1,581,064 sequences and a newer version as of April 7, 2004 with 2,745,128 sequences. The different results of two-stage SVMs on two NR databases were not significant (see Table 3.9).

A web server for RSA prediction using two-stage SVM method has been developed and is available at [177]. A set of 30 proteins containing 7545 residues (see Table 3.1) was selected for training two-stage SVM method presented on the web server.

3.4.5 Time and Space Complexities

Table 3.10 shows the training time and memory space of the two-stage approach on the Manesh training set of 30 proteins and the RS126 set of 126 proteins. For training phase, the program implementing the two-stage method for RSA prediction was run on a supercomputer, Compaq AlphaServer SC45 with 44 nodes, each node comprising of four 1GHz Alpha processors with 1GB memory. For the first stage, the sizes of Manesh and RS126 training sets are 14MB and 43MB, respectively. It takes 1 and 12 minutes for training on Manesh and RS126 sets, respectively. For the second stage, the sizes of Manesh and RS126 training sets are 2MB and 6MB, respectively. It takes 20 seconds and 2.5 minutes for training on Manesh and RS126 sets, respectively. For

testing phase, the web server for RSA prediction was run on Itanium server consisting of 16 processors with 733Mhz and 1TB storage. The NR database with 2,745,128 sequences occupies 1.4GB disc space. Like other methods using the evolutionary information from PSI-BLAST profiles, it averagely takes 10 minutes for searching on NR database to generate the PSSM profile of a query sequence. It requires the execution time of 5 seconds to predict the solvent accessibility of a query sequence from its PSSM profile on average.

In the training phase, the time complexity of SVM depends on the method used to solve the QP (quadratic programming) problem. A standard QP solver has the time complexity of order $O(Nm^2)$ where N is the number of training exemplars and m is the dimension of a feature vector. In order to solve the large scale QP problem, the decomposition methods or iterative methods have been suggested which break down the large scale QP problem into a series of smaller QP problems such as SMO (sequential minimal optimization) [133] and LIBSVM [25]. The time complexity of LIBSVM method used in the two-stage SVMs is (the number of iterations) $\times O(Nm)$ [25]. Unfortunately, so far we do not know much about the complexity of the number of iterations [25]. In the testing phase, if the SVM predictors have l support vectors we can build the hyperplane in $O(lm)$ time. Then, each prediction takes $O(m)$ time. Totally, the time complexity of the two-stage SVM approach is (the number of iterations) $\times O(Nm)$.

3.5 Summary

The existing bioinformatics techniques for RSA prediction are mostly single-stage approaches which predict the RSA types of residues based on only the information

available in amino acid sequences. We demonstrated a two-stage approach, by using SVMs, that utilizes the output predicted by single-stage prediction schemes to improve the accuracy of RSA prediction. In this way, the influence on the RSA value of a residue by those of its neighbors are accounted for. This is because the solvent accessibility at a particular position of the sequence depends on the structures of the rest of the sequence, i.e., it accounts for the fact that the buried or exposed type consists of at least two consecutive residues. Therefore, another layer of SVM classifier incorporating the contextual relationship among the solvent accessibility characteristics makes the prediction more realistic in terms of predicted mean lengths of solvent accessibility elements. The analysis of prediction results from single-stage and two-stage SVM methods showed that the second stage SVM ultimately cleans the output prediction of the first stage SVM, mostly by removing isolated buried or exposed residue.

SVMs are more suitable for prediction of RSA values because they minimize the generalization error in the prediction. We showed in Theorem 3.3.8 that the generalization error made in the first stage is further minimized by the second stage of the two-stage approach. SVMs are an optimal classifier for the second stage in terms of the margin of separation; they attempt to minimize not only the empirical risk of known sequences but also the actual risk for unknown sequences. Two stages of SVMs are sufficient to find an optimal classifier for RSA prediction as the second stage SVM minimizes the generalization error of the first stage by solving the optimization problem at the second stage.

Recently, Kim and Park [89] suggested to use the information of the position

specific scoring matrices generated by PSI-BLAST as inputs to SVMs for RSA prediction. By combining PSI-BLAST profiles, the present approach achieved better results than the methods using information from single sequences and multiple sequence alignments. Compared to the method of Kim and Park, our method showed a considerable improvement in the accuracy of prediction. By incorporating the state-of-the-art methods based on PSI-BLAST profiles and SVMs in a two-stage approach, we are able to report the best accuracies to date for RSA prediction on the tested datasets. The RSA elements of residues predicted by our approach could facilitate the prediction of the structure and function of amino acid sequences.

Chapter 4

Two-Stage Multi-Class SVMs to Protein Secondary Structure Prediction

As Support Vector Machines (SVMs) are known to behave well compared to other statistical or machine learning methods for many biological problems, we provide an account of Multi-class Support Vector Machine (MSVM) approach for the determination of Protein Secondary Structure (PSS). We argue that it is feasible to extend the single-stage MSVM approach by adding MSVM as the second stage to capture the contextual information among secondary structural elements, thereby improving the accuracy of the prediction of PSS. We demonstrate the performance of two-stage MSVM approach, using three datasets and report Q_3 accuracies of 78.0% and 76.3% on the RS126 dataset of 126 nonhomologous globular proteins and the CB396 dataset of 396 nonhomologous proteins, respectively, which are better than the highest score published on both datasets to date, and Q_3 accuracy of 79.4% on EVAsec dataset. A web server for PSS prediction using the two-stage MSVM method has been developed and is available at [178].

4.1 Introduction

Proteins are large molecules having a central role in coordinating living processes and constituting to the bulk of living organisms: enzymes, hormones and structural material. As such, understanding the mechanisms of protein's interaction and operation is vital to the study of many diseases. Key functional aspects of proteins depend on their three-dimensional (3-D) structure [29]. Therefore, the knowledge of a protein's structure and its components and their relation to its function are highly useful in life science applications. For example, if one protein is known to be involved in a certain disease, a detailed knowledge of its surface structure provides clues for finding new drugs that bind to activate or inactivate the protein as necessary.

Structural complexity of proteins can be largely described on three scales. Firstly, proteins are unbranched heterogeneous polymers comprising amino acids, each residue having the ability to append to a growing chain and side-chain with another having unique physical properties [7]. The amino acid sequence is necessary and sufficient to predict the 3-D structure of a protein. The sequential ordering of amino acids is referred to as the *primary structure* which level of information is relatively easy to determine. The next level of structural complexity is referred to as the *secondary structure* at which level short contiguous elements of amino acid sequences adopt to, primarily, α -helix (H), β -strand (E), or coil (C) forms. A protein may have a mixture of secondary elements packed together to form the next level of complexity, the *tertiary structure* which describes how the elements of secondary structure associate to form a globular 3-D structure.

The experimental approaches for protein 3-D structure prediction, such as NMR

Spectroscopy or X-Ray Crystallography techniques or by molecular dynamics simulations, are marred by long experimentation time, prone to difficulties, expensive, and, therefore, limited to small proteins or peptide sequences. Furthermore, out of approximately one million non-redundant protein sequences, only one thousand non-redundant structures have been experimentally determined to date. Although a large amounts of protein sequences exists, there are no comparably fast means to determine the structures. Without major innovations in efficiency and capacity of the experimental techniques, computational means of determining protein structure are being investigated. Insufficient knowledge of protein structure could be the limiting step in the development of modern medicine in the post-genome era [110].

Given that there is a physical mapping from sequence to structure, predicting structural aspects by using machine learning strategies has drawn considerable attention [110]. The trend from the studies suggests that as the structural properties being predicted move from the aspects of the primary structure through the secondary structure and on to the tertiary structure, there is a loss in prediction accuracy. This is due to the inherent difficulty in managing the increasing complexity and the increasing dominance of non-local effects of structural components as one moves towards the final 3-D structure. The problem of solving protein structure prediction based on searching through all accessible amino acid chain conformations is NP-complete [170]. Accurate and efficient prediction techniques of protein structure offer the possibility of shifting the focus from determination of structure to the interactions and mechanisms in control of biological processes. This chapter deals with the problem of predicting Protein Secondary Structure (PSS) elements of residues from their primary sequences, that provide important clues in predicting the protein's 3-D structure.

Many computational techniques have been proposed in the literature to solve the PSS prediction problem (referred in Chapter 2). Despite the existence of many approaches, the current success rates of existing approaches to PSS prediction are far from most biologist’s expectations. Most existing secondary structure techniques are single-stage approaches, which are unable to find complex relations (correlations) among structural elements in the sequence. This could be improved by incorporating the interactions or contextual information among the elements of the sequences of secondary structures. We argue that it is feasible in enhancing the present single-stage MSVM approach farther by augmenting with another prediction scheme at their outputs and propose to use MSVM as the second-stage. Previously, two stage approaches such as PHD method [149] have been proposed. But by cascading two MSVMs to minimize the generalization error, the two-stage MSVM method gives higher accuracies. By using the position specific scoring matrices generated by PSI-BLAST, the two-stage MSVM approach significantly achieves Q_3 accuracies of 78.0% and 76.3% on the RS126 and CB396 datasets based on a seven-fold cross validation, and Q_3 of 79.4% on EVAsec dataset.

4.2 Two-Stage MSVM Approach

In the two-stage MSVM approach, we use two MSVMs in cascade to predict secondary structures of residues in amino acid sequences.

Let $\mathbf{v}_i$ be the vector representing 21-dimensional coding of the residue r_i where 20 units are the values from raw matrices of PSI-BLAST profiles ranging from $[0, 1]$ and the other is used for the padding space to indicate the overlapping end of the sequence [84]. The padding component is set to 1 when padding is required for the

end of the sequence or 0, otherwise. Let the input pattern to the MSVM approach at site i be $\mathbf{r}_i = (\mathbf{v}_{i-h_1^1}, \mathbf{v}_{i-h_1^1+1}, \dots, \mathbf{v}_i, \dots, \mathbf{v}_{i+h_2^1})$ where $\mathbf{v}_i$ denote the center element, h_1^1 and h_2^1 denote the width of window on the two sides; $w_1 = h_1^1 + h_2^1 + 1$ is the neighborhood size around the element i .

4.2.1 First Stage

A MSVM scheme has been proposed by Crammer and Singer [31]. For PSS prediction, this method constructs three discriminant functions but all are obtained by solving one single optimization problem, which can be formulated as follows:

Minimize

$$\frac{1}{2} \sum_{k \in \Omega_T} (\mathbf{w}_1^k)^T \mathbf{w}_1^k + \gamma_1 \sum_{j=1}^N \xi_j^1,$$

subject to the constraints

$$\mathbf{w}_1^{t_j} \phi^1(\mathbf{r}_j) - \mathbf{w}_1^k \phi^1(\mathbf{r}_j) \geq c_j^k - \xi_j^1, \quad (4.2.1)$$

where t_j is the secondary structural type of residue r_j corresponding to the the training vector $\mathbf{r}_j$, $j = 1, 2, \dots, N$, ξ_j^1 is slack variable, and $c_j^k = \begin{cases} 0 & \text{if } t_j = k \\ 1 & \text{if } t_j \neq k \end{cases}$.

We find the minimization of the above formulation by solving the following quadratic programming problem [31]:

$$\max_{\alpha_j^k} -\frac{1}{2} \sum_{j=1}^N \sum_{i=1}^N \mathcal{K}^1(\mathbf{r}_j, \mathbf{r}_i) \sum_{k \in \Omega_T} \alpha_j^k \alpha_i^k - \sum_{j=1}^N \sum_{k \in \Omega_T} \alpha_j^k c_j^k, \quad (4.2.2)$$

$$\text{such that } \sum_{k \in \Omega_T} \alpha_j^k = 0 \text{ and } \alpha_j^k \leq \begin{cases} 0 & \text{if } t_j \neq k \\ \gamma_1 & \text{if } t_j = k \end{cases}, \quad (4.2.3)$$

where $\mathcal{K}^1(\mathbf{r}_i, \mathbf{r}_j) = \phi^1(\mathbf{r}_i)\phi^1(\mathbf{r}_j)$ denotes the kernel function and $\mathbf{w}_1^k = \sum_{j=1}^N \alpha_j^k \phi^1(\mathbf{r}_j)$.

Once the parameters α_j^k are obtained from the optimization, the resulting discriminant function f_1^k of a test input vector $\mathbf{r}_i$ is given by

$$\begin{aligned} f_1^k(\mathbf{r}_i) &= \sum_{j=1}^N \alpha_j^k \mathcal{K}^1(\mathbf{r}_i, \mathbf{r}_j), \\ &= \mathbf{w}_1^k \phi^1(\mathbf{r}_i). \end{aligned} \quad (4.2.4)$$

Let $f_1(\mathbf{r}_i) = \arg \max_{k \in \Omega_T} f_1^k(\mathbf{r}_i)$. In the single-stage MSVM method, the secondary structural type t_i corresponding to the residue at site i , r_i , is determined by

$$t_i = f_1(\mathbf{r}_i). \quad (4.2.5)$$

4.2.2 Second Stage

We extend the single-stage MSVM technique by cascading another MSVM at the output of the present single-stage approach to improve the accuracy of prediction. This is because the secondary structure at a particular position of the sequence depends on the structures of the rest of the sequence, i.e., it accounts for the fact that the strands span over at least three adjacent residues and helices consist of at least four consecutive residues [149]. This intrinsic relation cannot be captured by using single-stage approaches alone. Therefore, another layer of classifiers, which minimizes the generalization error of the output of single-stage methods by incorporating the sequential relationship among the protein structure elements, improves the prediction accuracy. Further, Crooks and Brenner use the analysis of information theory to indicate that correlations between neighboring secondary structures are much stronger than those of neighboring amino acids [34]. Figure 4.1 represents the architecture for PSS prediction by cascading two MSVM classifiers.

Figure 4.1: The two-stage Multi-class Support Vector Machine (MSVM) approach for protein secondary structure prediction

Consider a window of w_2 size at a site of the output sequence of the first stage; the vector at position i , $\mathbf{d}_i = (d_{i-h_1^2}^k, d_{i-h_1^2+1}^k, \dots, d_i^k, \dots, d_{i+h_2^2}^k)$ where $w_2 = 3(h_1^2 + h_2^2 + 1)$, $d_i^k = 1/(1 + e^{-f_1^k(r_i)})$, and f_1^k denotes the discriminant function of the first stage. The application of the logistic sigmoid function to the outputs of the first stage has the advantage of constraining the input units of the second stage to the (0,1) interval that is similar to the range of the input units of the first stage. The purpose of this choice is to easier determine parameters for optimal performance. The MSVM converts the input patterns, usually linearly inseparable, to a higher dimensional space by using

the mapping ϕ^2 with a kernel function $\mathcal{K}^2(\mathbf{d}_i, \mathbf{d}_j) = \phi^2(\mathbf{d}_i)\phi^2(\mathbf{d}_j)$.

As in the first stage, the hidden outputs in the higher dimensional space are linearly combined by a weight vector, $\mathbf{w}_2$, to obtain the prediction output. Let the training set of exemplars for the second stage MSVM be $\Gamma_{\text{train}}^2 = \{\mathbf{d}_j : j = 1, \dots, N\}$. The vector $\mathbf{w}_2$ is obtained by solving the following convex quadratic programming problem, over all the patterns seen in the training phase [31].

$$\begin{aligned} & \max_{\beta_j^k} -\frac{1}{2} \sum_{k \in \Omega_T} (\mathbf{w}_2^k)^T \mathbf{w}_2^k - \sum_{j=1}^N \sum_{k \in \Omega_T} \beta_j^k c_j^k, \\ & \text{such that } \sum_{k \in \Omega_T} \beta_j^k = 0, \quad j = 1, 2, \dots, N, \text{ and} \\ & \beta_j^k \leq \begin{cases} 0 & \text{if } t_j \neq k \\ \gamma_2 & \text{if } t_j = k \end{cases}, \\ & \text{where } \mathbf{w}_2^k = \sum_{j=1}^N \beta_j^k \phi^2(\mathbf{d}_j). \end{aligned} \tag{4.2.6}$$

The discriminant function, f_2^k , at the second stage is given by

$$\begin{aligned} f_2^k(\mathbf{d}_i) &= \sum_{j=1}^N \beta_j^k \mathcal{K}^2(\mathbf{d}_i, \mathbf{d}_j), \\ &= \mathbf{w}_2^k \phi^2(\mathbf{d}_i). \end{aligned} \tag{4.2.7}$$

Let $f_2(\mathbf{d}_i) = \arg \max_{k \in \Omega_T} f_2^k(\mathbf{d}_i)$. The secondary structural type t_i corresponding to the residue r_i is determined by

$$t_i = f_2(\mathbf{d}_i). \tag{4.2.8}$$

4.3 Generalization in Two-Stage Multi-Class SVMs

In this section, we prove that the second stage of two-stage MSVMs minimizes the generalization error made by the first stage MSVM for PSS prediction.

In the first stage MSVM, the function f_1 discriminates the types of secondary structure, based on the features or interactions among the residues in the input pattern. If the training and testing patterns, $(\mathbf{r}, t) \in \Gamma^1 \times \Omega_T$, are drawn independently and identically according to a probability distribution $\mathcal{Q}_1$, the generalization error $\text{err}_{\mathcal{Q}_1}(f_1)$ in the the first stage MSVM for PSS prediction is given by

$$\text{err}_{\mathcal{Q}_1}(f_1) = \int L(f_1(\mathbf{r}), t) d\mathcal{Q}_1(\mathbf{r}, t), \quad (4.3.1)$$

where the loss function

$$L(f_1(\mathbf{r}), t) = \begin{cases} 0 & \text{if } f_1(\mathbf{r}) = t \\ 1 & \text{if } f_1(\mathbf{r}) \neq t \end{cases},$$

where $f_1(\mathbf{r}) = \max_{k \in \Omega_T} f_1^k(\mathbf{r})$ and $t \in \Omega_T$ is the desired output for the input pattern $\mathbf{r}$.

The aim of two-stage MSVM approach is to find an optimal function, f_2 , in the second stage such that the output $f_2(\mathbf{d}) = \arg \max_{k \in \Omega_T} f_2^k(\mathbf{d})$, gives the desired output by incorporating the contextual relationships among secondary structure elements so that the generalization error in the prediction is further minimized.

$$\text{err}_{\mathcal{Q}_2}(f_2) = \int L(f_2(\mathbf{d}), t) d\mathcal{Q}_2(\mathbf{d}, t). \quad (4.3.2)$$

We consider the case where $h_1^2 = h_2^2 = 0$ and $f_2^{k*}(\mathbf{d}) = \ln \frac{d^k}{1-d^k}$. It follows that $\mathbf{d} = \mathbf{d}_i$ and $f_2^{k*}(\mathbf{d}) = \ln \left(e^{-f_1^k(\mathbf{r}_i)} \right)^{-1} = f_1^k(\mathbf{r})$. Thus $f_2^*(\mathbf{d}) = f_1(\mathbf{r})$ where $f_2^*(\mathbf{d}) = \max_{k \in \Omega_T} f_2^{k*}(\mathbf{d})$. The generalization error, $\text{err}_{\mathcal{Q}_2}(f_2^*)$, can now be written as

$$\begin{aligned} \text{err}_{\mathcal{Q}_2}(f_2^*) &= \int_{L=0} L(f_2^*(\mathbf{d}), t) d\mathcal{Q}_2(\mathbf{d}, t) + \int_{L=1} L(f_2^*(\mathbf{d}), t) d\mathcal{Q}_2(\mathbf{d}, t), \\ &= \int_{L=1} L(f_2^*(\mathbf{d}), t) d\mathcal{Q}_2(\mathbf{d}, t), \end{aligned}$$

$$\begin{aligned}
&= \int_{L=1} L(f_1(\mathbf{r}), t) d\mathcal{Q}_1(\mathbf{r}, t), \\
&= \int L(f_1(\mathbf{r}), t) d\mathcal{Q}_1(\mathbf{r}, t), \\
&= \text{err}_{\mathcal{Q}_1}(f_1).
\end{aligned}$$

As a result, there exists at least a discriminant function f_2^* such that $\text{err}_{\mathcal{Q}_2}(f_2^*) = \text{err}_{\mathcal{Q}_1}(f_1)$. Therefore, the optimal function f_2 ensures that the upper bound of $\text{err}_{\mathcal{Q}_2}(f_2) \leq \text{err}_{\mathcal{Q}_2}(f_2^*) = \text{err}_{\mathcal{Q}_1}(f_1)$.

Without loss of generality, we assume that the training set $\Gamma_{\text{train}}^2 \subset \Gamma^2$ and the testing set $\Gamma_{\text{test}}^2 \subset \Gamma^2$ for the second stage each contained N patterns. For the MSVM technique at the second stage, let $\mathbf{w}_2^{k/l}$ be the weight vector $\mathbf{w}_2^k - \mathbf{w}_2^l$, and, therefore $\mathbf{w}_2^{k/l} \phi^2(\mathbf{d}) = \mathbf{w}_2^k \phi^2(\mathbf{d}) - \mathbf{w}_2^l \phi^2(\mathbf{d}) = f_2^k(\mathbf{d}) - f_2^l(\mathbf{d})$. The secondary structure of a residue r is not l if $\mathbf{w}_2^{k/l} \phi^2(\mathbf{d}) \geq 0$ or not k otherwise.

Theorem 4.3.1. *Let $\mathcal{F} = \{f_2^{k/l} : \mathbf{d} \rightarrow \mathbf{w}_2^{k/l} \phi^2(\mathbf{d}); \|\mathbf{w}_2^{k/l}\| \leq 1; \mathbf{d} \in \Gamma^2; k, l \in \Omega_T\}$ be a set of discriminant functions defined on Γ^2 and restricted to points in a ball of m dimensions of radius R about the origin, that is $f_2^{k/l}(\mathbf{d}) = f_2^k(\mathbf{d}) - f_2^l(\mathbf{d})$, $\phi^2(\mathbf{d}) \in \mathbb{R}^m$, and $\|\phi^2(\mathbf{d})\| \leq R$. Then the fat-shattering dimension $\text{fat}_{\mathcal{F}}(\eta_2^{k/l})$ at scale $\eta_2^{k/l}$ is bounded:*

$$\text{fat}_{\mathcal{F}}(\eta_2^{k/l}) \leq \left(\frac{R}{\eta_2^{k/l}} \right)^2.$$

Proof. We use the same technique in Theorem 3.3.3 to prove this theorem. $\square$

Definition 4.3.1. Let $\mathcal{F} = \{f_2^{k/l} : \mathbf{d} \rightarrow \mathbf{w}_2^{k/l} \phi^2(\mathbf{d}); \|\mathbf{w}_2^{k/l}\| \leq 1; \mathbf{d} \in \Gamma^2; k, l \in \Omega_T\}$ be a set of discriminant functions defined on Γ^2 . The *decision directed acyclic graph* G on 3 classes H , E , and C , over $\mathcal{F}$ is a set of functions which can be implemented using a rooted binary directed acyclic graph with 3 leaves labeled by the classes H ,

Figure 4.2: A decision directed acyclic graph with discriminant functions $f_2^{k/l} \in \mathcal{F}$ at decision nodes k/l where $k, l \in \Omega_T$ and the leaves labeled by the classes H , E , and C .

E , and C , where each of 3 internal nodes is labeled with an element of $\mathcal{F}$ (see Figure 4.2).

Theorem 4.3.2. [135] Let G be a decision directed acyclic graph on 3 classes H , E , and C , with 3 decision nodes, H/E , E/C , and C/H , with margins $\eta_2^{k/l}$ and discriminant functions $f_2^{k/l} \in \mathcal{F}$ at decision nodes $k/l \in G$, where $\eta_2^{k/l} = \min_{\mathbf{d} \in \Gamma_{\text{train}}^2} \frac{|\mathbf{w}_2^{k/l} \phi^2(\mathbf{d})|}{\|\mathbf{w}_2^{k/l}\|}$. Then, the following probability is bounded:

$$\mathcal{Q}_2\{\Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists G \text{ such that } \text{err}_{\Gamma_{\text{train}}^2}(G) = 0 \text{ and } \text{err}_{\Gamma_{\text{test}}^2}(G) > \epsilon(\delta)\} < \delta \quad (4.3.3)$$

where $\epsilon(\delta) = \frac{1}{N} \left(\sum_{k/l \in G} h^{k/l} \log \frac{4eN}{h^{k/l}} \log(4N) + \log \frac{2^3}{\delta} \right)$, $h^{k/l} = \text{fat}_{\mathcal{F}} \left(\frac{\eta_2^{k/l}}{8} \right)$, $\text{err}_{\Gamma_{\text{train}}^2}(G)$

and $\text{err}_{\Gamma_{\text{test}}^2}(G)$ are the fractional error of G on the training set Γ_{train}^2 and a random testing set Γ_{test}^2 , respectively.

Theorem 4.3.3. *Let G be a decision directed acyclic graph with discriminant functions $f_2^{k/l} \in \mathcal{F}$ at nodes k/l . Then, the generalization error of f_2 where $f_2(\mathbf{d}) = \arg \max_{k \in \Omega_T} f_2^k(\mathbf{d})$ in the probability distribution $\mathcal{Q}_2$ is*

$$\text{err}_{\mathcal{Q}_2}(f_2) = \text{err}_{\mathcal{Q}_2}(G).$$

Proof. This is equivalent to proving that, for an arbitrary example $\mathbf{d} \in \Gamma^2$, $f_2(\mathbf{d})$ equals to the secondary structural type of $\mathbf{d}$ predicted by the decision directed acyclic graph G .

Firstly, we consider the case $f_2(\mathbf{d}) = \arg \max_{k \in \Omega_T} f_2^k(\mathbf{d}) = H$. It follows that $f_2^H(\mathbf{d}) > f_2^E(\mathbf{d}) > f_2^C(\mathbf{d})$ or $f_2^H(\mathbf{d}) > f_2^C(\mathbf{d}) > f_2^E(\mathbf{d})$.

- In the first case, $f_2^H(\mathbf{d}) > f_2^E(\mathbf{d}) > f_2^C(\mathbf{d})$. Starting at the root node H/E of G , the binary decision function of the classifier H/E is evaluated for the input pattern $\mathbf{d}$. The node is then exited via the right edge because $f_2^{H/E}(\mathbf{d}) = f_2^H(\mathbf{d}) - f_2^E(\mathbf{d}) > 0$. In the decision node C/H , the discriminant function $f_2^{C/H}(\mathbf{d})$ is evaluated. Since $f_2^{C/H}(\mathbf{d}) = f_2^C(\mathbf{d}) - f_2^H(\mathbf{d}) < 0$, the input $\mathbf{d}$ reaches to the leaf labeled H . Therefore, the secondary structural type of $\mathbf{d}$ predicted by G is H .
- In the second case, if $f_2^H(\mathbf{d}) > f_2^C(\mathbf{d}) > f_2^E(\mathbf{d})$ then we use an identical argument.

Similar proofs hold for cases $f_2(\mathbf{d}) = \arg \max_{k \in \Omega_T} f_2^k(\mathbf{d}) = E$ or C . $\square$

Theorem 4.3.4. [172] Let $\text{err}_{\mathcal{Q}_2}(G)$ be the generalization error of G at the output of the first stage. Then

$$\mathcal{Q}_2 \left\{ \Gamma_{\text{train}}^2 : \exists G \text{ such that } \text{err}_{\Gamma_{\text{train}}^2}(G) = 0 \text{ and } \text{err}_{\mathcal{Q}_2}(G) > 2\epsilon(\delta) \right\} \leq 2\mathcal{Q}_2 \left\{ \Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists G \text{ such that } \text{err}_{\Gamma_{\text{train}}^2}(G) = 0 \text{ and } \text{err}_{\Gamma_{\text{test}}^2}(G) > \epsilon(\delta) \right\}.$$

Theorem 4.3.5. Suppose N patterns in the training set Γ_{train}^2 are classified by using the two-stage MSVM method at second stage with optimal values of weight vectors $\mathbf{w}_2^k$, $k \in \Omega_T$. Then, the pac bound of the generalization error $\text{err}_{\mathcal{Q}_2}(f_2)$ is equal to

$$\epsilon(\delta) = \frac{1}{N} \left(390R^2 \sum_{k \in \Omega_T} \|\mathbf{w}_2^k\|^2 \log(4eN) \log(4N) + 2 \log \frac{2(2N)^3}{\delta} \right). \quad (4.3.4)$$

Proof. Since the margin $\eta_2^{k/l}$ is the minimum value of the distances from the instances labeled k or l to the hyperplane, $\mathbf{w}_2^{k/l} \phi^2(\mathbf{d}) = 0$, at the second stage, we have, $\eta_2^{k/l} = \min_{\mathbf{d} \in \Gamma_{\text{train}}^2} \frac{|\mathbf{w}_2^{k/l} \phi^2(\mathbf{d})|}{\|\mathbf{w}_2^{k/l}\|} = \min_{\mathbf{d} \in \Gamma_{\text{train}}^2} \frac{|(\mathbf{w}_2^k - \mathbf{w}_2^l) \phi^2(\mathbf{d})|}{\|\mathbf{w}_2^k - \mathbf{w}_2^l\|} \geq \frac{1}{\|\mathbf{w}_2^k - \mathbf{w}_2^l\|}$. Therefore, the quantity $\sum_{k/l} \frac{1}{(\eta_2^{k/l})^2} \leq \sum_{k/l} \|\mathbf{w}_2^k - \mathbf{w}_2^l\|^2 = 3 \sum_k \|\mathbf{w}_2^k\|^2 - (\sum_k \mathbf{w}_2^k)^2 \leq 3 \sum_k \|\mathbf{w}_2^k\|^2$. Solving the optimization problems at the second stage results in the minimization of the quantity $\sum_{k \in \Omega_T} \|\mathbf{w}_2^k\|^2$ which is directly related to the margin of the classifier. Plugging the binary classifiers induced by $\mathbf{w}_2^{k/l}$ results a stepwise method for calculating the maximum among $\{f_2^k(\mathbf{d}) = \mathbf{w}_2^k \phi^2(\mathbf{d}); k \in \Omega_T\}$ that is similar to the process of finding the secondary structure in the decision directed acyclic graph G . Let us apply the result of Theorem 4.3.2 for G with specified margin $\eta_2^{k/l}$ at each node to bound the generalization error $\text{err}_{\mathcal{Q}_2}(G)$. Since the number of decision nodes is 3 and the largest allowed value of $h^{k/l} = \text{fat}_{\mathcal{F}}\left(\frac{\eta_2^{k/l}}{8}\right)$ is N , the number of all possible patterns of $h^{k/l}$'s over the decision nodes is bounded by $(N)^3$. We let $\delta_i = \delta/(N)^3$ so that

$$\begin{aligned}
\sum_{i=1}^{(N)^3} \delta_i &= \delta. \text{ By choosing } \epsilon(\frac{\delta_i}{2}) \\
&= \frac{1}{N} \left(195R^2 \sum_{k \in \Omega_T} \|\mathbf{w}_2^k\|^2 \log(4eN) \log(4N) + \log \frac{2(2N)^3}{\delta} \right), \\
&\geq \frac{1}{N} \left(65R^2 \sum_{k/l \in G} \frac{1}{(\eta_2^{k/l})^2} \log(4eN) \log(4N) + \log \frac{2(2N)^3}{\delta} \right), \\
&> \frac{1}{N} \left(\sum_{k/l \in G} \frac{R^2}{(\eta_2^{k/l}/8)^2} \log \frac{4eN}{h^{k/l}} \log(4N) + \log \frac{2^3}{\delta_i/2} \right), \\
&\text{from Theorem 4.3.1} \\
&> \frac{1}{N} \left(\sum_{k/l \in G} h^{k/l} \log \frac{4eN}{h^{k/l}} \log(4N) + \log \frac{2^3}{\delta_i/2} \right).
\end{aligned}$$

Theorem 4.3.2 ensures that the probability of any of the statements failing to hold is less than $\delta/2$

$$\mathcal{Q}_2\{\Gamma_{\text{train}}^2, \Gamma_{\text{test}}^2 : \exists G \text{ such that } \text{err}_{\Gamma_{\text{train}}^2}(G) = 0; \text{err}_{\Gamma_{\text{test}}^2}(G) > \epsilon(\frac{\delta_i}{2})\} < \frac{\delta_i}{2} < \frac{\delta}{2}.$$

By using the result of the Theorem 4.3.4, the probability $\mathcal{Q}_2\{\Gamma_{\text{train}}^2 : \exists G \text{ such that } \text{err}_{\Gamma_{\text{train}}^2}(G) = 0; \text{err}_{\mathcal{Q}_2}(G) > 2\epsilon(\delta_i/2)\}$ is bound to be less than δ . From Theorem 4.3.3, the pac bound of the generalization error $\text{err}_{\mathcal{Q}_2}(f_2)$ is therefore equal to $2\epsilon(\frac{\delta_i}{2}) = \frac{1}{N} \left(390R^2 \sum_{k \in \Omega_T} \|\mathbf{w}_2^k\|^2 \log(4eN) \log(4N) + 2 \log \frac{2(2N)^3}{\delta} \right)$. $\square$

From Eq. (4.3.4), minimizing the quantity $\sum_{k \in \Omega_T} \|\mathbf{w}_2^k\|^2$ results in the minimization of the generalization error at the output of the first stage MSVM method. Since the MSVM at the second stage minimizes the error of the output of the first stage by solving the optimization problem, two stages are sufficient to find an optimal classifier for PSS prediction with minimal generalization error taking into account the contextual information of secondary structures.

4.4 Vapnik Chervonenkis (VC) Dimension

Vapnik and Chervonenkis have developed the analysis of generalization error based on the VC dimension [173, 174]. In this section, we discuss the difference between the VC dimension and the margin analysis.

Definition 4.4.1. Let U be a set of functions mapping from Γ^2 to $\{-1, +1\}$. The *VC dimension* z of U is the maximal number of patterns that can be separated into two classes in possible 2^z ways by U .

From this definition, we can therefore think of the fat-shattering dimension at scale $\eta_2/8$, $\text{fat}_{\mathcal{F}}(\eta_2/8)$, as the effective VC dimension when a margin of η_2 is observed.

Theorem 4.4.1. [173, 174] *Let U be a set of functions mapping from Γ^2 to $\{-1, +1\}$ and having VC dimension z . For any probability distribution $\mathcal{P}_2$ on $\Gamma^2 \times \{-1, +1\}$, with probability $1 - \delta$ over N patterns in the training set Γ_{train}^2 , any hypothesis $u \in U$ has generalization error no more than*

$$\text{err}_{\mathcal{P}_2}(u) \leq 2\text{err}_{\Gamma_{\text{train}}^2}(u) + \frac{4}{N} \left(z \log \frac{2eN}{z} + \log \frac{4}{\delta} \right) \quad (4.4.1)$$

provided $z \leq N$ and $\text{err}_{\Gamma_{\text{train}}^2}(u)$ is a fractional error of u on the training set Γ_{train}^2 .

By applying Theorem 4.4.1 to a sequence of hypothesis spaces (sets of classification functions) of increasing complexity, we can choose the hypothesis for which the bound of generalization error $\text{err}_{\mathcal{P}_2}(u)$ is tightest. The approach to trading off empirical error $\text{err}_{\Gamma_{\text{train}}^2}(u)$ with hypothesis space complexity is known as structural risk minimization (SRM). However, in order to apply SRM (using Theorem 4.4.1) we must define the hierarchy of hypothesis spaces before data is observed. Since the VC dimension also depends on the value of margin [172], maximizing the margin can perform SRM over

the hypothesis spaces defined by a sequence of increasing margin. The problem here is that we only measure the margin on the observed training data while the SRM inductive principle requires that the sequence of hypothesis spaces must be chosen before data is observed. By using the margin analysis based on the theory of Bartlett and Shawe-Taylor [11], this problem is overcome.

Furthermore, implementing SRM inductive principle of VC dimension to find the global minimum of the generalization error for the multi-class classification is still an open problem [72]. By using the margin analysis for the two-stage MSVM approach in PSS prediction, the bound on the generalization error is achieved (referred in Theorem 4.3.5).

4.5 Comparison of Two-Stage MSVM and Two-Stage MLP Methods for PSS Prediction

Rost and Sander have proposed the PHD approach using two Multi-Layer Perceptrons (MLPs) in cascade to improve the accuracy of prediction by capturing the contextual relations among secondary structures [149]. The first neural network, called sequence-to-structure network, predicts secondary structure at a particular element of the input sequence by using a symmetric window of size 13, and has three output neurons, each neuron representing one secondary structure type of protein. The second neural network, called structure-to-structure network, uses a window of 17 secondary structure predictions of the first stage as input. The predicted secondary structure is chosen as the largest of the three outputs of the second stage.

Recall Eq. (4.3.1), the generalization error of MSVM and MLP methods, $\text{err}_{Q_1}(f_1)$

where $f_1(\mathbf{r}) = \arg \max_{k \in \Omega_T} f_1^k(\mathbf{r})$, is given by

$$\text{err}_{\mathcal{Q}_1}(f_1) = \int L(f_1(\mathbf{r}), t) d\mathcal{P}_1(\mathbf{r}, t).$$

Note that in the single-stage MLP method, f_1^k is the transfer function of the network, including the activation function of the output neuron representing the secondary structure $k \in \Omega_T$.

The aim of the single-stage approaches is to find an optimal function, f_1 , taking into account the sequential relationships among amino acid residues to minimize the generalization error $\text{err}_{\mathcal{Q}_1}(f_1)$. Since minimizing $\text{err}_{\mathcal{Q}_1}(f_1)$ is not a trivial problem, the single-stage MLP method instead minimizes the following so-called the *empirical risk functional* [174]

$$\text{err}_{\Gamma_{\text{train}}^1}(f_1) = \frac{1}{N} \sum_{i=1}^N L(f_1(\mathbf{r}_i), t_i), \quad (4.5.1)$$

where $\mathbf{r}_i \in \Gamma_{\text{train}}^1$, t_i denotes the corresponding output for the training input pattern $\mathbf{r}_i$, and N is the number of patterns of Γ_{train}^1 .

By using the back-propagation algorithm, the single-stage MLP method can achieve a local minimum of the empirical risk functional $\text{err}_{\Gamma_{\text{train}}^1}(f_1)$, however, it is not guaranteed to find the global minimum of the generalization error $\text{err}_{\mathcal{Q}_1}(f_1)$. Similarly, the second stage MLP is not guaranteed to find the global minimum of the generalization error $\text{err}_{\mathcal{Q}_2}(f_2)$ at the output of the first stage MLP. Therefore, two-stage MLPs are not sufficient to achieve an optimal classifier for PSS prediction. As shown, two-stage MSVM method is optimal for PSS prediction because it minimizes both the generalization error $\text{err}_{\mathcal{Q}_1}(f_1)$ based on interactions among amino acids as well as the generalization error $\text{err}_{\mathcal{Q}_2}(f_2)$ of the output of the MSVM by capturing the contextual information of secondary structure. Hence, the prediction accuracy of two-stage MSVMs outperforms the result of two-stage MLPs for PSS prediction.

Set 0	4cpai 1fc2c	2or1l 2gn5	256ba 7rsa	9wgaa 1gp1a	3tima 3blm	4tsla 4cms	8adh 1lap	2tgpi 2sns	2pcy 1bmvl
Set 1	9insb 1crn	3ebx 3icb	1fkf 1eca	2ilb 2gcr	1mcpl 1pyp	2gbp 5ldh	6cpp 5gr1	2mev4 3cln	5hvpa 9pap
Set 2	1cbh 4rhv4	1il8a 1ubq	2hmza 1azu	6dfr 3hmgb	1r092 1s01	6tmne 1gd10	1bmvl 2glsa	4sgbi 1sdha	2wrpr 3cla
Set 3	2mhu 6hir	3ait 3b5c	4cpv 2ccya	1bbpa 1rbp	1wsya 2tsca	2cyp 5er2e	4xiaa 7icd	1fdx 1tnfa	5cytr 2stv
Set 4	1ppt 1bds	2utga 1hip	2paba 5lyz	2lh4 2ltna	4rhv3 2cab	2fnr 9apia	2aat 6acn	4rxn 4fxn	3rnt 1ak3a
Set 5	1mrt 2ltnb	1csei 1cc5	2rspa 4bp2	2tmvp 1cd4	3pgm 1rhd	8abp 3hmga	2phh 6cts	1ovoa 2sodb	1lrd3 2alp
Set 6	9apib 1sh1	1cdta 2fxb	1acx 1paz	1l58 1etu	1fdlh 4rhv1	6cpa 4pfk	1wsyb 7cata	1tgsi 2lhb	1fxia 3gapa

Table 4.1: The dataset of 126 nonhomologous globular protein chains from Rost and Sandar (1993) used for the seven-fold cross validation.

4.6 Experimental Results

4.6.1 Dataset 1 (RS126)

The set of 126 nonhomologous globular protein chains, used in the experiment of Rost and Sander [149] and referred to as the RS126 set, was used to evaluate the accuracy of the classifiers. The dataset contained 23,349 residues with 32% α -helix, 23% β -strand, and 45% coil. Many current generation secondary structure prediction

methods have been developed and tested on this dataset. The RS126 set is available at [41]. Since outputs of the RS126 dataset for PSS prediction were sequences of three secondary structure elements, a two-stage MSVM approach was proposed to directly predict three types of PSS. The single-stage and two-stage MSVM approaches were implemented, with the position specific scoring matrices generated by PSI-BLAST, and tested on the dataset, using a seven-fold cross validation to estimate the prediction accuracy. In order to avoid the selection of extremely biased partitions, the RS126 set was divided into seven subsets with each subset having similar size and content of each type of secondary structure. Table 4.1 shows the division of the dataset into seven subsets (set 0 - set 6). The SVMs were therefore trained on the values of six subsets and tested on the remaining subset. A more reliable estimate of the prediction accuracy could be achieved by using a full jack-knife type where one protein was left out while training on the rest, but this would lead to very large computational requirements.

4.6.2 Dataset 2 (CB396)

The second dataset generated by Cuff and Barton [35] at the European Bioinformatics Institute (EBI) consisted of 396 nonhomologous protein chains and was referred to as the CB396 set. Cuff and Barton used a rigorous method consisting on the computation of the similarity score to derive their nonredundant dataset. The CB396 set is available at [44]. The single-stage and two-stage MSVM approaches have been used to predict PSS based on the position specific scoring matrices generated by PSI-BLAST.

1aab	1ad0A	1afi	1ag8A	1ah6	1aihA	1al3	1alo	1ao5A
1aqb	1aq0A	1aqdA	1aru	1auyA	1avk	1awe	1aym1	1bco
1broA	1blj	1bp1	1cauA	1cb2A	1cg2A	1cgmE	1cgt	1ciu
1djaA	1ctn	1cx2A	1cyg	1dmr	1dpgA	1eapA	1ecpA	1extA
1hjaA	1frb	1gdhA	1ghsA	1gnd	1hlpA	1hna	1ift	1igcL
1lfaA	1igjB	1igtB	1jbc	1ksiA	1lcl	1mhlC	1mioB	1mn1
1pamA	1nhp	1nls	1npk	1oatA	1ojt	1ordA	1pax	1pmaA
1tca	1pmaB	1pnkB	1pth	1qba	1rom	1scuB	1sgt	1tib
1tsy	1ubi	1waj	1ycqA	1ycrA	1yge	2dri	2ack	2ctc
2ebn	2rn2	2sicI	3cla	3gar	3lip	3pmgA	3sicI	3tgl

Table 4.2: The list of 90 proteins extracted from PSIPRED dataset, in 1999, used for training the two-stage MSVM approach.

4.6.3 Dataset 3 (EVAsec)

Two-stage MSVM method has been trained on 90 protein chains (see Table 4.2) of PSIPRED dataset [40] and tested on a set of 64 proteins provided by EVAsec (see Table 4.3) based on the position specific scoring matrices generated by PSI-BLAST. By using the PSI-BLAST profiles directly, the PSIPRED method [84] achieved the highest published score for any previous methods. This testing contains sequences with no homology to the proteins of the training set extracted from PSIPRED dataset in 1999 [40]. The above training and testing sets were used to fairly compare the two-stage MSVM approach to PSIPRED method.

4.6.4 Protein Secondary Structure Definition

The state for each secondary structure in the training and testing sets was assigned from DSSP [87] that is the most widely used secondary structure definition. The

1hk9A	1i85A	1izmA	1j0wA	1lmmA	1mwqA	1nngA	1nnvA
1nslA	1nxhA	1oj5A	1oyiA	1p57A	1p5hA	1p94A	1pc0A
1pd3A	1pg6A	1pjuA	1pv6A	1pw4A	1px5A	1pzqA	1pztA
1q3jA	1q3kA	1q68A	1q7sA	1q90L	1q90M	1q90N	1qw2A
1r2mA	1r4gA	1rh5B	1rhzA	1rifA	1rjiA	1rklA	1rocA
1rpuA	1rqtA	1rwsA	1s0yB	1s4zC	1s5lI	1s5lJ	1s5lL
1s5lM	1s5lX	1s5lZ	1s68A	1s6cB	1s7bA	1uf3A	1ufiA
1uhwA	1ujxA	1usmA	1v74A	1v74B	1vjqa	1ocsA	1rf8B

Table 4.3: The list of 64 proteins of the EVAsec dataset used for testing the two-stage MSVM approach.

eight states, H(α -helix), G(3_{10} -helix), I(π -helix), E(β -strand), B(isolated β -bridge), T(turn), S(bend), and -(rest), were reduced to three classes, α -helix (H), β -strand (E), and coil (C), by using the following method: H and G to H; E and B to E; all others states to C. This reduction method is now used for many current secondary structure prediction methods. We adopted the reduction method to provide objective comparison of the prediction accuracy of two-stage MSVMs to the results of other methods.

4.6.5 Prediction Accuracy Assessment

Several different measures of prediction accuracy have been suggested in the literature. The most common measure is the overall three-state prediction percentage Q_3 defined as the ratio of correctly predicted residues to the total number of residues in the database under consideration [149, 153]

$$Q_3(\%) = \frac{\sum_{t \in \Omega_T} a_t}{\sum_{t \in \Omega_T} b_t} \times 100, \quad (4.6.1)$$

where a_t is the number of correctly predicted residues in structure t , and b_t is the number of residues observed in type t .

A measure of the performance on secondary structure class $t \in \Omega_T$ is the percentage Q_t of correctly predicted residues observed in class t

$$Q_t(\%) = \frac{a_t}{b_t} \times 100. \quad (4.6.2)$$

These measures can be very helpful in detecting over and under-prediction of one or more types of secondary structures.

A complementary measure of prediction accuracy is obtained from the Matthews' correlation coefficients [106] for each of the three secondary structures: ρ_H , ρ_E , and ρ_C

$$\rho_t = \frac{a_t m_t - u_t o_t}{\sqrt{(a_t + u_t)(a_t + o_t)(m_t + u_t)(m_t + o_t)}}, \quad (4.6.3)$$

where a_t is mentioned above, m_t is the number of correctly predicted residues in the other structure, u_t is the number of underpredicted residues, and o_t is the number of overpredicted residues. Let M_{tl} be the number of residues observed in structure t and predicted in structure l , where $t, l \in \Omega_T$. The above values can be written as

$$\begin{aligned} a_t &= M_{tt}, & m_t &= \sum_{l \neq t} \sum_{k \neq t} M_{kl}, \\ o_t &= \sum_{l \neq t} M_{lt}, & u_t &= \sum_{l \neq t} M_{tl}. \end{aligned}$$

The correlation coefficients provide the success of predicting residue for each type of secondary structure, which are $+1$ if the predictions are all correct or -1 if all the prediction are false. The advantage of the correlation coefficients is seen in the case of a random or trivial prediction. The Matthews' correlation coefficients have been widely used, and the exact definitions can also be found [140, 149].

Although Matthews' correlation coefficients give more reliable estimates of the prediction accuracy, they do not express how realistic the prediction is [169, 13, 152, 155, 47]. Segment overlap measure (Sov) can avoid this problems [153, 181]. It measures the accuracy by counting predicted and observed segments, and measuring their overlap. The Sov score is computed by summing over segments of secondary structure instead of individual residues. As with residue-by-residue measure, Sov can measure the accuracy of predicting a single secondary structure class, or the accuracy of an entire prediction consisting of several classes of secondary structure

$$\text{Sov} = \frac{1}{n} \sum_s \frac{\text{minov}(s_{\text{obs}}; s_{\text{pred}}) + \delta(s_{\text{obs}}; s_{\text{pred}})}{\text{maxov}(s_{\text{obs}}; s_{\text{pred}})} \times \text{len}(s_{\text{obs}}), \quad (4.6.4)$$

where n is the total number of residues, s_{obs} and s_{pred} are the observed and predicted secondary structure segments respectively, and $\text{len}(s_{\text{obs}})$ is the number of residues in the segments s_{obs} . The sum is taken over all segment pairs $s = \{s_{\text{obs}}; s_{\text{pred}}\}$. The actual overlap between the two segments is minov , i.e., the number of residues for which both segments have, while maxov is the total extent of both segments, i.e., the number of residues for which either of the two has. The accepted variation δ assures a ratio of 1.0 when there are only minor deviations at the ends of segments. Thus, δ is chosen to be smaller than minov and half the length of segment s_{obs} . The ratio $\text{minov}/\text{maxov}$ is constrained to a maximum value of 1.0, i.e., the allowance cannot lead to a more than perfect value of fractional overlap.

4.6.6 Results

Extensive experiments were performed to find the optimal window sizes, kernel, and kernel and MSVM parameters by first determining the window sizes of the first stage and second stage MSVM classifiers for PSS prediction. Once, the optimal window

Dataset	Accuracy	Window w_1									
		9	11	13	15	17	19	21	23	25	27
RS126	Q_3	75.8	76.0	76.1	76.2	76.2	76.0	75.7	75.5	75.2	75.0
	Sov	68.2	68.5	68.7	68.8	68.9	68.4	67.0	66.7	66.2	66.0
CB396	Q_3	74.2	74.3	74.5	74.5	74.3	74.2	73.9	73.7	73.5	73.2
	Sov	68.6	69.0	69.4	69.5	69.3	69.0	68.7	68.4	68.2	67.8

Table 4.4: Accuracies of PSS prediction by the single-stage MSVM approach on the RS126 and CB396 datasets. The window length indicates the size of neighborhood taken as the input for the single-stage approach.

sizes were found, we further investigated the influence of input window sizes on the performance of PSS prediction.

The selected window sizes of 13 and 15 of previous methods for PSS prediction [149, 176, 89, 75] were used as the initial selection of the two-stage MSVM approach. Then, the window size was tried in the extended range of the initial selection, i.e. [9, 27], to find the optimal value.

Tables 4.4 and 4.5 show the performance of the single-stage and two-stage MSVM approaches against different neighborhood windows on the RS126 and CB396 datasets. As seen, the neighborhood windows, w_1 of size 15 ($h_1^1 = h_2^1 = 7$), and w_2 of size 21 ($h_1^2 = 2$ and $h_2^2 = 4$), gave the optimal accuracy of the first stage and the second stage, respectively. Using window lengths in the interval [11, 19] at the first stage and [17, 25] at the second stage, the variation of the prediction accuracy was small, less than 0.3%. Table 4.6 shows that the different results of two-stage MSVM with window lengths w_1 in the [11, 19] range and the selected window $w_2 = 21$ on the RS126 and

Dataset	Accuracy	Window w_2									
		3	6	9	12	15	18	21	24	27	30
RS126	Q_3	76.4	76.9	77.2	77.5	77.7	77.9	78.0	77.9	77.8	77.8
	Sov	68.9	70.4	71.1	71.6	72.1	72.4	72.6	72.4	72.1	72.0
CB396	Q_3	74.6	75.2	75.5	75.8	76.0	76.1	76.3	76.3	76.2	76.0
	Sov	69.7	71.8	72.5	72.7	72.9	73.1	73.2	73.1	73.2	73.0

Table 4.5: Accuracies of PSS prediction by the two-stage MSVM approach on the RS126 and CB396 datasets at different neighborhood window sizes of the second stage. The size of the first stage neighborhood was taken as 15.

Manesh datasets were not significant. These results indicate that the selected optimal window size and parameters in both learning stages were not biased by the test data chosen. The results from Tables 4.4 and 4.5 confirm that the different results of two-stage MSVM without the contextual information of secondary structures, that is the window size w_2 of width 3 ($h_1^2 = 0$ and $h_2^2 = 0$), and the single-stage MSVM were not significant. By capturing the contextual relations among secondary structures, the prediction of two-stage MSVM outperformed the result of classical MSVM method. Further, Rost and Zemla [181] have indicated that the SOV accuracy can express how realistic the prediction is because SOV measures the prediction accuracy by secondary structure segment rather than individual residues. By introducing two-stage MSVM to take into account the contextual information of secondary structures, the proposed method achieved 3.8% and 3.7% improvement in SOV measure on the RS126 and CB396 datasets, respectively. The result shows that two-stage MSVM method improved significantly unrealistic prediction from the sing stage method.

Dataset	Method	Window w_1									
		9	11	13	15	17	19	21	23	25	27
RS126	Single-stage	75.8	76.0	76.1	76.2	76.2	76.0	75.7	75.5	75.2	75.0
	Two-stage	77.6	77.8	77.8	78.0	78.0	77.9	77.7	77.6	77.4	77.2
CB396	Single-stage	74.2	74.3	74.5	74.5	74.3	74.2	73.9	73.7	73.5	73.2
	Two-stage	76.1	76.1	76.3	76.3	76.2	76.1	75.9	75.8	75.6	75.4

Table 4.6: Accuracies of PSS prediction by the single-stage and two-stage MSVM approaches with different window lengths w_1 and the selected window $w_2 = 21$ on the RS126 and CB396 datasets.

We provide the biological reason why the window size of width 21 gave the optimal accuracy of the second stage technique. Kabsch and Sander [87] showed that a residue in an α -helix is hydrogen bonded to the approximate fourth residue above and the fourth residue below in the primary sequence, and it takes 3.6 residues to make a turn in an α -helix. This periodic structure is essential for the characterization of an α -helix. β -strands are formed by main-chain hydrogen bonds between a pair of three adjacent residues. These characteristics are all of local nature and can, therefore, be built into the second stage MSVM that predicts helices and strands from the window of size 7-9 of amino acid sequence. Since coils do not have such a locally described periodic structure, the MSVM model for coils can use only the local encoding scheme with a small window size. As a result, the window size of width 7 for each class, i.e. $7 \times 3 = 21$ for the whole window size of three classes, can give the optimal accuracy of the second stage.

Like the two-stage SVMs for RSA prediction, the optimal values of the kernel and

Dataset	Method	Kernel Function				
		Gaussian $\sigma_1 = 0.05$ $\sigma_2 = 0.01$	Linear	Polynomial $d = 2$	Polynomial $d = 3$	Polynomial $d = 4$
RS126	Single-stage	76.2	73.5	74.1	74.8	75.4
	Two-stage	78.0	76.4	76.8	77.1	77.3
CB396	Single-stage	74.5	72.6	72.9	73.1	73.4
	Two-stage	76.3	74.8	75.2	75.4	75.7

Table 4.7: Comparison of Q_3 accuracy of the single-stage and two-stage MSVM approaches with different type of kernel functions on RS126 and CB396 datasets with $\gamma_1 = \gamma_2 = 0.5$. The neighborhood windows of size 15 and 21 were used in the single-stage and two-stage MSVM approach, respectively.

Parameter γ Gaussian σ	0.5				1.0	1.5	2.0
	0.01	0.05	0.1	0.15	0.01	0.01	0.01
Single-stage	74.2	74.5	74.1	73.9	73.8	73.6	73.5
Two-stage	76.3	76.2	76.0	75.8	76.3	76.1	75.9

Table 4.8: Comparison of Q_3 accuracy of the single-stage and two-stage MSVM approaches with different parameters of Gaussian kernel on the CB396 dataset.

other parameters of the two-stage MSVM approach were empirically determined. The selected values of the kernel and SVM parameters of the previous SVM methods for PSS prediction [176, 89, 75] were used as the initial selection of the proposed method.

Table 4.7 shows PSS prediction accuracies of the MSVM method with the Gaussian RBF (Radial Basis Function) $\mathcal{K}(\mathbf{x}, \mathbf{y}) = e^{-\sigma \|\mathbf{x} - \mathbf{y}\|^2}$, linear kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$, and polynomial kernels $\mathcal{K}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d$ with different $d = 2, 3, 4$, on the RS126 and CB396 datasets with $\gamma_1 = \gamma_2 = 0.5$. The performance of PSS prediction by the single-stage and two-stage MSVM approaches with different parameters of Gaussian kernel on the CB396 dataset is shown in Table 4.8. The best performances were given by the Gaussians with parameters: $\sigma_1 = 0.05$ for the first stage and $\sigma_2 = 0.01$ for the second stage. We used BSVM library [77], which leads to faster convergence for large optimization problem, to implement the multi-class technique.

In Tables 4.9 and 4.10, the results of Zpred, NNSSP, PREDATOR, DSC and Jpred methods on the RS126 and CB396 datasets were obtained from Cuff and Barton [35], and the results of the refined neural network proposed by Riis and Krogh, SVM method of Hua and Sun, dual-layer SVM of Guo, BRNN, and PHD methods were published in their original publications [144, 78, 75, 10, 150]. In Table 4.11, the results of PHD, PHDpsi, Prof_King, and PSIPRED methods were taken from EVA [8].

Figure 4.3 and Table 4.9 show the performance of different secondary structure predictors and two-stage MSVM approach on the RS126 set. The best algorithm was found to be the cascade of two MSVMs with the PSI-BLAST profiles, which achieved 78.0% of Q_3 accuracy while the prediction accuracy made by single-stage MSVM method was 76.2%. On the RS126 set, the Q_3 accuracy of two-stage MSVMs based on the PSI-BLAST profiles was significantly higher than results of Zvelebil *et al.*'s Zpred

Figure 4.3: Comparison of Q_3 accuracy of different predictors in protein secondary structure prediction on RS126 dataset of 126 nonhomologous globular proteins. The notation (ext) indicates that the corresponding method uses position specific scoring matrices generated by PSI-BLAST instead of multiple sequence alignments.

(66.7%) [184], Salamov and Solovyev's NNSSP (72.7%) [156], Frishman and Argos's PREDATOR (70.3%) [60], King and Sternberg's DSC (71.1%) [92], the refined neural network proposed by Riis and Krogh (71.3%) [144], Baldi *et al.*'s BRNN (72.0%) [10], and Cuff and Barton's Jpred (74.8%) [35]. Comparing two-stage MSVMs over two multi-layer perceptron networks of PHD method proposed by Rost and Sander [150], a substantial gain of 7.2% of Q_3 accuracy was observed. Compared to SVM method of Hua and Sun [78], the two-stage MSVM method obtained 6.8% higher Q_3 score. Moreover, the Q_3 accuracy of two-stage MSVMs outperformed the result of the single-stage SVM method of Kim and Pard [89] that was the highest published scores for

Figure 4.4: Bar graph showing the distribution of Q_3 scores obtained by two-stage MSVMs for the benchmark RS126 set based on PSI-BLAST profiles.

Figure 4.5: Comparison of Q_3 accuracy of different predictors in protein secondary structure prediction on CB396 dataset of 396 nonhomologous proteins.

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
Zvelebil <i>et al.</i> (Zpred)	66.7	-	-	-	-	-	-	-
Rost and Sander (PHD)	70.8	-	72.0	66.0	72.0	0.60	0.52	0.51
Salamov <i>et al.</i> (NNSSP)	72.7	-	-	-	-	-	-	-
Frishman <i>et al.</i> (PREDATOR)	70.3	-	-	-	-	-	-	-
King and Sternberg (DSC)	71.1	-	-	-	-	-	-	-
Riis and Krogh	71.3	-	68.9	57.0	79.2	0.59	0.52	0.50
Baldi <i>et al.</i> (BRNN)	72.0	-	-	-	-	-	-	-
Cuff and Barton (Jpred)	74.8	-	-	-	-	-	-	-
Hua and Sun (SVM)	71.2	-	73.0	58.0	75.0	0.61	0.51	0.52
Kim and Park (SVM)	76.1	72.0	77.2	63.9	81.5	-	-	-
Single-stage MSVM	76.2	68.8	69.6	63.5	84.1	0.67	0.60	0.56
Two-stage MSVMs	78.0	72.6	73.1	65.7	83.8	0.69	0.62	0.59

Table 4.9: Comparison of performances of single-stage and two-stage MSVM approaches with other methods in protein secondary structure prediction on RS126 dataset with position specific scoring matrices generated by PSI-BLAST and the following reduction: H, G to (H); E, B to (E); the remainder to (C). The notation - indicates that the corresponding results were not available from the literature.

any previous methods on the RS126 set.

Figure 4.4 presents the distributions of Q_3 scores obtained by two-stage MSVMs for the benchmark RS126 dataset based on PSI-BLAST profiles.

Figure 4.5 and Table 4.10 show the performance of two-stage MSVMs and other methods with the CB396 dataset based on multiple sequence alignments and PSI-BLAST profiles. Two-stage MSVMs with PSI-BLAST profiles achieved 76.3% of Q_3 accuracy that is the highest published scores on the CB396 set to date. Compared to the newest method of Guo *et al.* using dual-layer SVM [75], the two-stage MSVM

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
Zvelebil <i>et al.</i> (Zpred)	64.8	-	-	-	-	-	-	-
Salamov <i>et al.</i> (NNSSP)	71.4	-	-	-	-	-	-	-
Frishman <i>et al.</i> (PREDATOR)	68.6	-	-	-	-	-	-	-
King and Sternberg (DSC)	68.4	-	-	-	-	-	-	-
Cuff and Barton (Jpred)	72.9	-	-	-	-	-	-	-
Guo <i>et al.</i> (Dual-Layer SVM)	74.0	-	79.3	69.3	72.0	-	-	-
Single-stage MSVM	74.5	69.5	68.5	62.0	82.4	0.61	0.59	0.55
Two-stage MSVMs	76.3	73.2	70.6	63.4	83.4	0.63	0.62	0.57

Table 4.10: Comparison of performances of single-stage and two-stage MSVM approaches in PSS prediction on the CB396 dataset with PSI-BLAST profiles.

method significantly obtained 2.3% higher Q_3 score. A direct comparison of Sov accuracy between the two methods was not possible because we used the Sov measure of Zemla *et al.* [181] which makes evaluation of secondary structure prediction more structurally meaningful than Sov94 measure [153] used in the Guo *et al.*'s method. On the CB396 set, the improvement of prediction accuracy of two-stage MSVMs was much more than results of Zvelebil *et al.*'s Zpred (11.5%) [184], Salamov and Solovyev's NNSSP (4.9%) [156], Frishman and Argos's PREDATOR (7.7%) [60], King and Sternberg's DSC (7.9%) [92], and Cuff and Barton's Jpred (3.4%) [35]. As shown, the prediction accuracy of two-stage MSVMs outperformed the result of single-stage MSVM method for PSS prediction.

Table 4.11 shows the performance of two-stage MSVMs on a set of 64 proteins provided by EVAsec with PSI-BLAST profiles. The best prediction was achieved to be

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
Rost and Sander (PHD)	72.3	68.9	65.4	37.8	67.7	0.65	0.67	0.51
Przybylski and Rost (PHDpsi)	72.7	69.1	65.8	37.8	67.8	0.65	0.67	0.51
Ouali and King (Prof_King)	71.5	69.1	60.1	39.6	73.2	0.62	0.68	0.50
Jones (PSIPRED)	77.4	75.8	73.4	36.0	72.3	0.70	0.69	0.55
Single-stage MSVM	78.0	70.1	82.8	73.6	80.5	0.70	0.70	0.60
Two-stage MSVMs	79.4	73.8	83.8	74.5	80.2	0.72	0.71	0.61

Table 4.11: Comparison of performances of two-stage MSVM approach based on position specific scoring matrices generated by PSI-BLAST, with other methods on the EVAsec dataset.

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
Jones (PSIPRED)	78.3	-	-	-	-	-	-	-
Two-stage MSVMs	79.4	76.2	79.0	62.2	82.1	0.67	0.60	0.61

Table 4.12: Comparison of performances of two-stage MSVM approach with PSIPRED method on PSIPRED dataset.

Amino acid	Occurrence (%)	Error in PSS prediction (%)	Helix (%)	Strand (%)	Coil (%)
<i>Non-polar R group (hydrophobic)</i>					
Gly G	7.8	21.3	14.0	8.9	77.1
Ala A	8.8	21.0	47.9	14.0	38.1
Val V	6.9	20.2	30.9	42.7	26.4
Leu L	8.6	21.4	47.1	24.9	28.0
Ile I	5.5	20.6	36.4	37.8	25.8
Met M	2.1	20.5	45.3	20.9	33.8
Pro P	4.6	21.9	14.0	6.8	79.2
<i>Aromatic R group (hydrophobic)</i>					
Phe F	3.9	23.6	33.6	30.9	35.5
Trp W	1.5	26.2	33.0	27.0	40.0
Tyr Y	3.7	25.1	34.6	28.2	37.2
<i>Polar, uncharged R group (hydrophilic)</i>					
Ser S	6.1	24.3	27.2	13.8	59.0
Thr T	5.9	25.3	25.4	24.3	50.3
Cys C	1.5	28.0	26.0	25.8	48.2
Asn N	4.7	22.4	24.1	9.9	66.0
Gln Q	3.7	22.4	45.1	13.7	41.2
<i>Positively R charged (hydrophilic)</i>					
Lys K	5.8	23.3	38.6	14.7	46.7
Arg R	4.7	23.3	42.0	18.2	39.8
His H	2.2	27.0	28.3	19.1	52.6
<i>Negatively R charged (hydrophilic)</i>					
Asp D	6.0	22.7	28.1	8.5	63.4
Glu E	6.1	22.6	49.3	12.5	38.3

Table 4.13: The properties of 20 amino acids: their average occurrences, the error in PSS prediction, and probabilities of α -helix, β -strand, and coil on the CB396 dataset.

Dataset	Number of proteins	Memory space (MB)	Training time (minute)
<i>First stage</i>			
RS126	30	49	15
CB396	396	132	60
PSIPRED	2245	975	2700
<i>Second stage</i>			
RS126	30	6	3
CB396	396	15	10
PSIPRED	2245	115	120

Table 4.14: The training time and memory space of the first stage and the second stage MSVM for the RS126, CB396, and PSIPRED datasets on a supercomputer, Compaq AlphaServer SC45.

two-stage MSVMs: 79.4% of Q_3 while the accuracy of single-stage MSVM was 78.0%. On the testing set, the Q_3 accuracy of two-stage MSVMs was substantial higher than results of PHD (72.3%), Prof_King (71.5%), PHDpsi (72.7%), and PSIPRED (77.4%) methods. By using the MSVM technique to enhance the prediction of the MSVM secondary structure scheme at the first stage, the new prediction schemes achieved $\sim 1\%$ - 2% of improvement in the Q_3 accuracy and $\sim 4\%$ of improvement in the Sov accuracy, compared to results of the single-stage MSVM.

As shown in Table 4.12, we performed an extensive experiment on the PSIPRED dataset for PSS prediction using two-stage MSVMs to address the issue of scalability. For our knowledge, the PSIPRED dataset with 2245 protein chains is the largest dataset used for training and testing of PSS prediction, which is available at [40].

On this dataset, two-stage MSVM approach achieved 79.4% of Q_3 accuracy while the accuracy of PSIPRED method was previously reported to be 78.3% [84].

Table 4.13 lists the properties of 20 amino acids and their average occurrence and error in PSS prediction and probabilities for α -helix, β -strand, and coil on the CB396 dataset. According to the statistical data, amino acids, Val, Ile, and Met were easy to predict while Trp, Cys, and His were difficult to predict by two-stage MSVMs. The statistical data shows that the amino acid residues in the non-polar R group (hydrophobic), Gly, ALa, Val, Leu, Ile, Met, and Pro, were predicted with higher accuracies than the other ones. The results from Table 4.13 suggest that five amino acids, Ala, Met, Gln, Arg, and Glu strongly tend to be α -helix, and Val and Ile tend to be β -strand, while Gly, Pro, Ser, Asn, and Asp are strong coil formers.

4.6.7 Time and Space Complexities

Table 4.14 shows the training time and memory space of the two-stage MSVM approach on the RS126 training set of 126 proteins, the CB396 set of 396 proteins, and PSIPRED set of 2245 proteins. For training phase, the program implementing the two-stage method for PSS prediction was run on the supercomputer. For the first stage, the sizes of RS126, CB396, and PSIPRED training sets are 49MB, 132MB, and 975MB, respectively. It takes 15, 60, and 2700 minutes for training on RS126, CB396, and PSIPRED sets, respectively. For the second stage, the sizes of RS126, CB396, and PSIPRED training sets are 6MB, 15MB, and 115MB, respectively. It takes 3, 10, and 120 minutes for training on RS126, CB396, and PSIPRED sets, respectively. The time of training at the second stage is much faster than that at the first stage because the dimension vectors at the second stage have smaller sizes than those used

at the first stage. For testing phase, the web server for PSS prediction was run on the Itanium server. It takes the execution time of 5 seconds to predict the secondary structure of a query sequence from its PSSM profile on average.

In the training phase, the time complexity of BSVM method used to solve the QP (quadratic programming) problem in the two-stage MSVM approach is (the number of iterations) $\times O(Nm)$ where N is the number of training exemplars and m is the dimension of a feature vector [77]. In the testing phase, if the MSVM predictors have l support vectors then the hyperplane is built in $O(lm)$ time. Then, each prediction takes $O(m)$ time. Totally, the time complexity of the two-stage MSVM approach is (the number of iterations) $\times O(Nm)$.

The first generation methods for PSS prediction such as Chou-Fasman [28] and GOR I [64] have the time complexity of order $O(N)$. In the second generation methods, the time complexity of GOR III [70] and GOR IV [63] approaches is also $O(N)$. The nearest-neighbor methods [94, 158, 179] have the time complexity of order $O(N(m + k))$ where m is the dimension of a feature vector, and k is the number of selected nearest neighbors. Therefore, the running time of the two-stage MSVM approach is slower than those of the first and second generation methods for PSS prediction. In PHD method using two MLPs in cascade [149], the back-propagation algorithm has an algorithmic complexity as the order of W computations for a network having W weights to learn one training sequence. Totally, the time complexity of PHD method is (the number of iterations) $\times O(NW)$. The back-propagation algorithm, though a very popular training procedure, has been known as a slow and uncertain training process, and thus the number of iterations cannot be known exactly. As a result, the comparison of the time complexity of the two-stage MSVMs

with that of PHD method is not possible.

4.7 Summary

We have introduced a two-stage MSVM approach to PSS prediction. With two-stage approaches, the accuracy of prediction is improved because secondary structure at a particular position of a sequence depends not only on the amino acid residue at a particular location but also on the structural formations of the rest of the sequence. Single-stage methods cannot capture this intrinsic relation [149]. Therefore, another layer of classifiers, which predicts the output of single-stage methods, improves the accuracy of prediction. This has also been shown in PHD approach with MLPs but with less prediction accuracy. MLPs are not optimal for this because they cannot generalize the prediction for unseen patterns. As shown in Theorem 4.3.5, the MSVM method was an optimal classifier for the second stage because it minimized not only the empirical risk of known sequences but also the actual risk of unknown sequences. Additionally, two stages were proven to be sufficient to find an optimal classifier for PSS prediction as the MSVM minimized the generalization error of the output of single-stage by solving the optimization problems at second stage.

Furthermore, we have compared two-stage SVM techniques for PSS problem: one method based on binary classifications of Guo [75] and the other approach for multi-class problem by solving one single optimization problem. We found that the two-stage MSVM method is more suitable for secondary structure prediction because of its capacities to lead faster convergence for large and complex training sets of PSS problem and to solve the optimization problem in one step.

Recently, Meiler and Baker [108] have used information of 3-D structures and the

position specific scoring matrices generated by PSI-BLAST as inputs to a neural network for PSS prediction. By combining 3-D structure information, the new approach reported 4%-5% higher accuracies than the methods using only the local information of amino acid sequences. We cannot compare the performance of our method because the results on RS126 and CB396 datasets with this approach are not available. As an avenue for future research, the combination information of 3-D structure and PSI-BLAST profiles with two-stage SVMs may achieve higher prediction accuracies.

Chapter 5

Discussion and Conclusion

This study started by defining the problem of protein structure prediction, which is currently one of the major goals of bioinformatics research, and then rendering it into Relative Solvent Accessibility (RSA) and Protein Secondary Structure (PSS) predictions. Determining a protein's structure solely from its amino acid sequence would greatly help us understand which structure is responsible for what function. As a result, we could add or remove functions in existing proteins by changing their structure, or synthesize new proteins to obtain desired functions. For example, determining the structures of viral proteins would help us design drugs for specific viruses. However, we still know only a few thousand protein structures, and the current theories of how amino acid sequences fold into 3-D structures are incomplete and insufficient to explain the sequence-to-structure mappings. For the past 30 years, a major focus in this research has been on secondary structures, which are the most regular substructures and occur in almost every protein.

Now, we are in a position to return to our initial proposition that in particular, Support Vector Machines (SVMs) are powerful and generally applicable tools

in computational biology. This is because many biological problems involve high-dimensional and noisy data, for which SVMs are known to behave well compared to other statistical or machine learning methods. Moreover, SVMs have the capacity to minimize the risk of an already prediction. In the previous chapters, we showed that the two-stage SVM approaches have better performances than the existing techniques for RSA and PSS predictions. In what follows, we propose to use SVMs as a second stage of the existing approaches for RSA and PSS predictions to improve the prediction accuracy.

5.1 Two-Stage Approaches to Protein Secondary Structure Prediction

Bioinformatics techniques to PSS prediction are mostly single-stage approaches; they predict secondary structures of the protein by taking into account only the information available in amino acid sequences. In this section, we show that it is feasible to extend a variety of existing single-stage approaches for PSS prediction problem by introducing Multi-class Support Vector Machine (MSVM) as a second stage, thereby improving the prediction accuracy. The purpose of the second stage is to capture the contextual relationship of secondary structure elements in a neighborhood in determining the secondary structure element at a particular site. We demonstrate our approach by introducing MSVM to the output of single-stage classifiers, such as GOR and Bayesian approaches.

In both techniques, GOR and Bayesian (referred in Chapter 2), the probabilities of a particular site belonging to secondary structural elements are obtained and the

Figure 5.1: Illustration of two-stage approaches with Multi-class Support Vector Machine (MSVM) at the second stage for protein secondary structure prediction.

secondary element is selected based on the maximum probability. The error or risk made in making such a decision is the Bayesian risk [49]. However, in bioinformatics, the sequences that need to be predicted are usually unseen before and therefore a procedure that has a good generalization ability is more useful. Above all, the probabilities that obtained in these techniques may not be accurate: the GOR techniques may suffer from the limited window length and the errors may be introduced in the Bayesian approach due to imperfect selection of the prior models. It is possible to obtain a higher prediction accuracy if the Bayesian risk is further minimized. This

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
<i>Single-Stage</i>								
GOR I	66.4	59.2	69.8	44.1	72.2	0.49	0.47	0.43
GOR III	66.9	50.7	74.4	45.1	71.2	0.50	0.47	0.42
GOR IV	65.3	61.4	71.3	56.3	66.1	0.57	0.42	0.41
Bayesian	66.3	59.7	74.1	51.1	68.6	0.55	0.41	0.42
MSVM	76.2	68.8	69.6	63.5	84.1	0.67	0.60	0.56
<i>Two-Stage</i>								
GOR I - MSVM	69.2	62.5	60.4	53.1	80.0	0.56	0.50	0.46
GOR III - MSVM	72.7	66.0	66.3	57.6	80.7	0.60	0.55	0.51
GOR IV - MSVM	71.3	64.9	66.8	56.1	80.4	0.61	0.50	0.48
Bayesian - MSVM	70.9	65.2	77.0	57.7	73.4	0.60	0.51	0.47
MSVM - MSVM	78.0	72.6	73.1	65.7	83.8	0.69	0.62	0.59
PHD	70.8	-	72.0	66.0	72.0	0.60	0.52	0.51

Table 5.1: Comparison of performances of single-stage and two-stage approaches with MSVM at the second stage in PSS prediction on the RS126 dataset.

risk could be minimized by incorporating the contextual information of the output sequences, which is the tenet of our approach. Therefore, we choose MSVM to connect to the output of the existing predictors because of the ability of MSVM to guarantee the minimum risk detection and to generalize for unseen data.

Consider a window of w_2 size at a site of the output sequence of the first stage; the vector at position i , $\mathbf{d}_i = (d_{i-h_1^2}^k, d_{i-h_1^2+1}^k, \dots, d_i^k, \dots, d_{i+h_2^2}^k)$ where $k \in \Omega_T$, $w_2 = 3(h_1^2 + h_2^2 + 1)$, and d_i^k is the probability output, $P(t_i = k|\mathbf{r}_i)$, of GOR methods or $P(t_i = k|\mathbf{r}, \theta)$ of the Bayesian technique. The second stage combines the output of the first stage methods to predict PSS. Figure 5.1 illustrates two-stage approach with

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
<i>Single-Stage</i>								
GOR I	64.3	58.5	73.0	41.6	66.1	0.44	0.43	0.43
GOR III	66.0	52.3	75.9	43.2	66.4	0.46	0.45	0.44
GOR IV	65.2	62.9	71.3	55.0	62.2	0.53	0.42	0.44
MSVM	74.5	69.5	68.5	62.0	82.4	0.61	0.59	0.55
<i>Two-Stage</i>								
GOR I - MSVM	67.2	63.3	63.6	51.7	75.5	0.50	0.47	0.47
GOR III - MSVM	71.3	67.1	68.6	55.1	77.8	0.57	0.52	0.41
GOR IV - MSVM	70.9	66.5	68.1	54.0	78.9	0.57	0.51	0.51
MSVM - MSVM	76.3	73.2	70.6	63.4	83.4	0.63	0.62	0.57
PHD	71.9	-	-	-	-	-	-	-

Table 5.2: Comparison of performances of single-stage and two-stage approaches with MSVM at the second stage in PSS prediction on the CB396 dataset.

MSVM at the second stage. In the MSVM technique, the input vectors, $\mathbf{d}_i$'s, are directly used as inputs to another MSVM classifier at the second stage. The outputs of second stage MSVM are combined by a winner-take-all scheme.

Tables 5.1 and 5.2 show the performance of single-stage and two-stage approaches with MSVM at the second stage in PSS prediction on the RS126 dataset of 126 non-homologous globular proteins and the CB396 dataset of 396 nonhomologous proteins, respectively. By using MSVM to predict the single-stage methods, the new prediction schemes significantly achieve $\sim 2\%$ - 6% and $\sim 3\%$ - 15% of improvements in Q_3 and Sov accuracies, respectively, on the RS126 and CB396 datasets. From these experiments, we infer that the two-stage MSVM approach is still the best prediction method on the benchmark amino acid sequences.

5.2 Two-Stage SVM Approaches for RSA and PSS Predictions

We introduced two-stage SVMs for RSA and PSS prediction of residues from amino acid sequences. The accuracy of prediction is improved with two-stage approaches because the solvent accessibility or secondary structure at a particular residue of a sequence depends not only on the amino acid residues in the neighborhood but also on the information of structural formations, such as solvent accessibilities or secondary structures, of the residues in the neighborhood. These intrinsic relations cannot be captured by using single-stage approaches alone. Therefore, another classifier predicting the RSA or PSS types from the output of a single-stage approach enhances the accuracy of the prediction.

For our knowledge, Kim and Park [89] reported the highest accuracies previously by using the position specific scoring matrices generated by PSI-BLAST as inputs to SVM for RSA prediction. By combining PSI-BLAST profiles, this approach achieved better results than methods using information from single sequences and multiple sequence alignments. Compared to the method of Kim and Park, two-stage SVMs showed considerable improvement in the prediction accuracy [120].

We compared two-stage SVM techniques for PSS prediction: one based on binary SVM [75] and the other based on MSVM which solves the multi-class classification in a single optimization step. We demonstrated that the two-stage MSVM approach is more suitable for PSS prediction because of its capacity to solve the optimization problem in one step. By incorporating the state-of-the-art single-stage prediction

technique, MSVM, the two-stage MSVMs achieved better accuracies than the previously reported accuracies to PSS prediction on the tested data [121].

As proved analytically, two-stage SVMs have the best generalization ability for RSA and PSS predictions, by minimizing the generalization error made in the first stage SVMs. Therefore, two-stage SVMs should accurately predict solvent accessibilities or secondary structures for previously unknown amino acid sequences. Two-stage approach for PSS prediction was first introduced in PHD approach which uses two MLPs in cascade for PSS prediction. MLPs are not optimal classifiers in terms of generalization capabilities over unseen input patterns. Further, a cascade of two MLPs effectively increases the input window size of sequences for prediction. On the other hand, SVMs provide classifiers with optimal margins and hence have the best generalization capabilities among the classifiers. Additionally, SVMs are optimal classifiers for the second stage, and two stages are proven to be sufficient to find an optimal classifier for RSA or PSS prediction.

However, since this scenario could not be compared with the other techniques as they stick to the seven-fold cross validation for evaluation, which does not test true generalization capabilities. Further, our comparisons with the other techniques were not complete due to the inaccessibility of previously used data and programs. Also, the kernels and SVM parameters were empirically determined as there does not exist any simple method to find them otherwise. Investigation on two-stage SVMs parameters could further enhance accuracies.

Since many methods for RSA and PSS predictions have used different datasets for evaluation as well as many authors have not published their datasets, it is actually difficult to provide an objective comparison of the prediction of the two-stage

SVM approach with such methods. Furthermore, all RSA techniques only classify amino acid residues into buried and exposed types based on different RSA thresholds. Thus, it is difficult to compare how important the accuracies obtained from different methods for any subsequent applications.

The performance of the two-stage SVM approach based on PSI-BLAST profiles for a novel amino acid sequence suffers from the lack of homologous structures of that sequence in the training set. For a very novel protein whose homologous proteins are not involved in training, the two-stage SVM method predicts its solvent accessibilities and secondary structures at a low accuracy. For our knowledge, Rost and Adamczak [151, 1] concluded that the overall performance of any method, based on evolutionary profiles as inputs, suffers when very remote or no homologues are included. Moreover, the two-stage SVM approach takes a considerable time to train large datasets for RSA and PSS predictions.

5.3 Directions of Future Work

As a framework, the two-stage SVM techniques can be extended further by incorporating more knowledge and requirements from many different aspects of data representation and analysis to build new applications for the prediction of the structure and function of amino acid sequences. This section expounds promising avenues to continue the present research.

5.3.1 Classification of Protein Domains

The classification of 3-D structures now plays a central role in understanding the principles of protein structure, function, and evolution. Classification of new structures can provide functional details through comparison to others, which is of growing importance as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy can now produce structures in advance of biochemical characterization [125]. More generally, structure classifications themselves provide an excellent source of data for analyses of all kinds. Therefore, a strategy for classifying protein structures is really necessary.

After assignment of secondary structures and domains, structural class can be assigned to domains. Structural classes divide proteins according to secondary structure content and organization. Globular proteins are first grouped into four principal classes [95]: all α (or α/α), all β (or β/β), α/β , and $\alpha + \beta$. However, a fifth class, small or irregular, is now generally used to group those proteins with few secondary structures.

Class assignment is usually straightforward for domains with predominantly α -helices or β -sheets. Small elements of secondary structure, such as 3_{10} -helices, or small β -hairpins are usually ignored in the assignment. Class is somewhat subjective, and may be based on the structure of the protein core rather than the abundance of α or β residues. Consider, e.g., staphylococcal nuclease (PDB ID 1snc), which contains an equal proportion of residues in α -helices and β -strands, but it is generally classed as all β because the core of the fold comprises an oligonucleotide/oligosaccharide binding-fold β barrel.

Protein domains that contain a mixture of α -helices and β -sheets are more difficult

to classify. Historically [95], α/β proteins are those containing both α -helices and β -sheets, where there is an intimate association of helices and strands. In contrast, $\alpha+\beta$ proteins define those consisting of segregated regions of helix and sheet. Recently, most exemplified by the Structural Classification of Proteins (SCOP) database, α/β proteins tend to refer to those structures containing many $\beta\alpha\beta$ units, which consist of two adjacent β -strands connected by a single α helix in a right-handed connection, whereas $\alpha+\beta$ proteins are those not falling easily into this definition.

5.3.2 3-D Structure Prediction from Predicted Secondary Structure and Solvent Accessibility

Given good secondary structure and solvent accessibility predictions, the next question to ask is how the secondary structures and/or solvent accessibilities might be arranged into the 3-D structure. *Ab initio* methods for folding secondary into 3-D structure search for possible arrangements of secondary structures that obey general packing rules [167]. These methods have been applied into numerous blind predictions [83, 79] with varied results. A limitation of the techniques is that the number of packing combinations must be considered.

The most successful predictions of protein 3-D structure in the absence of clear sequence similarity to a protein of known structure have been those where secondary structure predictions and experimental information are combined to suggest resemblance to an already known fold. Correct 3-D structures have been predicted in this way for the α -subunit of tryptophan synthase [32], a family of cytokines [12], and recently, for the von Willebrand factor type A domain [51], and the synaptotagmin

C2 domain [66]. Although the details of these studies differed, all methods used predicted secondary structures from evolutionary information, combined with the careful application of protein structural principles (often together with experimental data) to suggest a protein fold. Furthermore, the solvent accessibility prediction gives insight into the organization of 3-D structure: the information of solvent accessibility has improved the prediction of protein subcellular location [5]. The burial of core residues is discovered as a strong driving force in protein folding [22].

We now show how secondary structure and solvent accessibility predictions with the basic rules of protein structure may be used to find the correct fold within a database of protein structural domains. To be successful, a strategy for protein 3-D structure prediction will be able to allow for the following:

1. All possible matches (referred to as maps) between query and database secondary structure and solvent accessibility patterns are generated.
2. Maps are filtered by a series of structural criteria to arrive at a collection of sensible template structures.
3. The best possible template for homology modeling is selected from the set of sensible template structures.

5.3.3 Prediction of Protein-Protein Interactions

Protein-protein interactions play a central role in numerous processes in a cell and are one of the main fields of proteomics [6]. The structures and properties of contact surfaces, forces involved in protein-protein interactions, kinetic and thermodynamic parameters of these reactions provide useful information of functional proteomics.

The properties of protein contact surfaces depend on their functions [52]. The contact surfaces of permanent complexes resemble the domain contacts or the protein core, and, hence it is reasonable to consider such complex formation as a continuation of protein folding. Characteristics of contact surfaces of temporary protein complexes share some similarities with active sites of enzymes. The contact surfaces of the temporary protein complexes have unique structure and characteristics that are more conservative in comparison with active sites of enzymes [6]. Therefore, protein-protein interactions represent prospective targets for a new generation of drugs. The successful prediction of solvent accessibility is helpful in providing necessary information for protein-protein interactions [141].

5.3.4 Prediction of Function of an Unknown Protein

The function of a protein can be defined at various levels, from biochemical through cellular to physiological function. Increasing evidence is presented that many gene products have more than one biochemical function, depending on their biological context [82]. The functions of these “moonlighting” proteins may be modulated by location, pH, ligand availability etc. Such multi-functional proteins will cause complications for genome annotation. During evolution, possibly following gene duplication, one of the gene products may be released from functional constraints, to evolve a new function by the accumulation of local mutations. While the native structure is an absolute requirement for activity, it remains very difficult to predict even the biochemical function from structure except by recognition of similarity to a protein of known function. With the advent of structural genomics as well as the advances in protein structure prediction, structural information will increasingly be used to

provide functional clues to guide experiments.

5.3.5 Summary

In conclusion, we proposed the two-stage SVM approaches for RSA and PSS predictions, which outperform earlier techniques as evidenced by the results on the tested datasets and have better generalization capabilities, that could be used to aid the prediction of the 3-D structure, protein-protein interactions, and function of proteins. Furthermore, two-stage SVM approach can be used for many applications in the field of bioinformatics, involving sequence, such as DNA sequence analysis [9], Human Leukocyte Antigen (HLA) sequence classification [104], solvent accessible surface area (ASA) prediction [114], protein β -turns prediction [132], and protein disorder prediction [50]. This is because computational techniques to these applications are mostly single-stage approaches; they predict output sequences by taking into account only the information available in amino acid or DNA sequences. These single-stage approaches are insufficient to capture the contextual relationships among the elements of the predicted structural sequence. For example, it accounts for the fact that the strands span over at least three adjacent residues and helices consist of at least four consecutive residues in PSS sequences, and the buried or exposed type consists of at least two consecutive residues in RSA sequences. Therefore, another layer of SVM classifier incorporating the contextual relationship among the elements of the output sequence enhances the prediction of single-stage approaches and makes the prediction more realistic in terms of predicted mean lengths of output elements. The analyses of RSA and PSS prediction results from single-stage and two-stage SVM methods showed that the second stage SVM attempts to clean the errors in the prediction

of the first stage SVM, mostly by removing isolated secondary structure or solvent accessibility residue.

Basic probability theory states that all of the evidence relevant to a prediction should be used in making that prediction [81], and, therefore, if a novel prediction method provides new relevant information it should be combined with existing methods to obtain improved predictions. As a result, one prediction system using a hybrid of two-stage SVMs and other approaches could perform better than individual methods.

Since two-stage SVM approach provides a greater biological understanding of the results than simple classification for RSA and PSS predictions, other data mining techniques can be useful in conjunction with the two-stage SVMs to extract relevant features from input sequences and determine more structural and biological information, e.g., constructing the surface of a protein and determining interacting residues in proteins and the prediction of protein-protein interactions.

Appendix

A Tutorial on Two-Stage Support Vector Machines for RSA and PSS Predictions

Introduction

Protein-protein interactions play a central role in numerous processes in biological cells and are one of the major areas of research in proteomics [141]. Understanding the mechanisms of protein-protein interactions is vital when addressing issues associated with the biological function and disease. In addition, protein 3D structure prediction directly from amino acid sequences still remains as an open and important problem in life sciences [110]. Bioinformatics approaches focus on first predicting RSA and/or PSS of a protein's structure which represents the one-dimensional projections of the complicated 3D structure of a protein [24, 110, 151].

Solvent accessibility of an amino acid indicates the level that the residue is accessible to a solvent molecule. The RSA percentage (%) of an amino acid residue is defined as the ratio of the solvent accessible surface area (ASA) of the residue observed in the 3D structure to that observed in an extended tripeptide (Gly-X-Gly or Ala-X-Ala) conformation [171]. The objective of RSA prediction is to classify a

pattern of residues in amino acid sequences to a pattern of RSA types: buried (B) and exposed (E) residues.

The usual goal of secondary structure prediction is to classify a pattern of residues in amino acid sequences to a pattern of protein secondary structure elements: an α -helix (H), β -strand (E) or coil (C, the remaining type).

Figure 5.2: Illustration of the two-stage SVM approach for (a) relative solvent accessibility prediction and (b) protein secondary structure prediction, from amino acid sequences.

Two-Stage SVM Approach

Let us denote the given amino acid sequence by $\mathbf{r} = (r_1, r_2, \dots, r_n)$ where $r_i \in \Omega_R$ and Ω_R is the set of 20 amino acid residues; n is the length of the sequence. The problem

of RSA prediction is, given an amino acid sequence, $\mathbf{r}$, to predict the corresponding sequence of solvent accessibility types, $\mathbf{a} = (a_1, a_2, \dots, a_n)$ where $a_i \in \Omega_A$ denotes the solvent accessibility type of the residue r_i and $\Omega_A = \{B, E\}$ is the set of solvent accessibility types: buried (B) and exposed (E). The problem of PSS prediction is, given an amino acid sequence, $\mathbf{r}$, to predict the corresponding sequence of secondary structures, say $\mathbf{t} = (t_1, t_2, \dots, t_n)$ where $t_i \in \Omega_T = \{H, E, C\}$ is the set of secondary structure elements: helix (H), strands (E), or the remaining type, coil (C).

Firstly, the values of raw matrices of PSI-BLAST [3] to use as inputs to the first stage SVM are obtained from NR database [39]. The low-complexity regions, transmembrane regions, and coil-coil segments are then filtered from the NR database by PFILT program [84]. And an E-value threshold of 0.001 and three iterations are used for searching the non-redundant sequence database to generate position specific scoring matrix (PSSM) profiles. Let $\mathbf{v}_i$ be the vector representing 21-dimensional coding of the residue r_i where 20 units are the values from PSSM profiles ranging from $[0, 1]$ and the 21st unit is used as the padding space to indicate the overlapping end of the sequence; the padding component is set to 1 when padding is required for the end of the sequence or 0, otherwise [84]. Let the input pattern to SVM at site i be $\mathbf{r}_i = (\mathbf{v}_{i-h_1^1}, \mathbf{v}_{i-h_1^1+1}, \dots, \mathbf{v}_i, \dots, \mathbf{v}_{i+h_2^1})$ where $\mathbf{v}_i$ denotes the center element, h_1^1 and h_2^1 denote the width of input segment on the two sides; $w_1 = h_1^1 + h_2^1 + 1$ is the size of the neighborhood window around the center residue r_i .

RSA Prediction

The first stage for RSA prediction consists of a binary SVM classifier, B/E , that maps the input patterns of class B to -1 and the patterns of class E to +1. The

input vectors are transformed to a hidden-space and compared to the support vectors via a kernel function $\mathcal{K}^1$ [173, 174]. Let $\{(\mathbf{r}_j, q_j) : j = 1, 2, \dots, N\}$ denote the set of all training exemplars where q_j denotes the desired classification for the input pattern $\mathbf{r}_j$ and N is the number of training patterns. The first stage SVM attempts to predict solvent accessibility at a particular element of the input sequence by using a neighborhood window, w_1 . The kernel is selected to be Gaussian that achieved better results over the linear and polynomial kernels for RSA prediction [89].

Once the parameters α_j are obtained by maximizing following quadratic function [174]

$$\sum_{j=1}^N \alpha_j - \frac{1}{2} \sum_{j=1}^N \sum_{i=1}^N \alpha_j \alpha_i q_j q_i \mathcal{K}^1(\mathbf{r}_j, \mathbf{r}_i) \quad (5.3.1)$$

subject to $0 \leq \alpha_j \leq \gamma_1$ and $\sum_{j=1}^N \alpha_j q_j = 0$, the resulting discriminant function, say f_1 , is given by

$$\begin{aligned} f_1(\mathbf{r}_i) &= \sum_{j=1}^N q_j \alpha_j \mathcal{K}^1(\mathbf{r}_j, \mathbf{r}_i) + b_1 \\ &= \mathbf{w}_1 \phi^1(\mathbf{r}_i) + b_1 \end{aligned} \quad (5.3.2)$$

where the bias b_1 is chosen so that $q_j f_1(\mathbf{r}_j) = 1$ for any j with $0 < \alpha_j < \gamma_1$, $\mathcal{K}^1(\mathbf{r}_i, \mathbf{r}_j) = \phi^1(\mathbf{r}_i) \phi^1(\mathbf{r}_j)$, and the weight vector $\mathbf{w}_1 = \sum_{j=1}^N q_j \alpha_j \phi^1(\mathbf{r}_j)$. In the classical SVM method, the solvent accessibility type $a_i \in \Omega_A$ corresponding to the residue r_i is determined by

$$a_i = \begin{cases} E & \text{if } f_1(\mathbf{r}_i) \geq 0 \\ B & \text{otherwise} \end{cases} . \quad (5.3.3)$$

The SVM approach takes only inputs from a window of amino acid residues, but does not take into account the interactions among the neighboring RSA elements in the prediction. Although the SVM has the capacity to minimize the generalization

error among all the classifiers, there still exists an error that is not captured by the SVM, which could represent the contextual relationships among the predicted RSA elements. A second SVM predictor is used in two-stage SVM to predict the RSA type of a residue by using the predictions from the first-stage to capture the sequential relationships among the RSA values of residues in the neighborhood.

The second stage SVM in the two-stage SVM processes the output of the discriminant functions of the first stage to enhance the prediction. At the site i , the input to the second SVM is given by a vector $\mathbf{d}_i = (d_{i-h_1^2}, d_{i-h_1^2+1}, \dots, d_i, \dots, d_{i+h_2^2})$ where h_1^2 and h_2^2 are widths of the neighborhoods on the two sides and $d_i = 1/(1 + e^{-f_1(r_i)})$. A neighborhood window, $w_2 = h_1^2 + h_2^2$, of solvent accessibility elements is used as the input to the second-stage SVM which converts the inputs to a higher dimensional space by using a kernel function. The resulting discriminant function at the second stage, f_2 , is computed by solving the convex quadratic programming problem, over all the patterns seen in the training phase, similar to the first stage. The solvent accessibility type a_i corresponding to the residue r_i is given by

$$a_i = \begin{cases} E & \text{if } f_2(\mathbf{d}_i) \geq 0 \\ B & \text{otherwise} \end{cases} . \quad (5.3.4)$$

PSS Prediction

A multi-class SVM scheme has been proposed by Crammer and Singer [31]. For PSS prediction, this method constructs three discriminant functions, each obtained by solving one single optimization problem, which can be formulated as the following quadratic programming problem [31]:

$$\begin{aligned} & \max_{\alpha_j^k} -\frac{1}{2} \sum_{j=1}^N \sum_{i=1}^N \mathcal{K}^1(\mathbf{r}_j, \mathbf{r}_i) \sum_{k \in \Omega_T} \alpha_j^k \alpha_i^k - \sum_{j=1}^N \sum_{k \in \Omega_T} \alpha_j^k c_j^k \\ & \text{such that } \sum_{k \in \Omega_T} \alpha_j^k = 0 \text{ and } \alpha_j^k \leq \begin{cases} 0 & \text{if } t_j \neq k \\ \gamma_1 & \text{if } t_j = k \end{cases} \end{aligned} \quad (5.3.5)$$

where $\mathcal{K}^1(\mathbf{r}_i, \mathbf{r}_j) = \phi^1(\mathbf{r}_i)\phi^1(\mathbf{r}_j)$, $\mathbf{w}_1^k = \sum_{j=1}^N \alpha_j^k \phi^1(\mathbf{r}_j)$, and $c_j^k = \begin{cases} 0 & \text{if } t_j = k \\ 1 & \text{if } t_j \neq k \end{cases}$.

In the single stage multi-class SVM approach, the secondary structural type t_i corresponding to the residue at site i , r_i , is determined by

$$t_i = \arg \max_{k \in \Omega_T} f_1^k(\mathbf{r}_i) \quad (5.3.6)$$

where the resulting discriminant function for secondary type k , $f_1^k = \sum_{j=1}^N \alpha_j^k \mathcal{K}^1(\mathbf{r}_i, \mathbf{r}_j) = \mathbf{w}_1^k \phi^1(\mathbf{r}_i)$ is obtained by solving the above optimization problem.

We extend the multi-class SVM approach by cascading another multi-class SVM at the output of the first stage to improve the accuracy of prediction because of the fact that the secondary structure at a particular position of the sequence depends on the structural elements of the rest of the sequence. For instance, the strands span over at least three adjacent residues and helices consist of at least four consecutive residues [149]. These intrinsic relations are not effectively captured by using single-stage approaches alone. Therefore, a multi-class SVM processes the output from the first stage to minimize the generalization error by incorporating the sequential relationships among the secondary structure elements.

Consider a window, w_2 , of the output sequence of the first stage. The second stage multi-class SVM receives a vector $\mathbf{d}_i$ as the input at site i : $\mathbf{d}_i = (d_{i-h_1^2}^k, d_{i-h_1^2+1}^k, \dots, d_i^k, \dots, d_{i+h_2^2}^k; k \in \Omega_T)$ where $w_2 = 3(h_1^2 + h_2^2 + 1)$ and $d_i^k = 1/(1 + e^{-f_1^k(\mathbf{r}_i)})$. The second

stage multi-class SVM converts the input pattern to a higher dimensional space by using a kernel function. As in the first stage, the resulting discriminant function, f_2^k , at the second stage is obtained by solving the convex quadratic programming problem [31]. The secondary structural type t_i corresponding to the residue r_i is given by

$$t_i = \arg \max_{k \in \Omega_T} f_2^k(\mathbf{d}_i). \quad (5.3.7)$$

Conclusions

With the introduction of two-stage SVM, the accuracies of RSA and PSS predictions of residues from amino acid sequences improved. The two-stage SVM took into account the contextual interactions among secondary structure elements and solvent accessibility values in the prediction and minimized the generalization error made at the first stage. The analyses of the results from two-stage SVM methods shows that the second stage SVM removes the isolated residues and minimizes the generalization error of prediction. The two-stage SVM approach outperformed previous techniques of RSA and PSS predictions as evidenced by the results on the tested datasets and has better generalization capabilities, which could be used to aid the prediction the 3-D structures of proteins and protein-protein interactions.

Bibliography

- [1] R. Adamczak, A. Porollo, and J. Meller, *Accurate prediction of solvent accessibility using neural networks based regression*, Proteins: Structure, Function, and Bioinformatics **56** (2004), 753–767.
- [2] S. Ahmad and M. M. Gromiha, *NETASA: neural network based prediction of solvent accessibility*, Bioinformatics **18** (2002), no. 6, 819–824.
- [3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *Basic local alignment search tool*, Journal of Molecular Biology **215** (1990), 403–410.
- [4] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, *Gapped BLAST and PSI – BLAST: a new generation of protein database search programs*, Nucleic Acids Research **25** (1997), 3389–3402.
- [5] M. A. Andrade, S. I. O’Donoghue, and B. Rost, *Adaptation of protein surfaces to subcellular location*, Journal of Molecular Biology **276** (1998), 517–525.
- [6] A. I. Archakov, V. M. Govorun, A. V. Dubanov, Y. D. Ivanov, A. V. Veselovsky, P. Lewi, and P. Janssen, *Protein-protein interactions as a target for drugs in proteomics*, Proteomics **3** (2003), no. 4, 380–391.
- [7] R. Aurora and G. D. Rose, *Helix capping*, Protein Science **7** (1998), no. 1, 21–38.
- [8] The automatic server EVA for evaluation of different PSS prediction methods, http://cubic.bioc.columbia.edu/eva/sec/res_sec.html.

- [9] P. Baldi and S. Brunak, *Bioinformatics: The machine learning approach*, MIT Press, Cambridge, MA, 1998.
- [10] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, *Exploiting the past and the future in protein secondary structure prediction*, *Bioinformatics* **15** (1999), 937–946.
- [11] P. L. Bartlett and J. Shawe-Taylor, *Generalization performance of support vector machines and other pattern classifiers*, *Advances in Kernel Methods - Support Vector Learning* (1999), 43–54, MIT Press.
- [12] J. F. Bazan, *Structural design and molecular evolution of a cytokine receptor superfamily*, *Proceedings of the National Academy of Sciences of the United States of America* **87** (1990), 6934–6938.
- [13] S. A. Benner and D. L. Gerloff, *Predicting the conformation of proteins: man versus machine*, *FEBS Letters* **352** (1993), 29–33.
- [14] H. M. Berman, J. Westbrook, and Z. Feng, *The protein data bank*, *Nucleic Acids Research* **28** (2000), 235–242.
- [15] V. Biou, J. F. Gibrat, J. M. Levin, B. Robson, and J. Garnier, *Secondary structure prediction: combination of three different methods*, *Protein Engineering* **2** (1988), 185–191.
- [16] E. R. Blout, C. de Loz, S. M. Bloom, and G. D. Fasman, *Dependence of the conformation of synthetic polypeptides on amino acid composition*, *Journal of the American Chemical Society* **82** (1960), 3787–3789.
- [17] E. R. Blout, C. de Lozé, S. M. Bloom, and G. D. Fasman, *The dependence of the conformation of polypeptides and proteins upon amino acid composition*, In *Polyamino Acids, Polypeptides, and Proteins* (1962), 275–279.

- [18] H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, and B. Lautrup, *Protein secondary structure and homology by neural networks*, FEBS Letters **241** (1988), 223–228.
- [19] J. U. Bowie, R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure*, Science **253** (1991), 164–170.
- [20] C. Bystroff, V. Thorsson, and D. Baker, *HMMSTR: a hidden markov model for local sequence-structure correlations in proteins*, Journal of Molecular Biology **301** (2000), 173–190.
- [21] O. Carugo, *Predicting residue solvent accessibility from protein sequence by considering the sequence environment*, Protein Engineering **13** (2000), 607–609.
- [22] H. S. Chan and K. A. Dill, *Origins of structure in globular proteins*, Proceedings of the National Academy of Sciences of the United States of America **87** (1990), 6388–6392.
- [23] J. M. Chandonia and M. Karplus, *Neural networks for secondary structure and structural class predictions*, Protein Science **4** (1995), 275–285.
- [24] ———, *New methods for accurate prediction of protein secondary structure*, Protein Engineering **35** (1999), 293–306.
- [25] C. C. Chang and C. J. Lin, *Libsvm : a library for support vector machines*, <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf> (2003).
- [26] C. C. Chen, J. P. Singh, and R. B. Altman, *Using imperfect secondary structure predictions to improve molecular structure computations*, Bioinformatics **15** (1999), 53–65.
- [27] H. Chen, H. X. Zho, X. Hu, and I. Yoo, *Classification comparison of prediction of solvent accessibility from protein sequences*, The 2nd Asia-Pacific Bioinformatics Conference (2004).

- [28] P. Y. Chou and G. D. Fasman, *Prediction of protein conformation*, Biochemistry **13** (1974), 222–245.
- [29] P. Clote and R. Backofen, *Computational molecular biology*, Wiley and Sons, Ltd., Chichester, 2000.
- [30] JPREP: A consensus method for protein secondary structure prediction, <http://www.compbio.dundee.ac.uk/~www-jpred/submit.html>.
- [31] K. Crammer and Y. Singer, *On the learnability and design of output codes for multiclass problems*, Machine Learning **47** (2002), 201–233.
- [32] I. P. Crawford, T. Niermann, and K. Kirschner, *Predictions of secondary structure by evolutionary comparison: application to the alpha-subunit of tryptophan synthase*, Proteins: Structure, Function, and Bioinformatics **1** (1987), 118–129.
- [33] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines*, Cambridge University Press, 2000.
- [34] G. E. Crooks and S. E. Brenner, *Protein secondary structure: entropy, correlations and prediction*, Bioinformatics **20** (2004), 1603–1611.
- [35] J. A. Cuff and G. J. Barton, *Evaluation and improvement of multiple sequence methods for protein secondary structure prediction*, Proteins: Structure, Function, and Bioinformatics **4** (1999), 508–519.
- [36] ———, *Application of multiple sequence alignment profiles to improve protein secondary structure prediction*, Proteins: Structure, Function, and Bioinformatics **40** (2000), 502–511.
- [37] J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton, *JPred: a consensus secondary structure prediction server*, Bioinformatics **14** (1998), 892–893.

- [38] C. A. Cumbaa, *Modeling protein secondary structure by products of dependent expert*, Master thesis, University of Waterloo (2001).
- [39] NR database for generating Position Specific Scoring Matrix (PSSM) profiles, <ftp://ftp.ncbi.nih.gov/blast/db/fasta/>.
- [40] PSIPRED dataset for Protein Secondary Structure (PSS) prediction, <ftp://bioinf.cs.ucl.ac.uk/pub/psipred/old/data/>.
- [41] RS126 dataset of 126 proteins for Protein Secondary Structure (PSS) prediction, http://www.compbio.dundee.ac.uk/~www-jpred/data/pred_res/126_set.html.
- [42] RS126 dataset of 126 proteins for Relative Solvent Accessibility (RSA) prediction, <http://www.rtc.riken.go.jp/~shandar/netasa/rvp-net/rs-126/>.
- [43] Manesh dataset of 215 proteins for Relative Solvent Accessibility (RSA) prediction, <http://www.rtc.riken.go.jp/~shandar/netasa/rvp-net/manesh-215/>.
- [44] CB396 dataset of 396 proteins for Protein Secondary Structure (PSS) prediction, <http://www.compbio.dundee.ac.uk/~www-jpred/data/>.
- [45] CB513 dataset of 513 proteins for Protein Secondary Structure (PSS) prediction, <http://www.compbio.dundee.ac.uk/~www-jpred/data/>.
- [46] D. R. Davies, *A correlation between amino acid composition and protein structure*, Journal of Molecular Biology **9** (1964), 605–609.
- [47] T. Defay and F. E. Cohen, *Evaluation of current techniques for ab initio protein structure prediction*, Proteins: Structure, Function, and Bioinformatics **23** (1995), 431–445.

- [48] D. Donnelly, J. P. Overington, and T. L. Blundell, *The prediction and orientation of helices from sequence alignments: the combined use of environment-dependent substitution tables, fourier transform methods and helix capping rules*, Protein Engineering **7** (1994), 645–653.
- [49] R. O. Duda, P. R. Hart, and D. G. Stork, *Pattern classification*, John Wiley and Sons, New York, 2001.
- [50] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic, *Intrinsic disorder and protein function*, Biochemistry **41** (2002), no. 21, 6573–6582.
- [51] Y. J. K. Edwards and S. J. Perkins, *The protein fold of the von willebrand factor type a is predicted to be similar to the open twisted beta-sheet flanked by alpha-helices found in human ras-p21*, FEBS Letters **358** (1995), 283–286.
- [52] L. Ehrlich, H. Bohr M. Reczko, and R. C. Wade, *Prediction of water-binding sites on proteins using neural networks*, Protein Engineering **11** (1998), 11–19.
- [53] F. Eisenhaber and P. Argos, *Improved strategy in analytical surface calculation for molecular systems-handling of singularities and computational efficiency*, Journal of Computational Chemistry **14** (1993), 1272–1280.
- [54] V. A. Eyrich, D. M. Standley, A. K. Felts, and R. A. Friesner, *Protein tertiary structure prediction using a branch and bound algorithm*, Proteins: Structure, Function, and Bioinformatics **35** (1999), 41–57.
- [55] V. A. Eyrich, D. M. Standley, and R. A. Friesner, *Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set*, Journal of Molecular Biology **288** (1999), 725–742.

- [56] A. V. Finkelstein and O. B. Ptitsyn, *Statistical analysis of the correlation among amino acid residues in helical, β -structural and non-regular regions of globular proteins*, Journal of Molecular Biology **62** (1971), 613–624.
- [57] D. Fischer and D. Eisenberg, *Fold recognition using sequence-derived properties*, Protein Science **5** (1996), 947–955.
- [58] G. B. Fogel and D. W. Corne, *Evolutionary computation in bioinformatics*, Morgan Kaufmann, 2002.
- [59] V. Di Francesco, J. Garnier, and P. J. Munson, *Improving protein secondary structure prediction with aligned homologous sequences*, Protein Science **5** (1996), 106–113.
- [60] D. Frishman and P. Argos, *Knowledge-based protein secondary structure assignment*, Proteins: Structure, Function, and Bioinformatics **23** (1995), 566–579.
- [61] ———, *Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence*, Protein Engineering **9** (1996), 133–142.
- [62] ———, *75% accuracy in protein secondary structure prediction*, Proteins: Structure, Function, and Bioinformatics **27** (1997), 329–335.
- [63] J. Garnier, J. F. Gibrat, and B. Robson, *GOR method for predicting protein secondary structure from amino acid sequence*, Methods in Enzymology **266** (1996), 541–553.
- [64] J. Garnier, D. J. Osguthorpe, and B. Robson, *Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins*, Journal of Molecular Biology **120** (1978), 97–120.

- [65] O. Gascuel and J. L. Golmard, *A simple method for predicting the secondary structure of globular proteins: implications and accuracy*, Computer Applications in the Biosciences **4** (1988), 357–365.
- [66] D. L. Gerloff, G. Chelvanayagam, and S. A. Benner, *A predicted consensus structure for the protein kinase c2 homology (c2h) domain, the repeating unit of synaptotagmin*, Proteins: Structure, Function, and Bioinformatics **22** (1995), 299–310.
- [67] D. L. Gerloff and F. E. Cohen, *Secondary structure prediction and unrefined tertiary structure prediction for cyclin A, B, and D*, Proteins: Structure, Function, and Bioinformatics **24** (1996), 18–34.
- [68] D. L. Gerloff, T. F. Jenny, L. J. Knecht, G. H. Gonnet, and S. A. Benner, *The nitrogenase mofe protein*, FEBS Letters **318** (1993), 118–124.
- [69] G. Gianese, F. Bossa, and S. Pascarella, *Improvement in prediction of solvent accessibility by probability profiles*, Protein Engineering **16** (2003), no. 12, 987–992.
- [70] J. F. Gibrat, J. Garnier, and B. Robson, *Further developments of protein secondary structure prediction using information theory*, Journal of Molecular Biology **198** (1987), 425–443.
- [71] M.H.M. Giorgi, S. Hazout, and P. Tuffery, *PredAcc: prediction of solvent accessibility*, Bioinformatics **15** (1999), 176–177.
- [72] Y. Guermeur, *Combining discriminant models with new multi-class SVMs*, Pattern Analysis and Applications **5** (2002), no. 2, 168–179.
- [73] Y. Guermeur, C. Geourjon, P. Gallinari, and G. Deleage, *Improved performance in protein secondary structure prediction by inhomogeneous score combination*, Bioinformatics **15** (1999), 413–421.

- [74] Y. Guermeur and D. Zelus, *Combining protein secondary structure prediction methods with a new multi-category SVM*, International conference on Intelligent Systems for Molecular Biology, San Diego (2000).
- [75] J. Guo, H. Chen, Z. Sun, and Y. Lin, *A novel method for protein secondary structure prediction using dual-layer SVM and profiles*, Proteins: Structure, Function, and Bioinformatics **54** (2004), 738–743.
- [76] H. L. Holley and M. Karplus, *Protein secondary structure prediction with a neural network*, Proceedings of the National Academy of Sciences of the United States of America **86** (1989), 152–156.
- [77] C. W. Hsu and C. J. Lin, *A comparison on methods for multi-class support vector machines*, IEEE Transactions on Neural Networks **13** (2002), 415–425.
- [78] S. Hua and Z. Sun, *A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach*, Journal of Molecular Biology **308** (2001), 397–407.
- [79] Z. Huang, J. Gabriel, M. A. Baldwin, R. J. Fletterick, S. B. Prusiner, and F. E. Cohen, *Proposed three-dimensional structure for the cellular prion protein*, Proceedings of the National Academy of Sciences of the United States of America **91** (1994), 7139–7143.
- [80] T. J. P. Hubbard and J. Park, *Fold recognition and ab initio structure predictions using hidden markov models and strand pair potentials*, Proteins: Structure, Function, and Bioinformatics **23** (1995), 398–402.
- [81] E. T. Jaynes, *Probability theory: The logic of science*, <http://omega.albany.edu:8008/JaynesBook.html>, 1994.
- [82] C. J. Jeffery, *Moonlighting proteins*, Trends in Biochemical Sciences **24** (1999), 8–11.

- [83] L. Jin, F. E. Cohen, and J.A. Wells, *Structure from function: screening structural models with functional data*, Proceedings of the National Academy of Sciences of the United States of America **91** (1994), 113–117.
- [84] D. T. Jones, *Protein secondary structure prediction based on position-specific scoring matrices*, Journal of Molecular Biology **292** (1999), 195–202.
- [85] D. Juretic, B. Lee, N. Trinajstic, and R. W. Williams, *Conformational preference functions for predicting helices in membrane proteins*, Biopolymers **33** (1993), 255–273.
- [86] E. A. Kabat and T. T. Wu, *The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: Comparison of predicted and experimental determination of β -sheets in concanavalin*, Proceedings of the National Academy of Sciences of the United States of America **70** (1973), 1473–1477.
- [87] W. Kabsch and C. Sander, *Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features*, Biopolymers **22** (1983), 2577–2637.
- [88] K. Karplus, C. Barrett, and R. Hughey, *Hidden markov models for detecting remote protein homologies*, Bioinformatics **14** (1998), 846C856.
- [89] H. Kim and H. Park, *Protein secondary structure prediction based on an improved support vector machines approach*, Protein Engineering **16** (2003), 553–560.
- [90] ———, *Prediction of protein relative solvent accessibility with support vector machines*, Proteins: Structure, Function, and Bioinformatics **54** (2004), no. 3, 557–562.

- [91] R. D. King, M. Ouali, A. T. Strong, A. Aly, and A. Elmaghraby, *Is it better to combine predictions?*, Protein Engineering **13** (2000), 15–19.
- [92] R. D. King and M. J. Sternberg, *Identification and application of the concepts important for accurate and reliable protein secondary structure prediction*, Protein Science **5** (1996), 2298–2310.
- [93] D. G. Kneller, F. E. Cohen, and R. Langridge, *Improvements in protein secondary structure prediction by an enhanced neural network*, Journal of Molecular Biology **214** (1990), 171–182.
- [94] J. M. Levin, B. Robson, and J. Garnier, *An algorithm for secondary structure determination in proteins based on sequence similarity*, FEBS Letters **205** (1986), 303–308.
- [95] M. Levitt and C. Chothia, *Structural patterns in globular proteins*, Nature **261** (1976), 552–558.
- [96] M. Levitt and R. Sharon, *Accurate simulation of protein dynamics in solution*, Proceedings of the National Academy of Sciences of the United States of America **85** (1988), no. 20, 7557–7561.
- [97] X. Li and X. M. Pan, *New method for accurate prediction of solvent accessibility from protein sequence*, Proteins: Structure, Function, and Bioinformatics **42** (2001), 1–5.
- [98] K. Lin, V. A. Simossis, W. R. Taylor, and J. Heringa, *A simple and fast secondary structure prediction method using hidden neural networks*, Bioinformatics (accepted).
- [99] X. Liu, L. M. Zhang, and W. M. Zheng, *Prediction of protein secondary structure based on residue pairs.*, Journal of Bioinformatics and Computational Biology **2** (2004), no. 2, 343–352.

- [100] Y. Liu, J. Carbonell, J. Klein-Seetharaman, and V. Gopalakrishnan, *Comparison of probabilistic combination methods for protein secondary structure prediction*, Bioinformatics (accepted).
- [101] C. D. Livingstone and G. J. Barton, *Secondary structure prediction from multiple sequence data: blood clotting factor XIII and versinia protein-tyrosine phosphatase*, International Journal of Peptide and Protein Research **44** (1994), 239–244.
- [102] A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, *Prediction of protein structure: the problem of fold multiplicity*, Proteins: Structure, Function, and Bioinformatics **Suppl 3** (1999), 199–203.
- [103] A. Lupas, M. Van Dyke, and J. Stock, *Predicting coiled coils from protein sequences*, Science **252** (1991), 1162–1164.
- [104] J. Ma, M. N. Nguyen, and J. C. Rajapakse, *Classification of HLA sequences through codon usage patterns with SVM*, Bioinformatics (submitted).
- [105] R. Maclin and J. W. Shavlik, *Using knowledge-based neural networks to improve algorithms: refining the chou-fasman algorithm for protein folding*, Machine Learning **11** (1993), 195–215.
- [106] B. Matthews, *Comparison of the predicted and observed secondary structure of t4 phage lysozyme*, Biochimica et Biophysica Acta **405** (1975), 442–451.
- [107] P. K. Mehta, J. Heringa, and P. Argos, *A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%*, Protein Science **4** (1995), 2517–2525.
- [108] J. Meiler and D. Baker, *Coupled prediction of protein secondary and tertiary structure*, Protein Science **100** (2003), no. 21, 12105–12110.

- [109] J. Meiler, M. Mueller, A. Zeidler, and F. Schmaeschke, *Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks*, Journal of Molecular Modelling **7** (2001), no. 9, 360–369.
- [110] D.W. Mount, *Bioinformatics: Sequence and genome analysis*, Cold Spring Harbor Laboratory Press, 2001.
- [111] H. Naderi-Manesh, M. Sadeghi, S. Araf, and A.A.M. Movahedi, *Predicting of protein surface accessibility with information theory*, Proteins: Structure, Function, and Bioinformatics **42** (2001), 452–459.
- [112] D. L. Nelson and M. M. Cox, *Lehninger principles of biochemistry*, Worth Publishers, New York, 2000.
- [113] J. T. Ngo and J. Marks, *Computational complexity of a problem in molecular structure prediction*, Protein Engineering **5** (1992), 313–321.
- [114] M. N. Nguyen and J. C. Rajapakse, *Two-stage SVR approach for predicting accessible surface areas of amino acids*, Proteins: Structure, Function, and Bioinformatics (submitted).
- [115] ———, *Optimal approach to sequence-to-sequence prediction: Applications in bioinformatics*, International Conference on Neural Information Processing (ICONIP), Singapore **5** (2002), 2254–2258.
- [116] ———, *Prediction of protein secondary structure using bayesian method and support vector machines*, International Conference on Neural Information Processing (ICONIP), Singapore **2** (2002), 616–620.
- [117] ———, *Multi-class support vector machines for protein secondary structure prediction*, Genome Informatics **14** (2003), 218–227.

- [118] ———, *Two-stage support vector machines for protein secondary structure prediction*, Neural, Parallel and Scientific Computations **11** (2003), 1–18.
- [119] ———, *Two-stage support vector machines to protein relative solvent accessibility prediction*, IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), San Diego, USA (2004).
- [120] ———, *Prediction of protein relative solvent accessibility with two-stage SVM approach*, Proteins: Structure, Function, and Bioinformatics **59** (2005), 30–37.
- [121] ———, *Two-stage multi-class SVMs for protein secondary structure prediction*, Pacific Symposium on Biocomputing (PSB), Hawaii, USA (2005).
- [122] M. N. Nguyen, J. C. Rajapakse, and S. L. Ho, *Combining GOR techniques with support vector machines for protein secondary structure prediction*, International Conference on Control, Automation, Robotics and Vision (ICARCV), Singapore **3** (2002), 1528–1533.
- [123] W. S. Noble, *Support vector machine applications in computational biology*, Kernel Methods in Computational Biology (2004), 71–92.
- [124] T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga, *Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides*, Proceedings of the National Academy of Sciences of the United States of America **84** (1987), 3086–3090.
- [125] C. A. Orengo, M. B. Swindells, A. D. Michie, M. J. Zvelebil, P. C. Driscoll, M. D. Waterfield, and J. M. Thornton, *Structural similarity between the pleckstrin homology domain and verotoxin: the problem of measuring and evaluating structural similarity*, Protein Science **4** (1995), 1977–1983.

- [126] A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick, *Ab initio folding of proteins using restraints derived from evolutionary information*, Proteins: Structure, Function, and Bioinformatics **Suppl 3** (1999), 177–185.
- [127] M. Ouali and R. D. King, *Cascaded multiple classifiers for secondary structure prediction*, Protein Science **9** (1999), 1162–1176.
- [128] R. H. Pain and B. Robson, *Analysis of the code relating sequence to secondary structure in proteins*, Nature **227** (1970), 62–63.
- [129] S. Pascarella, R. .D. Persio, F. Bossa, and P. Argos, *Easy method to predict solvent accessibility from multiple protein sequence alignments*, Proteins: Structure, Function, and Bioinformatics **32** (1999), 190–199.
- [130] L. Pauling and R. B. Corey, *Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets*, Proceedings of the National Academy of Sciences of the United States of America **37** (1951), 729–740.
- [131] L. Pauling, R. B. Corey, and H. R. Branson, *The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain*, Proceedings of the National Academy of Sciences of the United States of America **37** (1951), 205–234.
- [132] T. H. Pham, K. Satou, and T. B. Ho, *Prediction and analysis of β -turns in proteins by support vector machine*, Genome Informatics **14** (2003), 196–205.
- [133] J. C. Platt, *Fast training of support vector machines using sequential minimal optimization*, in *b. schölkopf, c. burges, a. smola*, Advances in Kernel Methods: Support Vector Machines, MIT Press, Cambridge, MA (1998).

- [134] ———, *Using sparseness and analytic QP to speed training of support vector machines*, Advances in Neural Information Processing Systems 11, MIT Press, Cambridge (1999).
- [135] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, *Large margin DAG's for multiclass classification*, Proc. Advances in Neural Information Processing Systems 12. Cambridge, MA:MIT Press **12** (2000), 547–553.
- [136] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, *Prediction of coordination number and relative solvent accessibility in proteins*, Proteins: Structure, Function, and Bioinformatics **47** (2002), 142–153.
- [137] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, *Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles*, Proteins: Structure, Function, and Bioinformatics **47** (2002), 228–235.
- [138] A. Porollo, R. Adamczak, M. Wagner, and J. Meller, *Maximum feasibility approach for consensus classifiers: Applications to protein structure prediction*, International Conference on Computational Intelligence, Robotics and Autonomous Systems, (CIRAS) (2003).
- [139] D. Przybylski and B. Rost, *Alignments grow, secondary structure prediction improves*, Proteins: Structure, Function, and Bioinformatics **46** (2002), 197–205.
- [140] N. Qian and T. J. Sejnowski, *Predicting the secondary structure of globular proteins using neural network models*, Journal of Molecular Biology **202** (1988), 865–884.

- [141] M. F. Raih, S. Ahmad, R. Zhengc, and R. Mohamed, *Solvent accessibility in native and isolated domain environments: general features and implications to interface predictability*, Biophysical Chemistry **114** (2005), 63–69.
- [142] J. C. Rajapakse and M. N. Nguyen, *Two-stage support vector machines for protein structure learning*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, (First revision submitted).
- [143] ———, *General framework for two-stage approaches to protein secondary structure prediction*, Asia Pacific Biotech News **7** (2003), no. 3.
- [144] S. K. Riis and A. Krogh, *Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignment*, Journal of Computational Biology **3** (1996), 163–183.
- [145] B. Robson and R. H. Pain, *Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of pormation of helical regions*, Journal of Molecular Biology **58** (1971), 237–259.
- [146] G. D. Rose, *Prediction of chain turns in globular proteins on a hydrophobic basis*, Nature **272** (1978), 586–590.
- [147] B. Rost, *PHD: predicting one-dimensional protein structure by profile based neural networks*, Methods in Enzymology **266** (1996), 525–539.
- [148] ———, *Rising accuracy of protein secondary structure prediction*, In Protein structure determination, analysis, and modeling for drug discovery (ed. D Chasman), New York: Dekker (2003), 207–249.
- [149] B. Rost and C. Sander, *Prediction of protein secondary structure at better than 70% accuracy*, Journal of Molecular Biology **232** (1993), 584–599.

- [150] ———, *Combining evolutionary information and neural networks to predict protein secondary structure*, *Proteins* **19** (1994), 55–72.
- [151] B. Rost and C. Sander, *Conservation and prediction of solvent accessibility in protein families*, *Proteins: Structure, Function, and Bioinformatics* **20** (1994), 216–226.
- [152] B. Rost, C. Sander, and R. Schneider, *Progress in protein structure prediction?*, *Trends in Biochemical Sciences* **18** (1993), 120–123.
- [153] ———, *Redefining the goals of protein secondary structure prediction*, *Journal of Molecular Biology* **235** (1994), 13–26.
- [154] B. Rost, R. Schneider, and C. Sander, *Protein fold recognition by prediction-based threading*, *Journal of Molecular Biology* **270** (1997), 471–480.
- [155] R. B. Russell and G. J. Barton, *The limits of protein secondary structure prediction accuracy from multiple sequence alignment*, *Journal of Molecular Biology* **234** (1993), 951–957.
- [156] A. A. Salamov and V. V. Solovyev, *Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignment*, *Journal of Molecular Biology* **247** (1995), 11–15.
- [157] ———, *Protein secondary structure prediction using local alignments*, *Journal of Molecular Biology* **268** (1997), 31–36.
- [158] S. Salzberg and S. Cost, *Predicting protein secondary structure with nearest-neighbor algorithm*, *Journal of Molecular Biology* **227** (1992), 371–374.
- [159] R. Samudrala, E. Huang, P. Koehl, and M. Levitt, *Constructing side chains on near-native main chains for ab initio protein structure prediction*, *Protein Engineering* **13** (2000), 453–457.

- [160] R. Samudrala, Y. Xia, E. Huang, and M. Levitt, *Ab initio protein structure prediction using a combined hierarchical approach*, Proteins: Structure, Function, and Bioinformatics **Suppl 3** (1999), 194–198.
- [161] H. A. Scheraga, *Structural studies of ribonuclease: A model for the secondary and tertiary structure*, Journal of the American Chemical Society. **82** (1960), 3847–3852.
- [162] S. C. Schmidler, J. S. Liu, and D. L. Brutlag, *Bayesian segmentation of protein secondary structure*, Journal of Computational Biology **7** (2000), no. 1, 233–248.
- [163] B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, *Comparing support vector machines with gaussian kernels to radial basis function classifiers*, IEEE Transactions on Signal Processing **45** (1997), 2758–2765.
- [164] J. Selbig, T. Mevissen, and T. Lengauer, *Decision tree-based formation of consensus protein secondary structure prediction*, Bioinformatics **15** (1999), 1039–1046.
- [165] V. V. Solovyev and A. A. Salamov, *Predicting helix and strand segments of globular proteins*, Computer Applications in the Biosciences **10** (1994), 661–669.
- [166] P. Stolorz, A. Lapedes, and Y. Xia, *Predicting protein secondary structure using neural net and statistical methods*, Journal of Molecular Biology **225** (1992), 363–377.
- [167] S. Sun, P. D. Thomas, and K.A. Dill, *A simple protein folding algorithm using a binary code and secondary structure constraints*, Protein Engineering **8** (1995), 769–778.

- [168] M. J. Thompson and R. A. Goldstein, *Predicting solvent accessibility: higher accuracy using bayesian statistics and optimized residue substitution classes*, Proteins: Structure, Function, and Bioinformatics **47** (1996), 142–153.
- [169] J. M. Thornton, T. P. Flores, D. T. Jones, and M. B. Swindells, *Prediction of progress at last*, Nature **354** (1992), 105–106.
- [170] R. Unger and J. Moult, *Finding the lowest free energy conformation of a protein is an NP-hard problem: Proof and implications*, Bulletin of Mathematical Biology **55** (1993), no. 6, 1183–1198.
- [171] U.Samanta, R.P. Bahadur, and P.Chakrabarti, *Quantifying the accessible surface area of protein residues in their local environment*, Protein Engineering **15** (2002), 659–667.
- [172] V. Vapnik, *Estimation of dependences based on empirical data*, Springer-Verlag, New York, 1982.
- [173] V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, New York, 1995.
- [174] ———, *Statistical learning theory*, Wiley and Sons, Inc., New York, 1998.
- [175] V. N. Viswanadhan, B. Denckla, and J. N. Weinstein, *New joint prediction algorithm (Q7 – JASEP) improves the prediction of protein secondary structure*, Biochemistry **30** (1991), 11164–11172.
- [176] J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones, *Secondary structure prediction with support vector machines*, Bioinformatics **19** (2003), no. 13, 1650–1655.
- [177] The web server for protein relative solvent accessibility prediction using two-stage SVM approach, <http://birc.ntu.edu.sg/~pas0186457/rsa.html>.

- [178] The web server for protein secondary structure prediction using two-stage MSVM approach, <http://birc.ntu.edu.sg/~pas0186457>.
- [179] T. Yi and E. S. Lander, *Protein secondary structure prediction using nearest-neighbor methods*, Journal of Molecular Biology **232** (1993), 1117–1129.
- [180] Z. Yuan, K. Burrage, and J. Mattick, *Prediction of protein solvent accessibility using support vector machines*, Proteins: Structure, Function, and Bioinformatics **48** (2002), 566–570.
- [181] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost, *A modified definition of sov, a segment based measure for protein secondary structure prediction assessment*, Protein Engineering **34** (1999), 220–223.
- [182] X. Zhang, J. P. Mesirov, and D. L. Waltz, *Hybrid system for protein secondary structure prediction*, Journal of Molecular Biology **225** (1992), 1049–1063.
- [183] W. M. Zheng, *Clustering of amino acids for protein secondary structure prediction*, Journal of Bioinformatics and Computational Biology **2** (2004), no. 2, 333–342.
- [184] M. J. J. M. Zvelebil, G. J. Barton, W. R. Taylor, and M. J.E. Sternberg, *Prediction of protein secondary structure and active sites using the alignment of homologous sequences*, Journal of Molecular Biology **195** (1987), 957–961.