Novel Feature Extraction for Predicting Gram-Positive and Gram-Negative Bacteria Protein Sub-cellular Localization

Md. Raihan Uddin
ID : 012 152 009

A Thesis
in
The Department
of
Computer Science and Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Science in Computer Science and Engineering
United International University
Dhaka, Bangladesh
November 2017

Declaration

This is to certify that the work entitled "Novel Feature Extraction for Predicting Gram-Positive and Gram-Negative Bacteria Protein Sub-cellular Localization" is the outcome of the research carried out by me under the supervision of Swakkhar Shatabda, Assistant Professor and Undergraduate Program Coordinator.

Md. Raihan Uddin, 012 152 009, MSc in CSE

In my capacity as supervisor of the candidates thesis, I certify that the above statements are true to the best of my knowledge.

Swakkhar Shatabda, Assistant Professor and Undergraduate Program Coordinator

Abstract

Protein sub-cellular localization is defined as predicting the functioning location of a given protein in the cell. It is considered an important step towards protein function prediction and drug design. In this study, we used structural and evolutionary based features to represent the sequences of gram-positive and gram-negative bacteria protein dataset. In recent years all the works done in this sector are based on mainly PSSM features, they have not used SPIDER feature. PSSM gives us mainly evolutionary features and SPIDER gives us mainly structural features. In our study, we have extracted features from both PSSM and SPIDER. We have extracted a lot of features (total 46 features using 10 categories for both dataset) from both PSSM and SPIDER and choosed best features (total 6 features among 46 features for both dataset) among them. Our achieved result shows that SPIDER feature along with PSSM feature gives best performance than previous result with gram-positive benchmark, but gives a little low result for gram-negative benchmark. In this study, we have tried different kinds of classifier but among them Support Vector Machine (SVM) gives the best result.

Acknowledgment

Without the help and support of some people, this work would have not been possible. I would like to express my gratitude to everyone who have helped me throughout the time.

First of all, I would like to thank my respected academic supervisor, Dr. Swakkhar Shatabda, Assistant Professor and Undergraduate Program Coordinator. Then I would like to thank Dr. Iman Dehzangi, Assistant Professor, Department of Computer Science, School of Computer, Mathematical & Natural Sciences, Morgan State University, Baltimore, MD 21251, USA. I am very much lucky to have them as my supervisor. I am very much grateful to them for their continuous guidance, advice and inspiration throughout the whole period of time. Their continuous guidance kept me on the right track and gave me the opportunity to finish my thesis successfully.

My sincere gratitude goes to United International University (UIU). The research could not have been done without the work environment of this university. I am very much thankful to Dr. Nurul Huda for his useful advice during this period of time. Last but not the least, I owe to my family including my parents, my sister, my brothers for their unconditional love and immense emotional support.

Contents

1 Introduction **7**
- 1.1 Protein Sub-Cellular Localization . . . 7
- 1.2 Motivation . . . 9
- 1.3 Research Aims . . . 9
- 1.4 Methodology . . . 10
- 1.5 Research Contribution . . . 10
- 1.6 Thesis Organization . . . 10

2 Background **12**
- 2.1 Biological Background . . . 12
  - 2.1.1 Proteins . . . 12
  - 2.1.2 Position Specific Score Matrices (PSSM) . . . 16
  - 2.1.3 Scoring Protein Interaction Decoys using Exposed Residues (SPIDER) . . . 17
- 2.2 Machine Learning Background . . . 17
  - 2.2.1 Support Vector Machine (SVM) . . . 18
- 2.3 Literature Review . . . 19

3 Materials and Methods **22**
- 3.1 Data Set . . . 22
  - 3.1.1 Gram-Positive Bacteria Protein Dataset . . . 22
  - 3.1.2 Gram-Negative Bacteria Protein Dataset . . . 22
- 3.2 Normalization of PSSM . . . 24
- 3.3 Constructing the Consensus Sequence . . . 25
- 3.4 Feature Extraction Method . . . 26
  - 3.4.1 Occurrence Feature . . . 27
  - 3.4.2 Composition Feature . . . 27
  - 3.4.3 Bi-Gram Feature . . . 28
  - 3.4.4 One-Lead Bi-Gram Feature . . . 30
  - 3.4.5 Tri-Gram Feature . . . 31
  - 3.4.6 Auto Covariance Feature . . . 32
  - 3.4.7 PSSM-AAO Feature . . . 33
  - 3.4.8 PSSM-SD Feature . . . 34

3.4.9 PSSM-SAC Feature 37
3.4.10 Dubchak Feature 38

4 Experimentation, Result and Discussion **41**
4.1 Feature Extraction 41
4.2 Choosing Classifier and Parameter Tunning 41
4.3 Choosing Validation Method 42
4.4 Choosing The Final Features Set 43
4.5 Single Label Classification 44
4.6 Multi Label Classification 44
4.7 Sensitivity, Specificity, MCC and Accuracy 45
4.8 Summary 60

5 Conclusion **61**
5.1 Summary 61
5.2 Limitations 62
5.3 Future Work 62

Chapter 1

Introduction

In this section, we present an overview of the motivation of our research and the aims and objectives identied during literature review. We also give a brief summary of the methodology applied to achieve the research aims and present a summary of the contributions made by our research reported in this thesis.

1.1 Protein Sub-Cellular Localization

Protein sub-cellular localization prediction or just protein localization prediction involves the prediction of where a protein resides in a cell. In general, prediction tools take as input information about a protein, such as a protein sequence of amino acids, and produce a predicted location within the cell as output, such as the nucleus, Endoplasmic reticulum, Golgi apparatus, extracellular space, or other organelles. The aim is to build tools that can accurately predict the outcome of protein targeting in cells. Prediction of protein sub-cellular localization is an important component of bioinformatics based prediction of protein function and genome annotation, and it can aid the identification of drug targets. Protein sub-cellular localization is considered an importent step towards protein function prediction and consequently, drug design.

The predication of protein sub-cellular localization is based on determining the location sites of unknown protein in a cell. A cell consists of many different compartments that are specialized to carry out different tasks [16]. One of the fundamental goals in cell biology is to identify the sub-cellular location site of proteins and their functions [16]. Information about sub-cellular location can provide useful characteristics of its functions. Of all proteins, bacteria proteins are the most important proteins to determine their functions because of its biological aspects which are both harmful and useful [88]. Some bacteria can cause a wide range of diseases while some oth-

ers play the role of catalyst in biological interactions. Somme bacteria are also widely used to produce antibiotics. Bacteria are categorized as a kind of prokaryotic microorganism that can be divided in two groups, gram-positive and gram-negative [71]. Gram-positive bacteria are those that are stained dark blue or violet by gram staining while gram-negative bacteria cannot retain the stain, instead taking up the counter-stain and appearing red or pink [88]. As pointed in a recent review [14], in the last decade or so, a number of web-servers were developed for predicting the subcellular localization of proteins with both single site and multiple sites based on their sequences information alone. They can be roughly classified into two series [14]. One is the PLoc series and the other is iLoc series. The PLoc series contains the six web-servers [71], [18], [72], [19], [73], [74] to deal with eukaryotic, human, plant, Gram positive, Gram negative, and virus proteins, while the iLoc series contains the seven web-servers [90], [21], [87] , [54], [88], [90], [89] to deal with eukaryotic, human, plant, animal, Gram positive, Gram negative, and virus proteins, respectively.

Most proteins can only function in one specific place in the cell (e.g nucleus, membrane) while some other proteins can function in several places in the cell. For a given protein to function properly it needs to be in one or a few specific locations int the cell as it malfunctions in all other places. Therefore, the newly synthesized proteins play a critical role, if only they are placed in their correct sub-cellular compartments [96]. The sub-cellular location of a protein can be determined by various biological experiments, but it is costly and time consuming. The exponential increase in the number of sequenced proteins and the much slower rate of determining protein structure using experimental approaches highlights the demands for a fast computational approach to address this problem as an alternative to experimental methods. Recently computational methods have become increasingly important and recognized. Researches prefer to use predication system to identify the sub-cellular localization of proteins [83], [84], [80], [81]. Fast computational approaches address the problems of costly and time consuming experimental methods. A wide range of pattern recognition approaches has been used to solve sub-cellular localization problem which have given the most promising result. These approaches either involves classifier development or feature extraction development. The performance of pattern recognition technique to address protein sub-cellular localization prediction problem depends on the classification technique as well as features being used. Several classifiers have been developed and analyzed which includes: Artificial Neural Network (ANN), K-Nearest Neighbor (KNN) [31], Bayesian classifiers, Linear Discriminant Analysis (LDA), Hidden Markov Model (HMM), Nave Bayes [29], Support Vector Machine (SVM) [30], [37], and ensemble of classifiers. Amongst these classifiers Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) based classifiers have given the most promising

results [37]. Studies have shown that most significant enhancement in predication system is achieved by developing feature extraction method rather than improving the classifiers.

1.2 Motivation

Previously used features were mainly sequence based or PSSM profile based. In this area of research the researcher did not used SPIDER information which gives structural information. In literature [59] structure based features were used. We believe structure based features play an important role in protein sub-cellular localization as structural based feature contains a lot of information about a protein.

There are several features used in different literatures, but comprehensive analysis of these features were not done. Most of the literatures did not used previous literature features or did not combined new features with old one. We believe combining new features with old one will increase the performance of prediction. We believe if we select best features from combining all the features used previously with all our new extracted features from evolutionary based information or PSSM and structural based information or SPIDER, it will increase the sub-cellular localization prediction rate.

1.3 Research Aims

The main aim of our research is to extract best features that gives best performance for predicting gram-positive and gram-negative bacteria protein sub-cellular localization. Our aim in this research is to extract features from SPIDER as well as extract new features from PSSM. In recent years all the research done in this sector are mainly based on PSSM matrix. In this sector there is no research done using SPIDER spd3 matrix. We think using SPIDER along with PSSM will give good performance. So we have extracted both old and new features from both PSSM and SPIDER, and combined them to get the best features. Finally we have got that, features extracted from PSSM along with features extracted from SPIDER gives higher result than previous literature published in this sector. The aim of our research is enumerated shortly in below:

1. Extract features from SPIDER spd3 matrix.
2. Extract new features from PSSM.
3. Extract the features that were used in different literatures in protein sub-cellular localization field.
4. Combine all the features and choose best features among them.

5. Do a comprehensive analysis.

1.4 Methodology

Considering the scope of the research and the objectives discussed in the last two sections, we focus our goal to develop effective strategies that improve the performance of predicting protein sub-cellular localization for gram-positive and gram-negative bacteria protein dataset. Throughout this research, we have focused on selecting best features. In this sub-cellular prediction problem we have used gram-positive and gram-negative bacteria protein dataset. We have extracted features from both PSSM and SPIDER for both dataset. We have used Naive Baise Classifier (SVM), used 10-fold cross validation and jackknife test for validation. We have calculated single label classification accuracy, locative and absolute accuracy for multi label protein. A conceptual framework for predicting gram-positive and gram-negative bacteria protein is shown in figure 1.1.

1.5 Research Contribution

The key contributions of the research are as follows:

1. We have extracted features from SPIDER, which was not done before.
2. We have extracted features from PSSM which are already reported in various literatures.
3. We have extracted new features from PSSM.
4. Combined all the features and choose best features among them.
5. We have done a comprehensive analysis.

1.6 Thesis Organization

The report is furnished as follows. In chapter 2, we have explained in short about biological background, machine learning background and literature review to understand about protein, protein structure, how sub-cellular locations are defined, how machine learning works in this sector and previous works in this area. In chapter 3, we explain our work and all the necessary definitions to understand the research and explain the related task. In chapter 4, we explain how we have run our experiment, how classifiers are selected, how when which validation method was selected, how final features are chosen and results. In chapter chapter 5, we conclude with a discussion, limitations of our work and a possible outlines for future work.

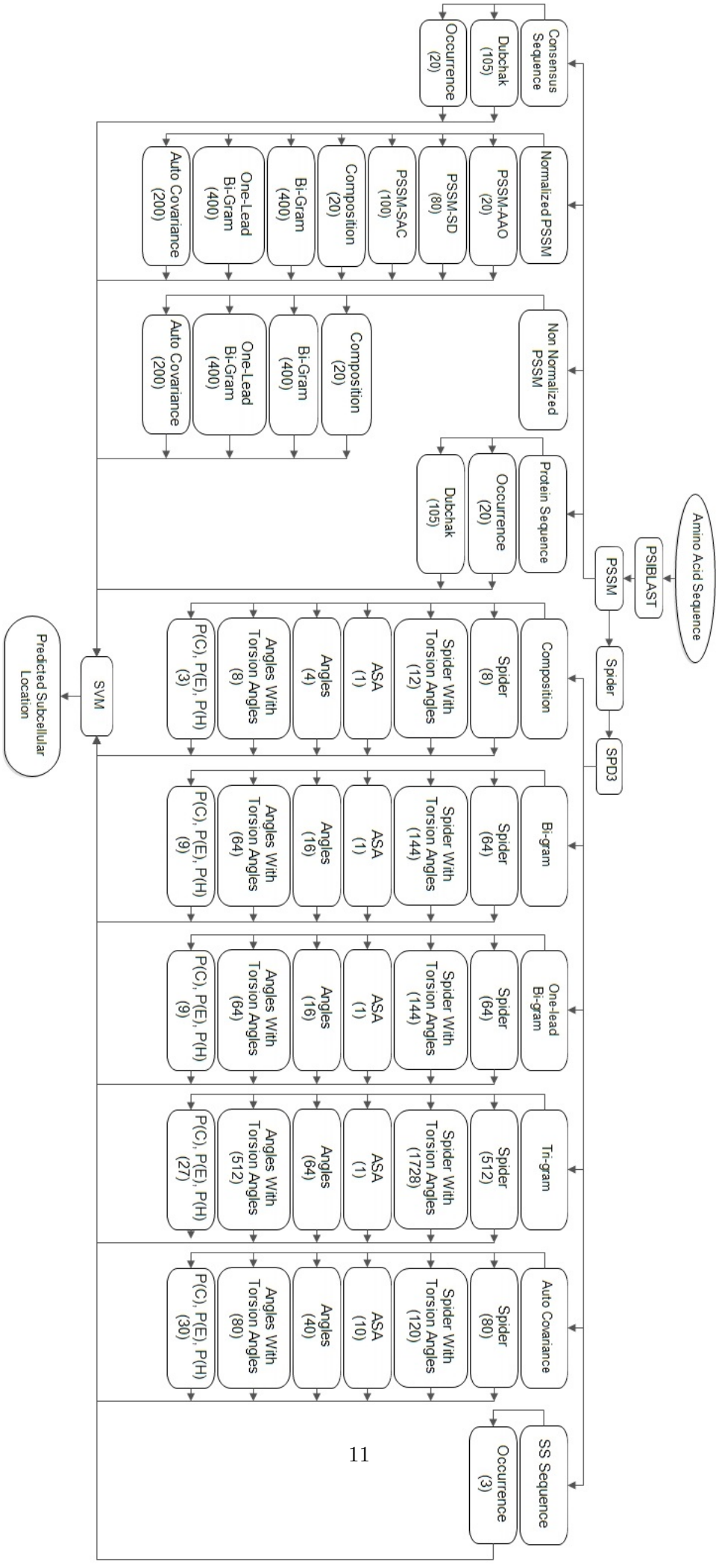

Figure 1.1: Conceptual framework for predicting bacteria proteins

Chapter 2

Background

2.1 Biological Background

2.1.1 Proteins

Proteins are one of the most important biological macromolecules, as they play important parts in almost all biological reactions. Some of the functions performed by proteins include determining the structure and shape of cells, bringing other proteins together so they can perform specific functions more efficiently, catalyzing chemical reactions, acting as signals, sensors and switches to control the activities of cells, and moving other proteins or even whole organisms around [58]. Proteins are polymers of amino acids linked together by peptide bonds. A diagram of an amino acid can be seen in Figure 2.1. An amino acid consists of 3 main groups: an amino group, a carboxyl group and an R group or side chain [58]. The R group determines the type of amino acid, of which there are 20 main types, shown in table 2.1.

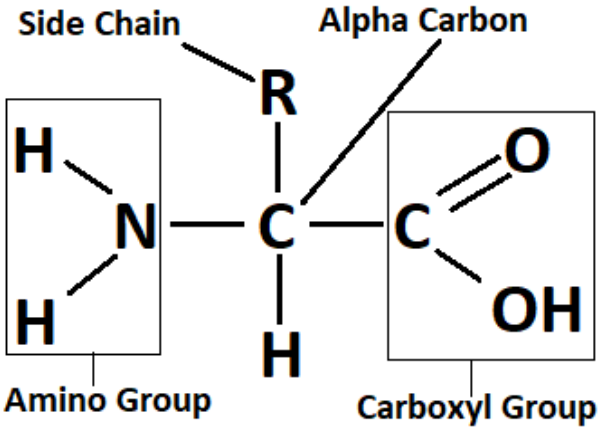

Figure 2.1: Structure of an amino acid

One of the most important concepts regarding proteins is that the function

Table 2.1: A list of the 20 standard amino acids

Amino Acid	3 Letter	1 Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

of a protein is derived from its three dimensional structure [9][54], and the three dimensional structure is determined primarily by the linear sequence of amino acids. Afinsen [2] found that proteins can fold into their native structures spontaneously, from this it was concluded that the folding is governed by the amino acid sequence itself. Understanding how proteins fold in space can reveal how they function in biological reactions. Determining the structure of a protein given its sequence is an open problem of much interest in computational biology.

Protein structure can be broken up into a hierarchy consisting of Primary, Secondary, Tertiary and Quaternary structure. For the sake of example, a short protein has been chosen so that examples of primary, secondary and tertiary structure can be shown.

An example of the primary structure of a protein sequence (Human Re-

laxin, 6rlxA) can be seen below:

$$ELYSALANKCCHVGCTKRSLARFC$$

This consists of a sequence of letters where each letter represents an amino acid from table 2.1. The next level of structure is the secondary structure, which primarily consists of the alpha helix, beta sheet and beta turn. These secondary structures are stable arrangements of short segments of a protein. The secondary structure of 6rlxA can be seen below (taken from PDB):

$$ELYSALANKCCHVGCTKRSLARFC$$

$$HHHHHHHHHHHT\ EEHHHHHTT$$

where H is an alpha helix, E is a beta strand, T is a turn and an empty spot means no secondary structure has been assigned.

Tertiary structure refers to the overall conformation of a ploy-peptide chain. The chemical properties of individual amino acid side chains influence the tertiary structure. The tertiary structure of the example protein 6rlxA is shown in figure 2.2

Figure 2.2: 3D structure of 6rlxA (taken from PDB)

The protein shown in figure 2.2 is a short, relatively simple protein. Proteins can be much longer with extremely complex tertiary structure. In addition, some proteins consist of multiple protein subunits; each subunit has its own tertiary structure, and the arrangement that the subunits assemble in is known as the quaternary structure.

For determining protein structure there are some methods. Among those some are explained here in short:

1. Genetic sequencing projects rapidly produce new protein sequences, but the number of sequences with known 3D structure increases much more slowly. This is because the laboratory methods for determining protein structure, namely Nuclear Magnetic Resonance (NMR) and X-ray crystallography are expensive and time consuming. To illustrate this point, the Universal Protein Resource (UniProt) contains about 8,000,000 protein sequences, while the number of protein structures stored in the Protein Data Bank (PDB) is around 95,000.

2. X-ray crystallography involves firing x-rays at a crystal and measuring the diffraction pattern that results. This allows a 3D picture of the crystal structure to be determined. The protein must be purified and concentrated, and a crystal of the protein formed, this requires a sufficient amount of protein to be present. The process is complex and time consuming, and some proteins are very difficult to crystallize.

3. NMR is another method of determining protein structure. In this case the proteins are in a solution of water. The sample is placed in a powerful magnetic field and radio waves are sent through the sample so that the absorption can be measured. Since different atomic nuclei absorb different frequencies, and the absorption spectra is perturbed by adjacent nuclei both the type and distances between nuclei can be determined. NMR is usually limited to shorter proteins due to limitations of the process.

4. Given the difficulty of experimentally determining protein structure, computational methods are desirable. Abinitio methods build the 3D structure of a protein from physical principles, these methods are based on various optimization schemes trying to minimize the protein free energy. This method is impractical for large proteins and high resolution models due to the large search space.

5. Homology based methods find proteins with known structure that have sequences similar to a protein with unknown structure. It is assumed that the structure of homologous proteins will be similar, and can at least be used as a starting point for inferring the structure of new proteins. Unfortunately, proteins with sequential similarities as low as 15% can have similar tertiary structure, but homology based approaches are unsuitable for these cases.

6. Protein Threading is a method of determining the fold of a protein by comparing it to a set of templates. Threading and homology

based methods are similar, however homology based methods are restricted to proteins with high sequential similarity, while threading is used for proteins with lower sequential similarity. Popular threading methods include HHsearch and SPARKS-X.

2.1.2 Position Specific Score Matrices (PSSM)

PSSM or Position-Specific Scoring Matrix, is a type of scoring matrix used in protein BLAST searches in which amino acid substitution scores are given separately for each position in a protein multiple sequence alignment. PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) [1] derives a position-specific scoring matrix (PSSM) or profile from the multiple sequence alignment of sequences detected above a given score threshold using proteinprotein BLAST [2].

BLAST (Basic Local Alignment Search Tool) is a sequence similarity search method that compares a protein to a set of protein sequences in a target database to identify regions of local alignment and report those alignments that score above a given score threshold [2]. The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. PSI-BLAST uses the results of BLAST run for several iterations, with the best matches from each iteration used in the next iteration.

The PSSM captures the conservation pattern in alignment and stores it as a matrix of scores for each position in the alignment - highly conserved positions receive high scores and weakly conserved positions receive scores near zero [2].

PSSM scores are generally shown as positive or negative integers. Positive scores indicate that the given amino acid substitution occurs more frequently in the alignment than expected by chance, while negative scores indicate that the substitution occurs less frequently than expected. Large positive scores often indicate critical functional residues, which may be active site residues or residues required for other intermolecular interactions.

The PSI-BLAST tool returns 2 matrices for each input protein. Each matrix is $L \times 20$ in size, where L is the length of the input protein. Both of these matrices store substitution probabilities, one contains linear probabilities and the other contains log-odds. The feature extraction algorithms introduced in this thesis uses log-odds probabilities of PSSM.

2.1.3 Scoring Protein Interaction Decoys using Exposed Residues (SPIDER)

SPIDER (Scoring Protein Interaction Decoys using Exposed Residues) is developed as a knowledge-based scoring function for protein-protein interaction decoys. SPIDER is a novel multi-body pose-scoring function that has no theoretical limit on the number of residues contributing to the individual interaction terms. SPIDER's score relies on the geometric similarity of interfacial residues between docking poses and naturally occuring (native) poses.

A coarse-grain representation of a protein-protein complex was used where each residue is represented by its side chain centroid. A computational geometry approach called Almost-Delaunay tessellation was applied that transforms protein-protein complexes into a residue contact network, or an un-directional graph where vertex-residues are nodes connected by edges. This treatment forms a family of inter facial graphs representing a dataset of protein-protein complexes. Then frequent subgraph mining approach was employed to identify common inter facial residue patterns that appear in at least a subset of native protein-protein interfaces. The geometrical parameters and frequency of occurrence of each "native" pattern in the training set are used to develop the new SPIDER scoring function.

SPIDER scoring function ranks native and native-like poses above geometrical decoys; SPIDER was ranked among the top 6 (out of 28) scoring functions in a recent round 21 of CAPRI (Critical Assessment of PRedicted Interactions) blind test of protein-protein docking methods [48], [36].

2.2 Machine Learning Background

Machine learning is a subfield of computer science and a fascinating field of artificial intelligence (AI) research where we investigate how computer agents can improve their perception, cognition, and action with experience. According to Arthur Samuel in 1959 machine learning is "computers the ability to learn without being explicitly programmed" [61]. Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data [50] such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. In simple words we can say that machine learning is the science of getting computers to act without being explicitly programmed. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or in-

feasible, for example email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR) [43] [77] [52], learning to rank, computer vision etc.

There are mainly two types of machine learning algorithms - one is supervised learning and another is unsupervised learning. Supervised learning is the machine learning task of inferring a function from labeled training data [60]. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). On the other hand unsupervised machine learning is the machine learning task of inferring a function to describe hidden structure from "unlabeled" data (a classification or categorization is not included in the observations).

Our experimental data sets are labeled data set. So here we have tried supervised learning algorithms. There are lots of supervised learning algorithms, such as - Linear Regression, Logistic Regression, Decision Tree, Support Vector Machine (SVM), Naive Bayes, KNN, K-Means, Random Forest etc. It has been shown in previous literatures [82], [69] that support vector machine (SVM) gives the promising result in protein sub-cellular localization area. So in this research we have adopted SVM as our classifier.

2.2.1 Support Vector Machine (SVM)

SVM is considered to be one of the best pattern recognition techniques [79]. It is also widely used in Bioinformatics and has outperformed other classifiers and obtained promising results for protein subcellular localization. It aims to reduce the prediction error rate by finding the hyperplane that produces the largest margin based on the concept of support vector theory. It transforms the input data to higher dimensions using the kernel function to be able to find support vectors (for non linear cases). The classification of some known points in input space x_i is y_i which is defined to be either -1 or $+1$. If x' is a point in input space with unknown classification then:

$$y' = \sin(\sum_{i=1}^{n} a_i y_i K(x_i, x') + b)$$

where y' is the predicted class of point x'. The function $K()$ is the kernel function, n is the number of support vectors and a_i are adjustable weights and b is the bias. In this study, the SVM classifier is implemented with the LIB SVM tool box using the Radial Basis Function(RBF) as its kernel [3]. RBF kernel is adopted in our experiments due to its better performance than other kernels functions (e.g.polynomialkernel, linearkernel, and sigmoid).

RBF kernel is defined as follows:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$$

where γ is the regularization parameter, x_i and x_j are input feature vectors. In this study, the γ in addition to the cost parameter C (also called the soft margin parameter) of the SVM classifier are optimized using a grid search algorithm which is also implemented in the LIB SVM package. Despite its simplicity, grid search has been shown to be an effective method to optimize these parameters.

2.3 Literature Review

With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in bioinformatics and system biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable sequence-order information or key pattern characteristic. This is because all the existing machine-learning algorithms can only handle vector but not sequence samples, as elucidated in a recent review [14]. However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid completely losing the sequence-pattern information for proteins, the pseudo amino acid composition [88], [47], [56], [12], [10], [49], [66] was proposed. Ever since then, the approach of PseAAC has penetrated into nearly all the computational proteomics [47], [49], [66] as well as a long list of references cited in [10]). Because it has been widely and increasingly used, recently three powerful open access soft-wares, called 'PseAAC-Builder', 'propy', and 'PseAAC-General', were established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC [12], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode, "Gene Ontology" mode, and "Sequential Evolution" or "PSSM" mode. Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, similar web-servers [7] were developed for generating various feature vectors for DNA/RNA sequences as well. Particularly, an extremely powerful web-server called Pse-in-One [56] and its very recently updated version Pse-in-One 2.0 [56] have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

In the case of developing features for protein subcellular localization, most of the feature extracting techniques started from using simple AAC feature which resulted in loss of sequence order information. To retain sequence order information, Chou presented PseAAC and since then it has been proven

to be one of the popular methods for feature extraction. The AAC has 20 features since it is derived from the 20 common amino acids present in the protein sequence; it is simply represented as its normalize occurrence frequency. To avoid losing sequence order information, PseAAC uses features where the first 20 elements of the features are the AAC components with additional elements which are used to incorporate the sequence order information. These elements are series of different rank of correlation factors and combination of factors. The concept of PseAAC has been widely used in predicting protein related problems. Several works have used the PseAAC feature with combination of the other features to predict protein subcellular localization.

Huang and Yuan analyzed series of classifiers for sub-cellular localization, but these were limited to single location site. For multi label prediction, Gpos-mplock and Gneg-mplock (predictor) are proposed [72], [73] to predict protein localization in gram positive and gram negative bacteria; and Plant-mploc (predictor) is developed [19] which uses top down strategy to predict single or multiple protein localization in plant protein. Virus-mploc (predictor) [74] was developed with fusion of classifiers and features of functional domain and gene ontology to predict virus proteins. To increase the quality of prediction, three revised version of the prediction systems were developed: iloc-Gpos (predictor) [88], iloc-plant (predictor) [87], iloc-virus (predictor) [89]. Huang and Yuan used AAC, evolution information and PseACC with backward propagation (BP) and radial basis function (RBF) neural network to predict both single and multi-site subcellular proteins.

A number of machine learning methods have been developed with many different combination and types of features along with different classifiers. For example, PSORT (predictor) [40] uses sequence features based on sorting signal, SubLoc (predictor) [5] uses SVM with AAC to obtain higher accuracy. TargetP (predictor) [34] uses ANN and N-terminal sequence to predict subcellular locations. Pierleoni et al. [65] used N-terminal, AAC and alignment profile to predict the subcellular localization. Similarly, Tamura and Akutsu [76] used alignment of block sequence. Chang et al. [4] developed and used gapped-dipeptide and probabilistic latent semantic analysis method for prediction of gram negative bacteria protein. Lee et al. [51] predicted protein localization by integrating an extensive set of protein physical characteristics over a proteins extended proteinprotein interaction neighborhood, using a classification framework called divide and conquer k-nearest neighbor (DC-KNN) to improve accuracy.

As demonstrated by a series of recent publications [6], [91], [53], [55], [44] and according to the Chou's 5-step rule [13], to establish a really useful sequence-based statistical predictor for a biological system, we should con-

sider the following five guidelines:

1. Construct or select a valid benchmark dataset to train and test the predictor.

2. Formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted.

3. Introduce or develop a powerful algorithm (or engine) to operate the prediction.

4. Properly perform cross validation tests to objectively evaluate the anticipated accuracy of the predictor.

5. Establish a user-friendly web-server for the predictor that is accessible to the public.

Below, we are to describe how to deal with these steps one by one. In this study, we attempted to predict the sub-cellular location of both gram-positive and gram-negative bacterial proteins using structural and evolutionary based features.

Chapter 3

Materials and Methods

3.1 Data Set

In this research we have used two dataset which have been used widely among this field of literature [41], [20], [17], [70], [63] for Gram-positive and Gram-negative sub-cellular localizations. The details of this two dataset are described below:

3.1.1 Gram-Positive Bacteria Protein Dataset

For gram-positive sub-cellular localizations we have used dataset that was introduced in the literature [41], [20], [17], [70]. This dataset contains total 523 protein samples which belongs to four gram positive sub-cellular localizations. Among this 523 samples there are total 519 different protein sample. Among 519 proteins there are total 515 protein samples which belongs to only one or single location while the rest 4 protein samples belongs to two location. Thus gram-positive bacteria protein dataset contains total 523 (515 + 4 * 2) protein samples. The four locations name and total number of protein samples each location contains are shown at Table 3.1. This dataset is available at the web-link http://www.csbio.sjtu.edu.cn/bioinf/Gpos-multi.

3.1.2 Gram-Negative Bacteria Protein Dataset

For gram-negative sub-cellular localizations we have used dataset that was introduced in the literature [63], [20], [17], [15]. This dataset contains total 1456 protein samples which belongs to eight gram negative sub-cellular localizations. Among this 1456 samples there are total 1392 different protein sample. Among 1392 proteins there are total 1328 protein samples which belongs to only one or single location while the rest 64 protein samples belongs to two location. Thus gram-negative bacteria protein dataset contains total 1456 (1328 + 64 * 2) protein samples. The eight locations name and total number of protein samples each location contains are shown at Table 3.2. This

dataset is available at the web-link http://www.csbio.sjtu.edu.cn/bioinf/Gneg-multi.

For classifying multi location proteins we have used the same method which was used in the literature [41], [63]. In gram positive and gram negative bacteria protein dataset which samples are in multi location means more than one location, we have used those multi labeled proteins as several single labeled protein samples based on their labels or classes which they belongs to. For example a protein sample which has two label or class location, we have used them in our experiment as two single labeled protein sample. Then we have performed our experimental classification task. Thus by making and adding extra protein samples, we have calculated the possible worst case scenario as we might not be able to predict the multi labeled protein samples or the extra label of a single protein that has more than one label. In this way, predicting a protein's sub-cellular location, we guarantee that we have considered all the worst case that can occur in performing my prediction task.

Table 3.1: Details of gram-positive bacteria protein dataset

No.	Sub-cellular location	Total protein samples
1	Cell membrane	174
2	Cell wall	18
3	Cytoplasm	208
4	Extracellular	123
Total number of locative proteins		523
Total number of different proteins		519

Table 3.2: Details of gram-negative bacteria protein dataset

No.	Sub-cellular location	Total protein samples
1	Cell inner membrane	557
2	Cell outer membrane	124
3	Cytoplasm	410
4	Extracellular	133
5	Fimbrium	32
6	Flagellum	12
7	Nucleoid	8
8	Periplasm	180
Total number of locative proteins		1456
Total number of different proteins		1392

3.2 Normalization of PSSM

In our experiment, we have made two main groups of PSSM matrix named as Normalized PSSM matrix and Non-Normalized PSSM matrix. Non-Normalized PSSM is the exact PSSM matrix which is the output matrix of PSSIBLAST software. Normalized PSSM matrix was used in the literature [69]. According to this [41] literature PSSM matrix can be represented as:

$$P = \begin{bmatrix} U_{1,1} & U_{1,2} & ... & U_{1,19} & U_{1,20} \\ U_{2,1} & U_{2,2} & ... & U_{2,19} & U_{2,20} \\ . & . & ... & . & . \\ . & . & ... & . & . \\ . & . & ... & . & . \\ U_{L,1} & U_{L,2} & ... & U_{L,19} & U_{L,20} \end{bmatrix}$$

The size of the PSSM matrix is $L \times 20$, here L is the length of the amino acid sequence or simply primary protein sequence, $U_{i,j}$ represents the score of amino acid which is located at the i-th location of the protein sequence which is changed into amino acid j during the process of evolution. In order to make the normalization of PSSM matrix, we have computed and formulated a new PSSM matrix N using the information from original PSSM matrix P. We are refering this new matrix N as my new normalized PSSM in this paper. The normalized matrix N is computed as follows:

$$N = \begin{bmatrix} V_{1,1} & V_{1,2} & ... & V_{1,19} & V_{1,20} \\ V_{2,1} & V_{2,2} & ... & V_{2,19} & V_{2,20} \\ . & . & ... & . & . \\ . & . & ... & . & . \\ . & . & ... & . & . \\ V_{L,1} & V_{L,2} & ... & V_{L,19} & V_{L,20} \end{bmatrix}$$

where, $V_{i,j} = \dfrac{U_{i,j} - Z_y}{Z_x - Z_y}$; i = 1,2,.....,L; $Z_x = max(P)$ & $Z_y = min(P)$
We have normalized all PSSM matrix one by one. First we have took one PSSM file and find out the highest score Z_x and lowest score Z_y of the matrix. Then putting this maximum and minimum value in my formula we have calculated the normalized score. Algorithm for converting a PSSM matrix to its corresponding Normalized PSSM matrix is shown at Algorithm 1.

From Normalized PSSM matrix we have extracted 7 features:

1. PSSM-AAO (Feature vector size 20)
2. PSSM-SD (Feature vector size 80)
3. PSSM-SAC (Feature vector size 100)

Algorithm 1: Normalization of PSSM Matrix

$Z_x \leftarrow \texttt{P}[0, 0]$;
$Z_y \leftarrow \texttt{P}[0, 0]$;
$L \leftarrow$ Length of PSSM Matrix;
$P \leftarrow$ Original PSSM Matrix;
$V \leftarrow$ Empty Array of Size $L \times 20$;
for $i = 0;\ i < L;\ i = i + 1$ **do**
    for $j = 0;\ j < 20;\ j = j + 1$ **do**
        if $\texttt{P}[\texttt{i}, \texttt{j}] > \texttt{Z}_\texttt{x}$ **then**
            $Z_x \leftarrow \texttt{P}[\texttt{i}, \texttt{j}]$;
        if $\texttt{P}[\texttt{i}, \texttt{j}] < \texttt{Z}_\texttt{y}$ **then**
            $Z_y \leftarrow \texttt{P}[\texttt{i}, \texttt{j}]$;
for $i = 0;\ i < L;\ i = i + 1$ **do**
    for $j = 0;\ j < 20;\ j = j + 1$ **do**
        $V_{i,j} = \dfrac{P_{i,j} - Z_y}{Z_x - Z_y}$;

4. Composition of Normalized PSSM (Feature vector size 20)
5. Bi-gram of Normalized PSSM (Feature vector size 400)
6. One-lead Bi-gram of Normalized PSSM (Feature vector size 400)
7. Auto Covariance of Normalized PSSM (Feature vector size 200)

From Non-Normalized PSSM matrix we have extracted 4 features:

1. Composition of Non-Normalized PSSM (Feature vector size 20)
2. Bi-gram of Non-Normalized PSSM (Feature vector size 400)
3. One-lead Bi-gram of Non-Normalized PSSM (Feature vector size 400)
4. Auto Covariance of Non-Normalized PSSM (Feature vector size 200)

3.3 Constructing the Consensus Sequence

The consensus sequence is also an amino acid sequence which is generated from a Position Specific Score Matrix (PSSM). For each amino acid in a protein, the corresponding row of the PSSM is extracted and the amino acid with the highest substitution probability value is substituted for the original amino acid. In this way a new sequence is generated of the same

length as the original sequence. Figure 3.1 shows how a consensus sequence is generated from a Position Specific Score Matrix generated by PSIBLAST. On the left, the original sequence is displayed. The amino acid with maximum probability value is identified with a box around the probability. The consensus sequence is listed on the right hand side of the image. Note that for some of the amino acids e.g. lines 1, 6 and 10 the original sequence and consensus sequence are both same. The others, however are different.

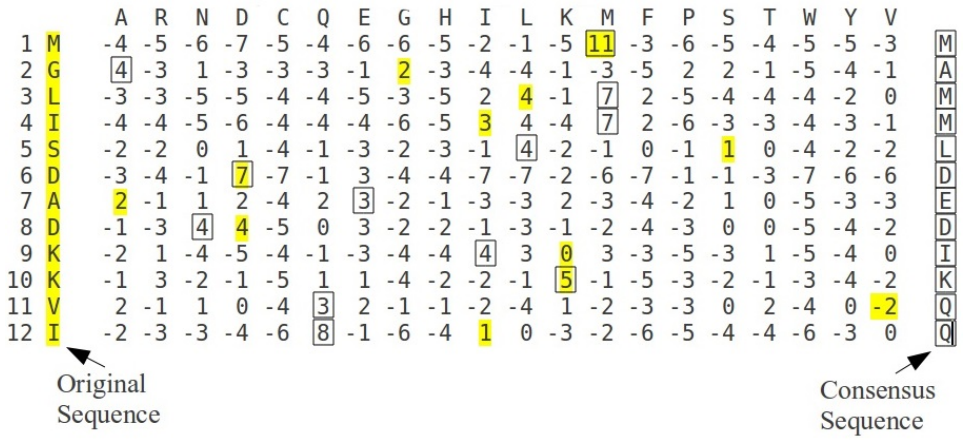

Figure 3.1: Extracting the consensus sequence from a PSSM file

From consensus sequence we have extracted only one feature:

1. Occurrence of Consensus Sequence (Feature vector size 20)

3.4 Feature Extraction Method

Feature extraction is the process of keeping information relevant to the classification task, and discarding other irrelevant information. Feature extraction involves reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power, also it may cause a classification algorithm to overfit to training samples and generalize poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Since most classifiers only accept fixed length feature vectors, the feature extraction step is also a way of creating a fixed length feature vector from the variable length protein sequences.

We have extracted total 46 features from both PSSM matrix and Spider SPD3 matrix. All the features and their corresponding feature vector size are shown in Table 3.4. We have extracted this 46 features from 10 feature extraction method which are described below:

3.4.1 Occurrence Feature

This feature is extracted from amino acid sequence or protein sequence, consensus sequence and Spider SS sequence. The dimensionality of this feature vector is 20 for amino acid sequence, consensus sequence and 3 for Spider SS sequence.

There are 20 distinct element in amino acid sequence or protein sequence and consensus sequence which are 'A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N', 'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y'. The length of the protein sequence vary from protein to protein, thus the sequence and element number in protein sequence vary from protein to protein. For extracting this feature we have taken this 20 elements one by one and counted how many times this element occur in the corresponding protein sequence or consensus sequence. Thus we have looped it for 20 times and found a feature of vector size 20.

There are 3 distinct element in Spider SS Sequence which are 'C', 'E', 'H'. The length of the Spider SS sequence vary from protein to protein, thus the SS sequence and element number in the sequence vary from protein to protein. For extracting this feature we have taken this 3 elements one by one and counted how many times this element occur in the corresponding Spider SS sequence. Thus we have looped it for 3 times and found a feature of vector size 3.

Using occurrence feature extraction method we have extracted 3 features:

1. Occurrence of Protein Sequence (Feature vector size 20)
2. Occurrence of Consensus Sequence (Feature vector size 20)
3. Occurrence of Spider SS Sequence (Feature vector size 3)

3.4.2 Composition Feature

This feature is extracted from both PSSM matrix and Spider SPD3 matrix. To calculate this feature we have taken one column by one column at a time from the respective matrix and summed up all the rows value of this particular column and finally divided it with the length of the protein. The equation for this feature is given below:

$$Composition_j = \frac{1}{L}\sum_{i=1}^{L} N_{i,j}$$

Here N is the corresponding matrix, L is the protein length and j is the respective column. The dimensionality of this feature vector will be ($Number\ of\ columns$). Algorithm for extracting composition feature is shown at Algorithm 2.

Algorithm 2: Composition Feature Extraction

$N \leftarrow$ Matrix from which feature will be extracted;
$L \leftarrow$ Length of the Protein;
$C \leftarrow$ Number of matrix column;
$V \leftarrow$ Empty array of size C;
for $j = 0;\ j < C;\ j = j + 1$ **do**
    $sum \leftarrow 0$;
    for $i = 0;\ i < L;\ i = i + 1$ **do**
        $sum = sum + N_{i,j}$;
    $V_j = \frac{sum}{L}$;

We have extracted total 8 composition features from both PSSM matrix and Spider SPD3 matrix. Using composition feature extraction method we have extracted 2 features from PSSM matrix:

1. Composition of PSSM (Feature vector size 20)
2. Composition of Normalized PSSM (Feature vector size 20)

Using composition feature extraction method we have extracted 6 features from Spider SPD3 matrix:

1. Composition of Spider (Feature vector size 8)
2. Composition of Spider with Torsion Angles (Feature vector size 12)
3. Composition of ASA (Feature vector size 1)
4. Composition of Angles (Feature vector size 4)
5. Composition of Angles with Torsion Angles (Feature vector size 8)
6. Composition of P(C), P(E), P(H) (Feature vector size 3)

3.4.3 Bi-Gram Feature

The equation for this feature is given below:

$$Bigram_{k,l} = \frac{1}{L} \sum_{i=1}^{L-1} N_{i,k} N_{i+1,l}$$

The dimensionality of this feature vector will be ($Number\ of\ columns$) $\times$ ($Number\ of\ columns$). Algorithm for extracting bi-gram feature is shown at Algorithm 3.

Algorithm 3: Bi-gram Feature Extraction

```
1  N ← Matrix from which feature will be extracted;
2  L ← Length of the Protein;
3  C ← Number of matrix column;
4  V ← Empty array of size C × C;
5  for k = 0; k < C; k = k + 1 do
6      for l = 0; l < C; l = l + 1 do
7          sum ← 0;
8          for i = 0; i < L − 1; i = i + 1 do
9              sum = sum + N_{i,k} N_{i+1,l};
10         V_{k,l} = sum / L;
```

We have extracted total 8 bi-gram features from both PSSM matrix and Spider SPD3 matrix. Using bi-gram feature extraction method we have extracted 2 features from PSSM matrix:

1. Bi-Gram of PSSM (Feature vector size 400)
2. Bi-Gram of Normalized PSSM (Feature vector size 400)

Using bi-gram feature extraction method we have extracted 6 features from Spider SPD3 matrix:

1. Bi-Gram of Spider (Feature vector size 64)
2. Bi-Gram of Spider with Torsion Angles (Feature vector size 144)
3. Bi-Gram of ASA (Feature vector size 1)
4. Bi-Gram of Angles (Feature vector size 16)
5. Bi-Gram of Angles with Torsion Angles (Feature vector size 64)
6. Bi-Gram of P(C), P(E), P(H) (Feature vector size 9)

3.4.4 One-Lead Bi-Gram Feature

The equation for this feature is given below:

$$OneLeadBigram_{k,l} = \frac{1}{L}\sum_{i=1}^{L-2} N_{i,k}N_{i+2,l}$$

The dimensionality of this feature vector will be $(Number\ of\ columns) \times (Number\ of\ columns)$. Algorithm for extracting one-lead bi-gram feature is shown at Algorithm 4.

Algorithm 4: One-Lead Bi-Gram Feature Extraction

```
N ← Matrix from which feature will be extracted;
L ← Length of the Protein;
C ← Number of matrix column;
V ← Empty array of size C × C;
for k = 0; k < C; k = k + 1 do
    for l = 0; l < C; l = l + 1 do
        sum ← 0;
        for i = 0; i < L − 2; i = i + 1 do
            sum = sum + N_{i,k} N_{i+2,l};
        V_{k,l} = sum / L;
```

We have extracted total 8 one-lead bi-gram features from both PSSM matrix and Spider SPD3 matrix. Using one-lead bi-gram feature extraction method we have extracted 2 features from PSSM matrix:

1. One-Lead Bi-Gram of PSSM (Feature vector size 400)
2. One-Lead Bi-Gram of Normalized PSSM (Feature vector size 400)

Using one-lead bi-gram feature extraction method we have extracted 6 features from Spider SPD3 matrix:

1. One-Lead Bi-Gram of Spider (Feature vector size 64)
2. One-Lead Bi-Gram of Spider with Torsion Angles (Feature vector size 144)
3. One-Lead Bi-Gram of ASA (Feature vector size 1)
4. One-Lead Bi-Gram of Angles (Feature vector size 16)
5. One-Lead Bi-Gram of Angles with Torsion Angles (Feature vector size 64)
6. One-Lead Bi-Gram of P(C), P(E), P(H) (Feature vector size 9)

3.4.5 Tri-Gram Feature

The equation for this feature is given below:

$$TriGram_{k,l,m} = \frac{1}{L}\sum_{i=1}^{L-2} N_{j,k}N_{j+1,l}N_{j+2,m}$$

The dimensionality of this feature vector will be $(Number\ of\ columns)^3$. Algorithm for extracting tri-gram feature is shown at Algorithm 5.

Algorithm 5: Tri-Gram Feature Extraction

```
1  N ← Matrix from which feature will be extracted;
2  L ← Length of the Protein;
3  C ← Number of matrix column;
4  V ← Empty array of size C × C × C;
5  for k = 0; k < C; k = k + 1 do
6      for l = 0; l < C; l = l + 1 do
7          for m = 0; m < C; m = m + 1 do
8              sum ← 0;
9              for j = 0; j < L − 2; i = i + 1 do
10                 sum = sum + N_{j,k} N_{j+1,l} N_{j+2,m};
11             V_{k,l,m} = sum / L;
```

Being large feature vector size while extracting features from PSSM matrix, we have extracted tri-gram features from only Spider SPD3 matrix. Using tri-gram feature extraction method we have extracted 6 features from Spider SPD3 matrix:

1. Tri-Gram of Spider (Feature vector size 512)
2. Tri-Gram of Spider with Torsion Angles (Feature vector size 1728)
3. Tri-Gram of ASA (Feature vector size 1)
4. Tri-Gram of Angles (Feature vector size 64)
5. Tri-Gram of Angles with Torsion Angles (Feature vector size 512)
6. Tri-Gram of P(C), P(E), P(H) (Feature vector size 27)

3.4.6 Auto Covariance Feature

A correlation factor coupling adjacent residues along the protein sequence [92] is known as Auto covariance (AC). It is also known as a kind of variant of auto cross covariance. It is a very powerful statistical tool which is used to analyze sequences of vectors [85], the Auto Covariance transformation has been widely applied in various fields of bioinformatics [38], [39], [32], [86], [93], [57]. Auto Covariance variables are able to avoid producing too many variants. The equation for this feature is given below:

$$AutoCovariance_{k,j} = \frac{1}{L}\sum_{i=1}^{L-k} N_{i,j}N_{i+k,j} \;\; (j = 1, ..., 20 \; and \; k = 1...DF)$$

where DF is the distance factor. Different values have been tested to find out the effective value of DF which gives the highest accuracy rate of prediction. In this research we have tested total 15 values for DF (DF = 1,2,3,4,.....,12,13,14,15) and took only one value which is DF = 10. We have observed that only DF = 10 gives the highest accuracy rate for my task. So, the effective value of DF is used as 10 for the employed benchmark since this value was investigated in other literature [23] which gives promising results for other benchmark datasets. The dimensionality of this feature vector will be $(Number \; of \; columns) \times DF$. Algorithm for extracting auto covariance feature is shown at Algorithm 6.

Algorithm 6: Auto Covariance Feature Extraction

```
DF ← 10;
P ← Matrix from which feature will be extracted;
L ← Length of the Protein;
V ← Empty array of size L× (Number of matrix column);
C ← Number of matrix column;
for k = 0; k < DF; k = k + 1 do
    for j = 0; j < C; j = j + 1 do
        sum ← 0;
        for i = 0; i < L − k; i = i + 1 do
            sum = sum + P_{i,j} P_{i+k,j};
        V_{k,j} = sum / L;
```

We have extracted total 8 auto covariance features from both PSSM matrix and Spider SPD3 matrix. Using auto covariance feature extraction method we have extracted 2 features from PSSM matrix:

1. Auto Covariance of PSSM (Feature vector size 200)

2. Auto Covariance of Normalized PSSM (Feature vector size 200)

Using auto covariance feature extraction method we have extracted 6 features from Spider SPD3 matrix:

1. Auto Covariance of Spider (Feature vector size 80)
2. Auto Covariance of Spider with Torsion Angles (Feature vector size 120)
3. Auto Covariance of ASA (Feature vector size 10)
4. Auto Covariance of Angles (Feature vector size 40)
5. Auto Covariance of Angles with Torsion Angles (Feature vector size 80)
6. Auto Covariance of P(C), P(E), P(H) (Feature vector size 30)

3.4.7 PSSM-AAO Feature

PSSM-AAO is amino acid occurrence based on PSSM. This feature group is directly extracted from PSSM matrix. It aims at capturing global discriminatory information regarding the substitution probabilities of the amino acids with respect to their positions in the protein sequence. This feature is extracted by summation of the substitution score of a given amino acid with all the amino acids along the protein sequence. The equation for this feature is given below:

$$PSSM - AAO_j = \sum_{i=1}^{L} N_{i,j} \;\; (j = 1, ..., 20)$$

Here N is the corresponding matrix, L is the protein length and j is the respective column. The dimensionality of this feature vector is 20. Algorithm for extracting composition feature is shown at Algorithm 7.

The main advantage of using the occurrence feature group over using the composition feature group (in both Occurrence and PSSM-AAO feature groups) is that it captures information regarding the length of the proteins whereas this information is disregarded in the composition feature group [75]. To extract the occurrence features, we have calculated the summation of the total substitution score for a given amino acids while in composition based features, we have divided this number with the length of the corresponding protein sequence. In fact, in composition method, we have normalized the occurrence of the amino acids, for a given protein, based on its length. In PSSM-AAO we do not do this normalization which maintain the general total occurrence of the amino acids based on their substitution scores. Since we do not normalize by dividing it by length, it becomes implicitly a part of that feature [75], [24], [67].

Algorithm 7: PSSM-AAO Feature Extraction

```
N ← PSSM Matrix;
L ← Length of the Protein;
C ← Number of matrix column;
V ← Empty array of size C;
for j = 0; j < C; j = j + 1 do
    sum ← 0;
    for i = 0; i < L; i = i + 1 do
        sum = sum + N_{i,j};
    V_j = sum;
```

3.4.8 PSSM-SD Feature

This method is specifically proposed to add more local discriminatory information about how the amino acids, based on their substitution probabilities (extracted from PSSM), are distributed along the protein sequence [25]. We propose this segmentation method in the manner where segments of a protein sequence are of unequal lengths and each segment is represented by a distribution feature which is computed as follows. First, for the jth column in PSSM, we calculate the total substitution probability $T_j = \sum_{i=1}^{L} P_i j$. Then, starting from the first row of PSSM, we calculate the partial sum S_1 of the substitution probabilities of the first i amino acids until reaching $F_p\%$ of the total sum $S_1 = \sum_{i=1}^{I_j^1} P_i j$. Using the distribution factor $F_p\%$, we calculate the I_j^1, where I_j^1 corresponds to the number of the amino acids such that the summation of their substitution probabilities is less than or equal to the $F_p\%$ of T_j. Similarly, we continue to calculate the partial sum of the first i amino acids (starting from the first row of PSSM) until reaching $n \times F_p\% = 50\%$ of the total sum ($S_n = \sum_{i=1}^{I_j^n} P_i j$) and calculate the I_j^n corresponding to the number of amino acids such that the summation of their substitution probabilities is less than or equal to 50% of T_j. Therefore, starting from the first row of PSSM, we extract n features $()I_j^1, I_j^2, ..., I_j^n)$ corresponding to the number of segments until reaching 50% of T_j.

We repeat the same process beginning from the last row of PSSM for the jth column. We calculate the partial sum of the substitution probabilities of the first i amino acids until reaching $n \times Fp\% = 50\%$ of the total sum which is $S_{n+1} = \sum_{i=1}^{I_j^{n+1}} P_i j$ until reaching $S_2 n = \sum_{i=1}^{I_j^2 n} P_i j$, respectively and calculate I_j^{n+1} until reaching $I_j^2 n . I_j^{n+1}$ and $I_j^2 n$ correspond to the number of amino acids such that the summation of their substitution probabilities are less than or equal to $Fp\%$ and $n \times Fp\% = 50\%$ of T_j, respectively (starting from the last row of PSSM). Therefore, starting from the last row

of the PSSM, we extract n features $(I_j^{n+1}, I_j^{n+2}, ..., I_j^{2n})$ corresponding to the number of segments until reaching 50% of T_j. In this manner we extract $2n$ segmented distribution features for each column in PSSM. The method used to calculat ePSSM-SD feature group from the jth column of PSSM is shown in Fig. 3.2. We repeat the same process for all 20 columns corresponding to 20 amino acids in PSSM.

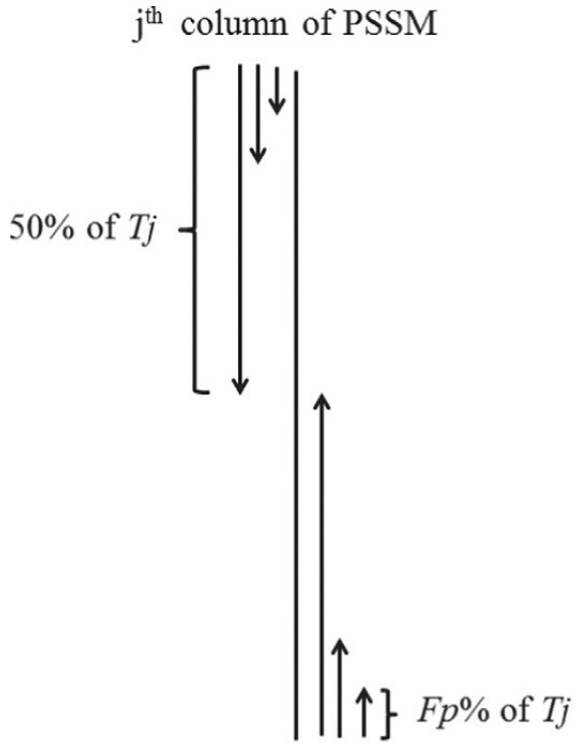

Figure 3.2: The segmentation method used to extract PSSM-SD feature group from the jth column of PSSM

In this study, we adopt three values for F_p (5, 10, and 25)to investigate the effectiveness of the number of segments on the achieved results and find the suitable number of segments to explore local discriminatory information embedded in PSSM. We have used other choices for F_p but these three remains the best representatives of all the choices. To maintain the generality and simplicity of the segmentation method, we avoid a very specific segmentation method as it might not be applicable for all cases. For PSSM-SD feature group, using $Fp = 5$, we divide the protein sequence into 20 segments ($n = 10$from each side) and extract 400 features in total in this feature group ($20 \times 20 = 400$). Similarly, using $F_p = 10$ ($n = 5$ from each side) we divide the protein sequence into 10 segments and extract 200 features in total ($10 \times 20 = 200$) and by using $F_p = 25$ ($n = 2$ from each

side), we extract 80 features in total ($4 \times 20 = 80$). General formula for feature vector size of PSSM-SD is
$(100 \div F_p) \times (Number\ of\ columns\ in\ the\ matrix)$.

However, in literature [22] has proven that $F_p = 25$ gives the best result and we also investigate the same result, so in our final experiment we have adopt $F_p = 25$ which gives 80 ($(100 \div 25) \times 20 = 80$) as feature vector size. Thus in this report we have only mentioned the feature vector size of PSSM-SD as 80. Algorithm for extracting PSSM-SD feature is shown at Algorithm 8.

Algorithm 8: PSSM-SD Feature Extraction

```
1  N ← PSSM Matrix;
2  L ← Length of the Protein;
3  C ← Number of matrix column;
4  F_p ← Desired value of F_p, e.g 5, 10, 25;
5  V ← Empty array of size (100 ÷ F_p) × C;
6  k ← 0 ;
7  for j = 0; j < C; j = j + 1 do
8      T_j ← Sum of jth column;
9      partialSum ← 0;
10     i ← 0;
11     for tp = F_p; tp <= 50; tp = tp + F_p do
12         while partialSum <= tp × (T_j ÷ 100) do
13             partialSum = partialSum + N_{i,j};
14             i = i + 1;
15         V_k = i;
16         k = k + 1;
17     partialSum ← 0;
18     i ← L;
19     index ← 0;
20     for tp = F_p; tp <= 50; tp = tp + F_p do
21         while partialSum <= tp × (T_j ÷ 100) do
22             partialSum = partialSum + N_{i,j};
23             i = i − 1;
24             index = index + 1;
25         V_k = index;
26         k = k + 1;
```

3.4.9 PSSM-SAC Feature

This feature was introduced in the literature [22]. It was shown that information about the interaction of neighboring aminoacids along the protein sequence can play an important role in providing significant local discriminatory information and enhancing protein subcellular localization prediction accuracy [11], [67]. To extract this information, the concept of auto covariance has been widely used in the literature in different ways (e.g.bigram([67]), tri-gram ([64]), auto correlation ([27], [24])). Among all these methods, pseudo amino acids composition has attained the best results to extract local information ([20], [82], [35]). In the present study, we extend the concept of segmented distribution features as described in the previous subsection to compute the auto covariance features from the segmented protein sequence. This is done to enforce local discriminatory information extracted from PSSM.

To extract this feature group, we calculate the auto covariance of the substitution probabilities of the amino acids using K_p as the distance factor in the following manner. Starting from the first row of PSSM, for the jth column of PSSM, we calculate K_p auto covariance features for the first I_j^1. Similarly, we calculate auto covariance until reaching the first I_j^n amino acids. Then starting from the last row of PSSM for the jth column of PSSM, We repeat the same process for I_j^{n+1} and until reaching $I_j^2 n$ (I_j^1 to I_j^n and I_j^{n+1} until reaching to $I_j^2 n$ are calculated in the way that is explained in the previous subsection). This process is repeated for all 20 columns of PSSM and corresponding features are calculated as follows:

$$PSSM - seg_{q,m,j} = \frac{1}{I_j^q - m} \sum_{i=1}^{I_j^q - m} (P_{i,j} - P_{ave,j})(P_{(i+m),j} - P_{ave,j}),$$

$$(q = 1, ..., 2n \ \& \ m = 1, ..., k_p \ \& \ j = 1, ..., 20)$$

Thus, we have extracted a total of $(nK_p + nK_p + K_p) = (2n + 1)K_p)$ auto covariance features in this manner (for the jth column of the PSSM). For all 20 columns of the PSSM, segmented auto covariance of the substitution probabilities of the amino acids are extracted and combined to build the corresponding feature group which will be referred to as PSSM-SAC (PSSM-seg + PSSM-AC which consists of $20 \times (2n + 1) \times K_p \ features \ in \ total$).

In the literature [22] the authors have tried different values for K_p starting from 1 to 10 (1,2,3,.....,8,9,10). They have reported that $K_p = 10$ gives the best result for PSSM-SAC. So in our experiment we have used 10 as a value for K_p. Thus our feature vector size for PSSM-SAC is 100. General formula for feature vector size for PSSM-SAC is $K_p \times 20 \times 5$.

3.4.10 Dubchak Feature

Dubchak feature is based on Physicochemical attributes. Physicochemical attributes refer to properties of the individual amino acids e.g. hydrophobicity or polarity. Dubchak et al. [33] proposed a set of features based on physicochemical properties of amino acids. The properties they chose were hydrophobicity, normalised van der Waals volume, polarity, polarizability and normalized frequency of alpha helix. The physicochemical attributes used in Dubchak method are shown in table 3.3. They used three descriptors: composition, transition and distribution. As an example, consider just the hydrophobicity attribute. These are split into polar, neutral, and hydrophobic. The composition descriptor records the percent compositions of the polar, neutral, and hydrophobic residues in the protein (3 numbers). The transition descriptor also consists of 3 numbers, the percent frequency of a polar $\rightarrow$ neutral or neutral $\rightarrow$ polar transition, the percent frequency of a polar $\rightarrow$ hydrophobic or hydrophobic $\rightarrow$ polar transition, or the percent frequency of a neutral $\rightarrow$ hydrophobic or hydrophobic $\rightarrow$ neutral transition. The distribution descriptor consists of 5 numbers for each of the 3 groups: the fraction of the entire sequence where the first residue of a given group is located, and where 25%, 50% 75% and 100% of those are contained. The total feature vector consists of 3 composition, 3 transition, and 15 distribution features (21 total) per physicochemical attribute. For 5 attributes this gives 105 features in total.

The features described above have been widely used for protein fold recognition [8], [46], [30], [26], [46], [45], [25], [9], [78]. Other physicochemical attributes explored include solvent accessibility [94], flexibilty [62], bulkiness [42], first and second order entropy [95], and size of the amino acid side chain [20]. Considering the number of possible physicochemical attributes, methods have also been proposed that systematically identify the attributes that work best for protein fold recognition [68].

Table 3.3: The physicochemical attributes used in Dubchak

Property	Group 1	Group 2	Group 3
Hydrophobicity	Polar R,K,E,D,Q,N	Neutral G,A,S,T,P,H,Y	Hydrophobic C,V,L,I,M,F,W
Normalised van der Waals volume	0 - 2.78 G,A,S,C,T,P,D	2.95 - 4.0 N,V,E,Q,I,L	4.43 - 8.08 M,H,K,F,R,Y,M
Polarity	4.9 - 6.2 L,I,F,W,C,M,V,Y	8.0 - 9.2 P,A,T,G,S	10.4 - 13.0 H,Q,R,K,N,E,D
Polarizability	0 - 0.108 G,A,S,T,D	0.128 - 0.186 C,P,N,V,E,Q,A,L	0.219 - 0.409 K,M,H,F,R,Y,W

Continued on next page

Table 3.3 – continued from previous page

Property	Group 1	Group 2	Group 3
Normalized frequency of alpha helix	0.57 - 0.83 G,P,N,Y,C,S,T	0.98 - 1.08 R,H,D,V,W,I	1.11 - 1.51 Q,F,K,L,A,M,E

Table 3.4: Feature Extraction Methods and Feature Vector Size

No.	Feature Extraction Method	Feature Vector Size
1	Occurrence of Protein Sequence	20
2	Occurrence of SS sequence	3
3	Occurrence of Consensus Sequence	20
4	PSSM-AAO	20
5	PSSM-SD	80
6	PSSM-SAC	100
7	Dubchak of Protein Sequence	105
8	Dubchak of Consensus Sequence	105
9	Composition of PSSM	20
10	Composition of Normalized PSSM	20
11	Composition of Spider	8
12	Composition of Spider With Torsion Angles	12
13	Composition of ASA	1
14	Composition of Angles	4
15	Composition of Angles With Torsion Angles	8
16	Composition of P(C) P(E) P(H)	3
17	Bi-Gram of PSSM	400
18	Bi-Gram of Normalized PSSM	400
19	Bi-Gram of Spider	64
20	Bi-Gram of Spider With Torsion Angles	144
21	Bi-Gram of ASA	1
22	Bi-Gram of Angles	16
23	Bi-Gram of Angles With Torsion Angles	64
24	Bi-Gram of P(C), P(E), P(H)	9
25	One-Lead Bi-Gram of PSSM	400
26	One-Lead Bi-Gram of Normalized PSSM	400
27	One-Lead Bi-Gram of Spider	64
28	One-Lead Bi-Gram of Spider With Torsion Angles	144
29	One-Lead Bi-Gram of ASA	1
30	One-Lead Bi-Gram of Angles	16

Continued on next page

Table 3.4 – continued from previous page

No.	Feature Extraction Method	Feature Vector Size
31	One-Lead Bi-Gram of Angles With Torsion Angles	64
32	One-Lead Bi-Gram of P(C), P(E), P(H)	9
33	Tri-Gram of Spider	512
34	Tri-Gram of Spider With Torsion Angles	1728
35	Tri-Gram of ASA	1
36	Tri-Gram of Angles	64
37	Tri-Gram of Angles With Torsion Angles	512
38	Tri-Gram of P(C), P(E), P(H)	27
39	Auto Covariance of PSSM	200
40	Auto Covariance of Normalized PSSM	200
41	Auto Covariance of Spider	80
42	Auto Covariance of Spider With Torsion Angles	120
43	Auto Covariance of ASA	10
44	Auto Covariance of Angles	40
45	Auto Covariance of Angles With Torsion Angles	80
46	Auto Covariance of P(C), P(E), P(H)	30

Chapter 4

Experimentation, Result and Discussion

4.1 Feature Extraction

In our research first step is to extract desired features from desired data sets. In **Chapter 3** a brief description has been given about dataset, manipulation of dataset and feature extraction methods. We have used two dataset named as Gram-Positive Bacteria Protein dataset and Gram-Negative Bacteria Protein dataset. We have implemented total 10 types (Occurrence, Composition, Bi-Gram, Obe-Lead Bi-Gram, Tri-Gram, Auto Covariance, PSSM-AAO, PSSM-SD, PSSM-SAC, Dubchak) of feature extraction methods from where we have extracted total 46 features per data set.

Note, in this research we have extracted 46 features for only Gram-positive dataset, not for Gram-negative dataset. We have run experiment using this 46 features and found that 6 features gives us our desired result. Then we have extracted those 6 features from Gram-negative dataset and reported our result.

4.2 Choosing Classifier and Parameter Tunning

For choosing appropriate classifiers firstly we have extracted those features which have been reported in these literature [22], [69]. These two literature have reported that they have used Support Vector Machine (SVM) with RBF kernel and used 3000 as a value for cost parameter (C), 0.005 as a value for gamma (γ). We have tried other classifier such as Naive Bayes, Nearest Neighbor, Decision Tree and Random Forest, but these classifiers did not gave the promising result. We have also tried a little to change SVM kernal and tune parameters (C & γ). But we did not get the promising result. So for our experimental purpose we have chosen Support Vector Machine

(SVM) as our classifier and values for γ and C are 0.005 and 3000. Using this classifier and parameters we have run our experiment to get final features set among 46 features. After getting the final features set we have tried to tune SVM parameters, but these did not give us the good result. So in this paper our reported classifier is Support Vector Machine (SVM), kernal is RBF, value of C is 3000 and value of γ is 0.005.

4.3 Choosing Validation Method

For our experiment we have adopted two types of validation method, one is 10-fold cross validation and another one is jackknife test also named as leave-one-out cross validation.

10-Fold Cross Validation: In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal sized subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged to produce a single estimation. The advantage of this method is all observations are used for both training and validation, and each observation is used for validation exactly once.

Jackknife Test: this is also known as leave one out cross validation method. In this method, the original sample is randomly partitioned into n equal sized subsamples where n is the total samples in the dataset. Of the n subsamples, a single subsample (means exactly one sample from the sample dataset) is retained as the validation data for testing the model, and the remaining n-1 subsamples are used as training data. The cross-validation process is then repeated n times, with each of the n subsamples used exactly once as the validation data. The n results from the folds can then be averaged to produce a single estimation. The advantage of this method is all observations are used for both training and validation, and each observation is used for validation exactly once. The main disadvantage of jackknife test is it takes more time to complete a full training and testing process.

For our primary experiment we have performed 10-fold cross validation and for our final experiment we have performed jackknife test. Our main target for this experiment is to find out the novel features among our selected 46 features that gives the best result. As there are lots of combination (about $2^{46} - 1$ combinations, though we have not performed all this combinations, but we have performed a lots of combinations by adopting some mechanism which will be discussed in later section which reduces the combination size)

for feature and 10-fold cross validation takes less time and is a remarkable validation method, therefor we have performed 10-fold cross validation for selecting best features. After selecting best features we have performed jackknife test as it has been widely used in the literature for this task and has been shown to be the most consistent and reliable method. In this thesis paper all our reported result is using jackknife test.

4.4 Choosing The Final Features Set

The main aim of this research is to extract novel features for gram-positive and gram-negative benchmark. In earlier sections we have discussed in details how we have extracted features, how many features are there, how many methods are there for extracting features. In this section we are going to describe how we have got our final 6 features.

There are total 46 feature set in our hand and we want to know which features combination gives us good result. So we have to run about $2^{46} - 1$ combinations which is totally impossible to run and it needs lots of time. So we have skipped all this features together and took about 10 features at a time. This gives about 1023 combinations. There are total 8 (4 + 4 = 8) features in the literature [22], [69]. For our first run or cycle we have took these 8 features plus another 2 features in total 10 features from our features set. We have run these combinations and got a good combination from these which gives the highest result. For second iteration we took the highest prediction resulted combination and add some other features from the rest of the features that has not been taken yet, so that total features become 10 (every time a good combination features number is less than 10, so we add extra features from the features set that has not been taken yet to make total features 10 for the next iteration). We repeat this process as long as there is no feature in the features set that has not been taken yet. Thus we have complete our first cycle. After 1 cycle means taking all 46 features in iteration, we have got a good combination features which gives the best result. Again we took the best combination features from first cycle and add some other features to make total feature 10 and run our experiment. Thus we have performed second cycle. We take the highest resulted features from second cycle which is our final features set. We have got total 6 features with total feature vector size 235 which gives us best result (note that, we have only run these two cycles upon gram-positive bacteria protein dataset, not for gram-negative bacteria protein dataset. We have only extracted final 6 features from gram-negative benchmark), they are:

1. PSSM-AAO (feature vector size 20)
2. PSSM-SD (feature vector size 80)

Table 4.1: Comparing of single label classification result of gram-positive & gram-negative bacteria protein dataset

	Gram-positive benchmark accuracy			Gram-negative benchmark accuracy		
	Overall		Average	Overall		Average
	K-Fold test	Jackknife test	Jackknife test	K-Fold test	Jackknife test	Jackknife test
Huang and Yuan [41]	83.7	-	-	-	-	-
Pacharawongsakda [63]	-	-	-	73.2	-	-
Dehzangi [28]	83.6	-	-	76.6	-	-
Dehzangi [22]	87.7	88.2	-	79.6	80	-
Ronesh [69]	84.3	85	89.8	85	86	95.1
This Paper	-	**91.01**	**95.4**	-	**81.66**	**95.1**

3. PSSM-SAC (feature vector size 100)
4. Composition of Angles (feature vector size 4)
5. Auto Covariance of P(C) P(E) P(H) (feature vector size 30)
6. One-Lead Bi-gram of ASA (feature vector size 1)

4.5 Single Label Classification

In this paper we have first calculated the single label classification result. For single label classification we have calculated two types of accuracy, one is overall accuracy and another one is average accuracy. To calculate overall accuracy we have used **sensitivity** which will be discussed in later section and to calculate average accuracy we have used **average accuracy**. Average accuracy is computed as follows:

$$Average\ Accuracy = \frac{1}{n}\sum_{j=1}^{n} accuracy_j$$

here n is the number of class in the dataset. In later section we will discuss how we have calculated **accuracy**. A comparison of single label classification result for both benchmark (gram-positive & gram-negative) among recently published result and our result is shown in the table 4.1. This table shows that for gram-positive benchmark our result is best result among other paper, for gram-negative benchmark our result is a little lower than other result.

4.6 Multi Label Classification

Since the two benchmark datasets used in this thesis contains multi labeled proteins, so besides single label classification we have also performed multi

Table 4.2: Comparing of multi label classification result of gram-positive & gram-negative bacteria protein dataset

	Gram-positive benchmark		Gram-negative benchmark	
	Locative Accuracy	Absolute Accuracy	Locative Accuracy	Absolute Accuracy
Ronesh [69]	84.8	85.16	86.3	86
This Paper	91.71	90.94	83.84	81.25

label classification. For calculating multi-label classification result, we have used overall locative accuracy and overall absolute accuracy. The overall locative accuracy and overall absolute accuracy are defined as follows:

$$Locative\ Accuracy = \frac{1}{N_{dif}} \sum_{i=1}^{N_{dif}} Z_i$$

$$Absolute\ Accuracy = \frac{1}{N_{dif}} \sum_{i=1}^{N_{dif}} C_i$$

here N_{dif} is the number of different proteins, $Z_i = 1$ if at least one sub-cellular locations of the $i-th$ protein are correctly predicted and 0 otherwise, $C_i = 1$ if all the sub-cellular locations of query protein are exactly predicted and 0 otherwise. Therefore the overall absolute accuracy is striker than overall locative accuracy. A comparison of multi label classification result for both benchmark (gram-positive & gram-negative) among recently published result and our result is shown in the table 4.2.

4.7 Sensitivity, Specificity, MCC and Accuracy

To provide more information about the statistical significance of our achieved results, we have also performed Sensitivity, Specificity, Matthew's Correlation Coefficient (MCC) and accuracy. Sensitivity, specificity, MCC and accuracy are statistical measures of the performance of a binary classification test, also known in statistics as classification function. **Sensitivity** (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified. **Specificity** (also called the true negative rate) measures the proportion of negatives that are correctly identified. The value of sensitivity and specificity varies between 0 and 1. Having specificity, and sensitivity equal to 1 represents a fully accurate model while 0 represents a fully inaccurate. On the other hand, **MCC** measures the prediction quality of the model. MCC takes into account true and false positives and negatives and is generally

regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between 1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and 1 indicates total disagreement between prediction and observation. The **accuracy** refers to the total correctly classified instances over the number of samples present in the dataset. The value of accuracy varies between 0 and 1. The equation for calculating sensitivity, specificity, MCC and accuracy are given below:

$$Sensitivity = \frac{TP}{TP + FN} \times 100$$

$$Specificity = \frac{TN}{TN + FP} \times 100$$

$$MCC = \frac{(TN \times TP) - (TN \times FP)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \times 100$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100$$

where TP is the number of correctly identified (true positive) samples, FN is the number of incorrectly rejected samples (false negative), TN is the number of correctly rejected (true negative) samples, and FP is the number of incorrectly accepted samples (false positive).

In earlier section we have told that we have got final 6 features. The total combination of these 6 features is 63. In below tables we have shown sensitivity, specificity and MCC for these 63 combinations. Sensitivity for gram-positive bacteria protein dataset is shown in table 4.3. Sensitivity for gram-negative bacteria protein dataset is shown in table 4.6. Specificity for gram-positive bacteria protein dataset is shown in table 4.4. Specificity for gram-negative bacteria protein dataset is shown in table 4.7. MCC for gram-positive bacteria protein dataset is shown in table 4.5. MCC for gram-negative bacteria protein dataset is shown in table 4.8.

Table 4.3: Sensitivity For Gram-Positive Bacteria Protein Dataset

No.	Feature Combinations	Class1	Class2	Class3	Class4
1	One-Lead Bi-gram of ASA	0.402	0.000	0.832	0.000
2	PSSM-SAC	0.632	0.167	0.933	0.618
3	PSSM-SD	0.529	0.111	0.933	0.894
4	PSSM-AAO	0.580	0.167	0.913	0.821
5	Auto Covariance of P(C) P(E) P(H)	0.586	0.000	0.793	0.553
6	Composition of Angles	0.534	0.000	0.841	0.285
7	Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.586	0.000	0.793	0.553

Continued on next page

Table 4.3 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4
8	PSSM-AAO, One-Lead Bi-gram of ASA	0.638	0.167	0.904	0.764
9	PSSM-AAO, PSSM-SAC	0.678	0.111	0.913	0.854
10	PSSM-AAO, PSSM-SD	0.805	0.333	0.928	0.919
11	PSSM-AAO, Auto Covariance of P(C) P(E) P(H)	0.655	0.222	0.938	0.748
12	PSSM-AAO, Composition of Angles	0.603	0.167	0.947	0.724
13	PSSM-SD, One-Lead Bi-gram of ASA	0.684	0.111	0.889	0.862
14	PSSM-SD, PSSM-SAC	0.753	0.222	0.928	0.935
15	PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.672	0.167	0.885	0.789
16	PSSM-SD, Composition of Angles	0.586	0.222	0.918	0.837
17	PSSM-SAC, One-Lead Bi-gram of ASA	0.678	0.167	0.889	0.699
18	PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.661	0.111	0.909	0.756
19	PSSM-SAC, Composition of Angles	0.632	0.167	0.942	0.724
20	One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.615	0.000	0.755	0.553
21	One-Lead Bi-gram of ASA, Composition of Angles	0.638	0.000	0.755	0.333
22	One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.638	0.000	0.784	0.545
23	PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.713	0.167	0.913	0.740
24	PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.690	0.167	0.923	0.780
25	PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.678	0.111	0.918	0.691
26	PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.672	0.222	0.894	0.813
27	PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.799	0.222	0.923	0.870
28	PSSM-SD, PSSM-SAC, Composition of Angles	0.805	0.222	0.933	0.911
29	PSSM-SD, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.701	0.222	0.870	0.797
30	PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles	0.695	0.222	0.894	0.837
31	PSSM-AAO, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.644	0.222	0.923	0.772
32	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.810	0.222	0.899	0.902
33	PSSM-AAO, PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.782	0.278	0.933	0.902
34	PSSM-AAO, PSSM-SD, Composition of Angles	0.799	0.333	0.947	0.919
35	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA	0.678	0.111	0.904	0.789
36	PSSM-AAO, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.707	0.111	0.928	0.772
37	PSSM-AAO, PSSM-SAC, Composition of Angles	0.690	0.111	0.928	0.789
38	PSSM-AAO, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.672	0.222	0.928	0.780
39	PSSM-AAO, One-Lead Bi-gram of ASA, Composition of Angles	0.661	0.167	0.938	0.715
40	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA	0.833	0.389	0.928	0.927
41	PSSM-AAO, PSSM-SD, PSSM-SAC	0.833	0.556	0.909	0.927
42	PSSM-AAO, PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.851	0.500	0.952	0.943
43	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles	0.833	0.556	0.942	0.935
44	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.816	0.278	0.947	0.878
45	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles	0.828	0.278	0.933	0.919
46	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.856	0.556	0.928	0.927
47	PSSM-AAO, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.661	0.167	0.918	0.772
48	PSSM-AAO, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.730	0.111	0.933	0.772
49	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.701	0.111	0.928	0.772
50	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.701	0.111	0.923	0.780

Continued on next page

Table 4.3 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4
51	PSSM-AAO, PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.787	0.278	0.938	0.894
52	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.833	0.333	0.928	0.886
53	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.793	0.278	0.938	0.902
54	PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.724	0.222	0.875	0.821
55	PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.793	0.278	0.913	0.894
56	PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.713	0.167	0.913	0.748
57	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.856	0.389	0.918	0.894
58	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.718	0.111	0.933	0.764
59	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.891	0.500	0.957	0.943
60	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.862	0.500	0.933	0.927
61	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.822	0.278	0.933	0.878
62	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.868	0.500	0.942	0.919
63	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.897	0.500	0.957	0.935

Table 4.4: Specificity For Gram-Positive Bacteria Protein Dataset

No.	Feature Combinations	Class1	Class2	Class3	Class4
1	One-Lead Bi-gram of ASA	0.751	1.000	0.389	1.000
2	PSSM-SAC	0.936	0.992	0.749	0.899
3	PSSM-SD	0.991	0.996	0.707	0.919
4	PSSM-AAO	0.954	0.996	0.836	0.854
5	Auto Covariance of P(C) P(E) P(H)	0.893	1.000	0.707	0.846
6	Composition of Angles	0.907	1.000	0.502	0.909
7	Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.872	0.998	0.720	0.851
8	PSSM-AAO, One-Lead Bi-gram of ASA	0.922	0.994	0.849	0.859
9	PSSM-AAO, PSSM-SAC	0.945	0.992	0.878	0.856
10	PSSM-AAO, PSSM-SD	0.971	0.998	0.887	0.914
11	PSSM-AAO, Auto Covariance of P(C) P(E) P(H)	0.942	0.994	0.846	0.876
12	PSSM-AAO, Composition of Angles	0.936	0.994	0.810	0.876
13	PSSM-SD, One-Lead Bi-gram of ASA	0.965	0.994	0.804	0.919
14	PSSM-SD, PSSM-SAC	0.971	0.994	0.871	0.922
15	PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.933	0.996	0.781	0.907
16	PSSM-SD, Composition of Angles	0.965	0.994	0.746	0.929
17	PSSM-SAC, One-Lead Bi-gram of ASA	0.887	0.992	0.846	0.869
18	PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.922	0.994	0.830	0.879
19	PSSM-SAC, Composition of Angles	0.907	0.994	0.810	0.884

Continued on next page

Table 4.4 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4
20	One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.875	1.000	0.714	0.861
21	One-Lead Bi-gram of ASA, Composition of Angles	0.823	1.000	0.621	0.902
22	One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.864	0.998	0.743	0.861
23	PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.919	0.990	0.830	0.881
24	PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.919	0.994	0.855	0.884
25	PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.904	0.994	0.846	0.881
26	PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.936	0.994	0.791	0.912
27	PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.942	0.986	0.875	0.944
28	PSSM-SD, PSSM-SAC, Composition of Angles	0.959	0.992	0.894	0.939
29	PSSM-SD, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.936	0.994	0.810	0.912
30	PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles	0.939	0.994	0.801	0.934
31	PSSM-AAO, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.930	0.994	0.833	0.876
32	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.933	0.988	0.900	0.929
33	PSSM-AAO, PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.959	0.994	0.871	0.939
34	PSSM-AAO, PSSM-SD, Composition of Angles	0.974	0.998	0.875	0.932
35	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA	0.916	0.992	0.881	0.864
36	PSSM-AAO, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.933	0.988	0.871	0.886
37	PSSM-AAO, PSSM-SAC, Composition of Angles	0.928	0.992	0.878	0.879
38	PSSM-AAO, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.925	0.992	0.842	0.884
39	PSSM-AAO, One-Lead Bi-gram of ASA, Composition of Angles	0.928	0.992	0.839	0.889
40	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA	0.962	0.998	0.894	0.922
41	PSSM-AAO, PSSM-SD, PSSM-SAC	0.968	0.992	0.878	0.952
42	PSSM-AAO, PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.974	0.986	0.900	0.967
43	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles	0.977	0.994	0.891	0.970
44	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.968	0.990	0.894	0.934
45	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles	0.962	0.998	0.897	0.932
46	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.971	0.994	0.900	0.955
47	PSSM-AAO, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.925	0.992	0.849	0.884
48	PSSM-AAO, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.930	0.984	0.871	0.874
49	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.928	0.990	0.881	0.879
50	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.919	0.992	0.881	0.879
51	PSSM-AAO, PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.971	0.986	0.887	0.934
52	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.939	0.990	0.894	0.949
53	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.951	0.992	0.887	0.942
54	PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.930	0.998	0.820	0.922
55	PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.942	0.986	0.878	0.937
56	PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.916	0.992	0.852	0.884
57	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.951	0.988	0.900	0.949

Continued on next page

Table 4.4 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4
58	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.919	0.984	0.891	0.874
59	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.968	0.982	0.916	0.970
60	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.971	0.996	0.910	0.957
61	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.965	0.986	0.897	0.932
62	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.974	0.986	0.900	0.965
63	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.974	0.988	0.923	0.967

Table 4.5: MCC For Gram-Positive Bacteria Protein Dataset

No.	Feature Combinations	Class1	Class2	Class3	Class4
1	One-Lead Bi-gram of ASA	-0.071	0.000	-0.034	0.000
2	PSSM-SAC	0.513	-0.087	0.410	0.269
3	PSSM-SD	0.619	0.000	0.345	0.556
4	PSSM-AAO	0.526	0.104	0.549	0.275
5	Auto Covariance of P(C) P(E) P(H)	0.356	0.000	0.247	0.047
6	Composition of Angles	0.351	0.000	0.049	-0.009
7	Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.305	-0.231	0.265	0.062
8	PSSM-AAO, One-Lead Bi-gram of ASA	0.475	0.000	0.567	0.249
9	PSSM-AAO, PSSM-SAC	0.576	-0.189	0.633	0.307
10	PSSM-AAO, PSSM-SD	0.756	0.440	0.666	0.554
11	PSSM-AAO, Auto Covariance of P(C) P(E) P(H)	0.549	0.088	0.587	0.293
12	PSSM-AAO, Composition of Angles	0.490	0.000	0.527	0.274
13	PSSM-SD, One-Lead Bi-gram of ASA	0.645	-0.103	0.470	0.532
14	PSSM-SD, PSSM-SAC	0.717	0.088	0.632	0.595
15	PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.536	0.104	0.427	0.430
16	PSSM-SD, Composition of Angles	0.569	0.088	0.394	0.555
17	PSSM-SAC, One-Lead Bi-gram of ASA	0.414	-0.087	0.549	0.231
18	PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.494	-0.103	0.533	0.307
19	PSSM-SAC, Composition of Angles	0.431	0.000	0.523	0.300
20	One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.335	0.000	0.229	0.091
21	One-Lead Bi-gram of ASA, Composition of Angles	0.237	0.000	0.114	0.017
22	One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.326	-0.231	0.291	0.084
23	PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.526	-0.163	0.537	0.303
24	PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.508	0.000	0.595	0.343
25	PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.459	-0.103	0.572	0.266
26	PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.545	0.088	0.451	0.467
27	PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.660	-0.209	0.634	0.642
28	PSSM-SD, PSSM-SAC, Composition of Angles	0.718	0.000	0.684	0.650
29	PSSM-SD, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.567	0.088	0.467	0.455
30	PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles	0.572	0.088	0.468	0.576
31	PSSM-AAO, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.505	0.088	0.550	0.311

Continued on next page

Table 4.5 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4
32	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.643	-0.146	0.670	0.602
33	PSSM-AAO, PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.701	0.164	0.635	0.644
34	PSSM-AAO, PSSM-SD, Composition of Angles	0.761	0.440	0.654	0.624
35	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA	0.491	-0.189	0.632	0.283
36	PSSM-AAO, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.563	-0.326	0.632	0.345
37	PSSM-AAO, PSSM-SAC, Composition of Angles	0.533	-0.189	0.645	0.332
38	PSSM-AAO, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.511	0.000	0.573	0.343
39	PSSM-AAO, One-Lead Bi-gram of ASA, Composition of Angles	0.510	-0.087	0.574	0.310
40	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA	0.749	0.494	0.680	0.589
41	PSSM-AAO, PSSM-SD, PSSM-SAC	0.768	0.373	0.630	0.716
42	PSSM-AAO, PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.800	0.116	0.713	0.797
43	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles	0.796	0.453	0.684	0.803
44	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.755	0.000	0.695	0.605
45	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles	0.745	0.380	0.691	0.624
46	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.794	0.453	0.694	0.727
47	PSSM-AAO, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.502	-0.087	0.578	0.337
48	PSSM-AAO, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.572	-0.437	0.635	0.303
49	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.542	-0.262	0.652	0.320
50	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.517	-0.189	0.648	0.326
51	PSSM-AAO, PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.743	-0.133	0.673	0.617
52	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.677	0.070	0.680	0.676
53	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.682	0.077	0.673	0.655
54	PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.568	0.312	0.488	0.512
55	PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.655	-0.133	0.633	0.628
56	PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.518	-0.087	0.581	0.318
57	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.730	0.064	0.686	0.682
58	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.530	-0.437	0.677	0.297
59	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Auto Covariance of P(C) P(E) P(H)	0.811	0.000	0.754	0.809
60	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles	0.799	0.492	0.719	0.739
61	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.750	-0.133	0.691	0.595
62	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.813	0.116	0.706	0.768
63	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.834	0.180	0.769	0.791

Table 4.6: Sensitivity For Gram-Negative Bacteria Protein Dataset

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
1	Composition of Angles	0.745	0.476	0.878	0.143	0.000	0.000	0.000	0.000
2	PSSM-SAC	0.833	0.435	0.959	0.489	0.031	0.917	0.000	0.572
3	PSSM-SD	0.806	0.226	0.961	0.451	0.000	0.000	0.000	0.461
4	PSSM-AAO	0.840	0.444	0.954	0.571	0.125	0.750	0.000	0.450
5	One-Lead Bi-gram of ASA	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	Auto Covariance of P(C) P(E) P(H)	0.829	0.524	0.827	0.444	0.094	0.083	0.000	0.100
7	Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.844	0.524	0.824	0.421	0.156	0.000	0.000	0.100
8	PSSM-AAO, Composition of Angles	0.840	0.516	0.937	0.571	0.750	0.750	0.000	0.506
9	PSSM-AAO, PSSM-SAC	0.837	0.492	0.946	0.586	0.594	0.917	0.000	0.667
10	PSSM-AAO, PSSM-SD	0.882	0.468	0.954	0.586	0.375	0.917	0.000	0.706
11	PSSM-AAO, One-Lead Bi-gram of ASA	0.856	0.468	0.941	0.571	0.344	0.750	0.000	0.550
12	PSSM-AAO, Auto Covariance of P(C) P(E) P(H)	0.858	0.524	0.920	0.647	0.813	0.750	0.000	0.539
13	PSSM-SD, Composition of Angles	0.813	0.492	0.898	0.534	0.438	0.167	0.000	0.394
14	PSSM-SD, PSSM-SAC	0.838	0.444	0.949	0.586	0.250	0.917	0.000	0.711
15	PSSM-SD, One-Lead Bi-gram of ASA	0.844	0.234	0.927	0.459	0.000	0.000	0.000	0.533
16	PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.846	0.492	0.917	0.556	0.688	0.583	0.000	0.439
17	PSSM-SAC, Composition of Angles	0.829	0.476	0.944	0.549	0.531	0.833	0.000	0.522
18	PSSM-SAC, One-Lead Bi-gram of ASA	0.855	0.452	0.944	0.541	0.063	0.917	0.000	0.594
19	PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.858	0.508	0.922	0.632	0.719	0.917	0.000	0.561
20	Composition of Angles, One-Lead Bi-gram of ASA	0.808	0.500	0.824	0.188	0.000	0.000	0.000	0.000
21	Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.837	0.500	0.824	0.436	0.594	0.000	0.000	0.106
22	Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.844	0.508	0.824	0.406	0.688	0.000	0.000	0.111
23	PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.858	0.524	0.934	0.647	0.781	0.917	0.000	0.583
24	PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.853	0.492	0.929	0.579	0.531	0.833	0.000	0.578
25	PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.862	0.492	0.934	0.609	0.781	0.917	0.000	0.578
26	PSSM-SD, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.864	0.492	0.915	0.564	0.719	0.583	0.000	0.439
27	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.864	0.468	0.937	0.586	0.250	0.917	0.000	0.717
28	PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.874	0.524	0.924	0.617	0.781	1.000	0.000	0.667
29	PSSM-SD, Composition of Angles, One-Lead Bi-gram of ASA	0.847	0.508	0.907	0.511	0.406	0.167	0.000	0.433
30	PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.851	0.484	0.910	0.534	0.813	0.500	0.000	0.483
31	PSSM-AAO, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.860	0.516	0.922	0.624	0.813	0.750	0.000	0.567

Continued on next page

Table 4.6 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
32	PSSM-SD, PSSM-SAC, Composition of Angles	0.846	0.516	0.941	0.571	0.688	0.917	0.000	0.667
33	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA	0.903	0.460	0.944	0.602	0.531	0.917	0.000	0.689
34	PSSM-AAO, PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.903	0.516	0.937	0.654	0.813	1.000	0.000	0.617
35	PSSM-AAO, PSSM-SAC, Composition of Angles	0.833	0.548	0.941	0.602	0.781	0.917	0.000	0.611
36	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA	0.856	0.500	0.944	0.579	0.656	0.917	0.000	0.650
37	PSSM-AAO, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.862	0.524	0.934	0.662	0.813	1.000	0.000	0.633
38	PSSM-AAO, Composition of Angles, One-Lead Bi-gram of ASA	0.856	0.516	0.929	0.541	0.750	0.750	0.000	0.522
39	PSSM-AAO, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.862	0.581	0.922	0.579	0.813	0.750	0.000	0.556
40	PSSM-AAO, PSSM-SD, Composition of Angles	0.887	0.532	0.941	0.617	0.813	0.917	0.000	0.667
41	PSSM-AAO, PSSM-SD, PSSM-SAC	0.896	0.484	0.946	0.624	0.625	1.000	0.000	0.756
42	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.916	0.492	0.944	0.602	0.656	1.000	0.000	0.733
43	PSSM-AAO, PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.899	0.516	0.939	0.654	0.813	1.000	0.000	0.661
44	PSSM-AAO, PSSM-SD, Composition of Angles, One-Lead Bi-gram of ASA	0.903	0.540	0.932	0.617	0.781	0.917	0.000	0.633
45	PSSM-AAO, PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.905	0.540	0.929	0.639	0.813	1.000	0.000	0.644
46	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles	0.896	0.532	0.944	0.594	0.844	1.000	0.000	0.722
47	PSSM-AAO, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.858	0.581	0.920	0.602	0.813	0.750	0.000	0.561
48	PSSM-AAO, PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.865	0.540	0.934	0.662	0.813	1.000	0.000	0.628
49	PSSM-AAO, PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.855	0.548	0.937	0.594	0.813	0.917	0.000	0.639
50	PSSM-AAO, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.862	0.524	0.937	0.662	0.813	1.000	0.000	0.672
51	PSSM-AAO, PSSM-SD, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.905	0.532	0.932	0.669	0.813	1.000	0.000	0.622
52	PSSM-SD, PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.865	0.556	0.937	0.564	0.750	0.917	0.000	0.689
53	PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.882	0.516	0.929	0.609	0.813	1.000	0.000	0.694

Continued on next page

Table 4.6 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
54	PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.855	0.492	0.910	0.541	0.813	0.500	0.000	0.506
55	PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.878	0.532	0.929	0.632	0.781	1.000	0.000	0.639
56	PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.867	0.500	0.932	0.602	0.781	0.833	0.000	0.561
57	PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.880	0.516	0.920	0.624	0.781	1.000	0.000	0.656
58	PSSM-AAO, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.858	0.516	0.939	0.654	0.813	1.000	0.000	0.650
59	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.903	0.565	0.937	0.609	0.813	1.000	0.000	0.711
60	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.898	0.540	0.934	0.669	0.813	1.000	0.000	0.694
61	PSSM-AAO, PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.907	0.540	0.929	0.669	0.813	1.000	0.000	0.622
62	PSSM-AAO, PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.907	0.532	0.934	0.647	0.813	1.000	0.000	0.689
63	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.905	0.524	0.937	0.647	0.812	1.000	0.000	0.683

Table 4.7: Specificity For Gram-Negative Bacteria Protein Dataset

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
1	Composition of Angles	0.879	0.968	0.587	0.976	1.000	1.000	1.000	0.999
2	PSSM-SAC	0.939	0.983	0.833	0.971	0.999	0.999	0.999	0.924
3	PSSM-SD	0.945	0.991	0.679	0.964	1.000	1.000	0.998	0.903
4	PSSM-AAO	0.921	0.988	0.823	0.959	0.995	0.996	0.999	0.930
5	One-Lead Bi-gram of ASA	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6	Auto Covariance of P(C) P(E) P(H)	0.855	0.951	0.739	0.943	0.998	1.000	1.000	0.969
7	Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.842	0.955	0.737	0.944	0.998	1.000	1.000	0.963
8	PSSM-AAO, Composition of Angles	0.933	0.975	0.851	0.963	0.983	0.997	0.999	0.949
9	PSSM-AAO, PSSM-SAC	0.957	0.983	0.883	0.959	0.994	0.999	0.999	0.931
10	PSSM-AAO, PSSM-SD	0.959	0.981	0.875	0.959	0.996	0.999	0.999	0.932
11	PSSM-AAO, One-Lead Bi-gram of ASA	0.920	0.984	0.849	0.956	0.992	0.996	0.999	0.935
12	PSSM-AAO, Auto Covariance of P(C) P(E) P(H)	0.926	0.982	0.882	0.952	0.988	0.997	0.999	0.947

Continued on next page

Table 4.7 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
13	PSSM-SD, Composition of Angles	0.927	0.973	0.769	0.961	0.993	1.000	0.999	0.942
14	PSSM-SD, PSSM-SAC	0.965	0.979	0.865	0.964	0.999	0.999	0.998	0.913
15	PSSM-SD, One-Lead Bi-gram of ASA	0.915	0.988	0.761	0.964	1.000	1.000	0.998	0.898
16	PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.920	0.978	0.838	0.945	0.989	0.999	0.999	0.939
17	PSSM-SAC, Composition of Angles	0.926	0.980	0.850	0.967	0.992	0.999	0.999	0.935
18	PSSM-SAC, One-Lead Bi-gram of ASA	0.909	0.983	0.859	0.968	0.999	0.999	0.999	0.929
19	PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.921	0.981	0.886	0.956	0.989	0.999	0.999	0.946
20	Composition of Angles, One-Lead Bi-gram of ASA	0.828	0.956	0.670	0.966	1.000	1.000	1.000	0.997
21	Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.846	0.970	0.741	0.939	0.989	1.000	1.000	0.956
22	Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.846	0.976	0.747	0.949	0.989	1.000	1.000	0.955
23	PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.921	0.979	0.889	0.955	0.990	0.999	0.999	0.946
24	PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.926	0.977	0.869	0.967	0.993	0.999	0.999	0.933
25	PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.922	0.979	0.887	0.958	0.990	0.999	0.999	0.946
26	PSSM-SD, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.920	0.978	0.846	0.947	0.989	0.999	0.999	0.942
27	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.949	0.979	0.883	0.963	0.996	0.999	0.999	0.922
28	PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.946	0.979	0.898	0.956	0.988	0.999	0.999	0.943
29	PSSM-SD, Composition of Angles, One-Lead Bi-gram of ASA	0.913	0.968	0.799	0.964	0.996	0.999	0.998	0.946
30	PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.916	0.977	0.833	0.952	0.991	0.999	0.999	0.941
31	PSSM-AAO, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.934	0.979	0.876	0.955	0.988	0.997	0.999	0.950
32	PSSM-SD, PSSM-SAC, Composition of Angles	0.958	0.976	0.885	0.961	0.988	0.999	0.999	0.931
33	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA	0.949	0.982	0.885	0.958	0.993	0.999	0.999	0.943
34	PSSM-AAO, PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.951	0.982	0.906	0.952	0.986	0.998	0.999	0.948
35	PSSM-AAO, PSSM-SAC, Composition of Angles	0.935	0.983	0.885	0.963	0.988	0.999	0.999	0.944
36	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA	0.939	0.980	0.894	0.960	0.993	0.999	0.999	0.942
37	PSSM-AAO, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.937	0.983	0.900	0.957	0.988	0.999	0.999	0.945
38	PSSM-AAO, Composition of Angles, One-Lead Bi-gram of ASA	0.938	0.972	0.874	0.967	0.986	0.997	0.999	0.948
39	PSSM-AAO, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.927	0.978	0.882	0.952	0.989	0.996	0.999	0.952

Continued on next page

Table 4.7 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
40	PSSM-AAO, PSSM-SD, Composition of Angles	0.958	0.979	0.884	0.963	0.988	0.999	0.999	0.948
41	PSSM-AAO, PSSM-SD, PSSM-SAC	0.976	0.979	0.902	0.960	0.993	0.999	0.999	0.935
42	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.963	0.978	0.910	0.959	0.992	0.999	0.999	0.941
43	PSSM-AAO, PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.962	0.982	0.914	0.956	0.988	0.999	0.999	0.948
44	PSSM-AAO, PSSM-SD, Composition of Angles, One-Lead Bi-gram of ASA	0.958	0.974	0.894	0.962	0.990	0.999	0.999	0.950
45	PSSM-AAO, PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.950	0.979	0.906	0.953	0.989	0.999	0.999	0.946
46	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles	0.969	0.983	0.901	0.960	0.989	0.999	0.999	0.942
47	PSSM-AAO, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.934	0.972	0.887	0.956	0.988	0.997	0.999	0.946
48	PSSM-AAO, PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.939	0.977	0.902	0.953	0.988	0.999	0.999	0.946
49	PSSM-AAO, PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.943	0.979	0.893	0.964	0.991	0.999	0.999	0.946
50	PSSM-AAO, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.937	0.979	0.897	0.958	0.990	0.999	0.999	0.946
51	PSSM-AAO, PSSM-SD, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.949	0.977	0.908	0.951	0.987	0.999	0.999	0.944
52	PSSM-SD, PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.958	0.974	0.889	0.965	0.989	0.999	0.999	0.936
53	PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.944	0.973	0.902	0.956	0.990	0.999	0.999	0.941
54	PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.926	0.977	0.843	0.951	0.991	0.999	0.999	0.941
55	PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.951	0.975	0.900	0.950	0.988	0.999	0.999	0.937
56	PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.928	0.976	0.891	0.954	0.990	0.999	0.999	0.949
57	PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.953	0.971	0.908	0.950	0.989	0.999	0.999	0.944
58	PSSM-AAO, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.941	0.979	0.901	0.954	0.988	0.999	0.999	0.941

Continued on next page

Table 4.7 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
59	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.963	0.976	0.908	0.959	0.991	0.999	0.999	0.947
60	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.958	0.976	0.910	0.955	0.990	0.999	0.999	0.943
61	PSSM-AAO, PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.958	0.977	0.908	0.952	0.988	0.999	0.999	0.947
62	PSSM-AAO, PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.966	0.978	0.918	0.952	0.988	0.999	0.999	0.950
63	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.970	0.977	0.911	0.949	0.989	0.999	0.999	0.941

Table 4.8: MCC For Gram-Negative Bacteria Protein Dataset

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
1	Composition of Angles	0.503	0.164	-0.060	-0.129	0.000	0.000	0.000	-0.070
8	PSSM-SAC	0.706	0.330	0.433	0.231	0.000	0.833	-0.352	0.055
3	PSSM-SD	0.700	0.235	0.119	0.121	0.000	0.000	-0.610	-0.168
4	PSSM-AAO	0.670	0.416	0.404	0.176	-0.158	0.308	-0.352	-0.021
5	One-Lead Bi-gram of ASA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	Auto Covariance of P(C) P(E) P(H)	0.520	0.023	0.138	-0.093	0.000	0.288	0.000	-0.184
7	Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.505	0.062	0.133	-0.102	0.124	0.000	0.000	-0.233
8	PSSM-AAO, Composition of Angles	0.698	0.283	0.464	0.226	0.025	0.399	-0.352	0.164
9	PSSM-AAO, PSSM-SAC	0.754	0.372	0.558	0.197	0.371	0.720	-0.352	0.177
10	PSSM-AAO, PSSM-SD	0.796	0.325	0.540	0.189	0.248	0.720	-0.352	0.213
11	PSSM-AAO, One-Lead Bi-gram of ASA	0.680	0.374	0.462	0.144	0.000	0.308	-0.352	0.104
12	PSSM-AAO, Auto Covariance of P(C) P(E) P(H)	0.695	0.389	0.535	0.179	0.271	0.399	-0.352	0.182
13	PSSM-SD, Composition of Angles	0.662	0.239	0.244	0.167	0.143	0.407	-0.498	0.006
14	PSSM-SD, PSSM-SAC	0.777	0.267	0.508	0.248	0.332	0.833	-0.610	0.105
15	PSSM-SD, One-Lead Bi-gram of ASA	0.659	0.182	0.247	0.129	0.000	0.000	-0.610	-0.128
16	PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.671	0.303	0.416	0.034	0.202	0.611	-0.498	0.028
17	PSSM-SAC, Composition of Angles	0.672	0.322	0.467	0.250	0.199	0.666	-0.352	0.079
18	PSSM-SAC, One-Lead Bi-gram of ASA	0.654	0.334	0.491	0.252	0.000	0.720	-0.352	0.105
19	PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.684	0.363	0.549	0.197	0.227	0.833	-0.352	0.196
20	Composition of Angles, One-Lead Bi-gram of ASA	0.450	0.047	0.021	-0.178	0.000	0.000	0.000	-0.139
21	Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.507	0.207	0.140	-0.133	0.120	0.000	0.000	-0.274

Continued on next page

Table 4.8 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
22	Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.512	0.286	0.151	-0.075	0.202	0.000	0.000	-0.276
23	PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.684	0.355	0.567	0.202	0.309	0.833	-0.498	0.216
24	PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.691	0.292	0.505	0.267	0.236	0.666	-0.352	0.118
25	PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.690	0.325	0.561	0.201	0.309	0.833	-0.352	0.205
26	PSSM-SD, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.686	0.303	0.435	0.055	0.227	0.611	-0.498	0.051
27	PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.754	0.302	0.551	0.231	0.146	0.833	-0.498	0.159
28	PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.757	0.344	0.586	0.192	0.216	0.880	-0.352	0.262
29	PSSM-SD, Composition of Angles, One-Lead Bi-gram of ASA	0.656	0.195	0.316	0.179	0.281	0.166	-0.610	0.069
30	PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.667	0.284	0.398	0.079	0.399	0.407	-0.498	0.089
31	PSSM-AAO, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.716	0.348	0.520	0.182	0.241	0.399	-0.352	0.226
32	PSSM-SD, PSSM-SAC, Composition of Angles	0.764	0.304	0.560	0.201	0.170	0.720	-0.498	0.177
33	PSSM-AAO, PSSM-SD, One-Lead Bi-gram of ASA	0.786	0.330	0.562	0.194	0.275	0.720	-0.352	0.280
34	PSSM-AAO, PSSM-SD, Auto Covariance of P(C) P(E) P(H)	0.792	0.382	0.620	0.177	0.183	0.670	-0.352	0.256
35	PSSM-AAO, PSSM-SAC, Composition of Angles	0.698	0.422	0.560	0.252	0.246	0.720	-0.352	0.221
36	PSSM-AAO, PSSM-SAC, One-Lead Bi-gram of ASA	0.725	0.344	0.589	0.199	0.347	0.720	-0.352	0.242
37	PSSM-AAO, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.724	0.413	0.600	0.238	0.241	0.771	-0.352	0.246
38	PSSM-AAO, Composition of Angles, One-Lead Bi-gram of ASA	0.722	0.252	0.519	0.243	0.133	0.399	-0.352	0.173
39	PSSM-AAO, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.701	0.383	0.537	0.113	0.301	0.308	-0.352	0.235
40	PSSM-AAO, PSSM-SD, Composition of Angles	0.797	0.351	0.557	0.265	0.271	0.720	-0.352	0.298
41	PSSM-AAO, PSSM-SD, PSSM-SAC	0.850	0.307	0.615	0.238	0.358	0.771	-0.352	0.275
42	PSSM-AAO, PSSM-SD, PSSM-SAC, One-Lead Bi-gram of ASA	0.832	0.303	0.638	0.203	0.310	0.880	-0.352	0.297
43	PSSM-AAO, PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H)	0.816	0.382	0.647	0.223	0.241	0.880	-0.352	0.294
44	PSSM-AAO, PSSM-SD, Composition of Angles, One-Lead Bi-gram of ASA	0.810	0.295	0.580	0.248	0.309	0.720	-0.352	0.283
45	PSSM-AAO, PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.790	0.358	0.615	0.180	0.301	0.771	-0.352	0.268

Continued on next page

Table 4.8 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
46	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles	0.832	0.408	0.610	0.212	0.325	0.771	-0.352	0.300
47	PSSM-AAO, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.715	0.300	0.550	0.171	0.271	0.399	-0.352	0.190
48	PSSM-AAO, PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.732	0.337	0.606	0.199	0.271	0.771	-0.352	0.254
49	PSSM-AAO, PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.732	0.377	0.580	0.254	0.399	0.720	-0.352	0.257
50	PSSM-AAO, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.724	0.344	0.592	0.246	0.365	0.771	-0.352	0.284
51	PSSM-AAO, PSSM-SD, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.787	0.329	0.623	0.182	0.211	0.771	-0.352	0.231
52	PSSM-SD, PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.780	0.309	0.568	0.236	0.253	0.720	-0.498	0.233
53	PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.757	0.262	0.602	0.177	0.333	0.880	-0.352	0.267
54	PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.692	0.292	0.423	0.071	0.399	0.544	-0.498	0.109
55	PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.772	0.298	0.596	0.143	0.216	0.880	-0.352	0.198
56	PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.708	0.278	0.571	0.155	0.309	0.782	-0.352	0.215
57	PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.779	0.232	0.614	0.137	0.277	0.880	-0.352	0.259
58	PSSM-AAO, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.732	0.348	0.607	0.200	0.271	0.771	-0.352	0.230
59	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, One-Lead Bi-gram of ASA	0.822	0.347	0.627	0.217	0.399	0.771	-0.352	0.328
60	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H)	0.806	0.315	0.631	0.220	0.365	0.880	-0.352	0.284
61	PSSM-AAO, PSSM-SD, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.813	0.337	0.621	0.197	0.271	0.771	-0.352	0.255
62	PSSM-AAO, PSSM-SD, PSSM-SAC, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.834	0.340	0.654	0.179	0.241	0.880	-0.352	0.335

Continued on next page

Table 4.8 – continued from previous page

No.	Feature Combinations	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
63	PSSM-AAO, PSSM-SD, PSSM-SAC, Composition of Angles, Auto Covariance of P(C) P(E) P(H), One-Lead Bi-gram of ASA	0.842	0.322	0.636	0.149	0.301	0.880	-0.352	0.263

4.8 Summary

This chapter explained the way we have performed our experiment and prepared result. We have mentioned here how we have extracted our features, how we have choose classifier, how we have done parameter tunning, how we have choose validation method and when we have used which validation method, how we have choose the final features set which is our ultimate goal. A detail comparison has been made with previous literatures for both single label and multi label classification.

Chapter 5

Conclusion

5.1 Summary

In this study, we have proposed some novel feature extraction techniques for predicting gram-positive bacteria protein and gram-negative bacteria protein sub-cellular localization. We have extract these features from PSSM matrix, Normalized PSSM matrix and Spider SPD3 matrix. We have extracted total 46 features using total 10 feature extraction methods from gram-positive bacteria protein dataset. We have found that among these 46 features combination only 6 features which total feature vector size is 235 gives us our desired result. These features are PSSM-AAO (feature vector size 20), PSSM-SD (feature vector size 80), PSSM-SAC (feature vector size 100), Composition of Angles (feature vector size 4), Auto Covariance of P(C) P(E) P(H) (feature vector size 30), One-Lead Bi-gram of ASA (feature vector size 1). We have extracted only these 6 features from gram-negative bacteria protein dataset and produced our result. In our final experiment means from which we have reported results in this paper, we have used these 6 features for both data set, Support Vector Machine (SVM) with RBF kernel, Cost parameter $(C) = 3000$, $\gamma = 0.005$ as classifier and jackknife test as validation method. As both datasets are multi-label dataset, so we have reported both single label and multi label prediction accuracy. For single label classification our reported result for gram-positive benchmark is 91.01% which is higher than any other literature published in this sector, for gram-negative benchmark is 81.66% which is little lower then previous study. For multi label classification our reported result for gram-positive benchmark is for locative accuracy 91.71%, for absolute accuracy 90.94% which is higher than any other literature published in this sector, for gram-negative benchmark is for locative accuracy 83.84%, for absolute accuracy 81.25 which is little lower then previous study. These enhancements highlight the effectiveness of the proposed method to explore the potential information embedded in the PSSM matrix and Spider SPD3 matrix.

5.2 Limitations

Though we have done a lot of experiments for producing the best result, there are some limitations in our work. These limitations are described below:

1. We have only extracted PSSM-SD feature from PSSM matrix not from Spider SPD3 matrix.
2. In PSSM-SD feature extraction method we have only tried three values (5, 10 and 25) for F_p.
3. We have only extracted PSSM-SAC feature from PSSM matrix not from Spider SPD3 matrix.
4. In PSSM-SAC feature extraction method we have only tried ten values (1,2,3,.....,8,9,10) for K_p.
5. In Auto Covariance feature extraction method we have only tested 15 values (1,2,3,.....13,14,15) for distance factor (DF) and took $DF = 10$ as it is reported in literature [69] which gives the best result and we have also investigated that this gives the best result among these 15 values.
6. In this literature we have only used Support Vector Machine (SVM) as our classifier.
7. In this experiment we have performed a little optimization for SVM parameters (γ and C).
8. In this experiment we have extracted 46 features using 10 extraction methods only for gram-positive bacteria protein dataset. We have got 6 best features from these 46 features and only extracted those 6 features for gram-negative bacteria protein dataset.
9. In this experiment we have only tried two types (gram-positive bacteria protein and gram-negative bacteria protein) of protein dataset.

5.3 Future Work

As there are some limitations in our current work, so we have planned to eliminate these limitations in future work. Besides this we have planned to do some extra things which may increase our accuracy rate. These future works are described below:

1. Try new different feature extraction methods.

2. We will try to extract PSSM-SD feature from Spider matrix.

3. In PSSM-SD feature extraction method we will try other value of F_p along with previous value to extract features from both PSSM matrix and Spider matrix.

4. We will try to extract PSSM-SAC feature from Spider matrix.

5. In PSSM-SAC feature extraction method we will try other value of K_p along with previous value to extract features from both PSSM matrix and Spider matrix.

6. In Auto Covariance feature extraction method we will try other valu es along with previous values which have been mentioned in previous section for distance factor (DF) to see the effectiveness of our result.

7. We will try other supervised learning algorithm for e.g. Decision Tree, Random Forest, Naive Base Classifier, K-Nearest Neighbor, AdaBoost etc along with SVM to see the effectiveness of our result.

8. We will try to optimize SVM parameters (γ and C) more to see whether it increases the accuracy or not.

9. In this paper our reported result for gram-negative bacteria protein dataset is little lower than previous literature result published in this sector. We think the main reason for getting lower result for this dataset is extracting preselected 6 features and produce the result using them. In future, we have planned to extract all these 46 features for gram-negative bacteria protein dataset and find out the best feature combinations to see whether this improves our current result or not.

10. We will try to make a web application for global users, so that they can give a protein or amino acids sequence as input and our web application will analyze it and predict which location(s) this protein sits as an output. This is our main goal in future.

11. We will try our new method and technique to predict other protein dataset such as eukaryotic, human, animal, virus protein dataset etc.

References

[1] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[2] Medha Bhagwat and L Aravind. Psi-blast tutorial. *Comparative Genomics*, pages 177–186, 2008.

[3] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[4] Jia-Ming Chang, Emily Chia-Yu Su, Allan Lo, Hua-Sheng Chiu, Ting-Yi Sung, and Wen-Lian Hsu. Psldoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. *Proteins: Structure, Function, and Bioinformatics*, 72(2):693–710, 2008.

[5] HU Chen, NI Huang, and Zhirong Sun. Subloc: a server/client suite for protein subcellular location based on soap. *Bioinformatics*, 2005.

[6] Wei Chen, Peng-Mian Feng, En-Ze Deng, Hao Lin, and Kuo-Chen Chou. itis-psetnc: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical biochemistry*, 462:76–83, 2014.

[7] Wei Chen, Tian-Yu Lei, Dian-Chuan Jin, Hao Lin, and Kuo-Chen Chou. Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition. *Analytical biochemistry*, 456:53–60, 2014.

[8] Arunkumar Chinnasamy, Wing-Kin Sung, and Ankush Mittal. Protein structure and fold prediction using tree-augmented naive bayesian classifier. *Journal of Bioinformatics and Computational Biology*, 3(04):803–819, 2005.

[9] WiesŁaw Chmielnicki, Katarzyna Sta, et al. A hybrid discriminative/generative approach to protein fold recognition. *Neurocomputing*, 75(1):194–198, 2012.

[10] KC Chou. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Current topics in medicinal chemistry*, 2017.

[11] Kuo-Chen Chou. Prediction of protein structural classes and subcellular locations. *Current protein and peptide science*, 1(2):171–208, 2000.

[12] Kuo-Chen Chou. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics*, 6(4):262–274, 2009.

[13] Kuo-Chen Chou. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology*, 273(1):236–247, 2011.

[14] Kuo-Chen Chou. Impacts of bioinformatics to medicinal chemistry. *Medicinal chemistry*, 11(3):218–234, 2015.

[15] Kuo-Chen Chou and Hong-Bin Shen. Large-scale predictions of gram-negative bacterial protein subcellular locations. *Journal of proteome research*, 5(12):3420–3428, 2006.

[16] Kuo-Chen Chou and Hong-Bin Shen. Recent progress in protein subcellular location prediction. *Analytical biochemistry*, 370(1):1–16, 2007.

[17] Kuo-Chen Chou and Hong-Bin Shen. Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nature protocols*, 3(2):153–162, 2008.

[18] Kuo-Chen Chou and Hong-Bin Shen. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mploc 2.0. *PLoS One*, 5(4):e9931, 2010.

[19] Kuo-Chen Chou and Hong-Bin Shen. Plant-mploc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PloS one*, 5(6):e11335, 2010.

[20] Kuo-Chen Chou, Hong-Bin Shen, et al. Cell-ploc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Natural Science*, 2(10):1090, 2010.

[21] Kuo-Chen Chou, Zhi-Cheng Wu, and Xuan Xiao. iloc-hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Molecular Biosystems*, 8(2):629–641, 2012.

[22] Abdollah Dehzangi, Rhys Heffernan, Alok Sharma, James Lyons, Kuldip Paliwal, and Abdul Sattar. Gram-positive and gram-negative

protein subcellular localization by incorporating evolutionary-based descriptors into chou s general pseaac. *Journal of theoretical biology*, 364:284–294, 2015.

[23] Abdollah Dehzangi, Kuldip Paliwal, Alok Sharma, Omid Dehzangi, and Abdul Sattar. A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. *IEEE/ACM transactions on computational biology and bioinformatics*, 10(3):564–575, 2013.

[24] Abdollah Dehzangi, Kuldip K Paliwal, Alok Sharma, James G Lyons, and Abdul Sattar. Protein fold recognition using an overlapping segmentation approach and a mixture of feature extraction models. In *Australasian Conference on Artificial Intelligence*, pages 32–43. Springer, 2013.

[25] Abdollah Dehzangi and Somnuk Phon-Amnuaisuk. Fold prediction problem: The application of new physical and physicochemical-based features. *Protein and Peptide Letters*, 18(2):174–185, 2011.

[26] Abdollah Dehzangi, Somnuk Phon Amnuaisuk, Keng Ng, and Ehsan Mohandesi. Protein fold prediction problem using ensemble of classifiers. In *Neural Information Processing*, pages 503–511. Springer, 2009.

[27] Abdollah Dehzangi and Abdul Sattar. Ensemble of diversely trained support vector machines for protein fold recognition. In *ACIIDS (1)*, pages 335–344, 2013.

[28] Abdollah Dehzangi, Sohrab Sohrabi, Rhys Heffernan, Alok Sharma, James Lyons, Kuldip Paliwal, and Abdul Sattar. Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features. *BMC bioinformatics*, 16(4):S1, 2015.

[29] Francesca Demichelis, Paolo Magni, Paolo Piergiorgi, Mark A Rubin, and Riccardo Bellazzi. A hierarchical naive bayes model for handling sample heterogeneity in classification problems: an application to tissue microarrays. *BMC bioinformatics*, 7(1):514, 2006.

[30] Chris HQ Ding and Inna Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–358, 2001.

[31] Yong-Sheng Ding and Tong-Liang Zhang. Using chous pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. *Pattern Recognition Letters*, 29(13):1887–1892, 2008.

[32] Qiwen Dong, Shuigeng Zhou, and Jihong Guan. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics*, 25(20):2655–2662, 2009.

[33] Inna Dubchak, Ilya B Muchnik, and Sung-Hou Kim. Protein folding class predictor for scop: approach based on global descriptors. In *Ismb*, pages 104–107, 1997.

[34] Olof Emanuelsson, Henrik Nielsen, Søren Brunak, and Gunnar Von Heijne. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *Journal of molecular biology*, 300(4):1005–1016, 2000.

[35] Maryam Esmaeili, Hassan Mohabatkar, and Sasan Mohsenzadeh. Using the concept of chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *Journal of theoretical biology*, 263(2):203–209, 2010.

[36] Sarel J Fleishman, Timothy A Whitehead, Eva-Maria Strauch, Jacob E Corn, Sanbo Qin, Huan-Xiang Zhou, Julie C Mitchell, Omar NA Demerdash, Mayuko Takeda-Shitaka, Genki Terashi, et al. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *Journal of molecular biology*, 414(2):289–302, 2011.

[37] Pradip Ghanty and Nikhil R Pal. Prediction of protein folds: extraction of new features, dimensionality reduction, and fusion of heterogeneous classifiers. *IEEE transactions on nanobioscience*, 8(1):100–110, 2009.

[38] Yanzhi Guo, Menglong Li, Minchun Lu, Zhining Wen, and Zhongtian Huang. Predicting g-protein coupled receptors–g-protein coupling specificity based on autocross-covariance transform. *Proteins: structure, function, and bioinformatics*, 65(1):55–60, 2006.

[39] Yanzhi Guo, Lezheng Yu, Zhining Wen, and Menglong Li. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids research*, 36(9):3025–3030, 2008.

[40] Paul Horton, Keun-Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, CJ Adams-Collier, and Kenta Nakai. Wolf psort: protein localization predictor. *Nucleic acids research*, 35(suppl_2):W585–W587, 2007.

[41] Chao Huang and Jingqi Yuan. Using radial basis function on the general form of chou's pseudo amino acid composition and pssm to predict subcellular locations of proteins with both single and multiple sites. *Biosystems*, 113(1):50–57, 2013.

[42] Ji-Tao Huang and Jing Tian. Amino acid sequence predicts folding rate for middle-size two-state proteins. *Proteins: Structure, Function, and Bioinformatics*, 63(3):551–554, 2006.

[43] S Impedovo, L Ottaviano, and S Occhinegro. Optical character recognitiona survey. *International Journal of Pattern Recognition and Artificial Intelligence*, 5(01n02):1–24, 1991.

[44] Jianhua Jia, Zi Liu, Xuan Xiao, Bingxiang Liu, and Kuo-Chen Chou. ippi-esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into pseaac. *Journal of theoretical biology*, 377:47–56, 2015.

[45] Kaveh Kavousi, Behzad Moshiri, Mehdi Sadeghi, Babak N Araabi, and Ali Akbar Moosavi-Movahedi. A protein fold classifier formed by fusing different modes of pseudo amino acid composition via pssm. *Computational biology and chemistry*, 35(1):1–9, 2011.

[46] Vojislav Kecman and Tao Yang. Protein fold recognition with adaptive local hyperplane algorithm. In *Computational Intelligence in Bioinformatics and Computational Biology, 2009. CIBCB'09. IEEE Symposium on*, pages 75–78. IEEE, 2009.

[47] Muslim Khan, Maqsood Hayat, Sher Afzal Khan, and Nadeem Iqbal. Unb-dpc: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into chou's general pseaac. *Journal of theoretical biology*, 415:13–19, 2017.

[48] Raed Khashan, Weifan Zheng, and Alexander Tropsha. Scoring protein interaction decoys using exposed residues (spider): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. *Proteins: Structure, Function, and Bioinformatics*, 80(9):2207–2217, 2012.

[49] Maede Khosravian, Fateme Kazemi Faramarzi, Majid Mohammad Beigi, Mandana Behbahani, and Hassan Mohabatkar. Predicting antibacterial peptides by the concept of chou's pseudo-amino acid composition and machine learning methods. *Protein and Peptide Letters*, 20(2):180–186, 2013.

[50] Ron Kohavi and Foster Provost. Glossary of terms. *Machine Learning*, 30(2-3):271–274, 1998.

[51] KiYoung Lee, Han-Yu Chuang, Andreas Beyer, Min-Kyung Sung, Won-Ki Huh, Bonghee Lee, and Trey Ideker. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic acids research*, 36(20):e136–e136, 2008.

[52] Esther Levin and Roberto Pieraccini. Dynamic planar warping for optical character recognition. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 3, pages 149–152. IEEE, 1992.

[53] Hao Lin, En-Ze Deng, Hui Ding, Wei Chen, and Kuo-Chen Chou. ipro54-pseknc: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic acids research*, 42(21):12961–12972, 2014.

[54] Wei-Zhong Lin, Jian-An Fang, Xuan Xiao, and Kuo-Chen Chou. iloc-animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Molecular BioSystems*, 9(4):634–644, 2013.

[55] Bin Liu, Longyun Fang, Fule Liu, Xiaolong Wang, Junjie Chen, and Kuo-Chen Chou. Identification of real microrna precursors with a pseudo structure status composition approach. *PloS one*, 10(3):e0121501, 2015.

[56] Bin Liu, Hao Wu, and Kuo-Chen Chou. Pse-in-one 2.0: An improved package of web servers for generating various modes of pseudo components of dna, rna, and protein sequences. *Natural Science*, 9(04):67, 2017.

[57] Taigang Liu, Xiaoqi Zheng, Chunhua Wang, and Jun Wang. Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. *Protein and peptide letters*, 17(10):1263–1269, 2010.

[58] Harvey Lodish. *Molecular cell biology*. Macmillan, 2008.

[59] Yosvany López, Abdollah Dehzangi, Sunil Pranit Lal, Ghazaleh Taherzadeh, Jacob Michaelson, Abdul Sattar, Tatsuhiko Tsunoda, and Alok Sharma. Sucstruct: Prediction of succinylated lysine residues by using structural properties of amino acids. *Analytical Biochemistry*, 527:24–32, 2017.

[60] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[61] Andres Munoz. Machine learning and optimization. *URL: https://www. cims. nyu. edu/˜ munoz/files/ml_optimization. pdf [accessed 2016-03-02][WebCite Cache ID 6fiLfZvnG]*, 2014.

[62] Rafael Najmanovich, Josef Kuttner, Vladimir Sobolev, and Marvin Edelman. Side-chain flexibility in proteins upon ligand binding. *Proteins: Structure, Function, and Bioinformatics*, 39(3):261–268, 2000.

[63] Eakasit Pacharawongsakda and Thanaruk Theeramunkong. Predict subcellular locations of singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of chou's pseaac. *IEEE transactions on nanobioscience*, 12(4):311–320, 2013.

[64] Kuldip K Paliwal, Alok Sharma, James Lyons, and Abdollah Dehzangi. A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. *IEEE transactions on nanobioscience*, 13(1):44–50, 2014.

[65] Andrea Pierleoni, Pier Luigi Martelli, Piero Fariselli, and Rita Casadio. Bacello: a balanced subcellular localization predictor. *Bioinformatics*, 22(14):e408–e416, 2006.

[66] Maryam Rahimi, Mohammad Reza Bakhtiarizadeh, and Abdollah Mohammadi-Sangcheshmeh. Oogenesis_pred: A sequence-based method for predicting oogenesis proteins by six different modes of chou's pseudo amino acid composition. *Journal of theoretical biology*, 414:128–136, 2017.

[67] Alok Sharma, James Lyons, Abdollah Dehzangi, and Kuldip K Paliwal. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. *Journal of theoretical biology*, 320:41–46, 2013.

[68] Alok Sharma, Kuldip K Paliwal, Abdollah Dehzangi, James Lyons, Seiya Imoto, and Satoru Miyano. A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. *BMC bioinformatics*, 14(1):233, 2013.

[69] Ronesh Sharma, Abdollah Dehzangi, James Lyons, Kuldip Paliwal, Tatsuhiko Tsunoda, and Alok Sharma. Predict gram-positive and gram-negative subcellular localization via incorporating evolutionary information and physicochemical features into chou's general pseaac. *IEEE Transactions on NanoBioscience*, 14(8):915–926, 2015.

[70] Hong-Bin Shen and Kuo-Chen Chou. Gpos-ploc: an ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins. *Protein Engineering Design and Selection*, 20(1):39–46, 2007.

[71] Hong-Bin Shen and Kuo-Chen Chou. Gpos-mploc: a top-down approach to improve the quality of predicting subcellular localization of gram-positive bacterial proteins. *Protein and peptide letters*, 16(12):1478–1484, 2009.

[72] Hong-Bin Shen and Kuo-Chen Chou. A top-down approach to enhance the power of predicting human protein subcellular localization: Hummploc 2.0. *Analytical biochemistry*, 394(2):269–274, 2009.

[73] Hong-Bin Shen and Kuo-Chen Chou. Gneg-mploc: a top-down strategy to enhance the quality of predicting subcellular localization of gram-negative bacterial proteins. *Journal of theoretical biology*, 264(2):326–333, 2010.

[74] Hong-Bin Shen and Kuo-Chen Chou. Virus-mploc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *Journal of Biomolecular Structure and Dynamics*, 28(2):175–186, 2010.

[75] YH Taguchi and M Michael Gromiha. Application of amino acid occurrence for discriminating different folding types of globular proteins. *BMC bioinformatics*, 8(1):404, 2007.

[76] Takeyuki Tamura and Tatsuya Akutsu. Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. *BMC bioinformatics*, 8(1):466, 2007.

[77] Øivind Due Trier, Anil K Jain, and Torfinn Taxt. Feature extraction methods for character recognition-a survey. *Pattern recognition*, 29(4):641–662, 1996.

[78] Ioannis K Valavanis, George M Spyrou, and Konstantina S Nikita. A comparative study of multi-classification methods for protein fold recognition. *International Journal of Computational Intelligence in Bioinformatics and Systems Biology*, 1(3):332–346, 2010.

[79] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[80] Shibiao Wan and Man-Wai Mak. *Machine learning for protein subcellular localization prediction*. Walter de Gruyter GmbH & Co KG, 2015.

[81] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. mgoasvm: Multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC bioinformatics*, 13(1):290, 2012.

[82] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. Goasvm: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of chou's pseudo-amino acid composition. *Journal of Theoretical Biology*, 323:40–48, 2013.

[83] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. mlasso-hum: A lasso-based interpretable human-protein subcellular localization predictor. *Journal of theoretical biology*, 382:223–234, 2015.

[84] Xiao Wang, Jun Zhang, and Guo-Zheng Li. Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. *BMC bioinformatics*, 16(12):S1, 2015.

[85] S Wold, J Jonsson, M Sjörström, M Sandberg, and S Rännar. Dna and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Analytica Chimica Acta*, 277(2):239–253, 1993.

[86] Jiang Wu, Meng-Long Li, Le-Zheng Yu, and Chao Wang. An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition. *The protein journal*, 29(1):62–67, 2010.

[87] Zhi-Cheng Wu, Xuan Xiao, and Kuo-Chen Chou. iloc-plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Molecular BioSystems*, 7(12):3287–3297, 2011.

[88] Zhi-Cheng Wu, Xuan Xiao, and Kuo-Chen Chou. iloc-gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. *Protein and peptide letters*, 19(1):4–14, 2012.

[89] Xuan Xiao, Zhi-Cheng Wu, and Kuo-Chen Chou. iloc-virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *Journal of Theoretical Biology*, 284(1):42–51, 2011.

[90] Xuan Xiao, Zhi-Cheng Wu, and Kuo-Chen Chou. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. *PloS one*, 6(6):e20592, 2011.

[91] Yan Xu, Xin Wen, Li-Shu Wen, Ling-Yun Wu, Nai-Yang Deng, and Kuo-Chen Chou. initro-tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PloS one*, 9(8):e105018, 2014.

[92] Li Yang, Yizhou Li, Rongquan Xiao, Yuhong Zeng, Jiamin Xiao, Fuyuan Tan, and Menglong Li. Using auto covariance method for functional discrimination of membrane proteins based on evolution information. *Amino Acids*, 38(5):1497–1503, 2010.

[93] Yu-hong Zeng, Yan-zhi Guo, Rong-quan Xiao, Li Yang, Le-zheng Yu, and Meng-long Li. Using the augmented chou's pseudo amino acid

composition for predicting protein submitochondria locations based on auto covariance approach. *Journal of theoretical biology*, 259(2):366–372, 2009.

[94] Hua Zhang, Tuo Zhang, Jianzhao Gao, Jishou Ruan, Shiyi Shen, and Lukasz Kurgan. Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility. *Amino acids*, 42(1):271–283, 2012.

[95] Tong-Liang Zhang, Yong-Sheng Ding, and Kuo-Chen Chou. Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. *Journal of theoretical biology*, 250(1):186–193, 2008.

[96] Tongliang Zhang, Yongsheng Ding, and Kuo-Chen Chou. Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. *Computational Biology and Chemistry*, 30(5):367–371, 2006.