Karachi Institute of Economics and Technology

MS Thesis

Machine Learning Approaches for the Prediction of Protein Sequences

Author:

Shujaat Khan

Supervisor:

Dr. Imran Naseem

A thesis submitted in fulfilment of the requirements

for the degree of Masters of Science

in the

Signal Processing Research Group

Graduate School of Science and Engineering

July 2015

Declaration of Authorship

I, Shujaat Khan, declare that this thesis titled, 'Machine Learning Approaches for the Prediction of Protein Sequences' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: _____

Date: _____

Karachi Institute of Economics and Technology

Abstract

Signal Processing Research Group

Graduate School of Science and Engineering

Masters of Science

Machine Learning Approaches for the Prediction of Protein Sequences

by Shujaat Khan

Structures and sequences of various proteins exhibit a high degree of heterogeneity, consequently prediction of the protein type is considered to be a challenging task. In this research, we have addressed two major problems on prediction in proteomics namely extracellular matrix (ECM) protein prediction and antifreeze protein (AFP) prediction. For the problem of ECM prediction we have proposed a novel sparse learning approach. Essentially the most discriminant features are selected by maximizing the class relevance and minimizing the redundancy (mRMR). The sparsity of these features is harnessed to employ the sparse representation classification (SRC) for prediction of the ECM proteins. The proposed ECMSRC approach has shown better results compared to the previous approaches. For the problem of AFP prediction we propose the notion of localized processing. In particular protein sequence is segmented into two sub-segments each of which is analyzed for sequence derived features amino acid, and di-peptide compositions. We propose to use only the most significant features using the concept of information gain (IG) followed by a random forest (RF) classification approach. The proposed RAFP-Pred has shown to achieved better classification results on a number of standard datasets. A new dataset consisting of 3572 sequences annotated as "antifreeze", obtained from the National Center for Biotechnology Information (NCBI) repository, is also reported in this work. The proposed RAFP-Pred achieved a high verificatin accuracy of 88.63% on this new dataset. The new dataset is made publicly available for the benchamrking.

Acknowledgements

I would like to sincerely thank my supervisor Dr. Imran Naseem, not only for his guidance and support in carrying out this thesis, but also for helping me understand the importance of true human values. Surely, he is the most courteous and gracious person, I have ever met. There are few good souls remaining on this planet (and unfortunately they too are decreasing with each passing day), I was fortunate enough to have one of them as my supervisor. The freedom and encouragement he gave me to explore new and different ideas during my tenure as a Research Assistant, helped me a lot in expanding my learning scope.

I am also extremely grateful to Dr. Mohammad Moinuddin, for re-instilling the self-belief and confidence which I had lost after my B.E. He is definitely one of the finest teachers I have known. Not studying more courses from him will remain one of my biggest regrets. I would also like to thank Dr. Ghazanfar Monir, Dr. Hussain Parvez, Dr. Kashif Ishaq, Dr. Muhammad Bilal Kadri, and Dr. Nadeem Ahsan for the time they served in teaching me and for making me able to do this work. Working with them has completely re-defined my concepts of terms like work-ethics and time-management. They made me realize that big things require big efforts.

This thesis would not have been possible without the support of my dear friends Muhammad Usman, Muhammad Sohail Ibrahim, Idrees Kayani and Salman Ali. I would also like to acknowledge the support of Iqra University, for providing me all the required resources needed in the completion of several cumbersome simulations.

Last but not the least, I am extremely grateful to my family for their patience, love, and support throughout my academic career.

Contents

List of Figures

List of Tables

Abbreviations

ECM	**E**xtra **C**ellular **M**atrix
SRC	**S**parse **R**epresentation **C**lassifier
AFP	**A**nti-**F**reeze **P**roteins
RF	**R**andom **F**orest
mRMR	**M**inimum **R**edundancy and **M**aximum **R**elevance
Info-Gain	**I**nformation **G**ain
MCC	**M**atthews **C**orrelation **C**oefficient
YI	**Y**ouden's **I**ndex

Chapter 1

Introduction

Various type of proteins are there such as extracellular matrix protein (ECM), antifreeze protein (AFP), bioluminescent protein (BLP), secretory and non-secretory proteins. The biological functions of these proteins are closely correlated with their cellular attributes such as which compartment of a cell they belong to and how they are associated with the lipid bilayer of an organelle. With the rapid increase in the number of protein sequences entering into data bank have motivated to develop a fast and precise system to envisage the cellular characteristic of a protein based on its amino acid sequence.

1.1 Proteomics

The proteins are the essential to life, they perform amazing functions in every biochemical reaction. Every biochemical reaction essential to life depends on the marvelous functions of proteins in one way or another. For example, proteins can serve as the beams and rafters of the cell; it can be a glue that join the body together; or an enzyme that build up and break down our energy reserves; and much more.

With the success of human genome project many challenging frontier were emerged : proteomics. Proteomics is the science of the cellular protein, it has became the essential element in the study of biology and medicine related to protein sequence analysis. Although proteins are generated according to their DNA code, they are far more complex and diverse than DNA. To comprehend the molecular underpinnings of life, it is essential to explore individual proteins, their complexes and cellular networks. This is what proteomics is all about.

1.1.1 ECM Proteins

Living tissues consist of cells and extracellular spaces filled with a complex network of macromolecules called extracellular matrix (ECM). The extracellular matrix is a dynamic, physiologically active component of all living tissues. In addition to providing the structural support for the cells embedded within a tissue, it is also a key factor in determining the functionality of the tissue. The ECM is made up of proteoglycans, water, minerals, and fibrous proteins. The mutation in ECM genes has shown to cause several diseases including macular degenerative disease [1], osteoarthritis [2] and congenital muscular dystrophy [3]. Several research works demonstrate that altered expression of a given ECM protein is linked to cancer [4].

The subcellular localization, in general, is of imperative significance [5]. Any meaningful study on unknown proteins cannot be achieved by overlooking the subcellular localization. Researchers have therefore been focusing on computational approaches for the identification of protein subcellular localization [6], [7], [8]. A comprehensive overview of several approaches for the prediction of protein subcellular localization is

provided in [9]. Several research works have carried out subcellular localization for eukaryotes, humans, plants, viruses, gram negative bacteria and gram positive bacteria [10],[11], [12], [13], [14], [15], [16], [17], [18]. Since the ECM proteins are closely related to secretory proteins, it is natural to focus on secretory proteins to explore the extensive properties of the ECM [19]. Although a number of researchers have investigated the secretory proteins [19], [20], [21], only few of them focus on the classification of ECM and non-ECM proteins.

1.1.2 AFP Proteins

Ice has an unusual property called recrystallization. When water starts to freeze, it forms many small crystals. Some of the small crystals soon dominate and continue to become large by stealing water molecules from the surrounding small crystals [22]. This phenomenon can prove to be particularly lethal for living organisms in extreme cold weather due to the intracellular formation of ice [23]. Antifreeze proteins (AFPs) neutralize this recrystallization effect by binding to the surface of the small ice crystals and retarding the growth into larger dangerous crystals [24][25] Therefore they are also called as ice structuring proteins (ISPs). The AFPs lower the freezing point of water without altering the melting point, this interesting property of the AFPs is called as thermal hysteresis [26].

The AFPs are critical for survival of living organisms in extremely cold environment. They are found in various insects, fish, bacteria, fungi and overwintering plants such as gymnosperms, ferns, monocotyledonous, angiosperms etc [22],[24],[26],[27], [28],[29],[30],[31],[32]. Several studies on various AFPs have shown that there is little structural and sequential

similarity for an ice-binding domain [24]. This inconsistency relates to the lack of common features in different AFPs and therefore reliable prediction of AFPs is considered to be an ardent task.

1.2 Motivation

The astounding success of the machine learning algorithms in the paradigm of protein classification, has intrigued several researchers to develop automated approaches for the identification of the ECM and AFPs. In machine learning, the difficult manifold learning problems can be effectively addressed using the localized processing approach compared to the holistic counterparts [33]. Considering the diversified structures of ECM and AFPs, it is intriguing to explore the localized processing of the protein sequences. We therefore propose to adopt a segmentation approach where each protein sequence is segmented into two sub-sequences each of which is analyzed for sequence derived features such as amino acid, di-peptide compositions, physicochemical properties and pseudo-amino acid compositions. To the best of our knowledge, this for the first time that the sparsity of sequence derived features is exploited for ECM prediction, we are also certain that it is the first time the localized processing is proposed to deal with the challenging problem of learning diversified structures of the AFPs.

1.3 Contribution

In this research we propose two novel machine learning approaches ECMSRC and RAFP-Pred for the prediction ECM and AFP proteins respectively from the protein sequences.

In ECMSRC essentially the most discriminant features are selected by maximizing the class relevance and minimizing the redundancy (mRMR) in an information theoretic sense. The sparsity of these discriminant features is harnessed to employ the sparse representation classification (SRC) for prediction of the ECM proteins. The proposed algorithm achieves a test-accuracy of 81.06% on a standard dataset which is superior compared to the EcmPred approach. For the case of prediction of the experimentally verified ECM proteins from humans, we report a verification accuracy of 80% which outperforms the EcmPred approach by a margin of 5%. In RAFP-Pred particular an AFP sequence is segmented into two sub-segments each of which is analyzed for amino acid and di-peptide compositions. We propose to use only the most significant features using the concept of information gain (IG) followed by a random forest classification approach. The proposed RAFP-Pred achieved excellent performance on a number of standard datasets. We report a high Youden's index (sensitivity+specificity-1) value of 0.75 on the standard independent test data set outperforming the AFP-PseAAC, AFP_PSSM, AFP-Pred and iAFP by a margin of 0.04, 0.06, 0.08 and 0.70 respectively. On the dataset attained from the Protein Data Bank (PDB), the proposed approach achieved the Youden's index of 0.36 comprehensively outperforming the AFP-PseAAC method by a margin of 0.31. The verification rate on the UniProKB dataset is found to be 83.19% which is substantially better than the 57.18% and 70.94% reported for iAFP and AFP-PseAAC respectively. A new dataset consisting of 3572 sequences annotated as "antifreeze", obtained from the National Center for Biotechnology Information (NCBI) repository, is also reported in this work.

1.4 Thesis Organization

The remainder of this thesis is organized as follows: In chapter 2, we discuss the proposed method for the prediction of extracellular proteins (ECMs) using sparse representation classifier (SRC), we also discuss the extraction of features from protein sequence and the selection of optimal feature subset from the orignal feature set using Minimum Redundancy and Maximum Relevance(mRMR). Chapter 3 is about the prediction of Antifreeze proteins, In this chapter we discuss the proposed method for the prediction of antifreeze protein using random forest classifier, we also discuss the different sequence derived features and the selection of optimal feature subset using the concept of information gain. Experimental results for ECMSRC and RAFP-Pred are provided in chapter 2 and chapter 3 respectively, while, chapter 4, covers the conclusion and the future work.

Chapter 2

Prediction of the Extracellular Matrix Proteins

2.1 Introduction

Living tissues consist of cells and extracellular spaces filled with a complex network of macromolecules called extracellular matrix (ECM). The extracellular matrix is a dynamic, physiologically active component of all living tissues. In addition to providing the structural support for the cells embedded within a tissue, it is also a key factor in determining the functionality of the tissue. The ECM is made up of proteoglycans, water, minerals, and fibrous proteins. The mutation in ECM genes has shown to cause several diseases including macular degenerative disease [1], osteoarthritis [2] and congenital muscular dystrophy [3]. Several research works demonstrate that altered expression of a given ECM protein is linked to cancer [4].

The subcellular localization, in general, is of imperative significance [5]. Any meaningful study on unknown proteins cannot be achieved by overlooking the subcellular localization. Researchers have therefore been focusing on computational approaches for the identification of protein subcellular localization [6], [7], [8]. A comprehensive overview of several approaches for the prediction of protein subcellular localization is provided in [9]. Several research works have carried out subcellular localization for eukaryotes, humans, plants, viruses, gram negative bacteria and gram positive bacteria [10],[11], [12], [13], [14], [15], [16], [17], [18]. Since the ECM proteins are closely related to secretory proteins, it is natural to focus on secretory proteins to explore the extensive properties of the ECM [19]. Although a number of researchers have investigated the secretory proteins [19], [20], [21], only few of them focus on the classification of ECM and non-ECM proteins.

ECMPP [34] can be regarded as the first predictor designed for the prediction of the ECM proteins. In ECMPP, ECM proteins are represented by augmenting five novel sequence-based features with the 91 commonly used features. EcmPred [35] used the random forest (RF) approach where proteins are represented by sequence-derived features such as the frequency of amino acid/amino acid groups and physicochemical properties.

In this research we propose a novel algorithm called ECMSRC for the prediction of ECM proteins. We propose to use the Minimum Redundancy and Maximum Relevance (mRMR) [36] method for the selection of the most significant sequence-derived features. At the prediction stage, we propose to use the state-of-art Sparse Representation Classification (SRC) [37] algorithm using the overcomplete dictionary matrix of

training samples. We will show that the proposed algorithm outperforms the EcmPred approach on a standard dataset.

The chapter is organized as follows: the mathematical framework of the proposed approach for the prediction of ECM is presented in Section 2.2 followed by the description of the data set and an extensive experimentation in Section 2.3. The analysis of the most discriminant features selected by the proposed approach is presented in Section 2.4 and 2.5 and the chapter is summarized in Section 2.6.

2.2 Proposed Approach

We propose a modular approach for the problem of ECM protein prediction. The first module selects the most significant features using the concept of Minimum Redundancy and Maximum Relevance (mRMR) feature selection [36]. The second module performs classification by developing an overcomplete dictionary matrix of the training samples using only the discriminant features selected by the first module [38].

2.2.1 Module 1: Feature Selection using Minimum Redundancy and Maximum Relevance(mRMR) Criteria

From the machine learning perspective, the task of ECM protein classification can be viewed as a problem of identifying class boundaries. For a given classification problem, irrelevant features tend to degrade the classifier's performance by inducing the Small Sample Size (SSS) effect. The resulting overfitting phenomenon makes the classifier work perfectly for the known samples and poorly for the testing (unknown) data. With

intelligent feature selection the cost of clinical prediction can also be largely reduced, it is much cheaper and efficient to focus on a few significant features compared to the entire population. It is therefore mandatory to retain only the most significant features and eliminate the irrelevant information before the classification module. We argue that the selected features, chosen from the entire pool of features, should be discriminant in an information theoretic sense. In particular, they should depict maximum relevant and minimum redundant information. We therefore propose to use the Minimum Redundancy and Maximum Relevance(mRMR) [36] approach for this purpose.

The mRMR approach uses the concept of mutual information and simultaneously optimizes the minimal redundancy and maximal relevance criterion to select the M most significant features. Given two random variables X and Y, the mutual information between X and Y can be defined as:

$$
\begin{aligned}
I(X;Y) &= \sum_{x}\sum_{y} \log \frac{f(x,y)}{f(x)f(y)} && (2.1)\\
&= E_{XY}\left[\log \frac{f(x,y)}{f(x)f(y)}\right] && (2.2)\\
&= D\left(f(x,y)\|f(x)f(y)\right)) && (2.3)
\end{aligned}
$$

where $f(.)$ and $f(.,.)$ denote marginal and joint distributions respectively, $D(.\|.)$ being the Kullback-Leibler distance between the two probability mass functions. Essentially mutual information $I(X;Y)$ is a measure of the information Y contains about X and therefore is an index of dependence. Consequently, mutual information is zero if and only if X and Y are independent.

Mutual information gives a sense of "similarity" between the features. The concept of minimum redundancy feature selection is primarily to select the features that are mutually different. The subset of the features obtained will therefore characterize complementary information by minimizing the redundancy between the features. Let there be a total of N number of features $\Psi_1, \Psi_2, \ldots, \Psi_N$ and our problem is to find the M most significant features ($M < N$). Let Ω be the set of selected M features with $|\Omega|$ the cardinality of Ω, then the minimum redundancy condition can be stated as:

$$\min U_I; \quad U_I = \frac{1}{|\Omega|^2} \sum_{i,j \in \Omega} I(\Psi_i, \Psi_j), \tag{2.4}$$

The relevance, of a given feature Ψ_i, for classification, is given by the mutual information between Ψ_i and c, $I(\Psi_i, c)$ [36]. Where c is the class-vector consisting of class-labels for all training samples. Consequently, the greater the value of $I(\Psi_i, c)$ for a particular feature, the more relevant the feature is for classification. The maximum relevance criterion therefore turns out to be:

$$\max V_I; \quad V_I = \frac{1}{|\Omega|} \sum_{i \in \Omega} I(\Psi_i, c), \tag{2.5}$$

The mRMR approach of feature selection is to simultaneously optimize the criteria in equations (2.4) and (2.5) by combining them to yield a single optimization problem [36]. The two most prevalent methods to achieve this are to either use the difference or

the ratio of the maximum relevance and minimum redundancy criteria:

$$\max(V_I - U_I) \tag{2.6}$$

$$\max(V_I / U_I) \tag{2.7}$$

Equation (2.6) is called the Mutual Information Difference (MID) criterion while equation (2.7) is called the Mutual Information Quotient (MIQ) criterion. Further details of these algorithms are discussed in [36].

2.2.2 Module 2: Overcomplete Dictionary Matrix for the Classification of the ECM Proteins

After the selection of the M most significant features, the proposed algorithm uses the concept of sparse representation for the ECM proteins classification. Formally, we label ECM proteins and Non-ECM proteins using class index k such that $k = 1$ and $k = 2$ correspond to ECM and Non-ECM proteins respectively. Let there be N training samples of both the classes such that $\mathbf{x}_p^{(k)} \in \mathbb{R}^{q \times 1}$ represents the p^{th} training sample from k^{th} class, $k = 1, 2$ with q number of features. We form a dictionary matrix $\mathbf{D} \in \mathbb{R}^{q \times 2N}$ by concatenating all training samples:

$$\mathbf{D} = [\mathbf{x}_1^{(1)} \mathbf{x}_2^{(1)} \dots \mathbf{x}_N^{(1)}, \mathbf{x}_1^{(2)} \mathbf{x}_2^{(2)} \dots \mathbf{x}_N^{(2)}] \tag{2.8}$$

An unknown test sample $\mathbf{y} \in \mathbb{R}^{q \times 1}$ is now represented as a linear combination of all training samples:

$$\mathbf{y} = \mathbf{D}\alpha; \tag{2.9}$$

where the unknown vector of coefficients $\alpha \in \mathbb{R}^{2N \times 1}$ is:

$$\alpha = [\alpha_1^{(1)} \alpha_2^{(1)} \dots \alpha_N^{(1)}, \alpha_1^{(2)} \alpha_2^{(2)} \dots \alpha_N^{(2)}] \tag{2.10}$$

If a given test sample $\mathbf{y}$ belongs to the k^{th} class, ideally all entries of the vector **alpha** are zero except $\alpha_1^{(k)} \alpha_2^{(k)} \dots \alpha_N^{(k)}$. It has been shown that given the matrix $\mathbf{D}$, the sparse vector α can be recovered [38], [39], [40]. In principle the sparsest α can be sought through the solution of the optimization problem:

$$\arg \underbrace{\min}_{\alpha} \|\alpha\|_0 , \quad \text{subject to} \quad \mathbf{y} = \mathbf{D}\alpha \tag{2.11}$$

where $\|\alpha\|_0$ is the l_0-norm of α and the problem in Equation (2.11) above is generally non-convex and NP-hard. Several alternate methods have been proposed in the literature to recover the sparse vector α. The Basis Pursuit (BP) algorithm, for instance, makes use of the l_1-norm to solve the convex optimization problem [40]:

$$\arg \underbrace{\min}_{\alpha} \|\alpha\|_1 , \quad \text{subject to} \quad \mathbf{y} = \mathbf{D}\alpha \tag{2.12}$$

Under certain conditions on the isometry constant of the matrix $\mathbf{D}$, the sparse vector α can be safely recovered using the BP algorithm [41], [42]. Ideally speaking, α will have high-value entries corresponding to the columns of $\mathbf{D}$ that are relevant to the class label of the probe $\mathbf{y}$. This embedded information about the class label of $\mathbf{y}$ can be used to identify y:

$$r_k(\mathbf{y}) = \|\mathbf{y} - \mathbf{D}\delta_k(\alpha)\|_2 ; \; k = 1, 2 \tag{2.13}$$

where the vector δ_k has all zero entries except at the locations corresponding to class k where the value is one. The decision is ruled in favor of the class with the minimum reconstruction error:

$$\text{class label}(\mathbf{y}) = \arg \underbrace{\min}_{k} r_k(\mathbf{y}) \tag{2.14}$$

2.3 Experimental Results

2.3.1 Evaluation Parameters

For any prediction framework, the Receiver Operating Characteristic (ROC) is considered to be the most comprehensive performance criterion. The proposed algorithm was

therefore extensively evaluated for true positive rate (sensitivity), true negative rate (specificity), prediction accuracy and the area under the curve (AUC). The proposed algorithm was also evaluated for Matthew's Correlation Coefficient (MCC). MCC, ranging from -1 to 1, which takes care of the unbalanced data and is considered to be an important statistical parameter. Values of MCC = 1 and MCC = -1 indicate the best possible and worst possible prediction respectively, MCC = 0 shows the case of a random prediction. Youden's index (or Youden's J statistics) is an interesting way of summarizing results of a diagnostic experiment [43]. Ranging from 0 to 1, 0 indicates worst performance while 1 shows perfect results with no false positives and false negatives. Youden's index is typically useful for the evaluation of highly imbalanced test data. A cross-validation evaluation protocol was adopted for all the experiments [35].

2.3.2 Dataset

Extensive experiments were conducted on the state-of-the-art dataset reported in [35]. The dataset containing 17,233 Metazoan secreted protein sequences, obtained from Swiss-Port release 67 [44], was used. Out of these 17,233 sequences, 1103 sequences are ECM proteins (class label $k = 1$) and the remaining 16,130 proteins are secreted proteins without ECM annotation (class label $k = 2$). Non-redundancy was ensured by allowing a sequence identity between any two proteins of not more than 70% [45]. The final dataset consisted of the 445 ECM proteins and 4187 Non-ECM proteins. Out of 445 ECM proteins 300 were randomly selected to form the training data of ECM proteins while the remaining 145 proteins were designated as the testing data. Similarly

TABLE 2.1: List of features [35]

Name of feature	Number of features
Frequency of 10 functional groups in partition 1 (first 30 residues from N-terminal)	10
Frequency of 24 physicochemical properties in partition 1 (first 30 residues from N-terminal)	24
Frequency of 10 functional groups in partition 2	10
Frequency of 24 physicochemical properties in partition 2	24
Total	68

for the case of the non-ECM proteins, 300 were randomly selected for training while the remaining 3887 proteins were used for testing [35].

Amino acids have traditionally been used as one of the fundamental features for the proteome in the context of sequence-based prediction [5]. Recently, the emerging concept of "pseudo amino acid composition" including the sequence-order information has shown better results for sequence-based prediction [46], [47], [48], [49]. PseAAC, a server based on the same concept provides a flexible way to generate various kinds of pseudo amino acid compositions for a given protein sequence [46], [47]. The general form of PseAAC for a given protein P is given as:

$$P = [\phi_1 \ \phi_2 \ \phi_3 \ \ldots \phi_\Delta]^T \tag{2.15}$$

where $\phi_v; \ v = 1, 2, \ldots, \Delta$ corresponds to each of the sequence derived features and $\Delta = 68$. To compare the proposed approach to the state-of-the-art methods, the same features were used as proposed in [35]. For the sake of completeness, the description of these derived features, as in [35], is given below.

Signal peptides are known to play a significant role in protein secretion [50]. Generally

signal peptides are found within the first 30 residues from the N-terminal. To incorporate the peptide information, each sequence is divided into two partitions. For a sequence with residue length L, the first 30 residues from the N terminal constitute partition 1 and the remaining residues (residues 31-L) form partition 2. 20 amino acids were categorized into 10 functional groups based on the presence of side chain chemical groups such as phenyl(F/W/Y), carboxyl(D/E), imidazole (H), primary amine(K), guanidino(R), thiol(C), sulfur(M), amido(Q/N), hydroxyl(S/T), and non-polar(A/G/I/L/V/P) [35]. The frequencies of these 10 functional groups were calculated for partition 1 and 2.

24 physicochemical properties from the UMBCAA Index database were chosen [35]. These physicochemical properties include molecular weight, hydrophobicity, hydrophilicity, refractivity, average accessible surface area, flexibility, melting point, side chain volume, side chain hydrophobicity, normalized frequency of beta-sheet and alphahelix, refractivity, membrane buriability, retention coefficient, sterichindrance, optical activity, polarity, heat capacity, and isoelectric point. For each sequence, 24 physicochemical property values were calculated by taking the sum of each physicochemical property value over all residues of the sequence and dividing it by the length of the sequence. The list of 68 features is provided in Table 3.1 [35].

2.3.3 Results

The proposed algorithm was trained using 300 ECM proteins and 300 non-ECM proteins [35]. The performance achieved by the proposed approach is compared to the state-of-art EcmPred [35], refer to Table 2.2. The proposed approach achieved a high training accuracy of 98.33% making use of only 29 significant features (Table 2.3). The EcmPred

method achieved a training accuracy of 83% (making use of 40 significant features) which lags the proposed approach by a margin of 15.33%.

For a comprehensive evaluation of the proposed approach, extensive experiments were conducted on a test dataset consisting of 145 ECM proteins and 3887 non-ECM proteins [35]. All results are compared to the EcmPred method in Table 2.2. The EcmPred method achieved 77% test accuracy utilizing a subset of 40 features. The proposed approach achieved a high test accuracy of 81.06% using only 29 features (refer to Table 2.3). This represents an improvement of 4.06% compared to the EcmPred method. The proposed approach requires a subset of only 29 features to attain this high accuracy. The high test accuracy of the proposed approach is well augmented by a sensitivity of 74.48%, a specificity of 81.30%, a Youden's index of 0.5579 and an MCC of 0.256331. The proposed approach comprehensively outperformed the EcmPred method with a performance gain of 9.48% in sensitivity, 4.31% in specificity, 0.1379 in Youden's index and 0.0650 in MCC. The Receiver Operating Characteristic curve of the proposed approach using the top 29 features is shown in Figure 2.1. The proposed approach achieved a high area under the curve (AUC) of 0.8671 which is significantly better compared to 0.7900 reported for EcmPred [35]. To demonstrate the efficacy of the proposed algorithm, an extensive comparative analysis was performed with the benchmark machine learning approaches. Results are shown in Table 2.4. For a fair comparison only the best reported results are presented. The proposed ECMSRC approach has shown to comprehensively outperform the benchmark machine learning approaches in all aspects of prediction performance.

The ability of true prediction of the proposed approach, was thoroughly investigated

TABLE 2.2: Performance of EcmPred [35] algorithm using different feature subsets.

Feature subset	**Sensitivity** (%)	**Specificity** (%)	**MCC**	**Test accuracy** (%)	**Youden's index**	**Training accuracy** (%)
10	51.00	75.00	0.1123	74.00	0.2600	73.00
20	48.00	77.00	0.1171	76.00	0.2500	80.00
30	53.00	78.00	0.1378	77.00	0.3100	81.00
40	**65.00**	**77.00**	**0.1906**	**77.00**	**0.4200**	**83.00**
50	57.00	77.00	0.1493	76.00	0.3400	82.00
60	60.00	77.00	0.1661	76.00	0.3700	83.00
All features (68)	63.00	76.00	0.1702	75.00	0.3900	82.00

TABLE 2.3: Performance of the proposed ECMSRC algorithm using different feature subsets.

Feature subset	**Sensitivity** (%)	**Specificity** (%)	**MCC**	**Test accuracy** (%)	**Youden's index**	**Training accuracy** (%)
20	58.62	77.04	0.154889	76.38	0.3566	88.83
25	67.59	82.02	0.231949	81.51	0.4961	90.83
29	**74.48**	**81.30**	**0.256331**	**81.06**	**0.5579**	**98.33**
35	80.69	76.83	0.246516	76.97	0.5752	99.83
40	80.69	74.13	0.227691	74.37	0.5482	100.00
50	80.69	70.38	0.204725	70.75	0.5107	100.00
All Features(68)	82.07	70.17	0.208988	70.60	0.5224	100.00

TABLE 2.4: Comparison of the proposed ECMSRC algorithm with the benchmark machine learning approaches.

Method	**Feature subset**	**Sensitivity** (%)	**Specificity** (%)	**MCC**	**Youden's index**	**Test accuracy** (%)
J4.8	40	57.00	66.00	0.097	0.2300	66.00
Bayesnet	40	57.00	76.00	0.149	0.3300	75.00
Adaboost	40	59.00	69.00	0.111	0.2800	59.00
Decision table	40	54.00	68.00	0.089	0.2200	55.00
Logistic	40	59.00	65.00	0.097	0.2400	65.00
SVM (polynomial)	40	56.00	68.00	0.100	0.2400	68.00
MLP	40	58.00	68.00	0.104	0.2600	59.00
EcmPred	40	65.00	77.00	0.191	0.4200	77.00
ECMSRC	**29**	**74.48**	**81.31**	**0.256**	**0.5579**	**81.06**

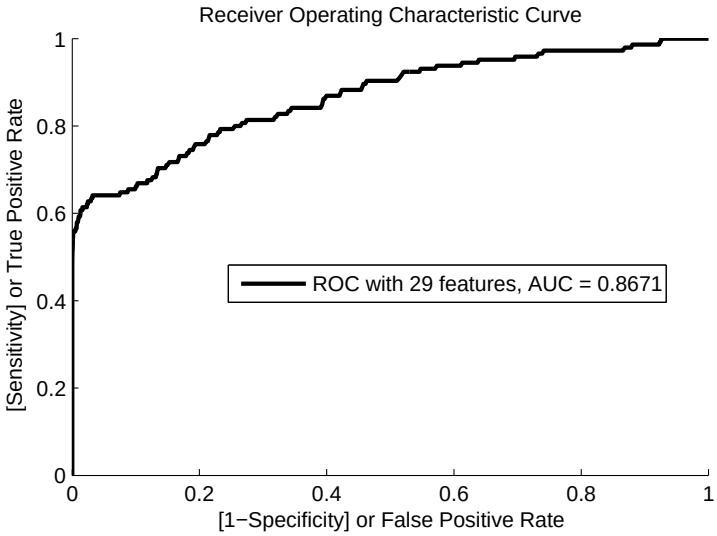

FIGURE 2.1: ROC of the proposed approach with the top 29 features

using 20 experimentally verified extracellular matrix proteins from human [35]. The results are compared to the ECMPP [34] and EcmPred [35] approaches. As shown in Table 2.5, ECMPP and EcmPred achieved a prediction accuracy of 30% and 75% respectively. The proposed ECMSRC method attained a high prediction accuracy of 80% by truly predicting 16 out of the 20 ECM proteins.

2.4 Discriminant Feature Subset

The proposed algorithm selected the 29 most discriminant features outperforming the contemporary methods. The selected feature subset contains 22 physicochemical properties which include normalized frequency of beta-sheet , flexibility indices(from both segments), steric parameter, transfer free energy to surface, normalized frequency of alpha-helix, signal sequence helical potential, heat capacity, buriability(from both segments), membrane-buried preference parameters, refractivity, molecular weight, average

TABLE 2.5: Prediction results for 20 experimentally verified ECM proteins. "+" indicates successful prediction of ECM proteins while "-" represents a failure.

SwissProt ID	Protein Annotation	ECMPRED	ECMPP	ECMSRC
Q9BY76	Angiopoietin-related protein	+	-	+
P07355	Annexin A2	+	-	+
Q9BXN1	Asporin	+	+	+
P01137	Transforming growth factor beta-1	-	-	-
Q8N6G6	ADAMTS-like protein 1	+	-	+
P27797	Calreticulin	+	-	+
Q76M96	Coiled-coil domain-containing protein	+	+	+
Q07654	Trefoil factor 3	-	+	+
O75339	Cartilage intermediate layer protein 1	+	-	+
Q15063	Periostin	-	-	-
O43405	Cochlin	+	-	+
Q96P44	Collagen alpha-1(XXI) chain	+	+	+
P01009	Alpha-1-antitrypsin	-	-	+
Q14118	Dystroglycan	+	-	-
Q12805	EGF-containing fibulin-like extracellular matrix protein 1	+	-	+
Q75N90	Fibrillin-3	+	+	+
P09382	Galectin-1	+	+	+
Q8N2S1	Latent-transforming growth factor beta- binding protein 4	+	-	+
P27487	Dipeptidyl peptidase 4	-	-	-
P08253	72 kDa type IV collagenase	+	-	+

membrane preference(from both segments), retention coefficient in TFA, optical rotation(from both segments), amphiphilicity index, polarity. In addition to these physicochemical properties, the proposed approach selected seven significant functional groups namely, "K" primary amine, "M" sulfur, "DE" carboxyl, "ST" hydroxyl, "AGILVP" non-polar, "C" thiol, "FWY" phenyl.

2.5 Analysis of the Selected Feature Subset

In order to appreciate the significance of the two types of features, the distribution of the features in the selected subset was further investigated. Figure 2.2 shows the proportion of the two types of features in the original feature set (black bar) and the selected subset (white bar). Clearly, physicochemical information plays a vital role in distinguishing

ECMs from non-ECMs, compared to the functional groups. The proposed approach selected 22 physicochemical properties (14 from individual segments and 4 from both segments) and 7 functional group frequencies. Out of the top 10 most discriminant features, 9 were found to be the physicochemical properties. However, the importance of the functional groups cannot be overlooked. Although they are ranked lower compared to the physicochemical properties, they still offer important complementary information. The proposed algorithm efficiently fuses these two pieces of complementary information to produce excellent results utilizing fewer number of features.

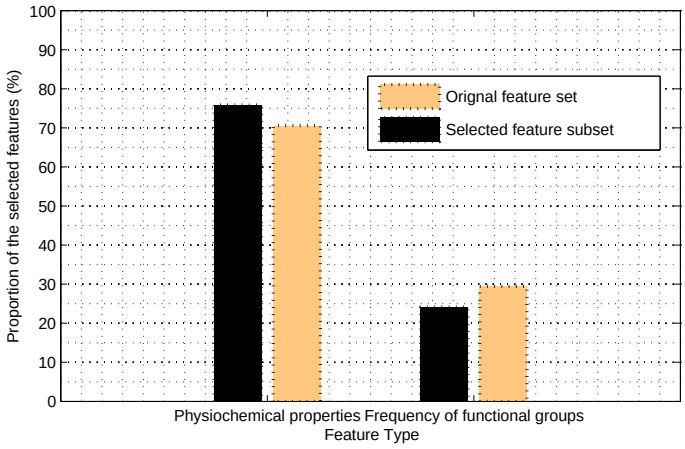

FIGURE 2.2: Comparison of proportion of physicochemical properties and functional group frequencies in the original and selected features.

2.6 Summary

The extracellular matrix (ECM) plays a vital role in the functionality of living tissues. A reliable prediction of the ECM is therefore of prognostic significance. In this research we propose a novel approach for the prediction of the ECM proteins from the protein sequences. Essentially the most discriminant features are selected by maximizing the

class relevance and minimizing the redundancy (mRMR) in an information theoretic sense. The sparsity of these discriminant features is harnessed to employ the sparse representation classification (SRC) for prediction of the ECM proteins. The proposed algorithm achieves a test-accuracy of 81.06% on a standard dataset which is superior compared to the EcmPred approach. For the case of prediction of the experimentally verified ECM proteins from humans, we report a verification accuracy of 80% which outperforms the EcmPred approach by a margin of 5%.

Chapter 3

Prediction of Antifreeze Proteins

3.1 Introduction

Ice has an unusual property called recrystallization. When water starts to freeze, it forms many small crystals. Some of the small crystals soon dominate and continue to become large by stealing water molecules from the surrounding small crystals [22]. This phenomenon can prove to be particularly lethal for living organisms in extreme cold weather due to the intracellular formation of ice [23]. Antifreeze proteins (AFPs) neutralize this recrystallization effect by binding to the surface of the small ice crystals and retarding the growth into larger dangerous crystals [24][25]. Therefore they are also called as ice structuring proteins (ISPs). The AFPs lower the freezing point of water without altering the melting point, this interesting property of the AFPs is called as thermal hysteresis [26].

The AFPs are critical for survival of living organisms in extremely cold environment. They are found in various insects, fish, bacteria, fungi and overwintering plants such as gymnosperms, ferns, monocotyledonous, angiosperms etc [22],[24],[26],[27], [28],[29],[30],[31],[32]. Several studies on various AFPs have shown that there is little structural and sequential similarity for an ice-binding domain [24]. This inconsistency relates to the lack of common features in different AFPs and therefore reliable prediction of AFPs is considered to be an ardent task.

Recent success of the machine learning algorithms in the paradigm of protein classification, has intrigued several researchers to develop automated approaches for the identification of the AFPs. AFP-Pred [51] is considered to be the earliest work in this direction. The work is essentially based on the random forest approach making use of sequence information such as functional groups, physicochemical properties, short peptides and secondary structural element. In AFP_PSSM [52] evolutionary information is used with the support vector machine (SVM) classification. In iAFP [53] n-peptide composition is used with limited experimental results. In particular amino acids, di-peptide and tri-peptide compositions were used. We argue that tri-peptide composition is computationally expensive (calculation of 20^3 combinations) resulting in redundant information. Consequently the selection of the most significant features using genetic algorithm (GA) has shown limited results [53]. The latest work in this regard is AFP-PseAAC [54] where pseudo amino acid composition is used with the SVM classifier to achieve a good prediction accuracy.

In machine learning, the difficult manifold learning problems can be effectively addressed using the localized processing approach compared to the holistic counterparts

[33]. Considering the diversified structures of AFPs, it is intriguing to explore the localized processing of the protein sequences. We therefore propose to adopt a segmentation approach where each protein sequence is segmented into two sub-sequences. The amino acids and di-peptide compositions are derived for each sub-sequence to constitute the relevant features. The most significant features are further selected using the concept of information gain and the random forest approach is used for the classification purpose. To the best of our knowledge, this for the first time that the localized processing is proposed to deal with the challenging problem of learning diversified structures of the AFPs. The proposed method has shown to comprehensively outperform all the existing approaches on standard datasets.

The chapter is organized as follows: the details and mathematical framework of the proposed approach is presented in Section 3.2 followed by the description of the data sets and experimental results in Section 3.3. Our conclusions are provided in Section 3.4.

3.2 Proposed Approach

Reliable prediction of proteins can only be achieved by robustly encoding the protein sequences into mathematical expressions. This ensures that the underlying structures of the protein sequences have been truly learned. In absences of robust learning of the protein sequences, the predictor is unlikely to perform well for the unseen test samples. From the machine learning perspective, the difficult manifold learning problems are effectively tackled using the localized processing approach [33, 55]. While holistic

methods deal with the training samples in a global sense, the localized learning focuses on various segments of the samples. Typically, features extracted from confined segments are efficiently fused. For challenging manifold learning problems, the localized learning has shown to outperform the holistic counterparts in various applications of machine learning [56–58]. We therefore propose to perform local analysis of AFPs for feature extraction.

3.2.1 Features

Structures of various AFPs are uncorrelated and lack similarity, the automated prediction of the AFPs is therefore considered to be a challenging task. Motivated by the robustness of the localized learning approaches, we propose to consider the localized segments of the AFP sequences. In particular, each protein sequence is segmented into two sub-sequences, each sub-sequence is individually analyzed for amino acid and di-peptide compositions.

Consider a protein chain of L amino acid residues:

$$\mathbf{P} = R_1 R_2 R_3 \dots R_L \tag{3.1}$$

where R_i represents the i^{th} residue of protein $\mathbf{P}$ [59]. According to the amino acid composition protein $\mathbf{P}$ can be expressed as an array of occurrence frequency of the twenty native amino acids:

TABLE 3.1: List of features

Features	**Number of attributes**
Segment 1	
Amino Acid Composition features	20
Dipeptide Composition features	400
Segment 2	
Amino Acid Composition features	20
Dipeptide Composition features	400
Total	840

$$\mathbf{P} = [f_1 f_2 f_3 \dots f_{20}]^T \tag{3.2}$$

where $f_j;\ j = 1, 2, 3, \dots, 20$ is the normalized occurrence frequency of the j^{th} native amino acid in $\mathbf{P}$, and T is the vector transpose operator. Accordingly, the amino acid composition of a protein can be readily derived once the protein sequencing information is known. This simple, but effective, amino acid composition (AAC) model has been widely used in a number of statistical methods for predicting protein structures [60], [61].

Dipeptide compositions are computed using 400 (20×20) dipeptides, i.e. AA, AC, AD,..., YV, YW, YY. Each component is calculated using the following equation:

$$\text{fraction of the } k^{th} \text{ dipeptide} = \frac{\text{total number of the } k^{th} \text{ dipeptide}}{\text{total number of all possible dipeptides}} \tag{3.3}$$

The 20 AACs and 400 dipeptide compositions are combined to form 420 attributes for

each segment of the AFP sequence. Finally the 420 attributes of individual sub-sequences are fused to form a single representative feature vector consisting of 840 attributes. Table 3.1 shows a list of derived features.

It is well established that the redundant information tends to degrade the classification results [62]. It is therefore customary to select the most relevant features for the classification purpose [63], [64]. Information gain (IG) or Info-Gain is considered to be an important criterion for the selection of the most significant features [65]. Given a training set S and an attribute A, the information gain with respect to the attribute A, can be defined as reduction in entropy of the training set once the attribute A is observed [66], mathematically:

$$IG(S, A) = H(S) - H(S/A) \tag{3.4}$$

where $H(S)$ is entropy of S and $H(S/A)$ is the entropy of S conditioned to the observation of attribute A. For the classical case of a dichotomizer:

$$H(S) = -\sum_{l=1}^{2} p_l \log_2 p_l \tag{3.5}$$

and

$$H(S/A) = \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{3.6}$$

where $Values(A)$ is a set of all possible values of the attribute A, S_v is the partition of the training set characterizing the value v of attribute A, $H(S_v)$ is the entropy of S_v and $|.|$ is the cardinality operator [66].

We propose to use the concept of the Info-Gain for the selection of the most significant features from the pool of 840 features discussed in section 3.2.1. The features are ranked using the above formulation of IG in a descending order such that the attribute with the highest IG is given the top priority.

3.2.2 Classification

Random forest approach has shown excellent results for various prediction problems in proteomics [56, 65, 67–72]. Random forest is an ensemble classification protocol which combines several weak classifiers (decision trees) to constitute a single strong classifier. The decision trees generated by random forest approach are combined using a weighted average scheme [73]. The approach harnesses the power of many decision trees, rational randomization, and ensemble learning to develop accurate classification models [73].

Random forest is a supervised learning approach consisting of two steps: (1) bagging, and (2) random partitioning. In bagging several decision trees are grown by drawing multiple samples (with replacement) from the original training data set. Although indefinite number of such trees can be grown, typically 200-500 trees are considered to be enough [66]. Random forest approach introduces randomness in tree-growing by first randomly selecting a subset of prospective predictors and then produce the split by

selecting the best available splitter. The approach is robust to overfitting and quite efficient on large datasets [73]. Random forest classifier was implemented using the WEKA tool [74]. The work-flow of the proposed RAFP-Pred is shown in Figure 3.1

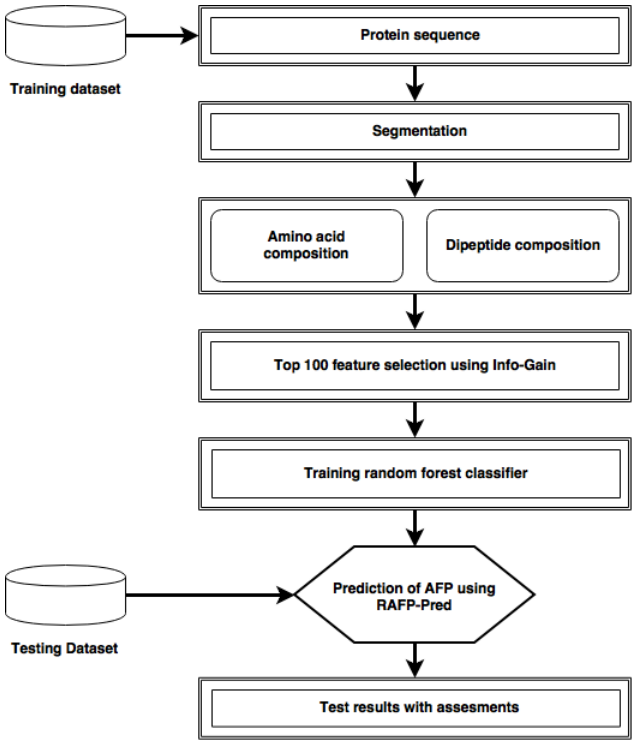

FIGURE 3.1: Work-flow of the proposed RAFP-Pred approach.

3.3 Experimental Results

3.3.1 Evaluation Parameters

For any prediction framework, the Receiver Operating Characteristic (ROC) is considered to be the most comprehensive performance criterion. The proposed algorithm was therefore extensively evaluated for true positive rate (sensitivity), true negative rate

(specificity), prediction accuracy and the area under the curve (AUC). The proposed algorithm was also evaluated for Matthew's Correlation Coefficient (MCC). MCC ranges from -1 to 1 with values of MCC = 1 and MCC = -1 indicating the best and the worst predictions respectively, MCC = 0 shows the case of a random guess. Youden's index (or Youden's J statistics) is an interesting way of summarizing results of a diagnostic experiment [43]. Ranging from 0 to 1, 0 indicates worst performance while 1 shows perfect results with no false positives and false negatives. Youden's index is typically useful for the evaluation of highly imbalanced test data.

3.3.2 Experimental Results

Extensive experiments were conducted on a number of state-of-the-art datasets reported frequently in literature [67], [75].

3.3.2.1 Dataset 1

Dataset 1 consists of 481 AFPs and 9493 non-AFPs reported in [67]. The dataset is further partitioned into training and testing sets. The training set characterizes 300 AFPs and 300 non-AFPs selected randomly from the pool of 481 AFPs and 9493 non-AFPs respectively. The remaining 181 AFPs and 9193 non-AFPs constitute the testing set. The proposed approach attained 100% accuracy on a randomly selected training set which outperforms the AFP-Pred method by a margin of 18.67% [67] and the AFP_PSSM method by a margin of 17.33% [52]. Average accuracy of three randomly selected training set, for the proposed method, was found to be 99.91% with a standard deviation

TABLE 3.2: Performance of the proposed RAFP-Pred on test dataset containing 181 AFPs and 9193 non-AFPs using different feature subsets.

Feature subset	**Sensitivity** (%)	**Specificity** (%)	**MCC**	**Accuracy** (%)	**Youden's index**
25 features	79.01%	89.24%	0.288	89.04%	0.68
50 features	82.32%	90.03%	0.314	89.88%	0.72
75 features	81.77%	89.83%	0.308	89.67%	0.72
100 features	**83.98%**	**91.07%**	**0.339**	**90.93%**	**0.75**
200 features	79.01%	90.10%	0.301	89.88%	0.69
400 features	80.11%	90.93%	0.320	90.72%	0.71
600 features	82.87%	90.20%	0.319	90.06%	0.73
800 features	82.87%	89.67%	0.310	89.54%	0.72
All features	83.43%	89.22%	0.306	89.11%	0.73

TABLE 3.3: Comparison of the proposed RAFP-Pred with different machine learning approaches.

Predictor	**Sensitivity** (%)	**Specificity** (%)	**Accuracy** (%)	**Youden's index**	**AUC**
iAFP [53]	7.18%	97.38%	95.46%	0.05	NA
AFP-Pred [51]	84.67%	82.32%	83.38%	0.67	0.89
AFP_PSSM [52]	75.89%	93.28%	93.01%	0.69	0.93
AFP-PseAAC [54]	86.19%	84.72%	84.75%	0.71	NA
RAFP-Pred	**83.98%**	**91.07%**	**90.93%**	**0.75**	0.95

of 0.16%. This prediction performance is 10.22% better compared to the AFP-PseAAC approach (standard deviation of 0.706%)[54].

Results for the test data set, using different feature subsets, are shown in Table 3.2. The proposed RAFP-Pred achieves best accuracy of 90.93% utilizing 100 most significant features. For comprehensive evaluation, the proposed approach is compared to the state-of-art methods reported in literature (refer to Table 3.3).

The test data set is highly imbalanced with 181 (AFPs) positive and 9193 (non-AFPs) negative examples. For such a highly imbalanced test data, there is a natural tendency of a predictor to be biased in favor of the class with more samples. In such

scenarios, evaluation parameters like AUC and Youden's index are more representative of a predictor's performance than the conventional sensitivity, specificity and accuracy measures. For instance in Table 3.3 iAFP achieves a very high specificity of 97.38% but a poor sensitivity of 7.18%. Therefore, although the overall accuracy of 95.46% appears to be the best reported accuracy, the predictor has a low Youden's index of 0.05 and therefore cannot be regarded as competitive. The proposed approach achieved a Youden's index of 0.75 which is better than all reported results in the literature. The highest AUC value of 0.95 is also reflective of the performance of the proposed RAFP-Pred approach, the receiver operating characteristics (ROC) are shown in Figure 3.2. The model generated using the training samples of dataset 1 is available online at http://sp.gsse.pafkiet.edu.pk/downloads, the same can be used for testing of any dataset.

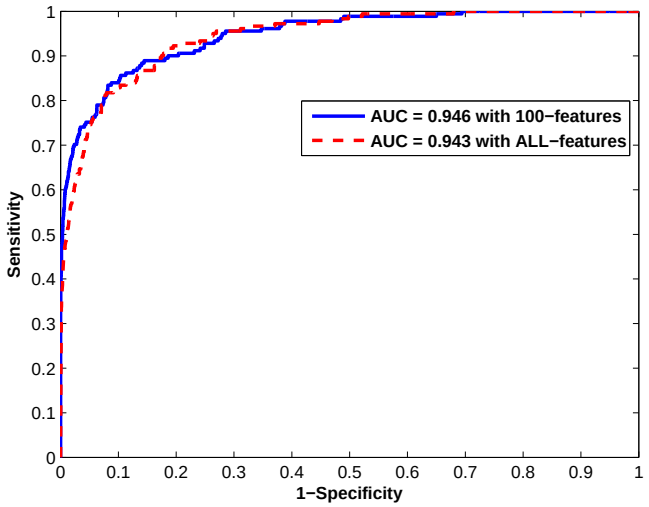

FIGURE 3.2: ROC curves for the proposed RAFP-Pred approach.

TABLE 3.4: Results of the proposed RAFP-Pred on dataset 2.

Feature subset	Sensitivity (%)	Specificity (%)	Accuracy (%)	Youden's index	MCC	AUC
AFP-PseAAC	29.55%	**75.07%**	74.54%	0.05	0.011	0.52
RAFP-Pred	**68.18%**	67.41%	**67.42%**	**0.36**	**0.081**	**0.74**

3.3.2.2 Dataset 2

Dataset 2 consists of 44 AFPs and 3762 non-AFPs collected from the Protein Data Bank (PDB) [76] and the PISCES server [77] respectively and reported in [75]. The model generated using dataset 1 was used for training the RAFP-Pred. The results are shown in Table 3.4.

Dataset 2 is also a highly imbalanced dataset with 44 positive and 3762 negative examples. Therefore, as discussed before, Youden's index and AUC are appropriate performance criteria. The proposed RAFP-Pred attains a high Youden's index value of 0.36 outperforming the AFP-PseAAC method by a comprehensive margin of 0.31. The MCC and AUC of the proposed approach are substantially better compared to the AFP-PseAAC method.

3.3.2.3 Dataset 3

Dataset 3 is an independent dataset representing an evolutionarily divergent group of organisms consisting of 369 AFPs obtained from the UniProKB database [78], [79]. Results on this dataset are reported only for the iAFP method [53] The proposed RAFP-Pred attained the highest verification of 83.19% which is substantially better than 57.18% reported for the iAFP.

3.3.2.4 Dataset 4

Dataset 4 was specifically created to extensively evaluate the sensitivity of the proposed approach on a large number of positive samples. The dataset was obtained from the NCBI data repository [80] consisting of 3572 AFPs annotated as "antifreeze" proteins. The proposed approach was able to identify 3166 out of 3572 AFPs achieving a verification rate of 88.63%. The dataset has been made available online at http://sp.gsse.pafkiet.edu.pk/downloads to facilitate benchmarking of the AFP predictors on this new dataset.

3.4 Summary

In extreme cold weather, living organisms produce Anti Freeze Proteins (AFPs) to counter the otherwise lethal intracellular formation of ice. Structures and sequences of various AFPs exhibit a high degree of heterogeneity, consequently prediction of the AFPs is considered to be a challenging task. In this research, we propose to handle this arduous manifold learning task using the notion of localized processing. In particular an AFP sequence is segmented into two sub-segments each of which is analyzed for amino acid and di-peptide compositions. We propose to use only the most significant features using the concept of information gain (IG) followed by a random forest classification approach. The proposed RAFP-Pred achieved excellent performance on a number of standard datasets. We report a high Youden's index (sensitivity+specificity-1) value of 0.75 on the standard independent test data set outperforming the AFP-PseAAC, AFP_PSSM, AFP-Pred and iAFP by a margin of 0.04, 0.06, 0.08 and 0.70 respectively.

On the dataset attained from the Protein Data Bank (PDB), the proposed approach achieved the Youden's index of 0.36 comprehensively outperforming the AFP-PseAAC method by a margin of 0.31. The verification rate on the UniProKB dataset is found to be 83.19% which is substantially better than the 57.18% and 70.94% reported for iAFP and AFP-PseAAC respectively. A new dataset consisting of 3572 sequences annotated as "antifreeze", obtained from the National Center for Biotechnology Information (NCBI) repository, is also reported in this work. The proposed RAFP-Pred achieved a high verificatin accuracy of 88.63% on this new dataset. The new dataset is made publicly available for the benchamrking.

Chapter 4

Conclusion and Future Work

4.1 Conclusion

ECM proteins not only provide structural support but also influence the functionality of living tissue. Reliable prediction of the ECMs is therefore imperative for diagnostic and therapeutic purposes. The experimental methods for identifying ECMs from protein sequences need a lot of time and resources. It is therefore imperative to develop machine learning algorithms for accurate and expeditious prediction of the ECMs. In this research we propose a novel approach for the classification of the ECM and non-ECM proteins called ECMSRC. The most representative features, in an information theoretic sense, are chosen by employing the mRMR approach [36]. These discriminant features are efficiently used in conjunction with the state-of-art SRC classifier for the identification of the ECM proteins. The proposed algorithm has shown superior performance compared to the state-of-art EcmPred [35] method. In particular, we report a test accuracy of 81.06% with 0.5579 Youden's index by making use of only 29 features. The proposed algorithm

outperforms the EcmPred methods by a margin of 4.06% and 0.1379 in test accuracy and Youden's index respectively. Noteworthy is the fact that the proposed ECMSRC algorithm utilizes fewer features compared to EcmPred (40 features) method to achieve this superior performance. For the case study of the experimentally verified ECM proteins [35], we report a 80% verification rate for the proposed ECMSRC algorithm which is 5% higher than the EcmPred method. The MATLAB implementation of the ECMSRC is made publicly available at `http://sp.gsse.pafkiet.edu.pk/downloads`.

The structural and sequential dissimilarity makes the prediction of the AFPs a difficult task. Previous sequence-based AFP predictors make use of the whole protein sequence. In this work we have proposed the novel concept of localized analysis of AFP sequences. Extensive experiments on a number of standard datasets have been conducted. The proposed RAFP-Pred approach has shown to perform better compared to the previous predictors such as AFP-PseAAC, AFP_PSSM, AFP-Pred and iAFP. We report a new independent dataset, cured from the NCBI repository, consisting of 3572 sequences annotated as "antifreeze". The proposed RAFP-Pred achieved a high verification accuracy of 88.63% on this new dataset. The new dataset and the model of the proposed approach have been made publicly available for the benchmarking purposes. Favorable results are suggestive of further exploration in this direction. For instance more extensive segmentation could be a possible area of the future research.

4.2 Future Work

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors, we shall make efforts in our future work to provide a web-server for the methods presented in this research. Currently, we have user-friendly and freely-available softwares for interested users at http://sp.gsse.pafkiet.edu.pk/downloads.

Bibliography

[35] Kumar Krishna Kandaswamy, Ganesan Pugalenthi, Kai-Uwe Kalies, Enno Hartmann, and Thomas Martinetz. EcmPred: Prediction of extracellular matrix proteins based on random forest with maximum relevance minimum redundancy feature selection. *Journal of Theoretical Biology*, 317(2013):377–383, 2013.

[1] PA. Klenotic et al. Tissue inhibitor of metalloproteinases-3 (TIMP-3) is a binding partner of epidermal growth factor-containing fibulin-like extracellular matrix protein 1 (EFEMO1). Implications for macular degeneration. *J. Biol. Chem.*, 279: 30469–30473., 2004.

[2] H Kizawa et al. An aspartic acid repeat polymorphism in aspirin inhibits chondrogenesis and increases susceptibility to osteoarthritis. *Nat. Genet.*, 37:138–144, 2005.

[3] T.E Hall et al. The zebrafish candyfloss mutant implicates extracellular matrix adhesion failure in laminin alpha-2-deficient congenital muscular dystrophy. *Proc. Natl. Acad. Sci. USA*, 104:7092–7097, 2007.

[4] J Hu et al. Matrix metalloproteinase inhibitors as therapy for inflammatory and vascular disease. *Nat. Rev. Drug Discov.*, 6:480–498, 2008.

[5] P Horton, KJ Park, T Obayashi, N Fujita, H Harada, CJ Adams-Collier, and K Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, 35: W585–W587, 2007.

[6] K.C Chou and H.B 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, e9931*, 2010.

[7] H.B Shen and K.C Chou. A top-down approach to enhance the power of predicting human protein subcellular localization:hum-mploc2.0. *Anal. Biochem.*, 394:269–274, 2009.

[8] K.C Chou. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology*, 273:236–247, 2011.

[9] E.W Klee and C.P Sosa. Computational classification of classically secreted proteins. *Drug Discovery Today*, 12:234–240, 2007.

[10] K.C Chou and H.B Shen. Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J.ProteomeRes.*, 6:1728–1734, 2007.

[11] H.B Shen and K.C Chou. Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem. Biophys. Res. Commun.*, 355:1006–1011, 2007.

[12] K.C Chou, Z.C. Wu, and X Xiao. iLoc-Hum:Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol.Biosyst.*, 8:629–641, 2012.

[13] K.C Chou and H.B Shen. Large-scale plant protein subcellular location prediction. *J.Cell.Biochem.*, 100:665–678, 2007.

[14] Z.C Wu, X Xiao, and K.C Chou. iLoc-Plant: A multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. Biosyst.*, 7:3287–3297, 2011.

[15] H.B Shen and K.C Chou. Virus-PLoc:a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. *Biopolymers*, 85: 233–240, 2007.

[16] X Xiao, Z.C Wu, and K.C Chou. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.*, 284:42–51, 2011.

[17] K.C Chou and H.B Shen. Large-scale predictions of gram-negative bacterial protein subcellular locations. *J.Proteome Res.*, 5:3420–3428, 2006.

[18] Z.C Wu, X Xiao, and K.C Chou. iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. *Protein Pept.Lett.*, 19:4–14, 2012.

[19] K.K Kandaswamy, G Pugalenthi, E Hartmann, K.U Kalies, S Möler, P.N Suganthan, and T Martinez. SPRED: A machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. *Biochem. Biophys. Res. Commun.*, 391:1306–1311, 2010.

[20] J.D Bendtsen, H Nielsen, G von Heijne, and S Brunak. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, 340:783–795, 2004.

[21] P Horton, K.J Park, T Obayashi, and K Nakai. Protein subcellular localisation prediction with WoLF PSORT. *APBC*, pages 39–48, 2006.

[22] Xiao-Ming Yu and Marilyn Griffith. Winter rye antifreeze activity increases in response to cold and drought, but not abscisic acid. *Physiologia Plantarum*, 112(1): 78–86, 2001.

[23] Marilyn Griffith, Mervi Antikainen, Wai-Ching Hon, Kaarina Pihakaski-Maunsbach, Xiao-Ming Yu, Jong Un Chun, and Daniel SC Yang. Antifreeze proteins in winter rye. *Physiologia Plantarum*, 100(2):327–332, 1997.

[24] Peter L Davies, Jason Baardsnes, Michael J Kuiper, and Virginia K Walker. Structure and function of antifreeze proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 357(1423):927–935, 2002.

[25] Garth L Fletcher, Choy L Hew, and Peter L Davies. Antifreeze proteins of teleost fishes. *Annual review of physiology*, 63(1):359–390, 2001.

[26] Maria E Urrutia, John G Duman, and Charles A Knight. Plant thermal hysteresis proteins. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 1121(1):199–206, 1992.

[27] PF Scholander, L Van Dam, JW Kanwisher, HT Hammel, and MS Gordon. Supercooling and osmoregulation in arctic fish. *Journal of Cellular and Comparative Physiology*, 49(1):5–24, 1957.

[28] M Moriyama, J Abe, M Yoshida, Y Tsurumi, and S Nakayama. Seasonal changes in freezing tolerance, moisture content and dry weight of three temperate grasses

[dactylis glomerata, lolium perenne, phleum pratense]. *Journal of Japanese Society of Grassland Science (Japan)*, 1995.

[29] John M Logsdon and W Ford Doolittle. Origin of antifreeze protein genes: a cool tale in molecular evolution. *Proceedings of the National Academy of Sciences*, 94 (8):3485–3487, 1997.

[30] KV Ewart, Q Lin, and CL Hew. Structure, function and evolution of antifreeze proteins. *Cellular and Molecular Life Sciences CMLS*, 55(2):271–283, 1999.

[31] Chi-Hing C Cheng. Evolution of the diverse antifreeze proteins. *Current opinion in genetics & development*, 8(6):715–720, 1998.

[32] Peter L Davies and Brian D Sykes. Antifreeze proteins. *Current opinion in structural biology*, 7(6):828–834, 1997.

[33] Artiom Kovnatsky, Klaus Glashoff, and Michael M Bronstein. Madmm: a generic algorithm for non-smooth optimization on manifolds. *arXiv preprint arXiv:1505.07676*, 2015.

[34] J Jung, T Ryu, Y Hwang, E Lee, and D Lee. Prediction of extracellular matrix proteins based on distinctive sequence and domain characteristics. *Journal of Computational Biology*, 17(1):97–105, 2010.

[36] C Ding and H Peng. Minimum redundancy feature selection from microarray gene expression data. *Computational Systems Bioinformatics*, 2003.

[37] J Wright, A.Y Yang, A Ganesh, S.S Sastry, and Ma Yi. Robust face recognition via sparse representation. *IEEE TPAMI*, 31:210–227, 2008.

[38] R Baraniuk, M Davenport, R DeVore, and M Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 2008.

[39] E Candès and T Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[40] D. L Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, June 2006.

[41] E Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci., Paris, Series I*, 346:589–592, 2008.

[42] E Candès, J Romberg, and T Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59 (8):1207–1223, 2006.

[43] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[44] B Boeckmann, A Bairoch, R Apweiler, MC Blatter, A Estreicher, E Gasteiger, M.J Martin, K Michoud, C O'Donovan, I Phan, S Pilbout, and M. Schneider. The SWISS-PORT protein knowledgebase and its supplement TrEMBL. *Nucleic Acids Res*, 31(1):365–370, 2003.

[45] W Li, L Jaroszewski, and A Godzik. Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics*, 17:282–283, 2001.

[46] K.C Chou. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins.*, 43(3):246–255, 2001.

[47] K.C Chou. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 21:10–19, 2005.

[48] K.C Chou and H.B Shen. Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, 22:1717–1722, 2006.

[49] K.C Chou and Y.D Cai. Prediction of membrane protein types by incorporating amphipathic effects. *J.Chem.Inf.Model*, 45:407–413, 2005.

[50] P Walter, R Gilmore, and G Blobel. Protein trans location across the endoplasmic reticulum. *Cell*, 38:5–8, 1984.

[51] K.K Kandaswamy, Kuo-Chen Chou, Thomas Martinetz, Steffen Möller, P. N Suganthan, S Sridharan, and G Pugalenthi. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived. *Journal of Theoretical Biology*, 270:56–62, 2011.

[52] Z Xiaowei, M Zhiqiang, and Y Minghao. Using support vector machine and evolutionary profiles to predict antifreeze protein sequences. *International Journal of Molecular Science*, 13:2196–2207, 2012.

[53] C-S Yu and C-H Lu. Identification of antifreeze proteins and their functional residues by support vector machine and genetic algorithms based on n-peptide compositions. *PLoS*, 2011.

[54] S Mondal and P.Pai Priyadarshini. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *Journal of Theoretical Biology*, 356: 30–35, 2014.

[55] HF Yang, Yong-mei CHENG, Shao-wu ZhANG, and Quan PAN. Prediction of protein subcellular localization using a novel feature extraction method: sequence-segmented pseudo amino acid composition. *Acta Biophysica Sinica*, 24(3):232–238, 2008.

[56] Krishna Kumar Kandaswamy, Ganesan Pugalenthi, Kai-Uwe Kalies, Enno Hartmann, and Thomas Martinetz. Ecmpred: Prediction of extracellular matrix proteins based on random forest with maximum relevance minimum redundancy feature selection. *Journal of theoretical biology*, 317:377–383, 2013.

[57] 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.

[58] Shao-Wu Zhang, Wei Chen, Feng Yang, and Quan Pan. Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. *Amino Acids*, 35(3):591–598, 2008.

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

[60] 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.

[61] Pufeng Du, Xin Wang, Chao Xu, and Yang Gao. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Analytical biochemistry*, 425(2):117–119, 2012.

[62] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

[63] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *In Proceedings of the Thirteenth International Conference*, pages 284–292. Morgan Kaufmann Publishers Inc., 1996.

[64] Pat Langley et al. *Selection of relevant features in machine learning*. Defense Technical Information Center, 1994.

[65] Krishna Kumar Kandaswamy, Ganesan Pugalenthi, Enno Hartmann, Kai-Uwe Kalies, Steffen Möller, PN Suganthan, and Thomas Martinetz. Spred: A machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. *Biochemical and biophysical research communications*, 391(3):1306–1311, 2010.

[66] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. ISBN 0070428077, 9780070428072.

[67] Krishna Kumar Kandaswamy, Kuo-Chen Chou, Thomas Martinetz, Steffen Möller, PN Suganthan, S Sridharan, and Ganesan Pugalenthi. Afp-pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *Journal of Theoretical Biology*, 270(1):56–62, 2011.

[68] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams, and Hongyu Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003.

[69] Jae Won Lee, Jung Bok Lee, Mira Park, and Seuck Heun Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005.

[70] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[71] K Krishna Kumar, Ganesan Pugalenthi, and PN Suganthan. Dna-prot: identification of dna binding proteins from protein sequence information using random forest. *Journal of Biomolecular Structure and Dynamics*, 26(6):679–686, 2009.

[72] Majid Masso and Iosif I Vaisman. Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. *Journal of Theoretical Biology*, 266(4):560–568, 2010.

[73] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[74] Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H Witten. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481, 2004.

[75] Chin-Sheng Yu and Chih-Hao Lu. Identification of antifreeze proteins and their functional residues by support vector machine and genetic algorithms based on n-peptide compositions. *PloS one*, 6(5):e20445, 2011.

[76] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, TN Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[77] Guoli Wang and Roland L Dunbrack. Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[78] Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 28(1):45–48, 2000.

[79] UniProt Consortium et al. The universal protein resource (uniprot) in 2010. *Nucleic acids research*, 38(suppl 1):D142–D148, 2010.

[80] National Center for Biotechnology Information protein database. http://http://www.ncbi.nlm.nih.gov/protein/. Accessed: 2015-06-30.