From the Institute for Neuro- and Bioinformatics
of the University of Lübeck
Director: Prof. Dr. Thomas Martinetz

Sequence function classification by machine learning methods

Dissertation for
Fulfillment of Requirements for the
Doctoral Degree of the University of Lübeck
from the Department of Computer Sciences/Engineering

Submitted by Krishna Kumar Kandaswamy
Born in Coimbatore - TN, India
Lübeck, 2011

First referee: Prof. Thomas Martinetz

Second referee: Prof. Enno Hartmann

Date of oral examination: 18.04.2012

Approved for printing, Lübeck : 23.04.2012

Acknowledgements

Many people contributed either directly or indirectly. It will be difficult to name all of them. To start, I express sincere gratitude to my thesis advisor Prof. Thomas Martinetz, I enjoyed every moment discussing many related things to my thesis. I always admired his perfect scientific judgment in solving problems, a clear wit and a rare ability to stay cool in the most critical situations. Despite his busy schedule, he was always there to advise and give support whenever required.

I convey my thanks and gratitude to Prof. Enno Hartmann, who introduced me to the field of protein secretion and translocation. In a budding phase, we had many discussions and I used to wonder his ability to put in tireless efforts in reading and discussing things out (which helped in speeding up the things). I am very grateful to Dr. Kai-Uwe Kalies and Dr. Steffen Möller for an incalculable number of suggestions, critical talk and motivational sessions.

A special thank you must go to Dr. Ganesan Pugalenthi for his valuable inputs in regards to the concept of sequence analysis. Periodic technical and non-technical guidance from him helped me to learn quite a number of things. I would also like to thank my colleagues and fellow doctoral students at Institute for Neuro- and Bioinformatics and Institute of Biology for the many friendly and enlightening discussions and the good times spent over the past years. These are Mehrnaz Hazrati, Vivica Stokes, Simon Dornseifer, Arne Weigenand, and Ingrid Braenne. It has been a pleasure to work with you.

I also thank my friends Balamurugan, Ramya, Madhan, Dilip and Merlin Veronika for their support, encouragement and timely suggestions. A special thanks goes to the technicians working in the biology lab.

I would like to acknowledge the support by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1].

Last but not the least, I am extremely grateful to my parents and brother for their affection and constant encouragement without which I would not have come to this position in my life.

Summary

A large number of protein sequences are being accumulated in genomic databases day by day. It has become a challenging task for researchers to identify the functions of these new proteins. Over the years, numerous methods to detect sequence similarity, such as BLAST and FASTA, have been developed and gained popularity due to their high success rates. However, these methods failed when the pairwise sequence similarities get lower and such an alignment-based method would rarely yield satisfactory prediction. Therefore, there is a need for alignment-free methods (machine learning models) for predicting functions of proteins and its inferences which remains as one of the most important research areas in bioinformatics [1-12].

Application of machine learning techniques has significantly benefited diverse areas including cell and molecular biology. Machine learning methods have been broadly classified into two main categories : supervised and unsupervised machine learning methods. The main objective of the proposed thesis work is to develop new sequence analysis tools utilizing machine learning algorithms for molecular cell biology problems [1-5, 10].

In this thesis, the usefulness of support vector machine and random forest for protein function prediction are extensively tested by applying it to the classification of a variety of functionally distinguished classes of proteins. These include classical and non-classical secretory proteins, extracellular matrix proteins and subcellular location of apoptosis proteins [1, 4, 5]. These classes of proteins are suitable for testing support vector machine and random forest as they represent proteins of diversely different functions that cover protein synthesis regulation, regulation of host cell infection, protein self-association, molecular signaling and drug discovery.

Some proteins may not have adequate sequence similarities although they share similar structures and biochemical functions. Identification of antifreeze and bioluminescent proteins from protein sequence is more interesting due to the low pairwise sequence similarity which often falls below the twilight zone [2, 3]. So far, no specific method has been reported to identify protein families (antifreeze and bioluminescent) from primary sequence. In this thesis work, we

have developed machine learning method to annotate hypothetical proteins of antifreeze and bioluminescent families.

We have developed a classification model to predict proteins pertaining to post-translational pathway and tested the top ranked predicted protein candidates experimentally. This work has identified putative signals on mammalian protein sequences that sign statistically responsible for the translocation pathway. An improved model for this sorting process has potential practical applications i.e. in gene therapies and the understanding of pathogen physiology since it allows assignment of subpathways of topogenesis to different proteins that are secreted.

In this thesis, various frequencies of amino acids were used to predict the cellular functions or location of proteins and protein functional families. Different feature selection and machine learning algorithms have been used to process the large amount of training and test data. The results obtained are quite good and validate the use of these machine learning methods. Thus, the tools developed by us will provide useful insights for both basic research and drug design.

Publications

1. **Kandaswamy, K. K**, Pugalenthi, G, Kalies, K, Hartmann, E, Martinetz, T (2011) EcmPred: Prediction of Extracellular matrix proteins based on Random Forest with maximum relevance minimum redundancy. *Journal of computational Biology*, (Under Review).

2. **Kandaswamy, K. K**, Pugalenthi, G, Hazrati, M. K, Kalies, K. U, Martinetz, T (2011) BLProt: prediction of bioluminescent proteins based on support vector machine and relieff feature selection. *BMC Bioinformatics*, 12, 345 (0 Citation).

3. **Kandaswamy, K. K**, Martinetz, T, Chou, K. C, Suganthan, P. N, Pugalenthi, G (2011) AFP-Pred: A random forest approach for the prediction of antifreeze proteins from sequence derived properties, *Journal of Theoretical Biology*, 270(1), 56-62 (25 Citation).

4. **Kandaswamy, K**. **K**, Pugalenthi, G, Möller, S, Hartmann, E, Kalies, K, Suganthan, P, Martinetz, T (2010) Prediction of apoptosis protein locations with Genetic Algorithms and Support Vector Machines through a new mode of pseudo amino acid composition. *Protein & Peptide Letters*, 17(12), 1473-9 (20 Citation).

5. **Kandaswamy, K. K**, Pugalenthi, G, Hartmann, E, Kalies, K. U, Möller, S, Suganthan, P. N, Martinetz T (2010) 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),1306-11 (5 Citation).

Additional papers

The papers below have been written or co-authored by me during the Ph.D. project, but are not considered part of the project:

6. Pugalenthi, G, **Kandaswamy, K**. **K**, Kolatkar, P (2011) RSARF: Prediction of residue solvent accessibility from protein sequence using random forest method. *Protein & Peptide Letters*, (In

Press)

7. Shameer, K, Pugalenthi, G, **Kandaswamy, K. K**, Sowdhamini R (2011) 3dswap-pred: Prediction of 3D Domain Swapping from Protein Sequence Using Random Forest Approach. *Protein & Peptide Letters*, 18 (1 Citation).

8. Pugalenthi, G, **Kandaswamy, K**. **K**, Suganthan, P. N, Sowdhamini, Martinetz, T, Kolatkar, P (2010) SMpred: a support vector machine approach to identify structural motifs in protein structure without using evolutionary information. *Journal of Biomolecular Structure and Dynamics*, 28(3), 405-14 (0 Citation).

9. Shameer, K, Pugalenthi, G, **Kandaswamy, K. K**, Suganthan, P. N, Archunan, G, Sowdhamini, R (2010) Insights in to protein sequence and structure derived features mediating 3D domain swapping mechanism using Support Vector Machine based approach. *Bioinformatics* and *Biology Insights*, 4, 33-42 (2 Citation).

10. Pugalenthi, G, **Kandaswamy, K. K**, Suganthan, P. N, Archunan, G, Sowdhamini R (2010) Identification of functionally diverse lipocalin proteins from sequence information using support vector machine. *Amino Acids*, 39(3), 777-83 (1 Citation).

11. **Kandaswamy, K. K**, Pugalenthi, G, Suganthan, P. N, Gangal, R (2010) SVMCRYS: An SVM approach for the prediction of protein crystallization propensity from protein, *Protein Peptide Letters*, 4, 423-430 (2 Citation).

12. **Kumar, K. K,** Pugalenthi, G, Suganthan, P. N (2009) Identification of DNA binding proteins from protein sequence information using random forest method, *Journal of Biomolecular Structure and Dynamics*, 26(6), 663-895 (4 Citation).

Table of contents

1 Biological Background

1.1 Motivation

The understanding of the biological function of proteins remains a prodigious task in biology. Due to remarkable growth in molecular biology, the genome sequences of many prokaryotic and eukaryotic organisms were obtained. With the new advancement in DNA sequencing techniques, new putative proteins are added to the databases in a much faster rate than they can be tested experimentally to determine the function. Thus a fundamental challenge in computational biology is to develop efficient and accurate algorithms to classify putative proteins and associate them with families of proteins with known functions.

Protein function classification is a vital aspect of genome annotation that primarily depends on sequence similarity (Lipman and Pearson, 1985). Over the years, numerous methods to detect sequence similarity, such as BLAST and FASTA, have been developed and gained popularity due to their high success rates (Pearson and Lipman, 1988; Altschul *et al.,* 1990; Altschul *et al.,* 1997). However, these methods may fail when the pairwise sequence similarities get lower and such an alignment-based method would rarely yield satisfactory prediction. Therefore, there is a need for alignment-free methods (machine learning models) for predicting functions of proteins and its inferences, which remains as one of the most important research areas in bioinformatics (Chou, 2011).

Application of machine learning techniques has significantly benefited diverse areas including cell and molecular biology (Chou, 2011). Many algorithms have been devised so far to solve problems that can be broadly classified into three different categories viz. pattern recognition/classification, regression/function approximation and density estimation (Vapnik, 1995). The main objective of the proposed thesis work is to develop new sequence analysis tools utilizing machine learning algorithms for molecular cell biology problems.

In this thesis, recent developments in machine learning and artificial intelligence are explored. Emphasis is mainly made on solving important pattern recognition problems, for which the support vector machine (SVM) and random forest (RF) are used extensively. Machine learning

techniques like data compression, feature selection, dimension reduction, etc. are used (as a preprocessing step to main algorithm) to process the large amount of data. These algorithms were applied to biological problems like classical and non-classical protein secretion, extracellular matrix prediction, subcellular location of apoptosis proteins, antifreeze and bioluminescent proteins (explained in detail in the following chapters). The results obtained were quite good and therefore justify the use of these modeling methodologies. Thus, the tools developed in this thesis provide insights for both fundamental research and drug design.

We developed a classification model to predict proteins pertaining to post-translational pathway. We tested the top ranked predicted protein candidates experimentally. This work has identified putative signals on mammalian protein sequences that sign statistically responsible for the translocation pathway (Rapoport, 2007). An improved model for this sorting process has potential practical applications i.e. in gene therapies and the understanding of pathogen physiology since it allows assignment of sub pathways of topogenesis to different proteins that are secreted.

1.2 Biological Background

1.2.1 From DNA to Proteins

Cells are the essential working units of every living system. The nucleus of every cell in eukaryotic organisms (including animals and plants) contains a large DNA (Deoxyribonucleic acid) molecule, which carries the genetic information of every organism (Nelson and Cox, 2005).

DNA consists of two long chains of nucleotides. Each nucleotide is composed of a nitrogenous base, one phosphate molecule and one sugar molecule. Four different bases are contained in DNA: adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the bases in any of the DNA strands is called the DNA sequence. The two DNA strands are complementary, which means that they contain the same genetic information (the information is duplicated) and are held together by weak hydrogen bonds (Berg *et al*., 2002).

The DNA sequence contains instructions for the synthesis of a protein. These are the specific sections of the DNA sequence usually called genes. The way how information stored in the DNA

is passed on for synthesis of proteins is called central dogma of molecular biology (Crick, 1958). A simplified scheme of this process is illustrated in Figure 1.1. This is commonly represented by two main steps as follows:

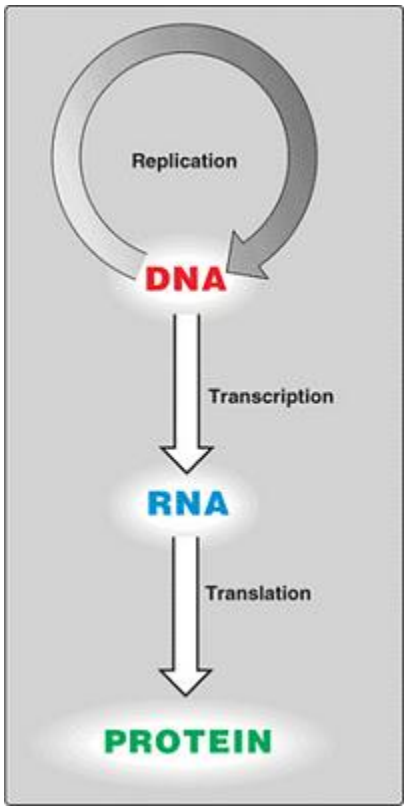


Figure 1.1: Scheme of the central dogma of molecular biology (taken from http://medicinexplained.blogspot.com/)

(i) Transcription (DNA → mRNA)

Transcription is the process by which information coded in a specific segment of the DNA sequence (or gene) is passed to a RNA molecule called messenger RNA (mRNA). RNA molecules are similar to DNA in composition. They also consist of a chain of nucleotides, but contain only one strand and use different nitrogenous bases and sugars. The process by which genes are transcribed into a RNA molecule is usually called gene expression (Berg *et al.*, 2002).

(ii) Translation (mRNA → Protein)

Translation is the process where genetic information now coded in the mRNA is used to synthesize a specific protein. This process is mediated by other macromolecules called

ribosomes and also other types of RNA molecules. The genetic information is translated from a chain of nucleotides (mRNA) to a chain of amino acids. This is done using the genetic code, where a nucleotide triplet (codon) is associated with a specific amino acid. The protein contains 20 different amino acids. The final sequence of amino acids generated corresponds to what we know as a protein (Champe *et al.*, 2004).

1.2.2 Molecular biology databases

Molecular biology databases play a vital role in Computational Biology (Baker *et al.*, 1999). Currently, there are 1330 molecular biology databases available for computational analyses (Galperin and Cochrane, 2011). These databases are extremely useful for investigation purposes since they give researchers access to huge amounts of data, which can be searched, inspected, and used for any analysis. Table 1.1 illustrates some of the commonly used databases.

Most molecular biology databases are very large: e.g. SWISS-PROT contains 529056 sequence entries comprising 187423367 amino acids abstracted from 198689 references (*Release 2011_06*). There is an exponential growth rate in these databases. The growth rate and actual size of most molecular biology databases have become a serious problem: without automated methods such as data mining and knowledge discovery algorithms, the data collected cannot be fully exploited.

Database Name	Description	Reference
EMBL	It maintains Europe's primary nucleotide sequence data resource	http://www.ebi.ac.uk/embl/ (Stoesser *et al.*, 2002)
SWISS-PROT	It contains high-quality annotation, non-redundant, and cross-referenced to several other databases (EMBL nucleotide sequence database, PDB and PROSITE pattern database)	http://www.ebi.ac.uk/swissprot/ (Bairoch and Apweiler, 2000)
NCBI/GenBank	It is a collection of publicly available annotated nucleotide sequences, including mRNA sequences with coding regions, segments of genomic DNA with a single gene or multiple genes, and ribosomal RNA gene clusters	http://www.ncbi.nlm.nih.gov /Genbank/index.html (Benson *et al.*, 2000)
PDB	It contains information about experimentally determined structures of proteins	http://www.rcsb.org/pdb/ (Berman, 2008)
IPI	It describes the proteomes of higher eukaryotic organisms (Human)	http://www.ebi.ac.uk/IPI/IPIhelp.html (Kersey *et al.*, 2004)

PROSITE	It consists information about biologically significant sites and patterns	http://www.expasy.org/prosite/ (Sigrist *et al.*, 2010)
Pfam	It is the collection of different protein families and domains	http://www.sanger.ac.uk/resources/databases/pfam.html (Finn *et al.*, 2010)
Gene Ontology	It standardizes the representation of gene and gene product attributes across various species and databases	http://www.geneontology.org/ (Gene Ontology Consortium, 2010)
MEDLINE	It is a bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, and the preclinical sciences	http://www.ncbi.nlm.nih.gov/PubMed/ (Roberts, 2001)

Table 1.1: Common molecular biology databases

1.3 Thesis Organization

Chapter 2 describes the conventional machine learning methods, feature selection and feature encoding methods that have been used in our work.

Chapter 3 presents the work on protein function classification. Knowledge about the function of proteins is vital in the understanding of biological processes. In this chapter, the usefulness of SVM and RF for protein function prediction is extensively tested by applying it to the classification of a variety of functionally distinguished classes of proteins. These include classical and non-classical secretory proteins, extracellular matrix proteins, and subcellular location of apoptosis proteins.

Chapter 4 presents the work on protein family classification. Identification of antifreeze and bioluminescent proteins from protein sequence is more interesting due to the poor sequence identity which often falls below the twilight zone. In this chapter, we present machine learning methods to annotate hypothetical proteins of antifreeze and bioluminescent families.

Chapter 5 describes co and post translocation. Interestingly, many proteins of the immune system seem to prefer the post-translational transport. We developed an in-silico model to predict proteins pertaining to the post-translational pathway. Top ranked candidate proteins were tested for their translocation behavior. Therefore, these proteins were analyzed in in vitro

translation/translocation assays suitable for co or post-translational protein translocation into ER-membranes, respectively.

Chapter 6 will give a conclusion of the thesis.

2 Machine learning algorithms and feature representation

2.1 Machine learning

Machine learning which deals with the study and analysis of large quantities of data in order to discover meaningful patterns. Most of the data mining algorithms need substantial amount of data in order to construct and train the models that will then be used to accomplish classification or regression. Machine learning methods have been applied broadly within the field of computational biology such as predicting protein subcellular location (Chou and Cai, 2002; Cai *et al.,* 2002a), membrane protein types (Cai *et al.,* 2003a; Cai *et al.,* 2004a), protein structural classes (Cai *et al.,* 2002b), specificity of GalNAc-transferase (Cai *et al.,* 2002c), HIV protease cleavage sites in proteins (Cai *et al.,* 2002d), beta-turn types (Cai *et al.,* 2002e), protein signal sequences and their cleavage sites (Cai *et al.,* 2003b), alpha-turn types (Cai *et al.,* 2003c), catalytic triads of serine hydrolases (Cai *et al.,* 2004b), micro-array and gene expression analysis (Brown *et al.,* 1999), Drug discovery (Burbidge *et al.,* 2001; Zernov *et al.,* 2003), biomarker discovery (Prados *et al.,* 2004) and among many others. Machine learning is organized into taxonomy, based on the desired outcome of the algorithm.

2.1.1 Supervised learning methods

Supervised learning typically consists of relating a series of attributes of the data to a specific class or numerical value known as a label of that specific instance. Wide spread algorithms used in supervised learning are Decision trees, Nearest neighbor classification, Artificial neural networks, Support vector machines, and Random forest (Tarca *et al.,* 2007).

2.1.2 Unsupervised learning methods

In unsupervised learning, the data do not have class label. The task is to group the given data into clusters based on the common features they share. Principally, one needs to explore the data and discover similarities between the objects. Wide spread algorithms in unsupervised learning are hierarchical clustering, K-means clustering, self-organizing feature maps (SOFM), and principal component analysis (Hinton *et al.,* 1999; Tarca *et al.,* 2007).

2.2 Classification Algorithms

2.2.1 Support Vector Machines

The support vector approach originally proposed by Glucksman (Glucksman, 1966) and Vapnik (Vapnik, 1995) and later developed and popularized by a number of authors is also known universally as a feed forward network (Cortes and Vapnik, 1995; Burges, 1998; Scholkopf *et al.*, 1999; Cristianini *et al.*, 2000). SVMs employ the structural risk minimization (SRM) principle. This principle is based on the fact that the error rate of a learning machine on test data is constrained by the sum of the training error rate and Vapnik-Chervonenkis (VC) dimension. The VC dimension is a measure of complexity of the decision space. It facilitates quantitative means of discriminating between the capacities of different classifiers. For non-linear and non-separable problems, the support vector methodology provides a decision space with a minimal VC dimension and training error so that the classifier has a low probability of generalization errors.

2.2.1.1 SVM classifier for linearly separable patterns

Consider a binary classification training sample $\{(\mathbf{x}_i, y_i)\}_{i=1, 2\ldots N}$ where $\mathbf{x}_i$ is the vector of the input pattern for the i^{th} example and y_i is the corresponding target output. The pattern represented by the subset $y_i = +1$ belongs to class 1 and the pattern represented by the subset $yi = -1$ belongs to class 2. For linearly separable data, there is a hyperplane defined by

$$\mathbf{w} \bullet \mathbf{x} + b = 0 \tag{2.1}$$

that separates the data into two different classes for optimal separation, the distance between the closest data points to the hyperplane must be maximal.

With the above constraint on ***w*** and *b*, it can be shown that (Figure 2.1) the distance between the closest point belonging to two different classes (margin) can be obtained as (Gunn, 1997)

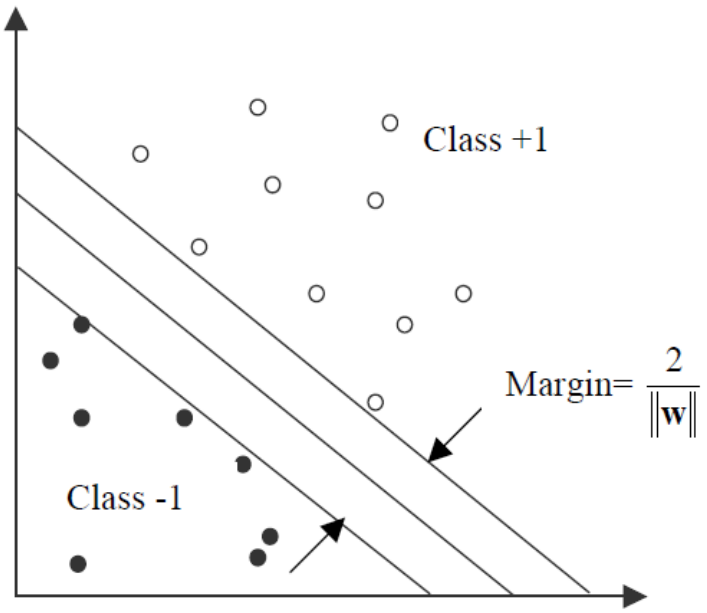


Figure 2.1: Maximum margin-minimum norm classifier

$$\rho(\mathbf{w}, b) = \frac{2}{||\mathbf{w}||} \tag{2.2}$$

and the separating hyperplane satisfies

$$y_i(\mathbf{w} \bullet \mathbf{x}_i + b) \geq 1 \quad \forall\, i. \tag{2.3}$$

From the above equation, it can be found that the maximum margin hyperplane is obtained by minimizing the norm of the weight vector. It can be shown that such a norm minimization is equivalent to minimizing an upper bound on the VC dimension. Thus the maximum margin hyperplane has a very high probability of simultaneously having the least training error as well as generalization error.

To maximize the margin, the task is therefore to minimize the function

$$\Phi(\mathbf{w}) = (1/2)||\mathbf{w}||^2 \tag{2.4}$$

such that $y_i\,(\mathbf{w} \bullet \mathbf{x_i} + b) \geq 1 \quad \forall\, i.$

The cost function (2.4) is a convex function of ***w*** and the constraints are linear in ***w***. Thus, we can solve the constrained optimization problem by constructing the augmented Lagrangian function (Bishop, 2006)

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}\bullet\mathbf{w} - \sum_{i=i}^{N} \alpha_i [y_i (\mathbf{w}\bullet \mathbf{x}_i + b) - 1] \tag{2.5}$$

The solution of the constrained optimization problem is determined by the saddle point of the Lagrangian function L (**w**, b, ***α***) which has to be minimized with respect to ***w*** and *b*. It is well known in the optimization literature, it would be useful to construct a dual problem. Expansion of the primal problem term-by-term and after suitable manipulation, the Wolfe dual Lagrangian can be written as

$$\mathbf{w}(\alpha) = \sum_{i=1}^{N} \alpha_i - (1/2) \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \bullet \mathbf{x}_j) \tag{2.6}$$

which must be maximized with respect to α_i subject to the constraints

$$\alpha_i \geq 0 \tag{2.7}$$

$$\sum_{i=i}^{N} \alpha_i y_i = 0 \tag{2.8}$$

So our task of binary classification reduces to the above maximization problem i.e. equations (2.6) to (2.8). It can be seen from equation (2.6) that the Lagrangian dual is cast entirely in terms of the training data. Also, we notice that the data points appear only inside the dot product. Solution of equations (2.6) - (2.8) determines the optimal separating hyperplane as

$$\bar{\mathbf{w}} = \sum_{i=1}^{N} \bar{\alpha}_i \mathbf{x}_i \, y_i \tag{2.9}$$

$$\bar{b} = \left(-\frac{1}{2}\right) \bar{\mathbf{w}} \bullet [\mathbf{x}_r + \mathbf{x}_s] \tag{2.10}$$

The points for which Lagrange multipliers α_i are non-zero and those which lie on the margin of the separation, are termed support vectors. For linearly separable data, all the support vectors will lie on the margin of separation and the number of support vectors will be small, the optimal hyperplane can be determined by a small subset of the training set. The hyperplane can be found out just by using these support vectors. Thus, a support vector machine summarizes information

content of a dataset using support vectors. Since no training errors are allowed, this optimization problem is called the hard margin. If the separating hyper plane is allowed to pass through the origin by selecting a fixed value $b = 0$, then in that case the SVM formulation is called the hard margin SVM without threshold.

2.2.1.2 Optimal Hyperplane for linearly non-separable patterns

For linearly non-separable data, it is not possible to construct a hyperplane without a certain amount of classification errors. We can find an optimal hyperplane that minimizes the probability of occurrence of classification errors, averaged over the training sets by introducing a set of non-negative scalar slack variables in the definition of the separating hyperplane in the form of a penalty function

$$y_i\ (\mathbf{w} \bullet\ \mathbf{x}_i + b) \geq 1 - \xi_i \quad i = 1, 2..., N \tag{2.11}$$

where $\xi_i \geq 0$. The generalized optimal separating hyperplane is determined by finding the vector **w** that minimizes the functional

$$\Phi(\mathbf{w}, \xi) = (1/2)||\mathbf{w}||^2 + C \sum_{i=1}^{N} \xi_i \tag{2.12}$$

(where, *C* , is a given value) subject to the constraints in equation (2.11).

The saddle point of the Lagrangian corresponds to the solution to the optimization problem of equation (2.12) under the constraints of equation (2.11). It can be shown that the dual solution can be obtained as

$$\max_\alpha W(\alpha)\ = \max_\alpha - (1/2) \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i\ \alpha_j y_i y_j \left(\mathbf{x}_i \bullet \mathbf{x}_j\right) + \sum_{i=1}^{N} \alpha_i \tag{2.13}$$

with the constraints,

$$0 \leq\ \alpha_i\ \leq C \qquad\qquad i=1..., N \tag{2.14}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{2.15}$$

Thus the problem for the case of linearly non-separable patterns is similar to that of the simple case of linearly separable patterns except that the constraints $\alpha_i \geq 0$ are now replaced by the

more stringent constraints $0 \leq \alpha_i \leq C$. The parameter C controls the tradeoff between complexity of the support vector machine and the number of non-separable points. It may therefore be necessary to view this parameter in the form of a regularization parameter. This parameter has to be selected by the user. In case of non-separable data one needs to allow for training errors and the algorithm is usually called the soft margin SVM algorithm.

2.2.1.3 Non-linear Support Vector Machines

The methods developed in the above sections are for linear classifiers and as such cannot deal with non-linearly separable data. The task of generalizing SVMs to handle nonlinearly separable data can be accomplished by mapping the data into a richer higher dimensional feature space and by subsequently using a linear classifier. The mapping of the input data x in the feature space $x \rightarrow \Phi(x)$ where they are linearly separable is shown schematically in Figure 2.2.

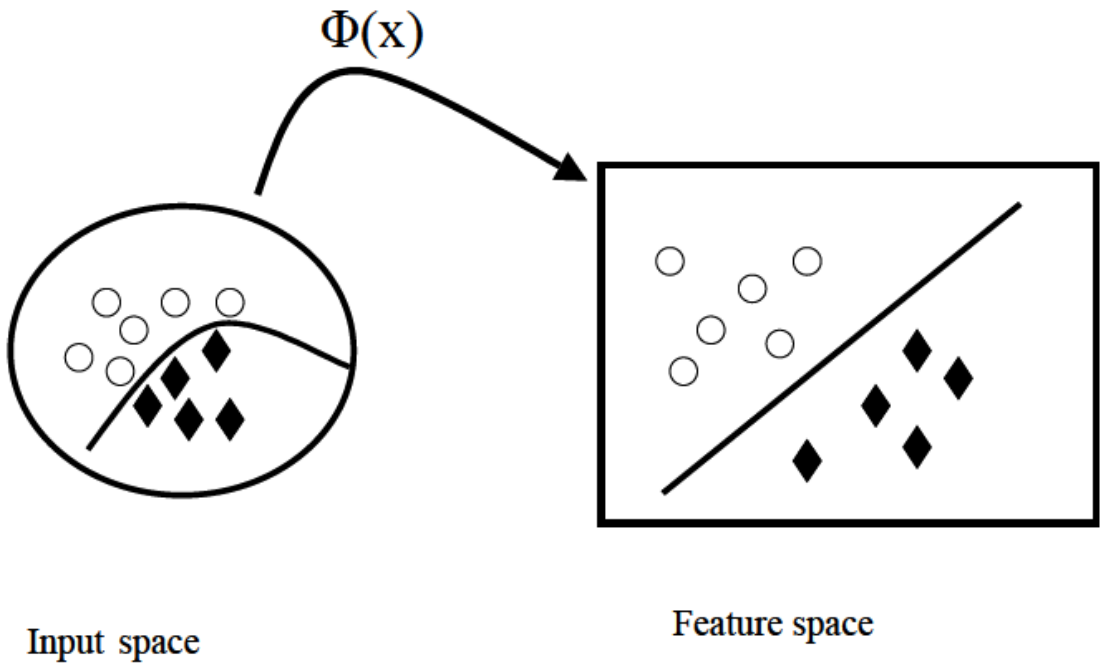


Figure 2.2: Mapping of data into feature space for linear separability

Therefore in the higher dimensional feature space equation (2.13) can be written as

$$\mathbf{w}(\alpha) = \sum_{i=1}^{N} \alpha_i - (1/2) \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i) \bullet \Phi(\mathbf{x}_j) \qquad (2.16)$$

Working in a higher dimensional feature induces a computational problem of having to deal with very large vectors. This problem can be solved by introduction of implicit mapping by kernels. A kernel function is defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \bullet \Phi(\mathbf{x}_j) \tag{2.17}$$

The idea of kernel functions is to perform operations in input space rather than the very high dimensional feature space (Burges, 1998). In other words, an inner product in the feature space has an equivalent kernel in the input space. Thus equation (2.16) can be written in the form of kernel functions in the low dimensional input space as

$$\mathbf{w}(\alpha) = \sum_{i=1}^{N} \alpha_i - (1/2) \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{2.18}$$

subject to the constraints

$$0 \leq \alpha_i \leq C \quad i=1 \ldots N \tag{2.19}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{2.20}$$

After the optimal values of α_i have been found, the decision function is based on the sign of

$$f(\mathbf{x}) = \sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_j) + b \tag{2.21}$$

Since the bias *b* does not feature in the dual formulation, it is found from the primal constraints

$$b = (-1/2)[\max_{\{i|y_i=-1\}} (\sum_{j \in \{sv\}}^{m} y_j \alpha_j K(x_i, x_j)) + \min_{\{i|y_i=+1\}} (\sum_{j \in \{sv\}}^{m} y_j \alpha_j K(x_i, x_j))] \tag{2.22}$$

where *m* are the number of support vectors.

It is the compact convex quadratic optimization form of SVM having a unique solution that has attracted researchers from different areas. One can thus map the SVM classification problem into a standard quadratic programming problem (QP) and solve it with the QP solvers. The kernel

function appearing in the problem can be selected by using the Mercer's theorem. The kernel matrix contains all the necessary information for the support vector machine learning algorithm and is generally known as the information bottleneck. A list of popular kernels is shown in Table 2.1. The definitions of hard margin SVMs and soft margin SVMs are valid for the non-linear SVMs also.

S. No	Name of the kernel	Expression
1	Polynomial	$K(\mathbf{x}_i,\mathbf{x}_j) = \left((\mathbf{x}_i \bullet \mathbf{x}_j) + 1\right)^p$
2	Gaussian Radial Basis Function	$K(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{-2\sigma^2}\right)$
3	Exponential Radial Basis Function	$K(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(\frac{\|\mathbf{x}_i-\mathbf{x}_j\|}{-2\sigma^2}\right)$
4	Perceptron	$K(\mathbf{x}_i,\mathbf{x}_j) = \tanh\left(b(\mathbf{x}_i \bullet \mathbf{x}_j) - c\right)$
5	Fourier Series	$K(\mathbf{x}_i,\mathbf{x}_j) = \frac{\sin\left(N+\frac{1}{2}\right)(\mathbf{x}_i-\mathbf{x}_j)}{\sin\left(\frac{1}{2}\right)(\mathbf{x}_i-\mathbf{x}_j)}$ N= dimension of the space
6	Tensor Product Splines	$K(\mathbf{x}_i,\mathbf{x}_j) = \prod_{m=1}^{n} K_m(\mathbf{x}_{im},\mathbf{x}_{jm})$

Table 2.1: Different types of kernel functions (Burges, 1998)

2.2.2 Random Forests

Random Forest is an ensemble decision tree classifier, which incorporates two effective machine learning techniques (bagging and selection of feature from random subspace) into a single method. Random forest is a collection of decision trees, where each tree is grown using a subset of the possible attributes in the input feature vector, instead of using complete features in all trees.

2.2.2.1 Introduction

The concept of random forest (RF) was proposed by Leo Breiman in 1999. It is an ensemble of randomly constructed independent decision trees (Breiman *et al.*, 1984). It is one of the most popular ensemble methods that are robust to noise, without overfitting, fast and offers

possibilities for explanation and visualization of its output. In the random forest method, a large number of classification trees are grown and combined. It uses bootstrap sampling technique, which is an improved version of bagging. It generally exhibits substantial performance improvement over single tree classifiers such as CART and C4.5 (Ross Quinlan, 1993; Breiman, 2001).

Decision trees provide an effective approach in a tree-like graph of model to predict the probable decision of a system by analysing various associated parameters. The random forest classification extends the concept of decision trees and has been successfully employed in developing solutions for a variety of problems in biology including tumor classification, microarray analysis, prediction of protein-protein interactions and classification of microRNAs etc (Dudoit *et al.*, 2002; Svetnik *et al.*, 2003; Wu *et al.*, 2003; Chen and Liu, 2005; Lee *et al.*, 2005; Qi *et al.*, 2005; Diaz-Uriarte and Alvarez de Andres, 2006; Jiang *et al.*, 2007; Statnikov *et al.*, 2008).

2.2.2.2 Algorithm

In random forest, each tree differs from all others owing to the randomness introduced in random forest algorithm in two ways: first, in the sample dataset for growing the tree and second, the choice of subset of attributes for node splitting while growing each tree (Breiman, 2001). Such a RF is grown in the following manner:

(i) From the training data of '*n'* instances, draw a bootstrap sample (i.e. randomly sample, with replacement, '*n*' instances).
(ii) For each bootstrap sample, a classification tree is grown with the following modification: choose the best split among a randomly selected subset of *m* (rather than all) features at each node. Each tree is grown to the maximum size.
(iii) Repeat the above steps until (a sufficiently large number) *N* such trees are grown.

For each tree, a bootstrap sample (with replacement) is drawn from the original training data set, i.e. a sample is taken from the training data set and is then replaced again in the data set before drawing the next sample. Likewise, '*n*' numbers of samples are taken to form 'In-Bag' data for a

particular tree, where '*n*' is the size of the training data set. The main advantage of bootstrap sampling is to avoid over fitting the training data. In each of the bootstrap training sets, about one-third of the instances are unused for making the 'In Bag' data on an average and these are called the Out-Of-Bag (OOB) data for that particular tree (Bylander, 2002). The classification tree is induced using this 'In-Bag' data using the CART (Classification and Regression Trees) algorithm.

In the CART algorithm for growing a single binary classification tree, each node is checked whether it is a leaf node or not (Breiman *et al.,* 1984). If it is not a leaf node, i.e. if all the data doesn't belong to a single class, Gini Index is first calculated for each of the attributes 'a_i' in the following manner: If a node contains dataset *T* with examples from '*n*' number of classes, gini index, gini (T) is defined as:

$$\mathrm{gini(T)} = 1 - \sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{p_j^2} \tag{2.23}$$

Where p_j is the relative frequency of class '*j*' in dataset *T*. Gini (T) is minimum if the classes in *T* are skewed. As an impurity measure, class sizes at the node are equal if gini index reaches its maximum value and then all cases in a node belong to the same class if the Gini index is equal to zero (Strobl, 2005). After splitting *T* into two subsets T_1 and T_2 with sizes N_1 and N_2, the gini index of the split data is defined as:

$$\mathrm{gini_{split}(T)} = \frac{\mathrm{N_1}}{\mathrm{N}}\mathrm{gini(T_1)} + \frac{\mathrm{N_2}}{\mathrm{N}}\mathrm{gini(T_2)} \tag{2.24}$$

The attribute that provides the smallest split gini is finally chosen for splitting the node. Gini impurity for node splitting is the default criterion. Alternative criteria like twoing rule, information gain etc. are also available. This procedure is done until all pure nodes are obtained (i.e. if all the examples in one node belong to a single class). Then the trees are pruned to prevent over fitting. In random forest, while inducing each tree from the CART algorithm, the following modifications are made: Instead of choosing among all the unused attributes for node splitting, earlier decided '*m*' number of unused attributes are chosen at random and the best split on these '*m*' is used to split the node. '*m*' is maintained constant for all the trees. The tree is grown to the maximum possible size (Breiman, 2001).

Pruning is not necessary in RF, since Bootstrap sampling takes care of the over fitting problem. This further reduces the computational load of the RF algorithm. There is no need for a separate test data in RF for checking the overall accuracy of the forest. It uses the OOB data for cross validation. After all the trees are grown, the k^{th} tree classifies the instances that are OOB for that tree (left out by the k^{th} tree). In this manner, each case is classified by about one third of the trees. A majority voting strategy is then employed to decide on the class affiliation of each case. The proportion of times that the voted class is not equal to the true class of case '*n*', averaged over all the cases in the training data set is called as the OOB error estimate. Now after growing the forest, if an unseen validation test dataset is given for classification, each tree in the random forest casts a unit vote for the most popular class in the test data. The output of the classifier is determined by a majority vote of the trees. The classification error rate of the forest depends on the strength of each tree and the correlation between any two trees in the forest. The key to accuracy is to keep low bias and low correlation among the trees. If the value of '*m*' is decreased, the strength of each tree decreases, but with increase in '*m*' the correlation among the trees increases and the computational load may also increase. The default value of '*m*' is chosen as $\sqrt{M}$, where '*M*' is the total number of attributes. The range of '*m*' employed is normally kept between $\sqrt{M}/2$ and $2*\sqrt{M}$. Minimizing the OOB error rate chooses the value of '*m*' and the number of trees (Breiman, 2001; Bylander, 2002).

2.2.2.3 Advantage of Random Forest

The important features of random forests are that they can handle any high dimensional and multi-class data easily and the threshold noise limit is more for random forest compared to other classification algorithms. It can be used even if the number of attributes is more than the number of examples. It can also be used to find the significance of each variable for classification. It can also handle missing data effectively using the proximities between the pairs of cases. The computed proximities can also be used in outlier detection and clustering (Breiman, 2001).

2.3 Data Encoding and Representation

The protein sequences are composed of alphabetic letters rather than arrays of numerical values. The sophisticated statistical analyses of sequences are not performed due to the alphabetic data. Data mining algorithms require the protein sequences to be characterized as fixed length vectors. In our work, we applied different feature encoding methods.

(i) Amino acid composition

Protein information can be encapsulated in 20 dimensional vectors by calculating the amino acid compositions of the given protein sequence. Amino acid composition is the frequency of each amino acid in the protein.

(ii) Frequency of function groups

We categorized amino acids 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). Further, we categorized 20 amino acids into three groups, namely hydrophobic (FIWLVMYCA), hydrophilic (RKNDEP) and neutral (THGSQ) amino acid groups.

The frequency of the 10 functional groups (number of occurrences of functional group ‘‘X” divided by length of the protein) and the frequencies of hydrophobic, hydrophilic, neutral, positively charged, negatively charged, polar and non-polar amino acids were computed for every sequence.

(iii) Frequency of tripeptides and short peptides

We utilized tripeptide information for the classification. The frequencies of these 27 tripeptides (three amino acid groups: hydrophobic, hydrophilic, and neutral) were calculated for every sequence. Additionally, we incorporated the frequencies of short peptides (10 residue length) which are rich in hydrophobic, hydrophilic, neutral, polar and non-polar amino acids. For example, a short peptide with more than five hydrophobic residues, we consider it as a hydrophobic peptide. Similarly, we calculated hydrophilic, neutral, polar and non-polar short peptides. In addition, we incorporated the frequencies of short peptides which are rich in the 10 functional amino acid groups.

(iv) Content of secondary structural element (SSE)

Secondary structure information for every sequence was assigned using PSIPRED (McGuffin *et al.,* 2000). PSIPRED provides two options for secondary structure prediction. The first option uses homologous sequence information and the second option predicts secondary structure from the query sequence without using homologous sequence information. We employed the second option of the PSIPRED method for all sequences. The overall composition of helix (H), beta sheet (E), coil (C) were calculated for each sequence.

(v) Frequency of amino acid groups at SSE

The frequencies of amino acid groups, hydrophobic, hydrophilic, neutral and polar amino acids at helix, sheet, and coil regions were calculated.

(vi) Frequency of short peptides

We incorporated the frequency of short peptides (10 residue length) which are rich in hydrophobic or hydrophilic or neutral amino acids. For example, if a short peptide has more than six hydrophobic residues, then we consider this peptide as hydrophobic rich short peptide. Similarly, we calculated hydrophilic, neutral rich short peptides. The frequency of hydrophobic rich peptides (contains at least 6 hydrophobic amino acids) or hydrophilic rich peptides (contains at least 6 hydrophilic amino acids) or neutral amino acid rich peptides (contains at least 6 neutral amino acids) were calculated for each sequence. Similarly, the frequencies of short peptides rich in polar, non-polar, positive, negative and neutral amino acids were computed.

(vii) Physicochemical properties

We took 544 physicochemical properties from the UMBC AAIndex database (Kawashima *et al.,* 2008). For each sequence, a physicochemical property value was calculated as the sum of those values of all amino acids in the given sequence, divided by the number of amino acids in the sequence.

(viii) Frequency of physicochemical groups

On the basis physicochemical properties, we classified 20 amino acids into 7 groups such as hydrophobic, hydrophilic, neutral, positive, negative, polar and non-polar amino acid groups. The frequencies of physicochemical groups were computed for each sequence.

(ix) Pseudo amino acid composition

This method represents the protein sequence into the fraction of amino acid in protein sequence along with sequence order and length information. The first 20 features of the sequence, derived from pseudo amino acid composition method represent the composition of the twenty amino acids. Remaining components represent the sequence order effect. These features can be calculated by using physical properties of amino acids. In our work, hydrophobicity, hydrophilicity and side chain mass of amino acids were used. These physical property values were taken from (Shen and Chou, 2008). The original values of these properties were converted into zero mean values (Chou, 2001). Then, other elements of pseudo amino acid composition were calculated using equations (2) - (6) of (Chou, 2001).

2.4 Feature Selection

Machine learning continually aims to handle an increased amount of data, which makes it necessary to extract the most important information from the huge amount of data abundant in irrelevant or low quality information. This irrelevant data not only increases the time complexity of the learning process, but also affects the accuracy of the process significantly. This makes the extraction of the relevant and the most useful information of paramount importance. Filter and wrapper methods serve as feature selection algorithms, which are essential for datasets containing large number of irrelevant attributes.

The filter method is a preprocessing step, which filters out the irrelevant features before the actual learning process. It uses only the intrinsic characteristics of the data to select the "good" features and exclude the "bad" ones, and thus, does not depend on the induction algorithm to be used. The "goodness" of a feature is usually calculated using empirical yet simple statistical relations. In filter feature selection, every feature is scored independently and the top *n* features are used by the classifier. Different scoring functions like correlation, mutual information, t-statistic, F-statistic, etc. were applied (Saeys *et al.*, 2007). The filter approach is quite easy and it can be applied to huge datasets.

The wrapper method on the other hand uses the induction algorithm as a method of extracting the relevant features from the dataset. The feature space is divided into various feature subsets, which are evaluated for percentage accuracy using the induction algorithm, and the feature

subset with the best accuracy is selected. Feature selection with the wrapper method can be thought of as a combinatorial problem, aimed at finding the smallest subset which gives the highest accuracy among all the large number of possible subsets. It is clear that for a large dataset, exhaustive search is not possible, thus heuristic search techniques were employed.

2.4.1 Information Gain

To identify the prominent features that separate the positive and negative classes, we used the Info Gain algorithm with the ranker method. This method was implemented using Weka 3.5 (Frank *et al.,* 2004). We calculated the information gain for each feature, and ranked them according to this measure, which indicated the gain of information (Haindl *et al.,* 2006).

2.4.2 ReliefF

ReliefF is used to choose the descriptors that discriminate between two classes (Zhang *et al.,* 2008). ReliefF is used as a feature subset selection method. The idea of ReliefF is to compute their nearest neighbors and give more weight to features that discriminate the instance from neighbors of different classes. This method was implemented using Weka 3.5 (Frank *et al.,* 2004).

2.4.3 Maximum Relevance Minimum Redundancy (mRMR)

The minimal-redundancy-maximal-relevance (mRMR) algorithm is a sequential forward selection algorithm developed by Peng *et al.* to analyze the importance of different features. mRMR uses the mutual information to select *M* features that best fulfill the minimal redundancy and maximal relevance criterion. More detailed description of the mRMR algorithm can be found in (Peng *et al.,* 2005).

2.4.4 Genetic Algorithm

Many feature subsets are scored based on classification performance (such as 5 fold cross-validation accuracy or LOOCV) in the wrapper approach, and the best feature subset is used. The different methods available for subset selection are forward selection, backward selection and genetic algorithm (GA) etc. (Kohavi and John, 1997). This method is computationally expensive and it may overfit. The GA combines the principle of a survival of the fittest of natural evolution with the genetic propagation of characteristics, to arrive at a robust search and optimization

technique. The fixed length string is evolved with a GA by using two operators (crossover and mutation) along with a fitness function. It will determine how likely individuals are to reproduce and survive in the population.

To select the most important features among different features, the evolution of the population was simulated. The population of the first generation was selected randomly. Each individual member in the population, defined by a chromosome of binary values, represented a subset of features. The number of the genes at each chromosome was equal to the number of the features. A gene was given the value of 1, if its corresponding feature was selected in the subset; otherwise, it was given the value of 0. The randomly initiated population of strings then evolves; its fittest members being selected to undergo mutation and crossover. More detailed information on GA can be found in (Goldberg, 1989; Holland, 2001).

2.5 Performance assessment of a classifier

The performance of various machine learning methods developed in this thesis was calculated by using threshold-based and threshold-independent parameters. In threshold-based parameters, we used sensitivity, specificity, overall accuracy and matthew's correlation coefficient (MCC) using the following equations. These measurements are expressed in terms of false negative (FN), true positive (TP), false positive (FP), and true negative (TN).

Sensitivity: Percentage of correctly predicted true proteins within the positive classifications.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.25}$$

Specificity: Percentage of correctly predicted false proteins within the negative classifications.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{2.26}$$

Accuracy: Percentage of correctly predicted true and false proteins.

$$\text{Accuracy} = \frac{(\text{TP+TN})}{(\text{TP+TN+FP+FN})} \tag{2.27}$$

Matthews's Correlation Coefficient (MCC): It is an important statistical factor to assess the quality of a classifier and to take care of the unbalancing in data. The Matthew's correlation coefficient ranges from $-1 \leq MCC \leq 1$. A value of MCC = 1 indicates the best possible prediction while MCC = -1 indicates the worst possible prediction (or anti- correlation). Finally, MCC = 0 would be expected for a random prediction scheme.

$$MCC = \frac{((TP*TN)-(FP*FN))}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \quad (2.28)$$

Area under the Curve (AUC): Most of the above measures have the common drawback that their value depends on the selected threshold. The so-called Receiver Operating Curve (ROC) provides a threshold independent measure. The ROC is a plot between the false positive rate (FP/FP+TN) and the true positive rate (TP/TP+FN) (Bradley, 1997).

2.6 Overview

In this thesis, various insilico models were developed for predictions of cellular function or location and protein functional families. Machine learning algorithms require that the protein sequences to be represented as fixed length vectors. In this work, we have used various frequencies of amino acids to predict the cellular function or location of proteins and protein functional families. Different feature selection and machine learning algorithms were used to process the large amount of training and test data. Table 2.2 shows different case studies used in the thesis. The dataset, different type of features, feature selection methods, and classifier were described in the Table 2.2.

Case studies	Dataset	Features	Total number of features	Feature selection	Classifier
Identification of classical and non-classical secretory proteins	Training dataset: 1200 protein sequences Test dataset: 1560 protein sequences	Frequency of function groups, frequency of tripeptides and short peptides, content of secondary structural element (SSE), frequency of amino acid groups at SSE, frequency of amino acid groups at SSE, physicochemical properties	119	Info Gain, Relief F, Maximum Relevance Minimum Redundancy (mRMR)	Random Forest
Prediction of Extracellular matrix proteins	Training dataset: 600 protein sequences Test dataset: 4032 protein sequences	Frequency of 10 functional groups, physicochemical properties	68	Info Gain, Relief F, Maximum Relevance Minimum Redundancy (mRMR)	Random Forest
Prediction of apoptosis protein locations	Training dataset: 317 protein sequences Test dataset: 98 protein sequences	Frequency of function groups, frequency of tripeptides and short peptides, content of secondary structural element (SSE), frequency of amino acid groups at SSE, frequency of amino acid groups at SSE, physicochemical properties	119	Info Gain, Relief F, Maximum Relevance Minimum Redundancy (mRMR), Genetic Algorithm	Support Vector Machine
Prediction of antifreeze proteins	Training dataset: 600 protein sequences Test dataset: 9374 protein sequences	Frequency of functional groups, frequency of physicochemical groups, frequency of short peptides, content of secondary structural element (SSE), physicochemical properties	119	Info Gain, Relief F, Maximum Relevance Minimum Redundancy (mRMR)	Random Forest

Prediction of bioluminescent proteins	Training dataset: 600 protein sequences Test dataset: 18343 protein sequences	Physicochemical properties	544	Info Gain, Relief F, Maximum Relevance Minimum Redundancy (mRMR)	Support Vector Machine
Prediction of co and post-translational translocation	Training dataset: 334 protein sequences Test dataset: 3853 protein sequences	Pseudo amino acid composition	50	Info Gain, Relief F, Maximum Relevance Minimum Redundancy (mRMR)	Support Vector Machine

Table 2.2: Overall workflow of different protein function/ family work

3 Protein Function Classification

3.1 Introduction

Proteins play a vital role in the living organism. Proteins are the molecular machinery that controls and accomplishes nearly every biological function (Krane and Raymer, 2006). Knowledge about the function of proteins is vital in the understanding of biological processes (Downward, 2001). In many cases, proteins concluded to share a remote common ancestor, amino acids determined to be homologous from aligned protein sequences not share strictly similar roles in function and stability, even though their correlation to an overall structural fold may be the same (Lau and, Chasman, 2004). Due to huge amount of data deposited in various databases, much attention has been devoted to the development of methods for the prediction of protein function (Fetrow and Skolnick, 1998; Koonin *et al.,* 1998) from sequence information. These methods regulate the function of a protein by categorizing into a specific protein family or functional class based on protein sequence similarity or the presence of conserved sequence motifs. In the absence of sequence or structural similarities, the criteria for inclusion of distantly related proteins into a protein functional class becomes increasingly random. Therefore, an alternative classification method may need to be explored in facilitating the study of protein function.

In this thesis, the usefulness of SVM and RF for protein function prediction is extensively tested by applying it to the classification of a variety of functionally distinguished classes of proteins. These include classical and non-classical secretory proteins, extracellular matrix proteins and subcellular location of apoptosis proteins. These classes of proteins are suitable for testing SVM and RF as they represent proteins of diversely different functions that cover protein synthesis regulation, regulation of host cell infection, protein self-association, molecular signaling, and drug discovery. Three different functionally diverse classes of proteins are explained below in detail.

3.2 Background of secretory proteins

After protein synthesis in cytoplasm, newly made polypeptides must be transported to their final destination in the cell. The process of protein transport to a particular cellular location is known as protein sorting (Palade, 1975; Walter *et al.,* 1984; Rothman and Wieland, 1996). Usually, eukaryotic protein secretion follows the classical secretory pathway that traverses the endoplasmic reticulum (ER) and golgi apparatus (Schatz and Dobberstein, 1996). Secretory proteins are usually characterized by short N-terminal signal peptides (14-60 amino acids) that have intrinsic signals for their transport and localization in the cell (Walter *et al.,* 1984; Heijne, 1990). Interestingly, several proteins have been found to be exported directly from the cytoplasm by molecular mechanisms that are independent from a signal peptide or any specific motif known to act as an export signal. The secretion of these proteins is referred to as non-classical or unconventional protein secretion (Müsch, 1990; Cleves, 1997; Hughes, 1999; Nickel, 2003). Some of the well-studied non-classical secretory proteins are fibroblast growth factors (FGF-1, FGF-2), interleukins (IL-1 alpha, IL-1 beta), galectins, thioredoxin, viral proteins and parasitic surface proteins potentially involved in the regulation of host cell infection (Mignatti and Rifkin, 1991; Rubartelli *et al.,* 1992; Mehul and Hughes, 1997; Denny *et al.,* 2000; Trotman *et al.,* 2003). Although the phenomenon of non-classical secretion in eukaryotes was discovered more than a decade ago, the molecular mechanisms are still unknown. However, it might be possible that this group contains proteins that leave the cell by cell disruption and not by a well-defined pathway.

Several methods have been proposed for the identification of secretory proteins that follow the classical secretory pathway (Bendtsen *et al.,* 2004a; Guda, 2006). The presence of the correct N-terminal end of the pre-protein is required for many of well-known prediction methods for correct classification. Sometimes in the large scale genome sequencing projects, the 5’ end of genes are assigned erroneously, and many proteins are annotated without the accurate N-terminal end which may lead to an incorrect subcellular localization annotation (Reinhardt and Hubbard, 1998). Further, signal peptides are completely absent in secretory proteins that follow non-classical secretion pathways. Therefore, an automated approach is required to predict classical and non-classical secretory proteins, irrespective of the N-terminal signal peptides.

Recently, a webserver SecretomeP has been developed to predict non-classically secreted proteins (Bendtsen *et al.*, 2004b). It is a neural network based method that uses several features of a protein such as the number of atoms, positively charged residues, propeptide cleavage sites, protein sorting, low complexity regions, and transmembrane helices as an input for a neural network. Despite considering a large number of protein features, the method has achieved a sensitivity of only 40% (Bendtsen *et al.*, 2004b). SRTPRED is another recently developed method which predicts secretory proteins irrespectively of N-terminal signal peptides. It achieves a sensitivity of 60.4% using hybrid modules (Garg and Raghava, 2008). In this work, we report a random forest method, SPRED, to identify classical and non-classical secretory proteins from protein sequence irrespective of N-terminal signal peptides. We scanned the entire human proteome by SPRED and predicted 566 proteins to be secreted by a non-classical secretory pathway.

3.3 Materials and Methods

3.3.1 Datasets

A set of 9890 extracellular mammalian proteins (positive dataset) were extracted from the UniProt database based on subcellular localization annotations in the comments block (Bairoch and Apweiler, 2000). The sequences without an experimental signal peptide annotation were not included in the dataset. Proteins with uncertain annotation labels such as ‘‘probable”, ‘‘potential” and ‘‘by similarity” were removed. 3131 extracellular proteins which are annotated with experimental observations were selected from the 9890 proteins. To make the dataset completely non-redundant, we applied the CD-HIT software (Li *et al.*, 2001) to remove sequences with greater than 40% sequence similarity to each other. Finally, 780 extracellular proteins were retained for the positive dataset. Similarly, a set of negative examples was constructed by extracting 20,610 mammalian proteins in UniProt which are annotated as residing in the cytoplasm and/or nucleus. 3891 proteins with experimental support were chosen from the 20,610 proteins after excluding membrane proteins, proteins with uncertain labels, and partial sequences. 1980 sequences remained for the negative dataset after removing redundant sequences which have >40% sequence similarity to each other using CD-HIT (Li *et al.*, 2001). Since non-classical secretory proteins lack N-terminal signal peptides, the method should have the capability to predict secretory proteins irrespective of N-terminal signal peptides. To achieve

this, we removed the signal peptides from the positive dataset. Finally, the training dataset consisted of 600 extracellular proteins that form the positive dataset and 600 cytoplasmic and/or nuclear proteins that form the negative dataset. The test dataset consisted of the remaining 180 extracellular proteins and 1380 cytoplasmic and/or nuclear proteins.

3.3.2 Human proteome screening

A human proteome database containing 86845 protein sequences was downloaded from the IPI database release 3.66 (http://www.ebi.ac.uk/IPI/) (Kersey *et al.,* 2004). Transmembrane proteins were removed using TMHMM (Krogh *et al.,* 2001). Finally, we obtained 65508 protein sequences for the prediction of novel putative classical or non-classical secretory proteins.

3.3.3 Features

In this work, each sequence was encoded by 119 features. The complete list is provided in Table 3.1.

Name of the feature	Number of features
Frequencies of 10 functional groups	10
Frequencies of hydrophobic, neutral, hydrophilic, positive, negative, polar and non-polar amino acids	7
Frequencies of secondary structurally elements (Helix, Strand and Coil)	3
Frequencies of 10 functional groups at Helix, Strand and Coil regions	30
Frequencies of hydrophobic, neutral, hydrophilic, positive, negative, polar and non-polar amino acids at Helix, Strand and Coil regions	21
Frequencies of short peptides rich in 10 functional groups	10
Frequencies of short peptides rich in hydrophobic, neutral, hydrophilic, positive, negative, polar and non-polar amino acids	7
Physicochemical properties	31
Total	119

Table 3.1: List of 119 features

3.3.4 Steps of the algorithm

1. Get the protein sequence data from the UniProt database.
2. Assign class labels: secretory proteins = +1 (positive class); non-secretory proteins = -1 (negative class).
3. Convert all the sequences to 119 features.
4. Get the top 50 features from Info Gain feature selection algorithm.
5. Partition the data into training and test sets.
6. Run the random forest classifier on the training set.
7. Run the random forest classifier on the test set to assess the generalization.
8. Screen the human proteome to find potential classical and non-classical secretory proteins.

3.4 Results and Discussion

3.4.1 Classification by SPRED

We trained our random forest model on the training dataset containing 600 extracellular proteins secreted via classical and non-classical pathways and 600 cytoplasmic and/or nuclear proteins. As shown in Table 3.2, on the training data, an overall prediction training accuracy of 85.67% was obtained using all features.

To identify the most prominent features, we carried out filter based feature selection methods: ReliefF, Info Gain, and mRMR. We selected five different feature subsets by decreasing the number of features, and the performance of each feature subset was evaluated (Table 3.2, Table 3.3, and Table 3.4).

In order to examine the performance of the newly developed model, we tested the trained model on a test dataset containing 180 extracellular proteins and 1380 cytoplasmic and/or nuclear proteins. As shown in Table 3.2, using the top 50 features, we obtained 82.18% accuracy with a sensitivity of 88.33% and a specificity of 81.38% (Info Gain).

The similar performance was obtained by ReliefF feature selection approach. This result suggests that our feature reduction approach selected useful features by eliminating uncorrelated and noisy features. The best performance was achieved with Info Gain selecting 50 features. Hence, this is chosen as the final model for screening human proteome to identify potential classical and non-classical secretory proteins.

Feature subset	**Sensitivity (%)**	**Specificity (%)**	**MCC**	**Test Accuracy (%)**	**Training Accuracy (%)**
10	79.44	80.51	0.4345	80.38	84.17
25	83.89	80.94	0.4691	81.28	85.58
50	**88.33**	**81.38**	**0.5036**	**82.18**	**85.92**
75	90.56	81.23	0.5163	82.31	85.58
100	89.44	81.16	0.5082	82.12	85.08
119	90.56	80.80	0.5109	81.92	85.67

MCC – Matthew's correlation coefficient

Table 3.2: Performance of random forest on the test dataset (180 positive and 1380 negative sequences) using different feature subsets (Info Gain)

Feature subset	**Sensitivity (%)**	**Specificity (%)**	**MCC**	**Test Accuracy (%)**	**Training Accuracy (%)**
10	74.44	83.84	0.4433	82.76	83.75
25	80.56	82.17	0.4624	81.99	85.84
50	88.33	81.23	0.5018	82.05	86.17
75	90.00	81.67	0.5182	82.63	85.69
100	88.89	81.23	0.5054	82.12	85.42
119	90.56	80.80	0.5109	81.92	85.67

MCC – Matthew's correlation coefficient

Table 3.3: Performance of random forest on the test dataset (180 positive and 1380 negative sequences) using different feature subsets (ReliefF)

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Test Accuracy (%)	Training Accuracy (%)
10	71.67	73.55	0.3106	73.33	76.50
25	78.89	74.64	0.3680	75.13	79.00
50	81.67	76.45	0.4042	77.05	82.25
75	83.89	78.55	0.4412	79.17	82.09
100	87.22	78.12	0.4581	79.17	83.50
119	90.56	80.80	0.5109	81.92	85.67

MCC – Matthew's correlation coefficient

Table 3.4: Performance of random forest on the test dataset (180 positive and 1380 negative sequences) using different feature subsets (mRMR)

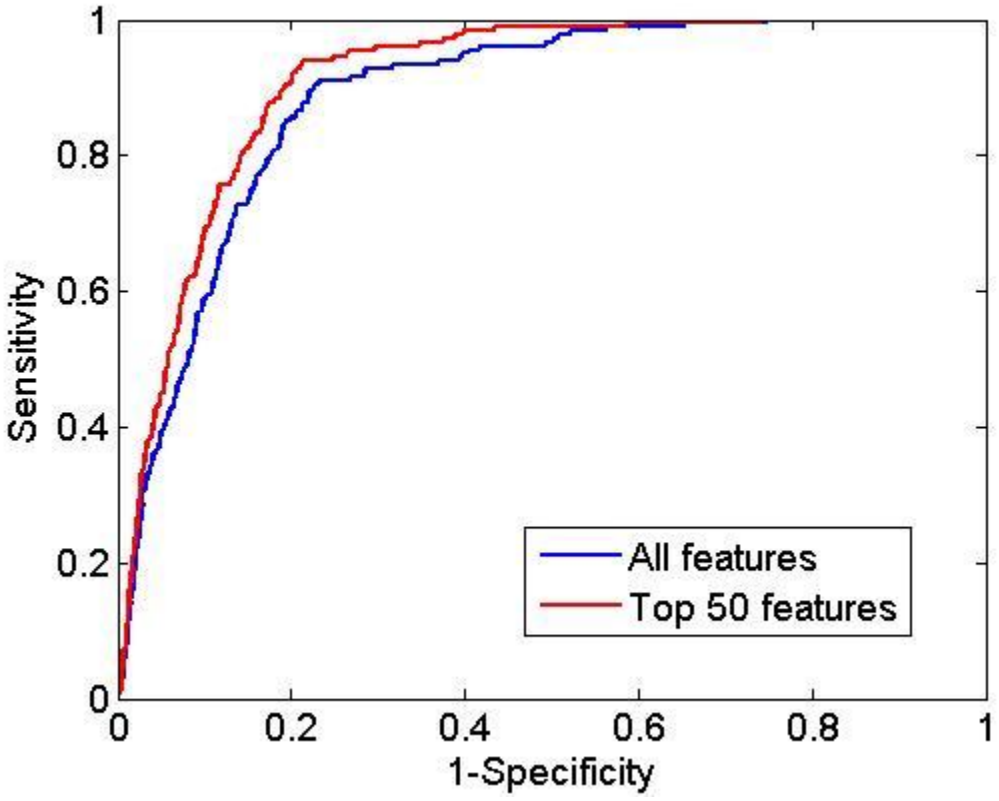


Figure 3.1: ROC Plot for random forest models using the top 50 features and all features (Info Gain)

We also plotted the sensitivity versus specificity chart, i.e. the receiver operator curve (ROC). The area under curve for all features was 0.89 and for the top 50 features (Info Gain) was 0.91, respectively (Figure 3.1).

3.4.2 Prediction result for known non-classical secretory proteins

For predicting non-classical secretory proteins, we did the following steps. First, SPRED tells us, whether the protein is secretory or non-secretory, and then we look whether the protein has a signal peptide or not. If not, we know that we have a non-classically secreted protein. As a final test we used 19 human proteins that are experimentally verified non-classical secretory proteins from various sources. Criteria for selection were clear experimental evidence from the literature for the given sequence entry. These secreted proteins without signal peptides are not found in any of the above datasets on which SPRED was trained or tested.

SWISS-PROT ID	Protein Annotation	SPRED	SecretomeP	SRTpred
P05230	Heparin-binding growth factor 1	+	+	+
P09038	Heparin-binding growth factor 2	+	+	+
P01584	Interleukin-1 beta	+	+	+
P01583	Interleukin-1 alpha	+	+	-
P17931	Galectin-3	+	+	-
P14174	Macrophage migration inhibitory factor	+	+	-
P26447	Protein S100-A4	+	+	-
P09211	Glutathione S-transferase P	+	+	-
Q06830	Peroxiredoxin-1	+	+	-
Q14116	Interleukin-18	+	+	-
P27797	Calreticulin	+	-	+
P62805	Histone H4	+	-	-
P29034	Protein S100-A2	+	-	-
P09382	Galectin-1	+	-	-
P10599	Thioredoxin	+	-	-
P26441	Ciliary neurotrophic factor	-	+	+
P19622	Homeobox protein engrailed-2	-	+	-
Q16762	Thiosulfate sulfurtransferase	-	+	-

P09429	High mobility group protein B1	-	-	-

Table 3.5: Prediction result for 19 experimentally verified non-classical secretory proteins using SPRED, SecretomeP and SRTpred. "+" denotes proteins correctly predicted as non-classical secretory proteins and "-" denotes proteins incorrectly predicted as non-classical secretory proteins

For comparison, we applied SPRED, SecretomeP (Bendtsen et al., 2004b) and SRTPRED (Garg and Raghava, 2008) to these 19 proteins. SPRED correctly predicted 15 proteins as non-classical secretory proteins whereas SecretomeP and SRTPRED predict 13 (with low score) and 5 proteins, respectively. The prediction results are given in Table 3.5.

3.4.3 Screening for classical and non-classical secretory proteins in the human proteome

To identify novel candidates in the human proteome for non-classical secretory proteins, we scanned the human proteome using SPRED (Figure 3.2). With SPRED, we classified these 65508 protein sequences into 44611 non-secreted proteins and 20897 proteins located outside of the nucleo-cytoplasm.

We removed all the classical secretory proteins (9542 protein sequences) using SignalP, leaving 11355 proteins which do not belong to the classical secretory pathway. Subsequently, we removed hypothetical proteins, fragmented proteins, mitochondrial proteins, peroxisomal proteins and false positive proteins. The remaining 566 protein sequences were finally classified as non-classical secretory proteins.

Our analysis shows that these 566 proteins include well studied non-classical secretory proteins such as Galectin (Hughes, 1999), Interleukin 1 alpha, Interleukin 1 beta (Nickel, 2003), thioredoxin (Ubartelli *et al.*, 1992), S100-A (Landriscina *et al.*, 2001), etc. which leave intact cells by defined pathways.

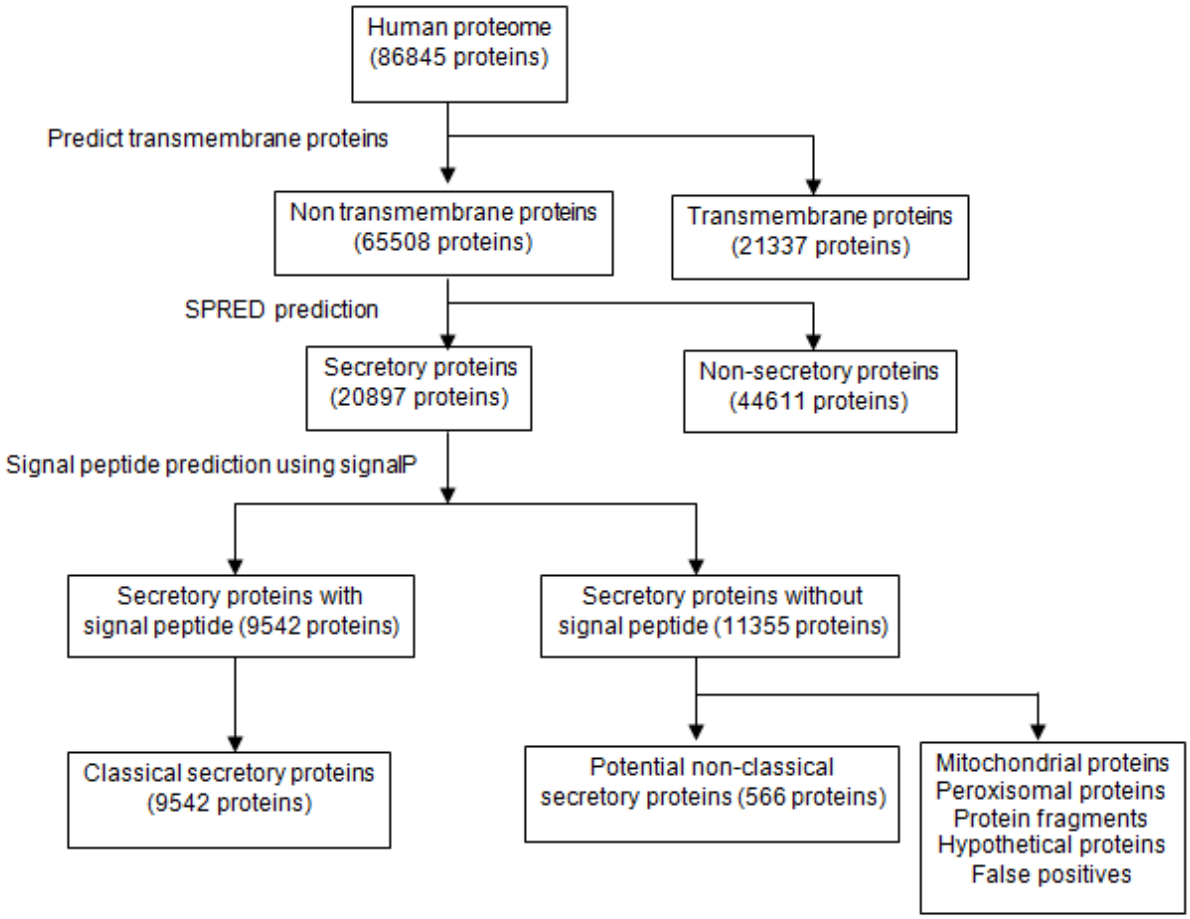


Figure 3.2: Screening for secretory proteins in human proteome

However, as the classification of proteins in the training dataset into the positive dataset ‘‘extracellular proteins” is often based on the detection of these proteins outside of cells without any knowledge about the export pathway, these predicted proteins may also include proteins that are released during cell disruption and are relatively stable in the extracellular environment.

3.4.4 Comparison of SPRED with other machine learning methods

The proposed SPRED method was compared with several state-of-the-art classifiers such as the naïve bayes classifier (George and Langley, 1995), instance learning based IBK algorithm (Aha and Kibler, 1991) and the support vector machine (linear and RBF kernel) (Vapnik, 1995). The

optimal values of the SVM parameters were obtained using a five-fold cross-validation on the training dataset. We compared the performance of SPRED with the other models using the same feature subsets that are mentioned in Table 3.2.

Method	Sensitivity (%)	Specificity (%)	MCC	Test Accuracy (%)
Naïve Bayes	70.00	78.28	0.2639	77.79
IBK	57.50	82.34	0.2344	80.88
SVM (Linear)	82.78	82.90	0.4867	82.88
SVM (RBF kernel)	78.89	80.87	0.4351	80.64
SPRED	**88.33**	**81.38**	**0.5036**	**82.18**

MCC – Matthew's correlation coefficient

Table 3.6: Comparison of SPRED with other machine learning methods using the top 50 features (Info Gain)

All models were tested on the test dataset containing 180 positive and 1380 negative sequences. With the top 50 features (Info Gain), SPRED and SVM (linear and RBF kernel) achieved comparable accuracy and specificity, however, the sensitivity of SPRED is still higher (Table 3.6).

3.4.5 Summary

Protein secretion is a universal process which occurs in all organisms and has tremendous importance to biological research. Identification of classical and non-classical proteins is an essential and also difficult task in protein function annotation. We implemented a random forest approach to predict protein secretion using sequence derived properties. The validation of SPRED on a test dataset showed 82.18% accuracy with a sensitivity of 88.33% and a specificity of 81.38%. SPRED performed better than SecretomeP and SRTPRED.

3.5 Background of Extracellular matrix proteins

The tissues of multicellular organisms are formed by cells and a network of macromolecules secreted by them, which is called extracellular matrix (ECM) (Lewin *et al.,* 2007). It consists of glycosaminoglycans, proteoglycans, fibrous proteins like collagenes, adhesive glycoproteins, enzymes involved in formation and remodelling of the ECM, like metalloproteases, and other factors (Lewin *et al.*, 2007). In the tissues, the ECM integrates the cells and provides structural support. In addition, it also influences the fate of cells during differentiation, morphogenesis, aging or pathogenesis (Schwartz *et al.,* 1995; Burridge and Chrzanowska-Wodnicka, 1996; Wary *et al.,* 1996). The ECM can coordinate cell functions by transducing signals across the plasma membrane. This can be achieved either directly by ECM molecules or indirectly by signal molecules, like growth factors, cytokines, chemokines, and hormones, which are sequestered in local depots within the ECM (Nelson and Bissell, 2006; Kim *et al.,* 2011). At first glance, the extracellular matrix seems to be a static structure with a slow turnover. However, it turned out that the ECM can easily adapt to changing conditions by a dynamic remodelling of its compounds (Green and Lund, 2005).

Malfunctions of ECM proteins lead to severe disorders that are linked to the structural functions of ECM molecules, such as the marfan syndrome, osteogenesis imperfecta, numerous chondrodysplasias, and skin diseases (Bruckner-Tuderman and Bruckner, 1998; Green and Lund, 2005; Aszódi *et al.,* 2006; Bateman *et al.,* 2009). Moreover, tumor growth, metastasis, inflammation, and other disorders can occur as a consequence of ECM malfunctions (Nelson and Bissell, 2006; Campbell *et al.,* 2010; Sorokin, 2010). Thus, extracellular matrix proteins promise great possibilities as therapeutic targets or diagnostic markers (Grønborg *et al.,* 2006).

Due to advances in sequencing technologies, tremendous amounts of DNA and protein sequences have accumulated in databases. Most of these sequences have unknown functions. It is very important to extract relevant biological information from sequences for functional annotation. Since the function of a protein is closely associated with its subcellular localization, the ability to predict the protein's subcellular localization will be useful in the characterization of the expressed sequences of unknown functions (Horton *et al.,* 2007; Chou and Shen, 2010a).

Various machine learning methods are available for predicting protein subcellular localization (Chou and Shen, 2007a, Chou and Shen, 2007b; Shen and Chou, 2009; Chou and Shen, 2010b). Protein subcellular localization prediction for human (Chou and Shen, 2006a), eukaryotes (Chou and Shen, 2006b), plants (Chou and Shen, 2006c), virus (Chou and Shen, 2006d) and gram negative bacteria (Chou and Shen, 2006e) have also been carried out. Several methods have been proposed for the identification of secretory proteins that follow the classical secretory pathway (Bendtsen *et al.,* 2004b) and non-classical secretory pathway (Kandaswamy *et al.,* 2010). Even though there are various tools available for predicting subcellular localization and protein secretion, there is no method with sufficient accuracy to predict ECM proteins among the secreted protein groups.

Recently, an in-silico model (ECMPP) has been developed to predict ECM proteins (Jung *et al.,* 2010). It uses SVM and RF to distinguish ECM proteins based on thirteen distinctive features. However, the performance of this method mainly depends on the PSSM profile, which needs sufficiently many sequence homologs to derive a sequence alignment. In this work, we present a random forest method, EcmPred, to identify extracellular matrix (ECM) proteins from sequence derived properties such as frequency of amino acid/amino acid groups and physico chemical properties. EcmPred achieves 83.00% and 77.52% accuracy on training and test data, respectively.

3.6 Materials and Methods

3.6.1 Datasets

We performed an extensive database and literature curation to collect sequences pertaining to extracellular matrix proteins. The dataset containing 17233 metazoan secreted protein sequences was obtained from SWISS-PROT release 67 (Boeckmann *et al.,* 2003). Out of these 17233 sequences, 1103 sequences are extracellular matrix proteins (positive dataset) and the remaining 16130 proteins are secreted proteins without extracellular matrix annotation (negative dataset). The positive and negative datasets were made completely non-redundant by allowing a sequence identity between any two proteins of not more than 70% (Li *et al.,* 2001). Finally, the training dataset consisted of 445 extracellular proteins that form the positive dataset and 4187 non-ECM proteins that form the negative dataset.

Training dataset:

300 ECM proteins were randomly selected from the 445 ECM proteins for the positive training dataset. Similarly, 300 non-ECM proteins were randomly taken from the 4187 non-ECM proteins for the negative training dataset.

Test dataset:

The remaining 145 ECM proteins served as positive dataset for testing. The remaining 3887 non-ECM proteins (after excluding 300 non-ECM proteins for training) were used as a negative dataset for testing.

Human proteome screening:

A human proteome database containing 86845 protein sequences was downloaded from the IPI database release 3.66 (http://www.ebi.ac.uk/IPI/) (Kersey *et al.,* 2004). Transmembrane proteins were removed using TMHMM (Krogh *et al.,* 2001). Finally, we obtained 65508 protein sequences for the computational screening and identification of novel ECM proteins.

3.6.2 Features

Each sequence is encoded by 68 sequence based features (frequency of 10 functional groups and physicochemical properties).

3.6.3 Steps of the algorithm

1. Get the metazoan secreted protein sequences from SWISS-PROT release 67.
2. Assign class labels: ECM proteins = +1 (positive class); non-ECM proteins = -1 (negative class).
3. Convert all the sequences to 68 features.
4. Get the top 40 features from mRMR feature selection algorithm.
5. Partition the data into training and test sets.
6. Run the random forest classifier on the training set.
7. Run the random forest classifier on the test set to assess the generalization.
8. Screen the human proteome to find potential ECM proteins.

3.7 Results and Discussion

3.7.1 Classification by EcmPred

We trained our random forest model on the training dataset containing 300 ECM proteins and 300 non-ECM proteins. Our model achieved 82% training accuracy using all the features (68 features). To identify the most prominent features, we carried out feature selection with mRMR, ReliefF and Info Gain. We selected six different feature subsets by decreasing the number of features, and the performance of each feature subset was evaluated. Using 40 features (mRMR), we obtained 83% training accuracy which is comparable to the accuracy obtained using 68 features. A similar performance was observed using 10, 20, 30, 50 and 60 features.

Feature subset	**Sensitivity (%)**	**Specificity (%)**	**MCC**	**Test Accuracy (%)**	**Training Accuracy (%)**
10	51.03	75.31	0.1123	74.44	73.00
20	48.97	77.63	0.1171	76.60	80.34
30	53.10	78.07	0.1378	77.17	81.84
40	**65.52**	**77.96**	**0.1906**	**77.52**	**83.00**
50	57.24	77.09	0.1493	76.38	82.67
60	60.69	77.40	0.1661	76.80	83.17
All features	63.45	76.24	0.1702	75.78	82.00

MCC – Matthew's correlation coefficient

Table 3.7: Performance of random forest using different feature subsets (mRMR)

In order to examine the performance of the newly developed model, we tested our training model on a test dataset containing 145 ECM proteins and 3887 non-ECM proteins. As shown in Table 3.7, we obtained 75.78% accuracy using all the features with a sensitivity of 63.45%, a specificity of 76.24%, and a MCC of 0.1702. Using 40 features, our model obtained 77.52% accuracy with 65.52% sensitivity, 77.96% specificity, and a MCC of 0.1906 (mRMR). Even though training accuracies of ReliefF and Info Gain (40 features) were better (Table 3.8, Table 3.9), the sensitivity and specificity values of mRMR was higher as compared to the other two approaches.

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Test Accuracy (%)	Training Accuracy (%)
10	68.97	69.79	0.1552	69.76	79.84
20	61.38	74.70	0.1520	74.22	81.84
30	61.38	75.21	0.1551	74.71	84.34
40	61.38	76.19	0.1612	75.66	84.67
50	62.76	75.62	0.1634	75.16	83.50
60	62.07	76.42	0.1655	75.90	82.67
All features	63.45	76.24	0.1702	75.78	82.00

MCC – Matthew's correlation coefficient

Table 3.8: Performance of random forest using different feature subsets (Info Gain)

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Test Accuracy (%)	Training Accuracy (%)
10	73.79	72.20	0.1878	72.26	82.00
20	66.21	77.42	0.1898	77.02	82.84
30	64.14	77.60	0.1822	77.12	84.34
40	59.31	76.99	0.1575	76.35	85.17
50	62.76	76.50	0.1690	76.00	84.84
60	62.07	76.73	0.1675	76.20	84.00
All features	63.45	76.24	0.1702	75.78	82.00

MCC – Matthew's correlation coefficient

Table 3.9: Performance of random forest using different feature subsets (ReliefF)

We also investigated the influence of the feature reduction by plotting Receiver Operating Characteristic (ROC) curves (Figure 3.3) derived from the sensitivity and specificity values for the classifiers using the top 40 features (mRMR) and all the features, respectively. The area under curve for all features was 0.76 and for the top 40 features was 0.79.

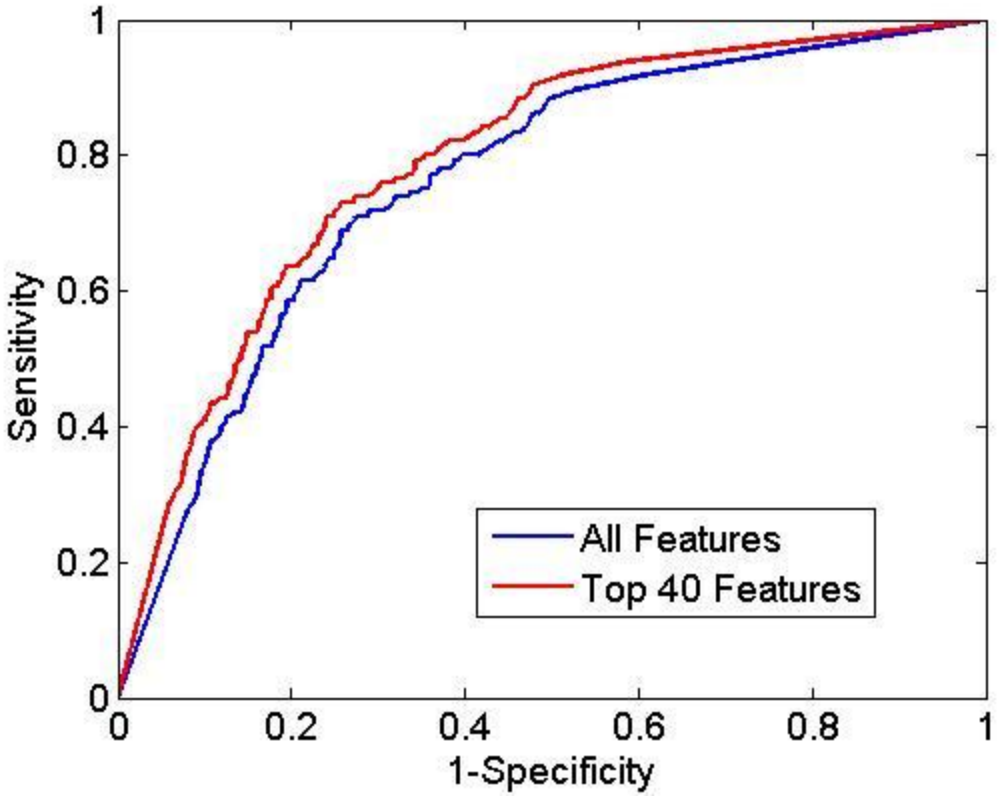


Figure 3.3: ROC plot for random forest with all and the top 40 features (mRMR)

3.7.2 Prediction result for known ECM proteins

We collected 20 experimentally verified extracellular matrix proteins from human. Criteria for selection were clear experimental evidence within the literature for the given sequence entry. We tested the efficiency of EcmPred and ECMPP (Jung *et al.*, 2010) using these 20 proteins (Table 3.10). As shown in Table 3.10, EcmPred correctly predicts 15 proteins as extracellular matrix proteins, whereas ECMPP predicts only 6 proteins.

SWISS-PROT ID	Protein Annotation	ECMPRED	ECMPP
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	+	-

Table 3.10: Prediction result for 20 experimentally verified extracellular matrix proteins using EcmPred and ECMPP. "+" represents proteins correctly predicted as extracellular matrix proteins and "-" represents proteins incorrectly predicted as extracellular matrix proteins

3.7.3 Screening for ECM in human proteome

To identify novel candidates in the human proteome as extracellular matrix proteins, we scanned the human proteome using SPRED (prediction of secretory protein) (Kandaswamy *et al.,* 2010) and EcmPred (Figure 3.4). With SPRED, we classified these 65508 protein sequences into 44611 non-secreted proteins and 20897 proteins located outside of the nucleo-cytoplasm. We predicted extracellular matrix proteins (6450) using EcmPred, leaving 14447 proteins which do not belong to the class of extracellular matrix proteins. Subsequently, we removed putative proteins, isoform sequences, hypothetical proteins, fragmented proteins and false positives. The remaining 2201 protein sequences were classified as extracellular matrix proteins.

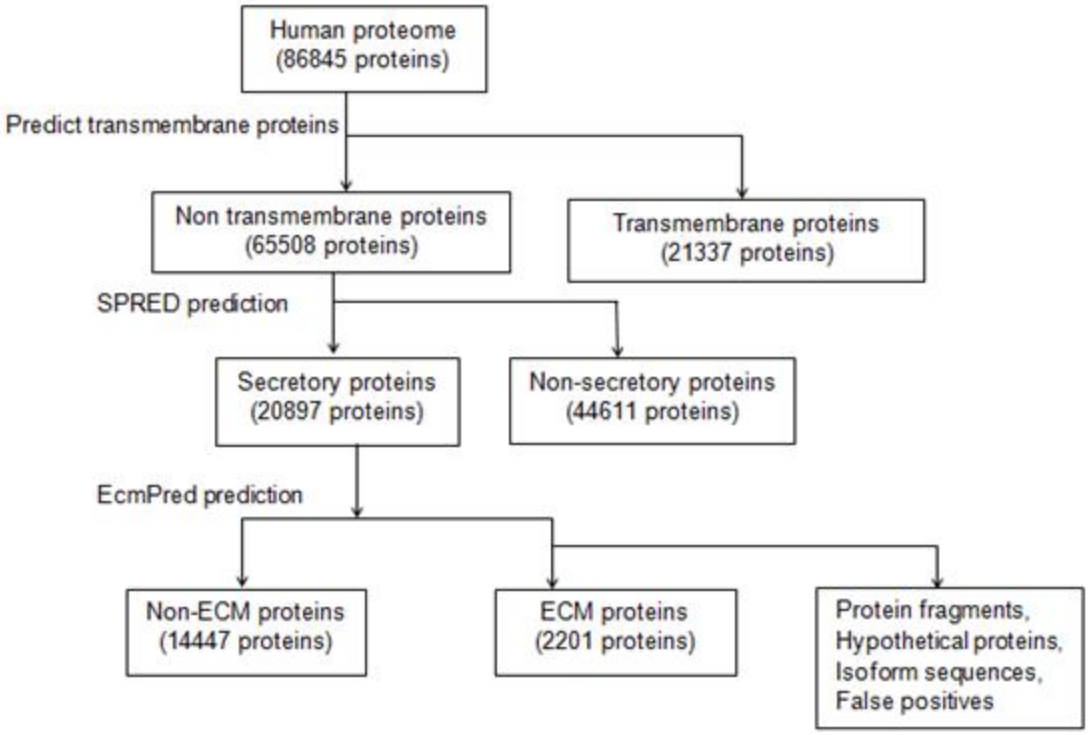


Figure 3.4: Screening for ECM proteins in the human proteome

We investigated the top listed putative ECM proteins using Interpro (Hunter *et al.*, 2009) and Gene ontology (GO) (Gene Ontology Consortium, 2010). Interpro annotation shows Collagen type XXI Alpha 1 and Adamts-like protein 2 as putative extracellular matrix proteins. Collagen, type V, alpha 1, Interphotoreceptor matrix proteoglycan 1, Protein Wnt, Galectin-1, and Galectin-7 were annotated with the Gene Ontology term ‘‘extracellular matrix’’. Thus, as could be expected by the composition of our training set we identified both, proteins forming the ECM network and more mobile proteins interacting transiently with the network.

3.7.4 Comparison of EcmPred with other machine learning methods

The proposed EcmPred method was compared with several state-of-the-art classifiers such as J4.8, SVM, Bayesnet, Logistic regression, Decision Table, Multi-Layer-Perceptron, and Adaboost (Quinlan, 1993; Bishop, 1995; Vapnik, 1995; Kohavi, 1995; Sumner *et al.*, 2005). The results based on 40 features (mRMR) are shown in Table 3.11.

Method	Sensitivity (%)	Specificity (%)	MCC	Test Accuracy (%)
J4.8	57.93	66.83	0.0973	66.51
Bayesnet	57.93	76.50	0.1485	75.83
Adaboost	59.63	69.66	0.1107	59.99
Decision table	54.95	68.97	0.0893	55.45
Logistic regression	59.31	65.62	0.0971	65.39
SVM	56.55	68.60	0.1001	68.17
MLP	58.63	68.97	0.1039	59.00
EcmPred	**65.52**	**77.96**	**0.1906**	**77.52**

MCC – Matthew's correlation coefficient

Table 3.11: Comparison of EcmPred with other machine learning methods

All models were tested on the test dataset containing 145 positive and 3887 negative sequences. The prediction accuracy of random forest is about 22% and 12% higher than Decision table and Logistic regression classifiers, respectively. The specificity of SVM is about 9% less than random forest. Although the performance of EcmPred and Bayesnet are comparable, sensitivity is 8% less than with our model.

3.7.5 Summary

The extracellular matrix (ECM) is a major component of tissues of multicellular organisms. It provides physical scaffolding for the cellular constituents and initiates critical biochemical and biomechanical signals required for tissue morphogenesis, differentiation, and homeostasis. The extracellular matrix proteins promise great possibilities as therapeutic targets or diagnostic markers. Identification of ECM proteins is vital for large scale genome annotation. We implemented a random forest approach to predict ECM proteins based on sequence derived properties. High prediction accuracies on the training and testing datasets show that EcmPred is a potentially useful tool for the prediction of extracellular matrix proteins from protein primary sequence. EcmPred performed better than ECMPP on experimental verified ECM proteins. The identification of ECM proteins will be helpful for the analysis of ECM-related functions and diseases.

3.8 Background of apoptosis protein subcellular locations

Apoptosis is an essential process for controlling tissue homeostasis, embryonic development and the immune system by regulating a physiological balance between cell proliferation and death (Jacobson *et al.,* 1997; Raff, 1998). Cell death and renewal are responsible for maintaining the proper turnover of cells, which ensures a constant controlled flux of fresh cells (Kerr *et al.,* 1972). Apoptosis can be triggered by internal or external signals and alteration in the subcellular localization of proteins may regulate it, e.g. Bcr-Abl sent to the nucleus causes apoptosis (Adams and Cory, 1998) and p53 dragged out of the nucleus is preventing it (Schulz *et al.,* 1999; Vogelstein *et al.,* 2000). Apoptosis entails not only protein degradation but also DNA fragmentation. As a result the destruction of a variety of cellular components occurs. Furthermore, the lipid composition of the plasma membrane changes and mitochondria become leaky. Programmed cell death and cell proliferation are tightly coupled (Jacobson *et al.,* 1997). A malfunction of apoptosis may cause or aggravate a variety of formidable diseases such as e.g. cancer, autoimmune diseases, ischemic damage, neurodegenerative diseases, and sepsis (Adams and Cory, 1998; Evan and Littlewood, 1998; Reed and Paternostro, 1999; Schulz *et al.,* 1999).

When observing apoptosis one can assign a defined subcellular localization of a series of otherwise shuttling proteins. This explains why regular sequence based predictors of a protein's subcellular location do not perform well with these proteins, as these predictors are not trained explicitly for this cellular condition. And it renders rather obvious that the comparison of the performance of a regular and a specialized predictor's results may be indicative of an active or passive involvement of a protein in the molecular physiology of apoptosis (Suzuki *et al.,* 2000; Dixon *et al.,* 2009).

Proteins contributing to apoptosis have been referred to as ‘apoptosis proteins’ (Zhou and Doctor, 2003). Various algorithms for protein subcellular localization prediction are available in the literature (Zhang *et al.,* 2006; Zhou *et al.,* 2007; Chou and Shen, 2007a). Covariant discriminant function (Zhou and Doctor, 2003), support vector machine (SVM) (Huang and Shi, 2005; Zhang *et al.,* 2006; Zhou *et al.,* 2007; Shi *et al.,* 2008), increment of diversity (ID) (Chen and Li, 2007a), increment of diversity combined with support vector machine (ID_SVM) (Chen and Li, 2007b) and fuzzy K-nearest neighbor (FKNN) (Ding and Zhang, 2008) have been

proposed to predict subcellular localization of apoptosis proteins based on various amino acid composition or pseudo amino acid composition. The pseudo amino acid composition (PseAAC) was first proposed by Chou to efficiently improve prediction of protein subcellular localization (Chou, 2001; Chou and Shen, 2007a; Chou, 2009).

In this work, we report a novel hybrid method that combines a genetic algorithm (GA) with a support vector machine (SVM) to predict the subcellular localization of apoptotic proteins on the basis of 119 sequence derived properties. A GA is used for feature selection to select a near-optimal subset of informative features that is most relevant for the classification. Jackknife cross-validation indicates the predictive capability of the proposed method on 317 apoptosis proteins. Our method achieved 89.91% accuracy for 25 features selected by the GA. Our models were examined by a test data of 98 apoptosis proteins. The predictive results of the proposed method has improved the predictive success rates, and therefore our current method plays an important role for the characterization of protein sequences of unknown proteins.

3.9 Material and Methods

3.9.1 Datasets

In our work, we used the dataset constructed by Chen and Li (Chen and Li, 2007a). The training dataset consisted of 317 apoptosis proteins divided into six subcellular locations: cytoplasmic proteins (112), mitochondrial proteins (34), nuclear proteins (52), secreted proteins (17), membrane proteins (55) and endoplasmic reticulum proteins (47). In addition, the 98 apoptosis proteins containing cytoplasmic proteins (43), plasma membrane-bound proteins (30), mitochondrial proteins (13) and other proteins (12) were also used to estimate the effectiveness of the method. The numbers in the brackets represent the total number of proteins in the respective class. To remove the homologous sequences from the benchmark dataset, a cut-off threshold of 25% was imposed in (Chou and Shen, 2008; Chou and Shen, 2010a) to exclude those proteins from the benchmark datasets that have equal to or greater than 25% sequence identity to any other in a same subset. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the numbers of proteins for some subsets would be too few to have statistical significance.

Some proteins can simultaneously exist at more than one location site. This kind of multiplex proteins may have special functions and hence are particularly interesting to both drug discovery

(Chou and Shen, 2008; Smith, 2008) and basic research (Chou and Shen, 2010a; Chou and Shen, 2010b; Chou and Shen, 2010c). However, the number of multiplex proteins in the existing apoptosis protein database is not large enough to allow us to construct a statistically meaningful benchmark dataset for studying multiplex nuclear proteins as done in (Chou and Shen, 2010a; Chou and Shen, 2010b) for the eukaryotic and plant protein systems. As a compromise, we studied the single-location apoptosis proteins.

3.9.2 Features

In this work, each sequence was encoded by 119 features. The complete list is provided in Table 3.1.

3.9.3 Multiclass SVM

Subcellular localization of the apoptotic proteins are divided into six classes. Hence, this becomes a multiclass prediction problem. Normally, a ‘‘One-against-one’’ or ‘‘One-against-all’’ approach is employed for multiclass SVM classifiers (Hsu and Lin, 2002). In the present study, the ‘‘One-against-one’’ approach was used. This method involves the construction of a binary SVM classifier corresponding to each pair of the classes. Hence, if there are K classes, a total of K (K-1)/2 classifiers will be constructed. Prediction of unseen test instances prediction follows the voting strategy. Predictions are made with each binary classifiers and the label is assigned to a class with maximum number of votes. In case of a tie, i.e. two classes have identical votes; the label assignment to the class is made on the basis of the smallest index. All the computations were performed using LIBSVM-2.81 (Chang and Lin, 2001). The various user-defined parameters, such as kernel parameter gamma (γ) and regularization parameter C were optimized on the training dataset.

3.9.4 Genetic Algorithm and Support Vector Machine (GASVM)

We report a novel hybrid method that combines the genetic algorithm (GA) and the support vector machine (SVM) approach to predict the subcellular localization of the apoptotic proteins using 119 sequence derived properties. A GA is used for selecting a near-optimal subset of informative features that are the most relevant for the classification. A hybrid GASVM system selects features from protein sequences and trains the SVM classifier simultaneously (Raymer *et al.,* 2000; Mohamad *et al.,* 2009). Figure 4.5 shows the model of the hybrid GASVM system,

where the feedback from the evaluation of the fitness function allows the GA to iteratively search for a feature subset that optimizes the fitness function value. Our aim was to optimize the following two objectives: minimization of the number of features used and maximization of the classification accuracy, which is a multi-objective optimization problem. The fitness function can be defined as:

$$\text{Fitness} = (w_1 * \text{LOOCV accuracy}) + \left(w_2 * \frac{(\text{TF}-\text{SF})}{\text{TF}}\right) * 100) \quad (3.1)$$

LOOCV- Leave One Out Cross Validation, TF- Total Features, SF- Selected Features, w_1, w_2 = weights given by the user, In our case $w_1 = 0.5$; $w_2 = 0.5$.

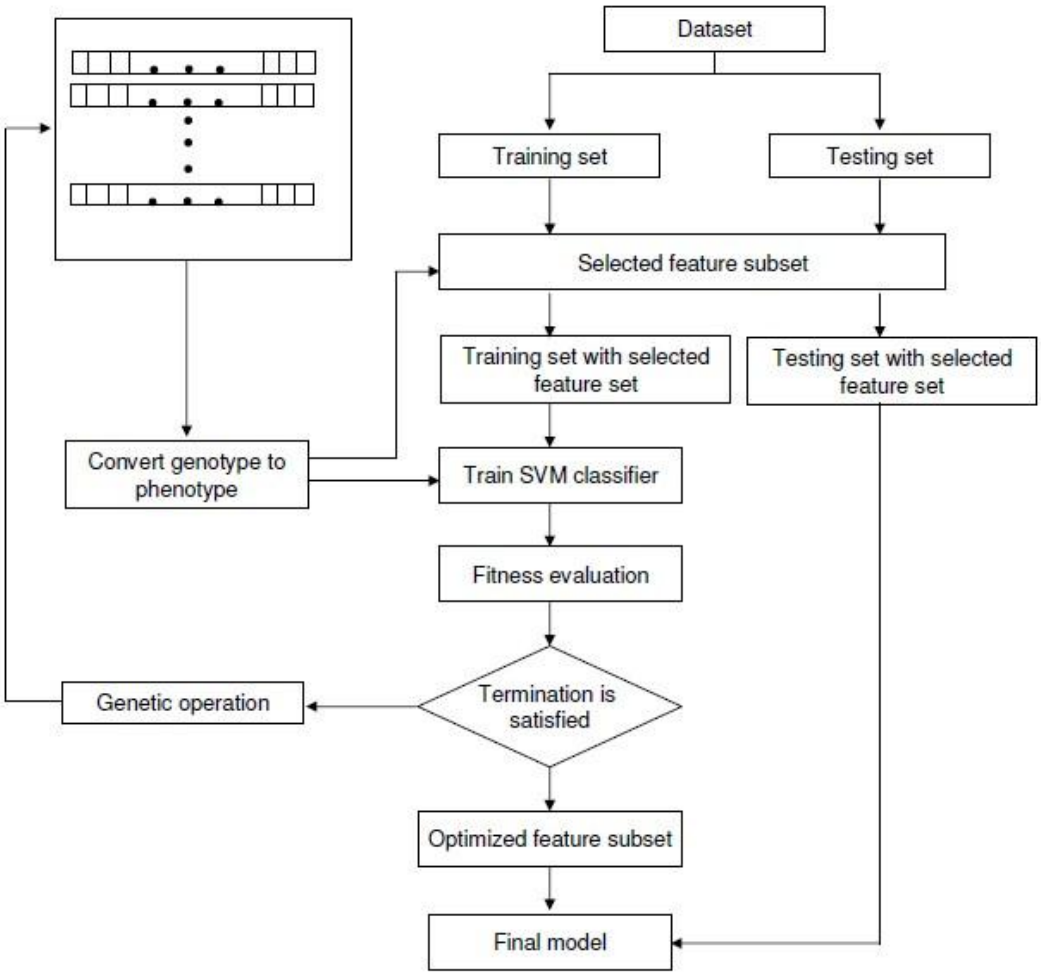


Figure 4.5: Architectures of the GA based feature selection for SVM

3.10 Results and Discussion

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent test dataset, sub-sampling test, and jackknife test. Of these three, the jackknife test is considered to be the most rigorous and objective one, and hence has been used frequently to determine the predictive power of various methods.

Jackknife cross-validation is applied to test the performance of the proposed method on 317 apoptosis proteins. Our method achieved 85.80% accuracy using all 119 features. Generally, not all features contribute equally to the classification; sometimes only few features play an important role in the classification model. In this work, we used filter and wrapper approaches to evaluate the performance of our model. The results are summarised in Table 3.12.

Feature Subset	**Jackknife test (%)**			
	GASVM	**InfogainSVM**	**ReliefFSVM**	**mRMRSVM**
10 Features	83.91	65.62	79.49	63.72
25 Features	**89.91**	79.81	88.95	72.87
50 Features	88.64	87.07	88.01	80.75
75 Features	89.59	**87.47**	**89.27**	86.11
100 Features	86.75	86.75	88.64	**89.27**
All Features	85.80	85.80	85.80	85.80

Table 3.12: The predictive results on the 317 apoptosis proteins using GASVM, InfogainSVM, ReliefFSVM and mRMRSVM (different feature subsets)

The filter approach is independent of the learning induction algorithm, computationally simple, fast and scalable. In this work, we have used Info Gain, ReliefF, and mRMR. Feature selection was performed by a five-fold cross-validation on the training dataset. Different models were built using the 10, 25, 50, 75 and 100 best features. The performance of the classification model is summarized in Table 3.12. It can be observed that all three approaches performed well. The feature selection generally does not deteriorate the classification performance much. With 75 features, we obtain a training accuracy of 89.27% (ReliefF). The individual accuracy for each class is summarized in Table 3.13. The success rate of the proposed approach (ReliefF) for cytoplasm is 90.17%, for membrane proteins 89.09%, for nuclear proteins 88.46%, for secreted

proteins 94.11%, for endoplasmic reticulum proteins 89.36%, and for mitochondria proteins 88.23%.

The wrapper approach uses the inductive algorithm to estimate goodness of a given feature subset. A GA is used for selecting a subset of useful features that is most appropriate for the classification. We developed five different models (10, 25, 50, 75 and 100 best features) using genetic algorithms as shown in Table 3.12. The training accuracy of 89.91% was obtained using 25 features. The success rate of the proposed approach (GASVM) for cytoplasm is 89.28%, for membrane proteins 92.72%, for nuclear proteins 86.53%, for secreted proteins 88.23%, for endoplasmic reticulum proteins 91.48% and for mitochondria proteins 91.17%.

Location	**Jackknife test (%)**			
	GASVM	**InfogainSVM**	**ReliefFSVM**	**mRMRSVM**
Cytoplasmic proteins	89.28	87.50	90.17	88.39
Mitochondrial proteins	91.17	88.23	88.23	85.29
Nuclear proteins	86.53	86.53	88.46	88.46
Secreted proteins	88.23	82.35	94.11	88.23
Membrane proteins	92.72	90.90	89.09	92.72
Endoplasmic reticulum proteins	91.48	89.36	89.36	91.48
Overall Accuracy	89.91	87.47	89.27	89.27

Table 3.13: Individual accuracies for each location using GASVM, InfogainSVM, ReliefFSVM and mRMRSVM (317 apoptosis proteins)

The overall accuracy of the proposed approach (GASVM) is 89.91% and about 2% higher than InfogainSVM and 0.5% higher than ReliefF and mRMR. We tested our model with a test dataset of 98 proteins (Table 3.15). Our model obtained an overall accuracy of 90.34%. The success rate of the proposed approach (GASVM) for cytoplasm is 90.70%, for membrane proteins 86.67%, for mitochondria proteins 92.31%, and for others 91.70%.

3.10.1 Comparison with other methods

We compared the performance of our method with the well-known methods in the literature. As shown in Table 3.13, GASVM achieved 89.91% accuracy for the 317 proteins in the jackknife test. We compared our results with WoLF PSORT which yielded an overall accuracy of 37.55%

on these 317 proteins. Table 3.14 shows the detailed comparison of our method with ID (Chen and Li, 2007a), ID_SVM (Chen and Li, 2007b), WoLF PSORT (Horton *et al.*, 2007) and IEPseAA (Shi *et al.*, 2007) on 317 proteins. We can observe that the overall accuracy of our method is higher than the other methods like ID, ID_SVM and IEPseAA. Moreover, in our method, the prediction accuracy for membrane proteins and mitochondria proteins are highest by 92.72% and 91.17%, respectively.

Model	**Sensitivity for each class (%)**						**Overall Accuracy (%)**
	Cyto	**Mito**	**Nucl**	**Secr**	**Memb**	**Endo**	
ID[a]	81.30	85.30	82.70	88.20	81.80	83.00	82.70
ID_SVM[b]	91.10	79.40	73.10	58.20	89.10	87.20	84.20
PSORT[c]	51.78	41.17	50.00	82.35	0.00	0.02	37.55
IEPseAA[d]	90.20	82.40	86.50	88.20	90.90	91.50	89.00
Our work	**89.28**	**91.17**	**86.53**	**88.23**	**92.72**	**91.48**	**89.91**

[a] Chen and Li, 2007a, [b]Chen and Li, 2007b, [c]Horton *et al.*, 2007, [d]Shi *et al.*, 2007

Cyto - Cytoplasmic proteins (112), Mito - Mitochondrial proteins (34), Nucl - Nuclear proteins (52), Secr - Secreted proteins (17), Memb - Membrane proteins (55) and Endo - Endoplasmic reticulum proteins (47)

Table 3.14: Prediction results with different models on 317 apoptosis proteins

Model	**Sensitivity for each class (%)**				**Test Accuracy (%)**
	Cyto	**Mito**	**PMemb**	**Other**	
Covariant[a]	97.70	30.80	73.30	25.00	72.50
Instab_SVM[b]	76.80	92.50	83.30	50.00	77.60
Dipep_Diver[c]	88.40	92.30	90.00	50.00	84.70
DWT_SVM [d]	95.40	53.90	93.30	91.70	88.80
Our work	**90.70**	**92.31**	**86.67**	**91.70**	**90.34**

[a]Zhou and Doctor, 2003, [b] Huang and Shi, 2005, [c]Chen and Li, 2004,[d]Qiu *et al.*, 2010

Cyto - Cytoplasmic proteins (43), PMemb - Plasma membrane-bound proteins (30), Mito - Mitochondrial proteins (13) and other - Other proteins (12)

Table 3.15: Prediction results with different models on 98 apoptosis proteins

The evaluation on the 98 proteins in test data confirms that the overall predicted successful rate of our model is higher than other methods. As shown in Table 3.15, the test accuracy obtained by the proposed approach is 90.34%, which is 17.84 % higher than the performance of covariant (Zhou and Doctor, 2003), 12.74% higher than Instab_SVM (Huang and Shi, 2005), 5.64% higher than Dipep_Diver (Chen and Li, 2004) and 1.54% higher than the result of DWT_SVM (Qiu *et al.*, 2009).

3.10.2 Summary

The comparison with different approaches on different datasets indicates that our method is effective and useful for predicting the subcellular localization of proteins during apoptosis. We hope that the encouraging results using novel features will improve the performance of protein subcellular location prediction of apoptosis proteins. This work will contribute to the understanding of the molecular physiology of apoptosis and with the hindsight of the associated diseases possibly contribute to the identification of targets for diagnostics or prognostic markers and therapeutic intervention.

3.11 Conclusion

The rapidly growing number of sequenced genomes needs an effective and reliable way of classifying the protein sequences into functional classes. In this work, we have analyzed and compared the accuracy of various protein classification and feature selection methods to classify an extremely diverged class of proteins (classical and non-classical secretory proteins, extracellular matrix proteins, and subcellular location of apoptosis proteins). We have not tested genetic algorithm for the prediction of secretory proteins and extracellular matrix proteins because genetic algorithm takes a long time to find an acceptable solution. Testing results on several diverged functional classes suggests that SVM and RF seems to be potentially useful tools for protein function prediction by means of classification of proteins into specific functional classes. We identified potential non-classical secretory proteins and extracellular matrix proteins by scanning the human proteome. Our models will contribute to functional annotation and drug discovery.

4 Protein Family Classification

4.1 Introduction

Protein family classification has several benefits as a basic approach for large-scale genomic annotation: (1) it helps the annotation of proteins, which is quite challenging to illustrate by pairwise sequence alignments (2) it assist an error free annotation and maintaining family based databases from various resources (3) it aids to recover substantial biological information from massive amounts of data (4) it exposes the essential gene families which is essential for the comparative genomics studies (Wu *et al.,* 2003).

So far a number of different classification methods have been developed to organize proteins. Among the variety of classification schemes are: (1) families/ superfamilies (Barker *et al.,* 1996) in the PIR-PSD (2) protein domain families : Pfam (Finn *et al.,* 2010) and ProDom (Bru *et al.,* 2005) (3) sequence motifs: PROSITE (Hulo, *et al.*, 2006) and PRINTS (Attwood *et al.,* 2002) (4) structural classes: SCOP (Andreeva, *et al.,* 2008) and CATH (Greene *et al.,* 2007) (5) combinations of various family classifications: iProClass (Wu *et al.,* 2004) and InterPro (McDowall and Hunter, 2011).

Although many protein families have a well conserved tertiary structure, their sequence identity is very low and in some cases falls within the twilight zone (< 25% sequence similarity) (Doolittle, 1986; Pearson, 1997). Assigning sequences to the respective family by sequence search methods is risky when the pairwise identity is below 25%. In above cases, alignment based methods will recognize the proteins incorrectly. With the rapid increase in newly found protein sequences entering into databanks, an efficient method is needed to identify protein families from the sequence databases.

Some proteins may not have adequate sequence similarities although they share similar structures and biochemical functions. Identification of antifreeze and bioluminescent proteins from protein sequence is more interesting due to the low pairwise sequence similarity which often falls below the twilight zone. So far, no specific method has been reported to identify protein families (antifreeze and bioluminescent) from primary sequence. In this work, we have

developed machine learning method to annotate hypothetical proteins of antifreeze and bioluminescent families.

4.2 Background of antifreeze proteins

The surrounding environment plays a key role in the survival of living organisms. Extremely cold temperature causes intracellular ice formation which is considered to be lethal to the cell. Initially, it was thought that the coldest regions like Antarctica are uninhabitable due to extremely cold temperature which is lower than the freezing point of body fluids. In 1957, Scholander *et al*. observed that certain fish species were able to survive in the conditions where the temperature is lower than the freezing point of their body fluids (Scholander *et al.,* 1957). Later it was reported that some overwintering plants such as Silenea caulis and Carex firma can survive at temperatures of less than -50^{o}C (Sakai and Larcher, 1987; Moriyama *et al.,* 1995; Yoshida *et al.,* 1997). These findings suggest that these organisms and plants have special antifreeze mechanisms to protect themselves against freezing stress. This antifreeze activity makes the organisms less sensitive to cold temperatures. Previous studies reported that the antifreeze effect is due to a group of proteins called ‘‘antifreeze proteins’’ (AFPs) (Davies and Sykes, 1997; Logsdon and Doolittle, 1997; Cheng, 1998; Ewart *et al.,* 1999).

Antifreeze proteins have the capacity to adsorb onto the surface of ice crystals. The interaction between AFPs and ice crystals has significant effects on the overall growth of ice (Davies *et al.,* 2002). Firstly, AFPs inhibit ice crystal growth and lower the freezing temperature of the water without altering the melting point. This process creates a difference between the freezing temperature and melting point which is known as thermal hysteresis (Urrutia *et al.,* 1992). Each antifreeze protein has its own characteristic values for thermal hysteresis. Secondly, AFPs obstruct the recrystallization of ice, which includes the growth of larger ice crystals instead of smaller ice crystals (Yu and Griffith, 2001). Larger ice crystals increase the possibility of physical damage within frozen plant tissues (Griffith *et al.,* 1997). Finally, AFPs also have the ability to interact with ice nucleators, which may result in either the inhibition or the enhancement of ice nucleation activity. Overwintering plants and animals adopt two strategies namely freeze tolerance and freeze avoidance to survive at low and subzero temperatures (Lewitt, 1980; Sformo *et al.,* 2009). Freeze tolerance involves the activation or synthesis of ice-

nucleating agents (INAs) on winter in freeze-tolerant species whereas freeze avoidance involves the inactivation or removal of ice-nucleating agents in freeze-avoiding species.

AFPs have been discovered in various fish, insects, bacteria, fungi and overwintering plants including ferns, gymnosperms, monocotyledonous, dicotyledonous, angiosperms, etc. (Scholander *et al.,* 1957; Urrutia *et al.,* 1992; Moriyama *et al.,* 1995; Davies and Sykes, 1997; Logsdon and Doolittle, 1997; Cheng, 1998; Ewart *et al.,* 1999; Yu and Griffith, 2001; Davies *et al.,* 2002). Analyses of AFPs from fish, insects and plants have shown that there is no consensus sequence or structure for an ice-binding domain. Some AFPs undergo structural changes at low temperatures (Davies *et al.,* 2002). One explanation for AFP diversity is that ice can present many different surfaces with different geometric arrangements of oxygen atoms (Davies *et al.,* 2002). The ice binding domains and their interaction with ice varies from species to species. For example, the ice binding domains of fish and insect AFPs are relatively hydrophobic and their adsorption onto ice is a hydrophobic interaction whereas plant antifreeze proteins have multiple, hydrophilic ice binding domains (Davies *et al.,* 2002).

In fish, AFPs are classified into five known types namely AFGPs, AFP I, AFP II, AFP III and AFP IV (Davies and Hew, 1990; Chou, 1992; Davies *et al.,* 2002). AFGPs are made up of 4 to more than 50 tandem repeats of Ala-Ala-Thr with a disaccharide attached to each Thr OH. It has an amphipathic polyproline type II helix fold. Type I AFPs are made up of alanine-rich, amphipathic helices. Type II AFPs are globular proteins with mixed secondary structure. Type III AFPs are made up of short beta-strands and one helix turn that gives it a unique flat-faced globular fold. Type IV AFPs are helix-bundle protein. Insect AFPs shows a beta helical structure (Graether *et al.,* 2000). So far, crystal structure is not available for plant AFPs.

AFPs have potential industrial, medical, biotechnological and agricultural application in different fields, such as food technology, preservation of cell lines, organs, cryosurgery and freeze-resistant transgenic plants and animals (Griffith and Ewart, 1995; Breton *et al.,* 2000). Identification of novel AFPs is important in understanding protein-ice interactions and also in creating novel ice-binding domains in other proteins. With the fast growth of protein sequences in various databases, the need for an automated and accurate tool to recognize AFP becomes increasingly important.

Encouraged by the overwhelming success of machine learning methods in engineering, medical and financial applications, many research groups have been using Neural networks, Support vector machines, KNN, Random forests and other machine learning algorithms in the biological field, especially in the classification and prediction of protein structures and functional profiles (Chou, 2001, 2005; Chou and Cai, 2005; Anand *et al.,* 2008; Chou and Shen, 2009; Huang *et al.*, 2009; Qiu *et al.,* 2009). So far, bioinformatics and statistical learning methods like SVM and RF have not been explored for the prediction of antifreeze proteins. In this work, we report a random forest approach to identify antifreeze proteins from sequence information, irrespective of the sequence similarity.

4.3 Materials and Methods

4.3.1 Datasets

We obtained 221 antifreeze protein sequences from seed proteins of the Pfam database (Sonnhammer *et al.,* 1997). To enrich the dataset, we performed PSI-BLAST search for each sequence against non-redundant sequence database with stringent threshold (E value - 0.001) (Altschul *et al.,* 1997). Each sequence was subjected to manual inspection to retain only antifreeze proteins. Proteins with incomplete sequences were excluded. The final positive dataset contained 481 non-redundant antifreeze proteins. The negative dataset was constructed from 9493 seed proteins (representative members) of Pfam protein families, which are unrelated to antifreeze proteins (Sonnhammer *et al.,* 1997). The sequences with >=40% sequence similarity were removed from the dataset using CD-HIT (Li *et al.,* 2001).

Training dataset:

300 antifreeze domains were randomly selected from 481 antifreeze proteins for the positive dataset. Similarly, 300 non-antifreeze proteins were randomly taken from 9493 non-antifreeze proteins for the negative dataset.

Test dataset:

The remaining 181 antifreeze proteins domains served as a positive dataset for testing. The remaining 9193 non-antifreeze proteins (after excluding 300 non-antifreeze proteins that were used for training) were used as a negative dataset.

4.3.2 Features

In this work, each sequence is encoded by 119 features (Table 3.1).

4.3.3 Steps of the algorithm

1. Get the protein sequence data from the Pfam database.
2. Assign class labels: antifreeze proteins = +1 (positive class); non-antifreeze proteins = -1 (negative class).
3. Encode each sequence into 119 features.
4. Get the top 25 features from ReliefF feature selection algorithim.
5. Partition the data into training and test sets.
6. Run the random forest classifier on the training set.
7. Run the random forest classifier on the test set to assess the generalization.
8. Annotate the hypothetical proteins using AFP-Pred.

4.4 Results and Discussion

4.4.1 Prediction using PSI-BLAST

PSI-BLAST is an widely used pair wise sequence search tool for recognizing homologous sequences (Altschul *et al.,* 1997). The performance of PSI-BLAST was evaluated using jackknife cross validation, where each sequence in the positive dataset (481 antifreeze proteins) was used as a BLAST query sequence and remaining sequences (480 antifreeze proteins) were used as a BLAST database. Three iterations of PSI-BLAST were carried out at E value - 0.001. It was observed that only 280 antifreeze proteins showed similarity (BLAST hit) with other antifreeze proteins (E value - 0.001) and no hits were obtained for the remaining 201 AFPs. The result suggests that pair wise sequence similarity methods alone may not be the good choice for the annotation of antifreeze proteins. Therefore, we decided to explore machine learning methods to predict AFPs from sequence derived features such as frequency of amino acid groups, secondary structural elements, and physiochemical properties, etc.

4.4.2 Prediction of antifreeze proteins by AFP-Pred

In this work, we report a random forest method for the prediction of antifreeze proteins from protein sequence using 119 sequence derived properties. We trained our random forest model on the dataset containing 300 antifreeze proteins and 300 non-antifreeze proteins. AFP-Pred achieved 82% training accuracy using all the features. In order to examine the performance of the newly developed model, we tested our training models on a dataset containing 181 antifreeze proteins and 9193 non-antifreeze proteins. In this work, we have used three different filter approaches namely ReliefF, Info Gain, and mRMR, to build an in-silico machine learning model. The number of features was stepwise reduced from 119 to 10 features. Table 4.1 shows the performance of our model on the test dataset using different feature subsets (10 to 119 features).

As seen in Table 4.1, feature selection improves the classification accuracy until the number of features decreases to 25. Using all features, our model achieved 83.38% accuracy with 84.67% sensitivity, 82.32% specificity and MCC of 0.6674. The prediction accuracy was slightly improved when the features were reduced from 119 to 25. Using 25 features (ReliefF), AFP-Pred obtained 84.29% accuracy with 84.67% sensitivity, 83.98% specificity, and MCC of 0.6846. With Info Gain and mRMR (Table 4.2, Table 4.3), the sensitivity and specificity values were less as compared to ReliefF. The results suggest that the ReliefF feature reduction approach selected useful features by eliminating the uncorrelated and noisy features.

Feature subset	**Sensitivity (%)**	**Specificity (%)**	**MCC**	**Test Accuracy (%)**	**Training Accuracy (%)**
10 Features	80.00	80.66	0.6052	80.36	80.17
25 Features	**84.67**	**83.98**	**0.6846**	**84.29**	**83.84**
50 Features	86.00	81.77	0.6749	83.69	83.34
75 Features	84.00	82.87	0.6667	83.38	83.83
100 Features	84.00	81.77	0.6553	82.78	80.67
All Features	84.67	82.32	0.6674	83.38	82.00

MCC – Matthew's correlation coefficient

Table 4.1: Performance of random forest on test data containing 181 AFPs and 9193 non-AFPs using different feature subsets (ReliefF)

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Test Accuracy (%)	Training Accuracy (%)
10 Features	74.03	81.22	0.5539	77.62	82.00
25 Features	74.59	80.11	0.5478	77.35	80.84
50 Features	72.38	76.80	0.4922	74.59	83.50
75 Features	73.48	81.22	0.5486	77.35	82.34
100 Features	72.38	82.87	0.5556	77.62	83.50
All Features	84.67	82.32	0.6674	83.38	82.00

MCC – Matthew's correlation coefficient

Table 4.2: Performance of random forest on test data containing 181 AFPs and 9193 non-AFPs using different feature subsets (Info Gain)

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Test Accuracy (%)	Training Accuracy (%)
10 Features	75.14	80.66	0.5589	77.90	81.50
25 Features	73.48	79.56	0.5314	76.52	81.00
50 Features	72.93	78.45	0.5146	75.69	83.17
75 Features	74.03	80.66	0.5482	77.35	82.34
100 Features	71.27	85.08	0.5690	78.18	83.34
All Features	84.67	82.32	0.6674	83.38	82.00

MCC – Matthew's correlation coefficient

Table 4.3: Performance of random forest on test data containing 181 AFPs and 9193 non-AFPs using different feature subsets (mRMR)

We also investigated the influence of the feature reduction by plotting Receiver Operating Characteristic (ROC) curves (Figure 4.1) derived from the sensitivity (true positive rate) and specificity (false positive rate) values for the classifiers using all features and the 25 best performing features (ReliefF), respectively. The area under curve for all features was 0.87 and for the top 25 features was 0.89.

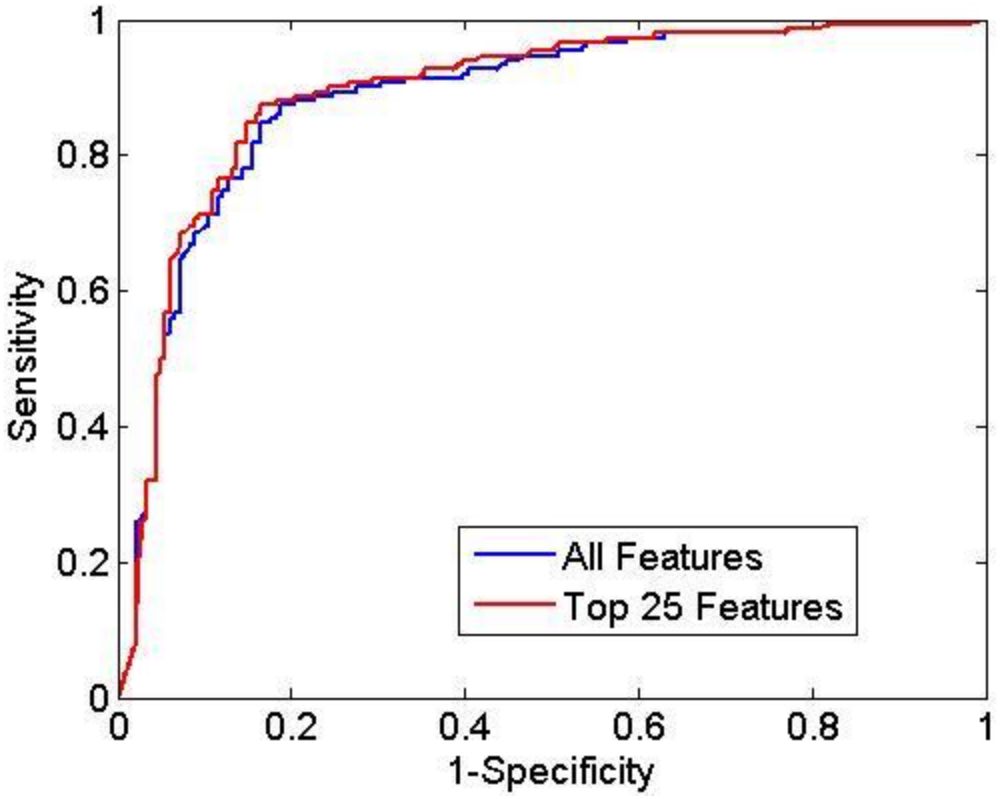


Figure 4.1: ROC Plot for random forest using all features and top 25 features **(**ReliefF**)**

4.4.3 Performance of AFP-Pred, BLAST and HMM

In the next step we evaluated our algorithm with an independent dataset obtained from INTERPRO and KEGG databases (Kanehisa and Goto, 2000; Hunter *et al.,* 2009). The sequences that are present in the positive training dataset were removed from the list. Finally, we got 16 proteins which are annotated as ‘‘antifreeze proteins’’(database annotation) (Table 4.4). Our approach correctly predicted 15 proteins as antifreeze proteins. The performance of our algorithm was compared with PSI-BLAST and HMM (Altschul *et al.,* 1997; Eddy, 1998).

PSI-BLAST search for each sequence was carried out against the SWISS-PROT database with an E value of 0.1. HMM search for each query sequence was performed against the HMM profile obtained from Pfam database (Pfam release 23) (Sonnhammer *et al.,* 1997). Out of 16 proteins, BLAST search retrieved antifreeze protein hits from SWISS-PROT database for only 9 proteins. No hits were found for the remaining 7 proteins. Similarly, HMM search against Pfam database returned no hits for 11 proteins. As seen in Table 4.4, AFP-Pred, BLAST and HMM predicted

15, 9 and 5 proteins, respectively. This result indicates that AFP-Pred is a useful approach to predict AFPs from sequence information in the absence of sequence similarity. Out of 16 proteins, 3 proteins are annotated as ‘‘anti-freeze proteins’’ (NCBI definition) and the remaining 14 proteins are annotated as ‘‘unnamed protein product’’ or ‘‘hypothetical proteins’’ in NCBI database. AFP-Pred correctly predicted all the hypothetical proteins as antifreeze proteins. This shows that AFP-Pred can be efficiently used to annotate hypothetical proteins.

GI Code	**AFP-Pred**	**BLAST**	**HMM**	**Source of annotation**	**NCBI definition**
26325086	---	AFP	AFP	INTERPRO	unnamed protein product
26344193	AFP	AFP	---	INTERPRO	unnamed protein product
74221639	AFP	AFP	---	INTERPRO	unnamed protein product
12843602	AFP	---	---	INTERPRO	unnamed protein product
257049854	AFP	---	---	KEGG	hypothetical protein
30249105	AFP	AFP	AFP	INTERPRO	Type I antifreeze protein
226941159	AFP	AFP	AFP	INTERPRO	Type I antifreeze protein
126464034	AFP	---	---	KEGG	Type I antifreeze protein
45435722	AFP	AFP	---	INTERPRO	hypothetical protein
281341260	AFP	AFP	AFP	INTERPRO	hypothetical protein
2315605	AFP	---	---	INTERPRO	hypothetical protein
260817607	AFP	AFP	AFP	INTERPRO	hypothetical protein
26388908	AFP	---	---	INTERPRO	unnamed protein product
26348120	AFP	---	---	INTERPRO	unnamed protein product

26333557	AFP	---	---	INTERPRO	unnamed protein product
26332695	AFP	AFP	---	INTERPRO	unnamed protein product

AFP - Antifreeze proteins

Table 4.4: Prediction result for 16 potential antifreeze proteins

4.4.4 Comparison with other machine learning methods

The proposed random forest method was compared with several state-of-the-art classifiers such as SVM, Naïve Bayes, MLP, and the K-nearest neighbor classifier (Aha and Kibler, 1991; George and Langley, 1995; Vapnik, 1995). We compared the performance of AFP-Pred with the other models using the same feature subsets (top 25 features from ReliefF) (Table 4.5). All models were tested on the test dataset containing 181 positive and 9193 negative sequences. The prediction accuracy of random forest is about 7% and 6% higher than Naïve Bayes and K-nearest neighbor classifiers (IBK), respectively. Although the performance of random forest, SVM and MLP is comparable, there is a slight improvement in the sensitivity and specificity values of random forest. This result shows that AFP-Pred can be used to predict antifreeze proteins with higher accuracy.

Method	**Sensitivity (%)**	**Specificity (%)**	**MCC**	**Test Accuracy (%)**
Naïve Bayes	66.60	84.53	0.5233	76.44
MLP	80.00	80.66	0.6052	80.36
IBK	78.67	75.69	0.5413	77.04
SVM	82.67	80.11	0.6254	81.27
AFP-Pred	**84.67**	**82.32**	**0.6674**	**83.38**

MCC – Matthew's correlation coefficient

Table 4.5: Comparison of AFP-Pred with other machine learning methods

4.4.5 Summary

Identification of antifreeze proteins from sequence databases is difficult due to poor sequence similarity. We reported a random forest based approach, AFP-Pred, for the prediction of antifreeze proteins from sequence using sequence derived properties. Very high prediction

accuracies on the training and testing datasets show that AFP-Pred is a potentially useful tool for the prediction of antifreeze from protein primary sequence. Because of its simplicity, this approach can be easily extended to recognizing other specific families and functions and should be a useful tool for the high-throughput and large-scale analysis of proteomic and genomic data.

4.5 Background of bioluminescent proteins

Bioluminescence is an enchanting process in which light is produced by a chemical reaction within an organism (Hastings, 1995; Wilson, 1995). Bioluminescence is found in various organisms like ctenophora, bacteria, certain annelids, fungi, fish, insects, algae, squid, etc. (Lloyd, 1978; Hastings, 1995; Haddock *et al.,* 2010). The bioluminescence mechanism involves two chemicals, namely luciferin, a substrate, and the enzyme luciferase (White *et al.,* 1971; Wilson, 1995). The oxidation of luciferin is catalyzed by the enzyme luciferase, resulting in light and an intermediate called oxyluciferin. Occasionally, the luciferin catalyzing protein and oxygen (co factor) are bound together to form a single unit called photoprotein. This molecule is triggered to produce light when a particular type of ion is added to the system. The proportionality of the light emission makes a clear distinction between a photoprotein and a luciferase (White *et al.,* 1971). Photoproteins are capable of emitting light in proportion to the amount of the catalyzing protein, but in luciferase-catalyzed reactions, the amount of light emitted is proportional to the concentration of the substrate luciferins (Hastings, 1995).

Different creatures produce different colors of light, from violet through red (Wilson and Hastings, 1998; Haddock *et al.,* 2010). The different colors of light produced are often dependent on the roles the light plays, the organism in which it is produced, and the varieties of chemicals produced.

Bioluminescence serves a diversity of functions, but many of those functions are not known. The known functions include camouflage, finding food, attraction of prey, attraction of mates, repulsion by way of confusion, signaling other members of their species, confusing potential predators, communication between bioluminescent bacteria (quorum sensing), illumination of prey, burglar alarm, etc. (Lloyd, 1978; Hastings, 1995; Haddock *et al.,* 2010).

The application of bioluminescence promises great possibilities for medical and commercial advances. Bioluminescent proteins serve as invaluable biochemical tools with applications in a variety of fields including gene expression analysis, drug discovery, the study of protein dynamics and mapping signal transduction pathways, bioluminescent imaging, toxicity determination, DNA sequencing studies, estimating metal ions such as calcium, etc. (Cormier *et al.,* 1975; Chalfie *et al.,* 1994; Gonzalez and Negulescu, 1998; Kain, 1999; Ward *et al.,* 2000; Contag and Bachmann, 2002; DiPilato *et al.,* 2004; Hayes *et al.,* 2004).

The detailed analysis of bioluminescence proteins helps to understand many of the functions which are still unknown and also helps to design new medical and commercial applications. Due to advances in sequencing technologies, huge amount of data is available in various databases (Schuster, 2008). Despite tremendous progress in the annotation of protein, there are no existing online tools available for the prediction of bioluminescent proteins using primary protein sequences.

A support vector machine is a supervised learning algorithm, which has been found to be useful in the recognition and discrimination of hidden patterns in complex datasets (Zhang *et al.,* 2006). SVM has been successfully applied in various fields of computational biology, e.g., protein sequence/structure analysis, micro-array and gene expression analysis (Zhang *et al.,* 2006; Zhou, *et al.,* 2007).

So far, bioinformatics and statistical learning methods like support vector machine and random forest have not been explored for the prediction of bioluminescent proteins. In this work, we present a novel prediction method that uses a support vector machine and physicochemical properties to predict bioluminescent proteins.

4.6 Materials and Methods

4.6.1 Datasets

We obtained 300 bioluminescent proteins from seed proteins of the Pfam database (Sonnhammer *et al.,* 1997). To enrich the dataset, we performed PSI-BLAST search against non-redundant sequence database with stringent threshold (E value - 0.01) (Altschul *et al.,* 1997). Redundant sequences that have >=40% sequence similarity were removed from the dataset using CD-HIT

(Li *et al.*, 2001). After careful manual examination, a total of 441 bioluminescent proteins were selected for the positive dataset.

Training dataset: 300 bioluminescent proteins were selected from 441 bioluminescent proteins for the positive dataset. 300 non-bioluminescent proteins for the negative set were randomly taken from seed proteins of Pfam protein families, which were unrelated to bioluminescent proteins.

Test dataset: The remaining 141 bioluminescent proteins served as a positive dataset for testing. The negative dataset was created from the seed proteins of non-bioluminescent proteins, which are selected from seed proteins of non-bioluminescent Pfam protein families (Sonnhammer *et al.*, 1997). The negative sequences present in the training dataset were removed. Furthermore, non-bioluminescent protein domains with less than 40 amino acids were excluded from the negative set. Finally, the test dataset consisted of 141 bioluminescent proteins and 18202 non-bioluminescent proteins.

4.6.2 Features

In this work, each sequence is encoded by 554 features (physicochemical properties).

4.6.3 Steps of the algorithm

1. Get the protein sequence data from the Pfam database.
2. Assign class labels: bioluminescent proteins = +1 (positive class); non-bioluminescent proteins = -1 (negative class).
3. Convert all the sequences to numerical equivalents based on physicochemical properties.
4. Get the top 100 features from ReliefF feature selection algorithm.
5. Partition the data into training and test sets.
6. Run the SVM classifier on the training set.
7. Run the SVM based classifier on the test dataset to assess the performance of the classifier.
8. Annotate the hypothetical proteins using BLProt.

4.7 Results and Discussion

4.7.1 Performance of similarity based search using PSI-BLAST

Similarity search methods play a vital role in the classification of proteins. PSI-BLAST is the most popular similarity based search method for searching sequence databases (Altschul *et al.*, 1997). PSI-BLAST search for each query sequence was performed against the database of 441 bioluminescent proteins that were used for the training and testing. PSI-BLAST was carried out at an E value of 0.001 with three iterations. It was observed that 280 bioluminescent proteins showed similarity (BLAST hit) with other bioluminescent protein sequences (E value - 0.001). The performance of the sequence similarity method drops when there is no significant sequence similarity between two proteins. Hence, such an alignment-based method would rarely yield satisfactory predictions. Therefore, there is a need for alignment-free methods (machine learning models) for predicting bioluminescent proteins.

4.7.2 Prediction of bioluminescent proteins by BLProt

A SVM classifier was applied to predict bioluminescent proteins. Each sequence was encoded by 554 features. The model was trained with a dataset containing 300 bioluminescent protein sequences and 300 non-bioluminescent protein sequences.

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Test Accuracy (%)	Training Accuracy (%)
75 features	69.50	77.13	0.4663	73.86	77.16
100 features	**74.47**	**84.21**	**0.5904**	**80.06**	**80.00**
200 features	68.09	81.58	0.5022	75.83	78.00
300 features	67.38	82.11	0.5017	75.83	78.67
400 features	64.54	86.32	0.5260	77.04	78.00
500 features	65.96	85.79	0.5323	77.34	78.00
All features	63.12	78.19	0.4182	71.73	75.16

MCC – Matthew's correlation coefficient

Table 4.6: Performance of the SVM using different feature subsets selected by ReliefF

BLProt achieved 75.16% training accuracy (5 fold cross-validations) with all of the 544 physicochemical properties (Table 4.6).

To identify the most prominent features, we carried out feature selection with three different filter approaches, ReliefF, Info Gain, and mRMR. We selected five different feature subsets by decreasing the number of features, and the performance of each feature subset was evaluated (Table 4.6, Table 4.7, and Table 4.8). The best performance of 80% training accuracy was achieved with ReliefF selecting 100 features. Hence, this is chosen as the final model for our work.

Feature subset	**Sensitivity (%)**	**Specificity (%)**	**MCC**	**Test Accuracy (%)**	**Training Accuracy (%)**
100 features	69.50	74.21	0.4351	72.21	74.83
200 features	76.60	75.79	0.5193	76.13	78.00
300 features	70.92	77.37	0.4821	74.62	78.33
400 features	68.09	77.89	0.4611	73.72	78.17
500 features	68.09	84.21	0.5326	77.34	78.33
All features	63.12	78.19	0.4182	71.73	75.16

MCC – Matthew's correlation coefficient

Table 4.7: Performance of the SVM using different feature subsets selected by Info Gain

Feature subset	**Sensitivity (%)**	**Specificity (%)**	**MCC**	**Test Accuracy (%)**	**Training Accuracy (%)**
100 features	65.96	84.21	0.5134	76.44	78.33
200 features	65.25	84.74	0.5132	76.44	78.5
300 features	65.96	83.68	0.5072	76.13	78.5
400 features	65.96	83.68	0.5072	76.13	78.33
500 features	65.96	83.68	0.5072	76.13	78.5
All features	63.12	78.19	0.4182	71.73	75.16

MCC – Matthew's correlation coefficient

Table 4.8: Performance of the SVM using different feature subsets selected by mRMR

After training, we tested our algorithm on the test dataset consisting of 141 bioluminescent protein sequences and 18202 non-bioluminescent proteins sequences. The maximum accuracy of 80.06% with 74.47 % sensitivity and 84.21% specificity was obtained using the top 100 features (ReliefF, Table 4.6).

Figure 4.2 presents a chart with the true positive rates and false positive rates on the test data at different thresholds for the classifiers using all the features and the top 100 features, respectively (ReliefF). The area under curve for all features was 0.79 and for the top 100 features was 0.87, respectively.

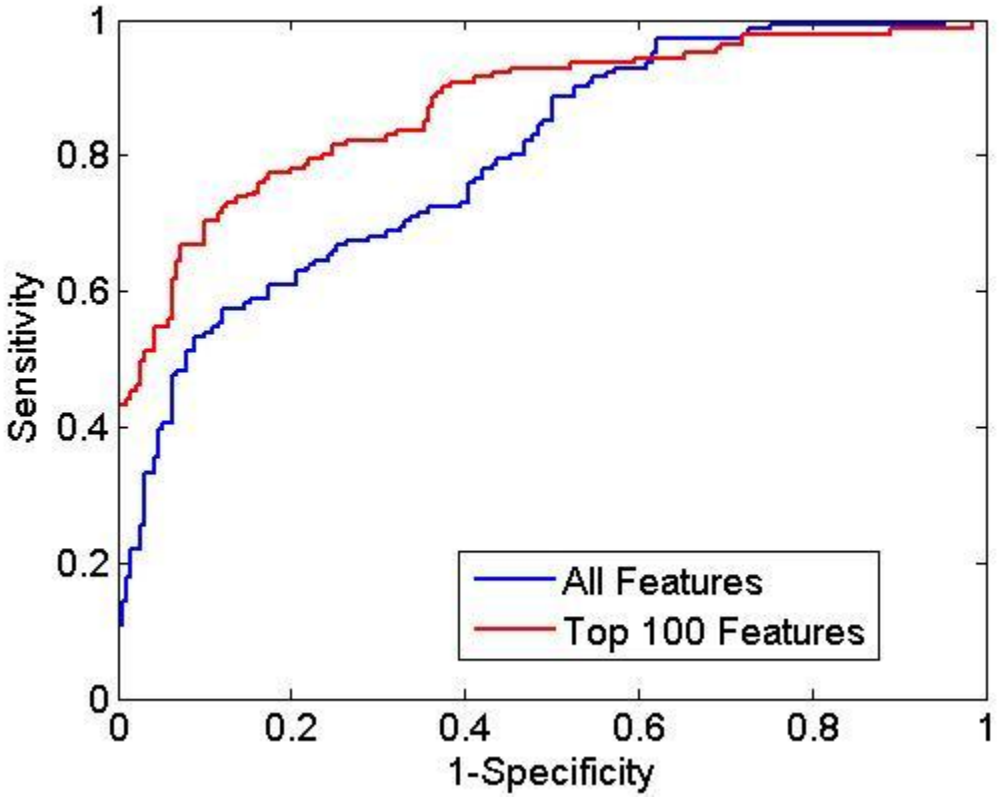


Figure 4.2: ROC Plot for SVM models using all and the top 100 features (ReliefF)

4.7.3 Comparison of BLProt with HMM and BLAST

The performance of BLProt was compared with other sequence search methods, namely HMM and PSI-BLAST using 141 bioluminescent proteins (Altschul *et al.,* 1997; Eddy, 1998). PSI-BLAST search for each sequence was carried out against the SWISS-PROT database with an E value of 0.1. HMM search for each query sequence was performed against the HMM profile obtained from the Pfam database (Pfam release 23) (Sonnhammer *et al.,* 1997). Out of 141 proteins, 114 proteins were correctly predicted by BLProt. PSI-BLAST and HMM correctly predicted 99 and 76 proteins, respectively.

Our algorithm was further evaluated by 9 hypothetical proteins obtained from the INTERPRO, CDD and KEGG databases (Kanehisa and Goto, 2000; Hunter *et al.*, 2009; Marchler-Bauer, *et al.*, 2011) (Table 4.9). Our approach correctly predicted all proteins as bioluminescent proteins. The performance of our algorithm was compared with PSI-BLAST and HMM (Altschul *et al.*, 1997; Eddy, 1998). Out of these 9 proteins, the PSI-BLAST search retrieved bioluminescent protein hits from the SWISS-PROT database for only 4 proteins. No hits were found for the remaining 5 proteins. Similarly, HMM search against the Pfam database returned no hits for 3 proteins. This result indicates that BLProt is a useful approach for predicting bioluminescent proteins from sequence information in the absence of sequence similarity.

GI	BLProt	PSI-BLAST	HMM	Source of annotation
156529049	BLP	Non-BLP	BLP	INTERPRO
37528019	BLP	BLP	Non-BLP	KEGG
37528018	BLP	BLP	BLP	CDD
45440453	BLP	Non-BLP	BLP	INTERPRO
45440453	BLP	Non-BLP	BLP	INTERPRO
153796564	BLP	Non-BLP	Non-BLP	INTERPRO
49257059	BLP	BLP	BLP	CDD
159576911	BLP	BLP	Non-BLP	CDD
49257059	BLP	Non-BLP	BLP	INTERPRO

BLP - Bioluminescent protein; Non-BLP - Non-bioluminescent protein; CDD - Conserved Domain Database

Table 4.9: Prediction result for 9 potential bioluminescent proteins

4.7.4 Comparison with other machine learning methods

The proposed SVM model was compared with several state-of-the-art classifiers such as J4.8, PART, Random forest, Adaboost and IBK (Aha and Kibler, 1991; Quinlan, 1993; Freund and Schapire, 1996; Frank and Witten, 1998; Breiman, 2001). We compared the performance of BLProt with the other approaches using the same feature subset (top 100 features from ReliefF). All models were tested on the test dataset containing 141 positive and 18202 negative sequences. The performance of different classifier on test dataset is shown in Table 4.10. The prediction accuracy of BLProt is about 7% and 12% higher than that of J4.8 and PART, respectively. Although the sensitivity of BLProt, random forest and IBK is comparable, BLProt is superior in specificity and concerning the MCC values.

Method	Sensitivity (%)	Specificity (%)	MCC	Test Accuracy (%)
J4.8	69.50	75.79	0.4518	73.11
PART	63.12	72.11	0.3519	68.28
IBK	76.60	69.47	0.4556	72.51
Random Forest	75.18	73.16	0.4787	74.02
AdaBoost	68.79	72.63	0.4117	71.00
BLProt	**74.47**	**84.21**	**0.5904**	**80.06**

MCC – Matthew's correlation coefficient

Table 4.10: Comparison of BLProt with other machine learning methods

4.7.5 Summary

Bioluminescence is a process in which light is emitted by a living organism. It is an important protein family which has wide medical and commercial values. In this study, we developed a method for predicting bioluminescent proteins from its primary sequence using ReliefF coupled with SVM. BLProt will help the experimental biologist to predict bioluminescence from a protein sequence and thus, help to avoid unnecessary experiments.

4.8 Conclusion

The classification of proteins into families is useful because it can suggest potential function of a unknown proteins. Support vector machine and random forest were used for classification of protein families with weak sequence similarities. We have developed machine learning method to annotate hypothetical proteins of antifreeze and bioluminescent families. Our tools, thus constitutes a fundamental bioinformatics resource for biologists who contemplate using bioinformatics as an integral approach to their genomic/proteomic research and scientific inquiries.

5 Experimental validations of predicted candidate proteins for post-translational translocation into the ER-Membranes

5.1 Introduction

In eukaryotic cells, where several cellular compartments have evolved to carry out specialized functions, the correct localization of their resident proteins is essential for cell viability. After protein synthesis in cytoplasm, newly made polypeptides must be transported to their final destination in the cell. The process of protein transport to a particular cellular location is known as protein sorting (Rothman, 1996). For secreted proteins or those that have to be transported into compartments along the secretory pathway, the journey starts with the translocation of the protein into the endoplasmic reticulum (ER) membrane (Palade, 1975; Walter *et al.,* 1984). Many experiments showed that translocation of protein are usually directed by "postal code" like targeting signals encoded within the amino acid sequences (Blobel and Sabatini, 1971).

The entry to the general secretory pathway is controlled by the signal sequence, an N-terminal part of the polypeptide chain, which is cleaved off while the protein is translocated through the ER-membrane. Signal sequences are short peptide chains of a length of 14-60 amino acids. They share certain common structural features: a net positive charge in the N-terminus, a hydrophobic core and a polar cleavage site (von Heijne, 1985).

5.1.1 Co-translational translocation of proteins into the ER

The translocation of secretory proteins across the ER-membrane can occur co or post-translationally. Co-translational translocation in eukaryotes is dependent on a cytoplasmic protein-RNA complex called SRP (Signal Recognition Particle). The mammalian SRP comprises six polypeptides (SRP9, SRP14, SRP19, SRP54, SRP68, and SRP72), and one RNA (7SL RNA/SRP RNA) (Walter and Blobel, 1980; Walter and Blobel, 1982).

The translation of the secretory protein initiates on a free ribosome in the cytoplasm, but as soon as the signal peptide emerges from the ribosome, it binds to SRP, which prevents folding of the nascent polypeptide chain and arrests the elongation step of translation (Schatz and Dobberstein,

1996). This elongation arrest is mediated by the Alu domain of eukaryotic SRP and is necessary for correct coupling of protein translation and translocation (Mason *et al.*, 2000).

SRP directs the ribosome complex, including mRNA and nascent protein, to the ER membrane by interacting with its membrane receptor. At the membrane the translation resumes. Thus, the remaining translation takes place on membrane-bound ribosomes, while the protein is translocated across the membrane (Figure 5.1) (Rapoport, 1990; Mothes *et al.*, 1997; Rapoport, 2007).

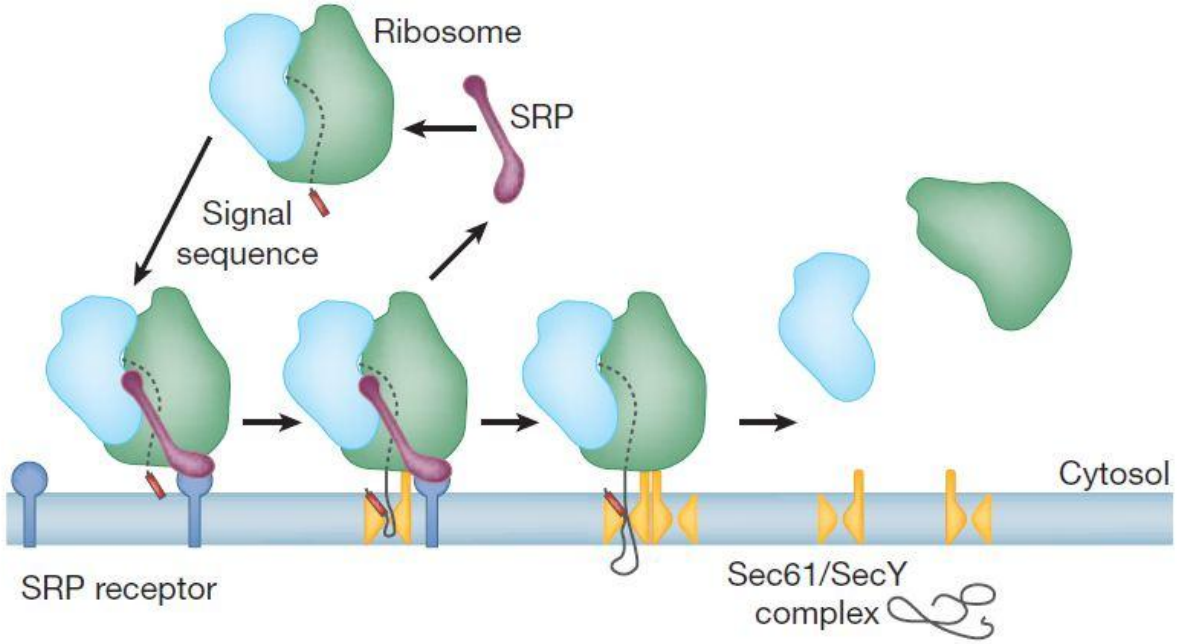


Figure 5.1: Model of co-translational translocation (Rapoport, 2007)

The actual protein-conducting channel in the ER-membrane is formed by the Sec61 complex. This membrane protein complex consists of 3 subunits: Sec61α, Sec61β and Sec61γ. The large α- subunit was discovered to be a homolog of Sec61p of Saccharomyces cerevisiae which was found earlier in genetic screens for mutants defective in translocation (Deshaies and Schekman, 1987). The β-and γ-subunits of the mammalian Sec6l complex are small membrane proteins, which are anchored in the membrane by C-terminal hydrophobic tails.

5.1.2 Post-translational translocation of proteins into the ER

Post-translational translocation pathway has been extensively studied in the yeast Saccharomyces cerevisiae, where most proteins are post-translationally translocated (Hann and Walter, 1991). In post-translational translocation, a protein does not interact with SRP during the protein synthesis. Although it has been shown that cytosolic chaperones (TRiC, HSP70) interact with the translocation substrate and keep it in a translocation competent conformation, little is known about the targeting phase in the post-translational translocation pathway. At the membrane the translocation substrate is inserted into the protein-conducting channel, which contains beside the trimeric Sec61 complex the proteins Sec62, Sec63, Sec71 and Sec72 (Panzner *et al.*, 1994).

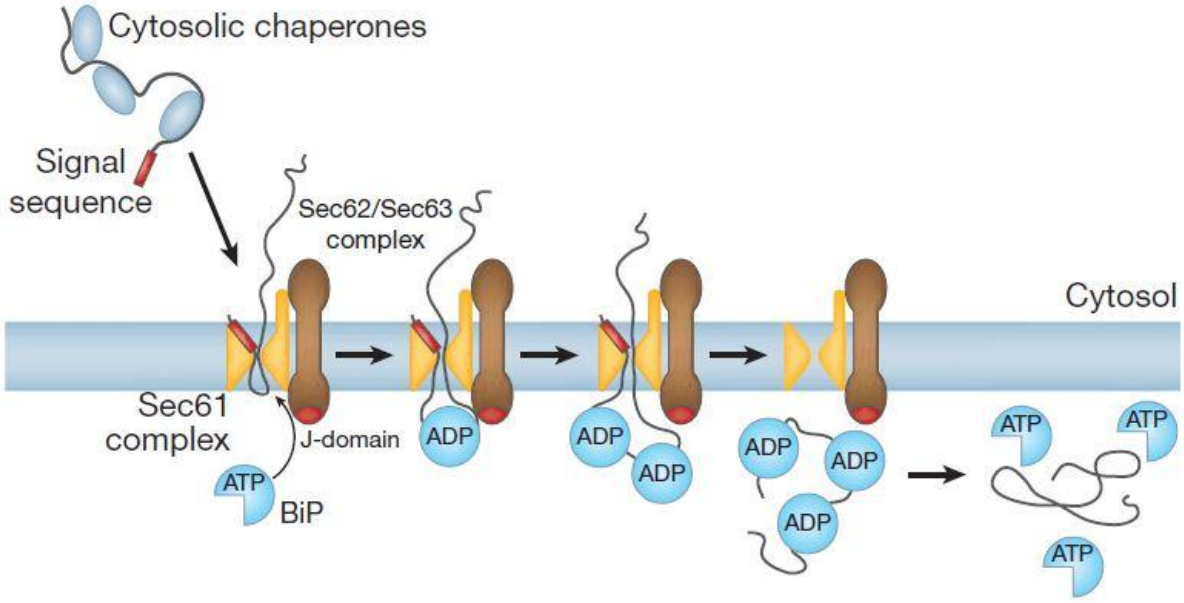


Figure 5.2: Model of post-translational translocation (Rapoport, 2007)

The actual translocation of the protein through the membrane occurs by a ratcheting mechanism and involves the luminal protein Bip, a member of the HSP70 family of ATPases (Figure 5.2) (Liebermeister *et al.*, 2000; Rapoport, 2007; Gouridis *et al.*, 2009). ATP is hydrolyzed and the peptide-binding pocket of Bip closes around the translocation substrate. The polypeptide chain in the channel can slide in either direction by Brownian motion, but its binding to Bip inside the lumen of the endoplasmic reticulum prevents movement back into the cytosol, resulting in net forward direction. The next Bip molecule binds with polypeptide chain, after it moved adequately and then the complete process is repeated until the entire polypeptide chain is

translocated. Finally, the Bip molecule is released from the polypeptide, when peptide-binding pocket opens (nucleotide exchange of ADP for ATP) (Rapoport, 2007).

5.1.3 Aim of the study

Little information is available about post-translational protein translocation into the ER of mammals. So far, only non mammalian substrate proteins were analyzed. Moreover, Sec71 and Sec72, which are part of the post-translational translocation channel in yeast, do not exist in mammals. Currently, only few mammalian substrate proteins (less than 90 amino acid length) were identified experimentally for post-translational protein translocation. Therefore, there is a need of computational methods to address this issue. We are interested to develop a database to list the potential co and post-translocation mammalian substrate proteins. Further, we intend to develop machine learning method (feature selection and classification) to predict potential post-translational substrate proteins. Finally, top ranked candidate proteins should be tested for their translocation behavior. Therefore, these proteins should be analyzed in in vitro translation/translocation assays suitable for co or post-translational protein translocation into ER-membranes, respectively.

5.2 Materials and Methods

5.2.1 Training and test dataset

A set of 6890 proteins (Drosophila melanogaster, Mus musculus, Homo sapiens, Xenopus laevis) were extracted from the UniProt database based on sequence annotations (signal peptide) (Bairoch and Apweiler, 2000). The secreted protein sequences with length less than 90 amino acids were taken as post-translational pathway proteins (positive dataset). Similarly, the secreted protein sequences with length more than 150 amino acids were taken as co-translational pathway proteins (negative dataset). We removed the signal peptides from the positive and negative dataset and kept mature part to make the dataset completely non-redundant. We applied the CD-HIT software (Li *et al.,* 2001) to remove sequences with greater than 40% sequence similarity to each other. We added the signal peptide to the corresponding mature part of positive and negative sequences. Finally, the training dataset consisted of 134 post-translational pathway proteins that form the positive dataset and 200 co-translational pathway proteins that form the

negative dataset. The test dataset consisted of the remaining 10 post-translational pathway proteins and 3843 co-translational pathway proteins.

5.2.2 Features

Each sequence was encoded by 50 features (pseudo amino acid composition) (Chou, 2001).

5.2.3 cDNA clones

3 cDNA clones were obtained from Imagenes GmbH, Berlin, Germany. The clone's details are shown in Table 5.1.

Gene	Clone Name	Host	Vector	Resistence
Bip	IRAUp969A0481D	GeneHogs DH10B	pOTB7	Chloramphenicol
Rspo2	IRAVp968B03107D	DHl0B	pSPORT1	Ampicillin
Tmem9	IRAVp968E0732D	DHl0B	pCMV- pSPORT6	Ampicillin

Table 5.1: Details of cDNA clones obtained from Imagenes GmbH, Berlin, Germany

5.2.4 DNA isolation

Plasmid DNA was extracted from E. coli overnight cultures using the Nucleospin plasmid kit according to the manufacturer's protocol. DNA was eluted in sterile water and stored at 4ºC. Quantification of extracted DNA was carried out by both UV spectrophotometry (λ=260 nm) and 1 % agarose gel electrophoresis.

5.2.5 Polymerase Chain Reaction

Primers

Primers were designed with the help of the software clone manager 7.0. The primer sequences and their Tm values are shown in Table 5.2.

cDNA	Primer sequences	Tm (ºC)
Bip_3_Kpn1	GGTCGGTACCTCAGTGTCTACAACTCATC	68.1
Bip_5_Xho1	CAGTCTCGAGTGGCAAGATGAAGCTCTCC	69.5
Rspo2_3_Kpn1	CGTGGTACCTGCCCCAGCTATTTCTTG	68.0
Rspo2_5_ Xho1	CAGTCTCGAGCGTCCAGATGCGTTTTTGC	69.5
Tmem9_3_Kpn1	CAGTCTCGAGGATAAGCATGAAGCTGCTG	68.1
Tmem9_5_ Xho1	GAATGGTACCGGCAACCATCTAACTGAGC	68.1

Table 5.2: List of primer sequences

The synthesized primers were dissolved in sterile double distilled water to get a concentration of 100 pmol/µl. The primer stock solutions were stored at -20ºC.

PCR

In order to clone the open reading frame (ORF) of the cDNA of interest in a suitable vector, it was necessary to amplify the appropriate DNA sequences by PCR. PCR reactions were carried out in T3 thermocycler in a volume of 25 µl. The composition of a standard reaction is shown in Table 5.3.

Ingredients	Amount (µl)
ddH_2O	15.5
10x buffer	2.5
dNTP's (2mM)	0.5
Primer 1 (FP) (1pmol/µl)	2.5
Primer 2 (RP) (1pmol/µl)	2.5
Taq DNA Polymerase (1U/µl)	0.5
Template (500ng/µl)	1
Total	25

Table 5.3: Standard PCR reaction mix per 25 µl of total volume

The PCR reaction include 95ºC with 3 min for denaturation of the double-stranded DNA molecule, 55ºC for hybridization of the primers (30 sec) to the single-stranded DNA template and 72ºC for enzymatic polymerization (extension) (1 min). This temperature sequence is then

cycled many times (~30) to provide an exponential amplification of the starting material. Finally, the PCR products were analyzed by 1 % agarose gel electrophoresis.

5.2.6 Agarose gel electrophoresis

1% agarose was prepared with TAE buffer (40 mM Tris-acetate, 1 mM EDTA), which contained ethidium bromide at a final concentration of 0.5 µg/ml. DNA samples were loaded onto the gel in loading buffer (50% [v/v] glycerol, 0.005% [w/v] bromophenol blue, 1x TAE). Electrophoresis was carried out at 100 to 120 V in TAE buffer and the DNA bands were subsequently visualized using an UV transilluminator and a CCD camera.

5.2.7 Gel elution of PCR fragments

After agarose gel electrophoresis, PCR products of interest were excised from low melting agarose gel with a sharp sterile scalpel blade under low UV intensity (70%). The agarose gel piece that contains the DNA was collected in a sterile pre-weighed micro-centrifuge tube. The DNA was eluted using Nucleospin extract II kit following the method described in the user's manual.

5.2.8 Digestion of DNA with restriction enzymes

Digestion with the used restriction enzymes generates overhangs that allow ligation of the DNA insert with the vector. Double digestion can be a single step process or a double step process depending on whether the respective restriction enzymes share the same digestion buffer or not.

Reagents	Volume (50 µl)
ddH_2O	39.5
10X buffer	5.0
BSA	0.5
DNA	2.0
Restriction enzyme	3.0

Table 5.4: Linearization of DNA reaction mix per 50 µl of total volume

The restriction enzymes were purchased from New England BioLabs Inc. XhoI and KpnI both used NEBuffer 1, therefore the vector and the PCR product could be digested in a single step

reaction. Double digest mixture was incubated at 37ºC for 1 hour (thermomixer). The reagents and their volumes are mentioned in Table 5.4. The samples were incubated in a thermomixer at 37°C for a period of 3 hours.

5.2.9 Ligation

Ligation is a two-step process, first the sticky ends generated by the digestion of the DNA insert and the vector hybridize or anneal and second, new phosphodiester bonds close the nicks that are left behind after annealing. The reaction mixture consists of the DNA insert, the vector, ligation buffer and the DNA ligase enzyme. The reagents and their volumes are mentioned in Table. 5.5.

Reagents	Volume (10 µl)
ddH_2O	3.1
10x ligase buffer	1.0
Vector (after double digestion)	4.0
PCR product (after double digestion)	1.4
T4 DNA ligase	0.5

Table 5.5: Standard Ligation reaction mix per 10 µl of total volume

The vector, PCR product and double distilled water were added and heated for 5 min at 45ºC to separate the DNA fragments and then the reaction mixture was cool down on ice for 5 min. T4 DNA ligase and 10x ligase buffers were added and gently mixed by pipetting. The reaction mixture was incubated at 16°C overnight.

5.2.10 Transformation

Transformation is the process by which bacteria take up foreign DNA. Competent DH5α cells were thawed on ice. 5 µl of the ligation product was added to the culture tube with the competent cells. The cells were mixed with the ligation mix by swirling and incubated on ice for 1 min. The cells were heat shocked at 42°C for 1 min and immediately placed on ice for 1 min. Subsequently, 900 µl of LB medium were added and the samples were mixed carefully by shaking the culture tubes with 800 rpm on thermomixer at 37°C for one hour. The transformed cells were plated on LB agar plates and incubated overnight at 37°C.

5.2.11 Colony PCR

Bacteria originating from a single colony can be used directly as template for PCR because the initial denaturation heat disrupts the cell walls and make DNA accessible. Each colony chosen for colony PCR was transferred to a new ampicillin plate with a pipette tip and remains of the cells from the tip were mixed into the PCR reaction. The PCR reaction mix was set up as described in Table 5.3, except that the total volume was adjusted to 50 µl without an additional DNA template. The amplified products were separated from the template DNA with 1 % agarose gel electrophoresis containing TBE buffer and visualized by UV transilluminator and a CCD camera.

5.2.12 DNA sequencing

DNA sequencing was used to verify the exact sequence of the fragments cloned into the pGEM vector. The samples were sequenced by GATC, Germany and the results were analyzed with clone manager 7.0 software.

5.2.13 Transcription

Before the cloned ORF's were transcribed into mRNA, the corresponding vectors were linearized by restriction enzymes. For Bip, Rspo2 and Tmem9 the enzyme NheI was used.

Reagents	Volume (50 µl)
ddH_2O	11.5
5x buffer	10.0
DTT (10 mg/ml)	5.0
RNAsin (40 U/µl)	2.5
DNA (1 µg)	18
BSA (10 mg/ml)	1.0
rNTPs (25 mM)	1.0
T7 Polymerase (19 U/µl))	1.0

Table 5.6: Transcription reaction mix per 50 µl of total volume

The vectors coding for the control proteins pPL and ppαF were cleaved with pst2 and BamH1 respectively. The linearized ORF sequences were transcribed into mRNA using standard protocols. The reagents and their volumes are mentioned in Table 5.6. The reaction mixture was

incubated for 2 hours at 37°C. Afterwards, 0.5 µl T7 Polymerase and 0.5 µl rNTPs were added and incubated for 2 hours at 37°C. Finally, the reaction was purified with phenol-chloroform extraction following the manufacturer's protocol.

5.2.14 Translation

The translation reaction was carried out with rabbit reticulocyte lysate, mRNA, amino acid (without methionine), ^{35}S radio labeled methionine and H_2O (Table 5.7).

Reagents	Volume (10 µl)
Lysate	6.0
Amino acid (without methionine)	0.2
^{35}S radio labeled methionine	0.3
mRNA	0.5/1.0/1.5
ddH_2O	3.0/2.5/2.0

Table 5.7: Translation reaction mix per 10 µl of total volume

The samples were incubated at 25ºC for 45 min. In order to reduce the globin concentration the samples were treated with 60% ammonium sulfate as indicated. 8 µ1 samples were mixed with 100 µ1 of 60% ammonium sulfate for 10 min on ice and centrifuged at 13000 rpm for 45 min at 4ºC. The supernatant was removed and around 55 µ1 of 60% ammonium sulfate was added. It was centrifuged at 13000 rpm for 45 min at 4ºC. The supernatant was taken and pellet was dissolved with 20 µ1 of sample buffer. All samples were resolved by SDS-PAGE (12.5% SDS gel) and analyzed with an image analyzer (FLA-3000, Raytest, Germany).

5.2.15 Co-translational translocation assay

In order to directly investigate the transport of our proteins into the endoplasmic reticulum (ER) membrane, we used an in vitro translation/translocation system. mRNA encoding our substrate proteins were translated in a reticulocyte lysate system in the presence of dog rough microsomes (dRM) or yeast rough microsomes (yRM) and ^{35}S radiolabeled methionine (Table 5.8).

Reagents	Volume (10 µl)
Lysate	6.0
Amino acid (without methionine)	0.2
^{35}S radio labeled methionine	0.3
mRNA	0.5
0.3 eq rough microsomes membrane (dog/yeast)	1.0
ddH_2O	2.0

Table 5.8: Co-translational translocation reaction mix per 10 µl of total volume

The translation translocation reaction was carried out at 25ºC for 45 min. Afterwards, one half of the samples were treated with proteinase K (0.5 mg/ml) and proteinase K buffer was added into another half of the sample. The samples were incubated on ice for 30 min. After adding 0.5 mM PMSF (5 min on ice) the samples were precipitated with 60% ammonium sulfate as described in 5.2.14. The pellet was dissolved with 20 µ1 of sample buffer. All samples were resolved by SDS-PAGE (12.5% SDS gel) and analyzed with an image analyzer (FLA-3000, Raytest, Germany).

5.2.16 Post-translational translocation assay

Post-translation assay was carried out in a reticulocyte lysate system in the presence of ^{35}S methionine and bovine pancreatic rough microsomes (bRM) or yeast rough microsomes (yRM), respectively (Table 5.9).

Reagents	Volume (10 µl)
Lysate	6.0
Amino acid (without methionine)	0.2
^{35}S radio labeled methionine	0.3
mRNA	0.5
rough microsomes membrane (bovine (2.5 eq)/yeast (3.2 eq))	1.0
ddH_2O	2.0

Table 5.9: Post-translational translocation reaction mix per 10 µl of total volume

The samples were translated at 25°C for 45 min. After addition of 1 mM cycloheximide the samples were incubated on ice for 5 mins. Samples were centrifuged for 20 min at 70,000 rpm in a TLA100 rotor (Beckman Instruments) using micro test tubes. The supernatant was collected and 2.5 eq bRM or 3.2 eq yRM were added in 10 µl sample and incubated on ice for 20 min and subsequently at 25ºC for 20 min. The samples were treated proteinase K and precipitated with 100 µ1 of 60% ammonium sulfate as described in 5.2.14. The pellet was dissolved with 20 µ1 of sample buffer. All samples were resolved by SDS-PAGE (12.5% SDS gel) and analyzed with an image analyzer (FLA-3000, Raytest, Germany).

5.3 Results and Discussion

5.3.1 Computational analysis of post-translocation pathway proteins

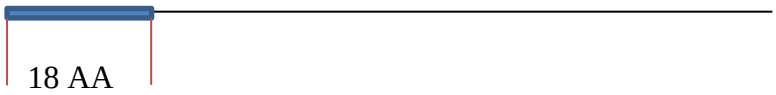
Little information was available about post-translational protein translocation into the ER of mammals. So far, only non mammalian substrate proteins were analyzed. Identification of mammalian substrate proteins for post-translocation pathway was quite difficult. We developed an insilico database for mammalian co/post-translocation pathway proteins. Identification of post-translocation pathway proteins using curated database by sequence similarity methods or motif approach was difficult. Most of the mammalian proteins do not have conserved patterns for post-translocation pathway. Therefore, there is a need for machine learning methods for predicting post-translocation pathway proteins.

We have developed a new SVM model to differentiate post- translocation pathway proteins from co-translocation pathway proteins. The model was trained on a training dataset containing 134 proteins from the positive dataset and 200 proteins from the negative dataset. The performance of the model was evaluated using the five-fold cross-validation method. An overall prediction accuracy of 76.33% was obtained by five-fold cross validation. In order to examine the performance of the newly developed model, we tested our training model on the test dataset consisting of 10 proteins from the positive dataset and 3843 proteins from the negative dataset. An insilico model achieved 73.81% accuracy with 73.02% sensitivity and 74.60% specificity using all features. We ranked all the proteins pertaining to post-translocation pathway using Info Gain algorithm. We predicted 150 proteins pertaining to post-translocation pathway proteins by our algorithm. Out of 150 proteins, we have selected 3 mammalian proteins for our experiments.

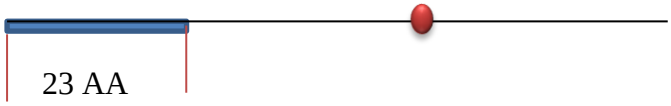
5.3.2 Cloning of ORF's of the potential post-translational pathway proteins into a vector for in vitro translation

Three proteins were selected from the list that was created by the algorithm. Bip is a luminal ER protein, Tmem9 is a membrane protein and Rspo2 belongs to a protein family of secreted proteins. These proteins are schematically illustrated in Figure 5.3.

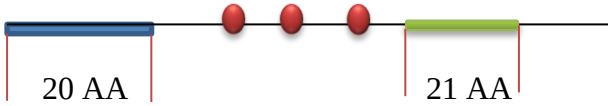
Heat shock 70 KDa protein 5 (Bip), Length of the protein sequence - 654 AA



R-spondin-2 (Rspo2), Length of the protein sequence - 243 AA



Transmembrane protein 9 (Tmem9), Length of the protein sequence - 183 AA



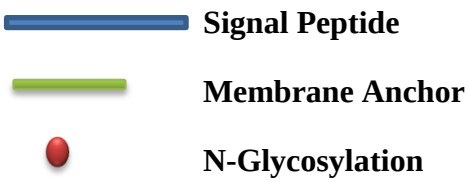


Figure 5.3: List of potential post-translocation pathway proteins

In order to clone the corresponding ORF for in vitro translation, plasmids containing the appropriate cDNA were used as templates for PCR. Under optimal PCR conditions, a single product was amplified. The amplified PCR products were separated on a 1 % agarose gel (Figure 5.4 A). As a control, Annexin V (AnxV) ORF was amplified with primers that were used before. The size of PCR products for AnxV was 1000 bp, for Bip was 1962 bp, for Rspo2 was 720 bp and for Tmem9 was 549 bp.

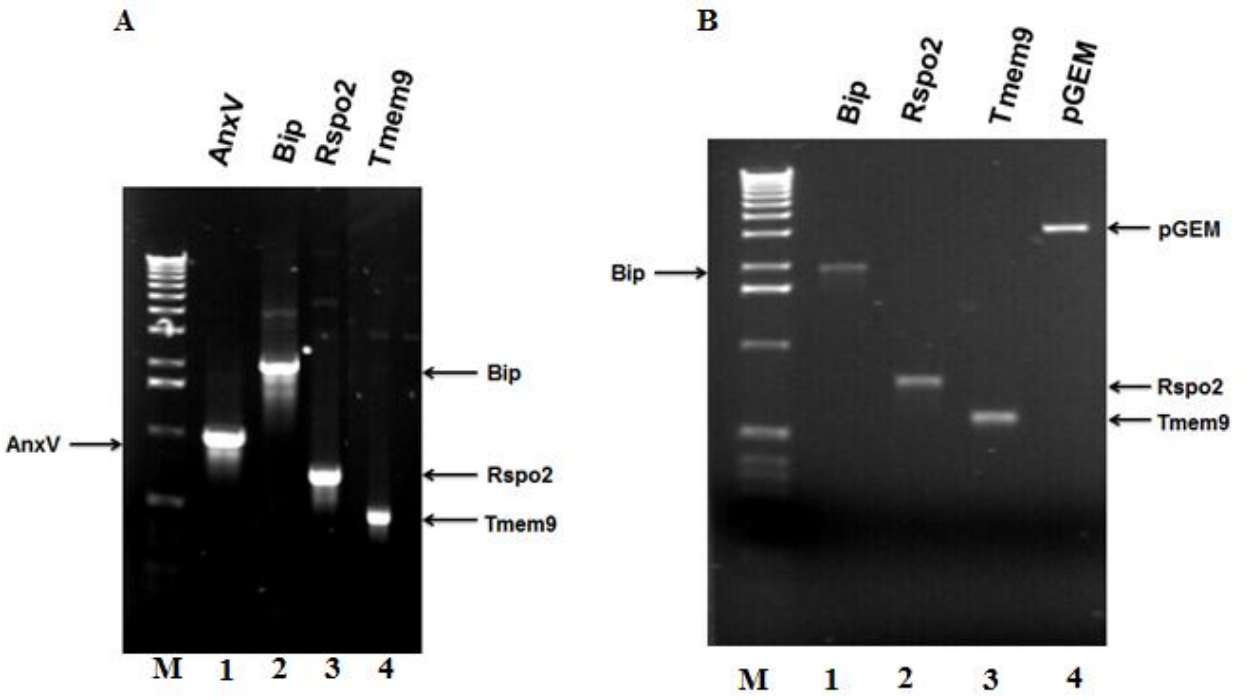


Figure 5.4: The ORF's of Bip, Rspo2 and Tmem9 are cloned into the pGEM vector. (A) cDNA coding for Bip, Rspo2 and Tmem9 were used as a template for PCR. AnxV cDNA was used as a positive control. Amplification was carried out with primers that contain restriction sites for XhoI and KpnI and the samples were separated on a 1 % agarose gel. (B) The purified PCR products from (A) and the vector pGEM were digested with the restriction enzymes XhoI and KpnI. The samples were analyzed on a 1 % agarose gel.

Next, the PCR products should be cloned into the pGEM vector. Since the PCR primer contained restriction sites for XhoI and KpnI, the PCR products and the pGEM vector were digested with these enzymes to create compatible ends (Figure 5.4 B). The amplified products were ligated with the cleaved pGEM Vector with the help of DNA ligase. It was transformed into chemocompetant E. coli DH5α cells. The transformation into E. coli cells gave good growth and 10 colonies were picked from the plates and grown over night. The presence of the pGEM vector containing the insert of interest in the colonies growing on ampicillin plates was verified by PCR using the cells directly as a DNA template. To test for plasmid DNA with correct insert, those DNA samples were send for sequencing. The sequences were analyzed by software clone manager 7.0. It turned out that cloned sequences were 100% correct.

5.3.3 The cloned test proteins Bip, Rspo2 and Tmem9 can be translated in vitro

In the next step, the test proteins should be translated in the reticulocyte lysate system. Therefore, the plasmids containing the appropriate ORF's were linearized and used for in vitro transcription with T7 polymerase. The corresponding mRNA's were isolated and translated in the reticulocyte lysate in the presence of ^{35}S methionine at 25°C for 45 min. To get the optimal mRNA concentration, three different amounts of mRNA were used (Figure 5.5). The size of the expressed proteins for Bip was 62 kDa, for Rspo2 was 34 kDa and for Tmem9 was 28 kDa.

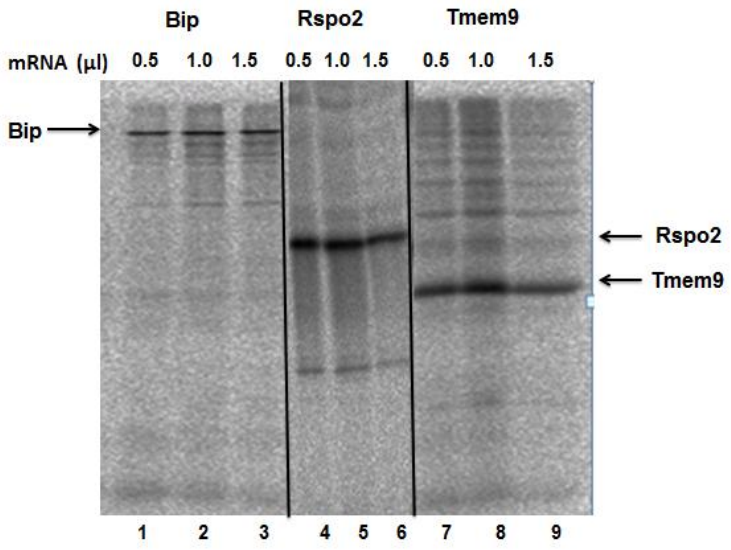


Figure 5.5: Bip, Rspo2 and Tmem9 can be translated in the reticulocyte lysate system in vitro. Three different concentrations of mRNA were translated in the reticulocyte lysate in the presence of ^{35}S methionine at 25°C for 45 min. The samples were analyzed by SDS-PAGE using 12.5% polyacrylamide gels and autoradiography.

To reduce the background and to increase the intensity of the signals, the experiment shown in Figure 5.5 was repeated but the samples were precipitated with 60% ammonium sulfate prior to the SDS-PAGE. This treatment clearly improved the quality of the corresponding autoradiogram (Figure 5.6).

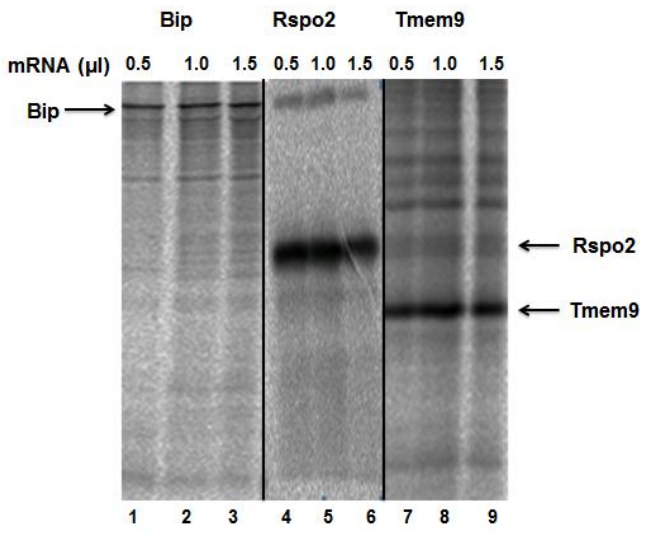


Figure 5.6: Ammonium sulfate precipitation increases the quality of the autoradiogram. Translation reaction was carried out as described in Figure 5.5. The samples were treated with 60% ammonium sulfate and analyzed by SDS-PAGE and autoradiography.

5.3.4 Only Bip is translocated into ER membranes under co-translational conditions in vitro

The assay for protein translocation into the ER relies on the ability of a radiolabeled protein precursor to transit across the membrane and thus become inaccessible to exogenously added protease. In order to directly investigate the integration of Tmem9 or translocation (Bip and Rspo2) of the candidate proteins into the ER-membrane, we used an in vitro translation/translocation system. First the protein transport should be analyzed under co-translational conditions and dog pancreatic microsomes were used.

To be sure that the used ER-membranes are active for protein translocation preprolactin (pPL) was employed as a control protein. mRNA coding for pPL was translated in a reticulocyte lysate system in the presence of ^{35}S methionine and ER-membranes as indicated. To test for proper translocation half of the samples were treated with proteinase K and finally analyzed by SDS-PAGE and autoradiography (Figure 5.7).

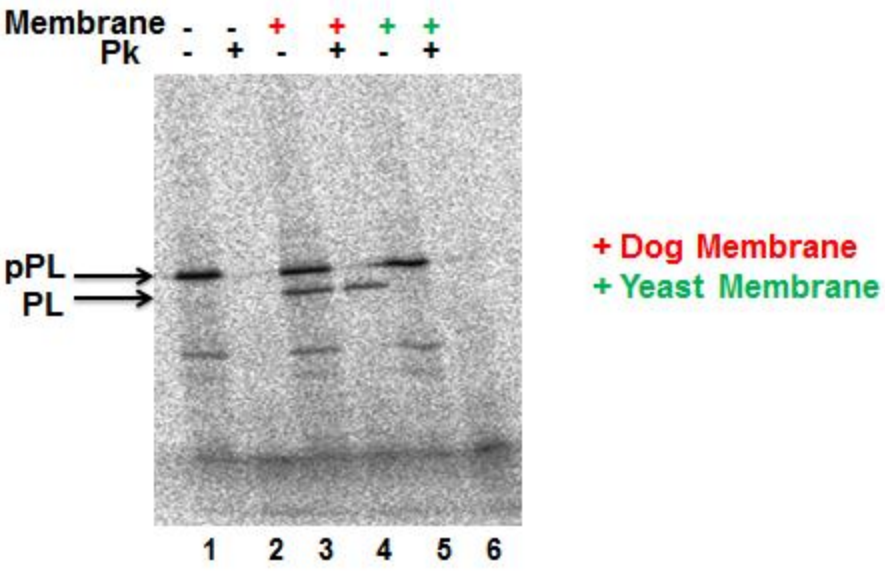


Figure 5.7: The control protein preprolactin (pPL) is specifically transported into dog rough microsomes (dRM). pPL was used as a positive control for co-translational translocation into dog pancreatic microsomes. pPL mRNA was translated in the reticulocyte lysate system in the presence of dog or yeast microsomes as indicated. Thereafter, half of the samples were treated with proteinase K (PK). Finally, the radiolabelled samples were analyzed by SDS-PAGE using 12.5% polyacrylamide gels and analyzed with an image analyzer (FLA-3000, Raytest, Germany).

Without microsomes a single band was detected (lane 1, pPL), which was totally degraded by the proteinase K treatment (lane 2). In the presence of dog rough microsomes (dRM) pPL was translocated into the membranes and the signal peptidase complex cleaved off the signal peptide, generating a smaller protein band (lane 3, PL). This cleaved prolactin was completely protected against the proteinase K treatment demonstrating that it was really translocated into the lumen of the membrane vesicles (lane 4). Since it is well known that pPL is not transported into yeast rough microsomes (yRM) in vitro these membranes were used as a negative control. As expected neither signal peptide cleavage nor proteinase K protection were observed (lane 5 and 6), demonstrating the specificity of the used translocation assay.

In the next step, the luminal ER protein Bip should be analyzed for its translocation into dRM or yRM, respectively (Figure 5.8), using the same translocation assay as described for the control protein pPL.

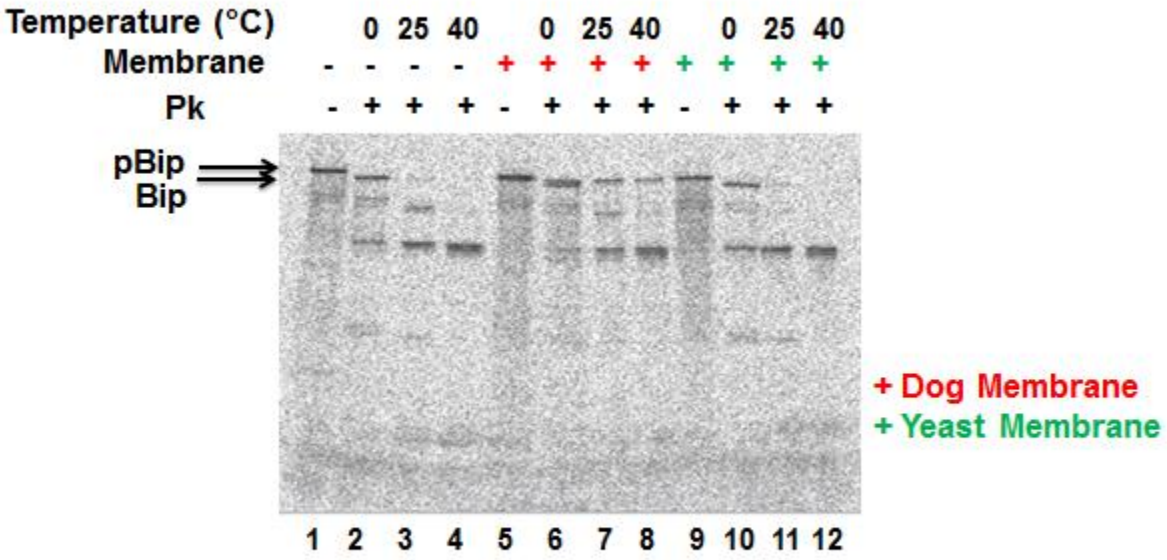


Figure 5.8: Translocation of Bip into the endoplasmic reticulum in vitro. mRNA encoding pBip was translated in a reticulocyte lysate system in the presence of dog rough microsomes (dRM) or yeast rough microsomes (yRM) and ^{35}S radiolabeled methionine. The samples were treated with proteinase K and analyzed by autoradiography.

A single protein band was detected in the autoradiogram when the microsomes were omitted during the translation reaction (lane 1). Surprisingly, the treatment with proteinase K under standard conditions did not degrade the protein, but produced only a slightly smaller protein band compared to the untreated protein (lane 2). This indicates that the main part of pBip is resistant against proteinase K treatment at 0°C. Unfortunately, the protease resistant part of pBip has almost the same size as the mature Bip with cleaved off signal peptide that could be expected after translocation into the microsomes. Therefore, the proteinase K treatment was repeated at 25°C or 40°C, respectively, to get a better degradation of Bip. At 40°C a larger part of Bip was degraded and the size of the remaining protein should not interfere with the size of the translocated Bip with cleaved off signal peptide.

In the presence of dog rough microsomes (dRM) Bip was translocated into the membranes and the signal peptidase complex cleaved off the signal peptide, generating a smaller protein band (lane 5, pBip). This cleaved Bip was completely protected against the proteinase K treatment (40°C) demonstrating that it was really translocated into the lumen of the membrane vesicles (lane 8, Bip). In the presence of yeast rough microsomes (yRM) Bip cannot translocate into the membranes and the signal peptide was not cleaved off (lane 12).

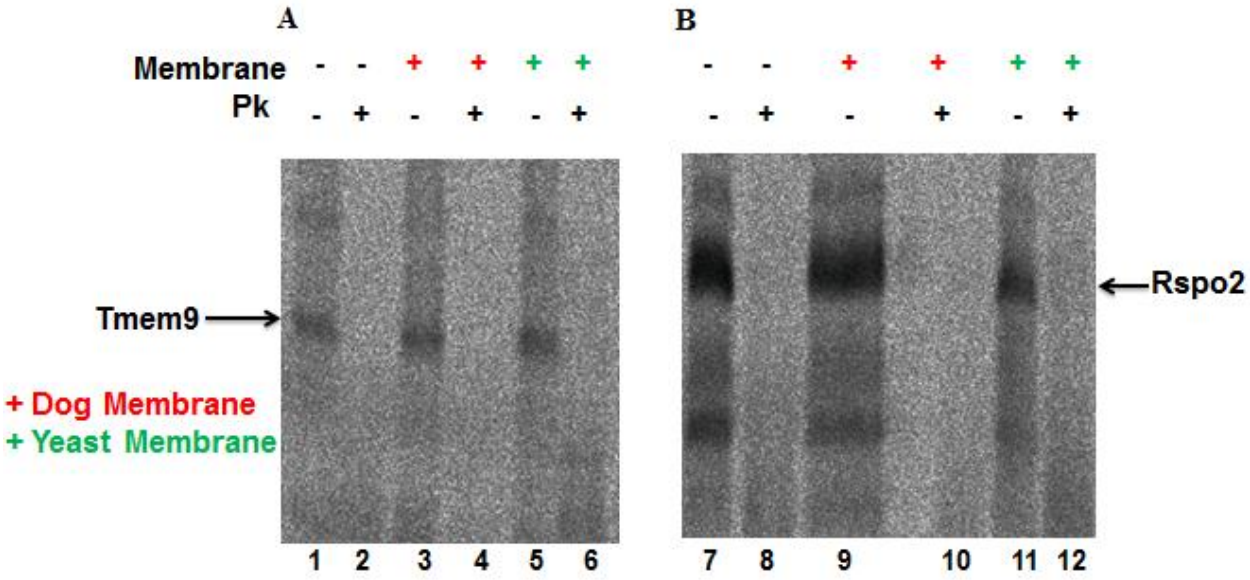


Figure 5.9: Tmem9 and Rspo2 cannot translocate into dRM or yRM. In vitro translation/translocation reactions were performed using reticulocyte lysate in the presence of dRM/yRM and mRNA encoding either Tmem9 (A) or Rspo2 (B). Microsomal membranes (dRM/yRM) were added as indicated. The samples were treated with proteinase K. Finally, the radiolabelled samples were analyzed by SDS-PAGE using 12.5% polyacrylamide gels and analyzed with an image analyzer (FLA-3000, Raytest, Germany).

Next, we asked whether the proteins Rspo2 and Tmem9 can be translocated in vitro. Translation in rabbit reticulocyte lysate was performed for 45 min at 37°C. The samples were analyzed by SDS-PAGE and autoradiography (Figure 5.9). Without microsomes a single band was detected (lane 1, 7) which was totally degraded by the proteinase K treatment (lane 2, 8). In the presence of dog or yeast rough microsomes (dRM or yRM), neither processing nor proteinase K protection were observed (lane 3-6, 9-12). Rspo2 and Tmem9 cannot translocate into the membranes and the signal peptide was not cleaved off as indicated. Further, we have optimized

the conditions (temperature, time, lysate and membrane concentration) to check the translocation of Rspo2 and Tmem9. It turned out again that Rspo2 and Tmem9 cannot translocate in vitro by the co-translational mechanism under the used conditions (data not shown).

5.3.5 None of the candidate proteins are translocated post-translationally in vitro

To analyze post-translational translocation or insertion of proteins into ER-membranes, a similar translocation assay was used as for co-translational translocation. The crucial difference is that now RMs were added into the reaction after termination of translation by cycloheximide and sedimentation of all ribosomes. This ensures that neither translation nor ribosomes are involved in the targeting or translocation process, respectively.

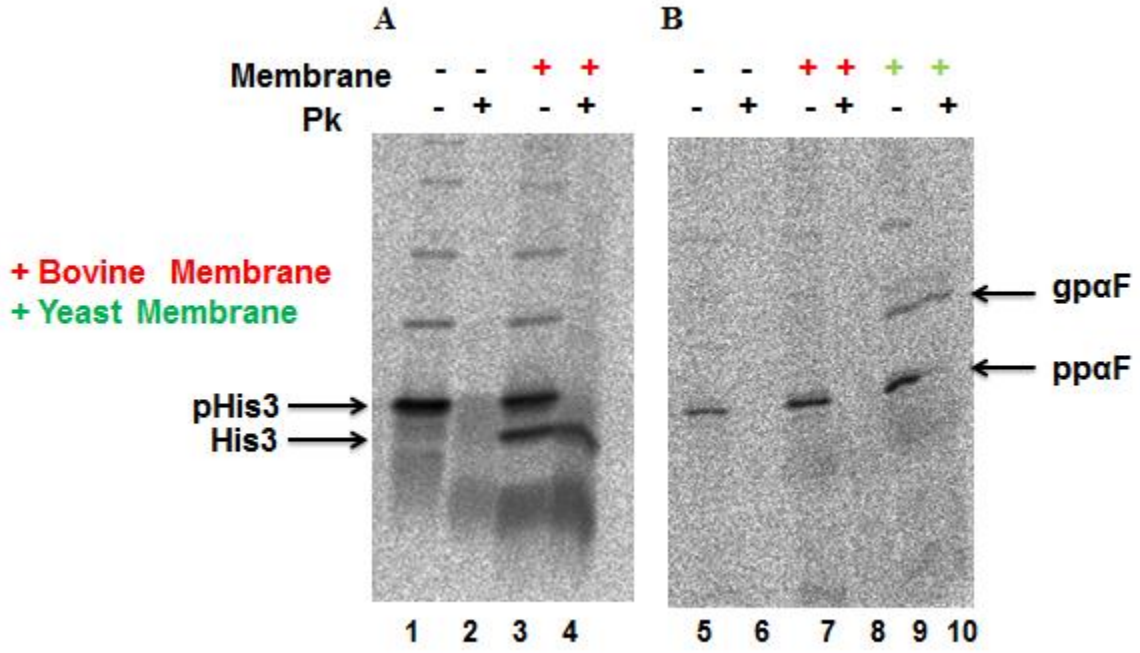


Figure 5.10: The control protein His3 and ppαF are transported post-translationally into bRM and yRM. Histatin 3 (His3) and yeast protein prepro-alpha-factor (ppαF) were used as positive controls for post-translational translocation into bovine pancreatic microsomes or yeast microsomes. His3 (A) and ppαF (B) mRNA was translated in the reticulocyte lysate system. The reaction was stopped by cycloheximide and ribosomes were sedimented. Afterwards bovine membranes or yeast membranes were added to the supernatant as indicated. The samples were analyzed by SDS-PAGE using 12.5% polyacrylamide gels and autoradiography.

Two control proteins were employed for the post-translational translocation assay. The small human secretory protein Histatin 3 (His3) is transported post-translationally into bovine rough

microsomes (bRM) (Vivica Stokes, personal communication and Figure 5.10 A). The signal sequence is cleaved off and the translocated mature part of the protein is protected against the externally added proteinase K (lane 3 and 4). The yeast protein prepro-alpha-factor (ppαF) can be translocated post-translationally into yRM. After translocation into the ER-membranes the signal sequence is cleaved off and the protein becomes glycosylated (Figure 5.10 B). This results in a larger protein band (gpαF), which is protected against proteinase K treatment (lane 9 and 10). A translocation of ppαF into bRM could not be observed (lane 7 and 8).

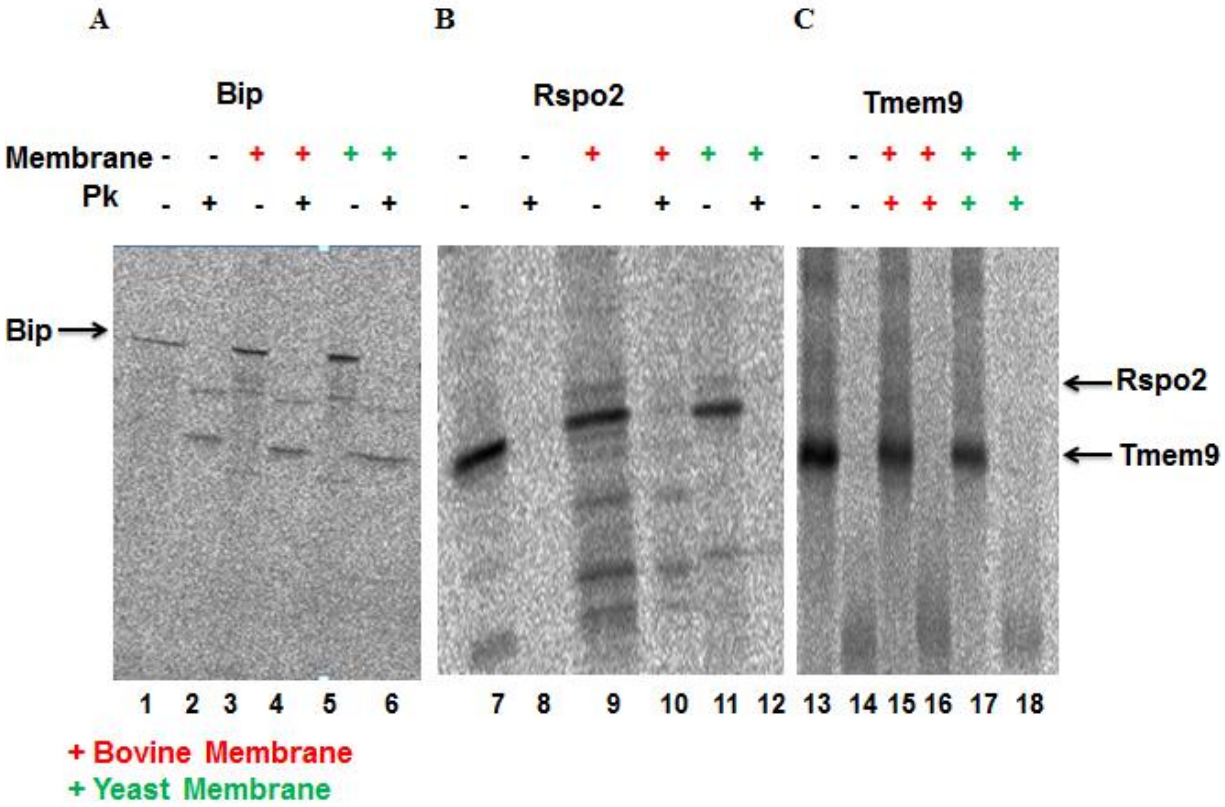


Figure 5.11: None of the substrate proteins (Bip, Rspo2 and Tmem9) are translocated post-translationally in vitro. In vitro transcribed Bip (A), Rspo2 (B) and Tmem9 (C) mRNA was translated in reticulocyte lysate and labeled by ^{35}S methionine. In vitro translation occurred either in the absence of microsomes or microsomes (bRM/yRM) were added after sedimentation of ribosomes. Thereafter, half of the samples were treated with proteinase K. The samples were analyzed by SDS-PAGE using 12.5% polyacrylamide gels and autoradiography.

In the next step, the luminal ER protein Bip should be analyzed for its post-translationally translocation into bRM or yRM, respectively (Figure 5.11 A), using the same translocation assay as described for the control proteins His3 and ppαF. After translation a single protein band was

detected in the autoradiogram (lane 1). The treatment with proteinase K led to the degradation of the protein (lane 2). After addition of microsomes neither processing nor proteinase K protection were observed (lane 3-6). This indicates that Bip was not translocated post-translationally into the membranes.

Unfortunately, also the other candidate proteins showed no post-translational translocation or integration into the microsomes under the used conditions (Figure 5.11 B and C). Rspo2 and Tmem9 were degraded by the proteinase K treatment regardless whether membranes were added after the translation reaction or not.

Taken together, our results show that the mammalian proteins Bip, Rspo2 and Tmem9 cannot translocate post-translationally into ER- membranes in vitro. The failure of our substrate proteins depends on two major reasons: computational methods and translation/translocation assay system.

The success of machine learning method depends on the quality of training dataset. In this study, we have not used experimentally determined protein sequences for co and post-translocation pathway proteins. We have hypothesized the positive and negative dataset (length of protein sequences and organism, etc.). Due to these reasons, our algorithm may fail to recognize post-translocation pathway proteins appropriately. We predicted 150 proteins pertaining to post-translocation pathway proteins by our algorithm. Out of 150 proteins, we have selected only 3 mammalian proteins for our experiments. Cross validation accuracy of our SVM model was only 76.33%, our three mammalian proteins can be false positive proteins.

We showed that our substrate proteins were not translocated post-translationally into bovine or yeast microsomes. It is possible that our candidate proteins may need unknown factors to translocate post-translationally. It could be that the lysate we used for translation/translocation does not contain these factors in the required concentration.

Our results also show that Rspo2 and Tmem9 cannot translocate co-translationally in vitro. However these proteins may translocate in another translation/translocation assay system (eg: HeLa cell lines) or if purified SRP is added into the assay.

5.4 Conclusion

Translocation of large pre-secretory proteins into the mammalian endoplasmic reticulum requires the ribonucleoparticles, signal recognition particle, and ribosomes and is tightly coupled to ongoing protein synthesis. In this study we developed a method, for the first time, for predicting post-translational translocation proteins from its primary sequence using pseudo amino acid composition coupled with support vector machine. We tested experimentally top ranked post-translocation pathway proteins in mammals (Bip, Rspo2 and Tmem9). Our analysis shows that Rspo2 and Tmem9 cannot translocate in vitro by the co and post-translational mechanism. We show that the luminal ER protein Bip protein translocate co-translationally (with the aid of signal recognition particle and ribosome) during its in vitro synthesis in the presence of dog pancreas microsomes. Our further interest is focused on Rspo2 and Tmem9 proteins translocate in vivo by the co and post-translational mechanism.

6 Conclusion

Modeling and simulation is used routinely in academia as well as industry. Conventionally, modeling is carried out employing various conservation laws (aka phenomenological models). But as the complexity of the phenomenon increases, it becomes difficult to keep track of the modeling parameters. For many complicated processes, for e.g. biological processes, it becomes difficult to build even an elementary model. In light of these facts, alternative modeling techniques, mainly data driven, have become popular recently in various science streams and has shown promising results in many real life applications. The main objectives of the proposed thesis work were to develop new sequence analysis tools for protein function and family classification. Chapters 3 to 5 summarize the efforts in achieving these objectives.

In the second chapter, the support vector machines and random forest for classification, important machine learning algorithms with many desirable properties, were introduced to classify putative proteins and associate them with families of proteins with known functions. Feature selection algorithms like ReliefF, Info Gain, Maximum Relevance Minimum Redundancy (mRMR) and Genetic algorithms were used to solve important biological problems.

A major difficulty in analysing the biological sequence data using machine learning methods is the nature of the data i.e. characters (amino acids of different length). In this thesis, various encoding methods were developed to represent biological sequences into arrays of numerical values.

The identification of proteins in the drug discovery process is quite important because it is responsible for many functions required for maintenance of life. Chapter 3 described the prediction of protein function. We solved three different protein function classification problems 1) prediction of classical and non-classical secretory proteins 2) prediction of extracellular matrix proteins 3) prediction of the subcellular locations of apoptosis proteins. The proposed approach can be quite effective in assigning putative functions to novel sequences. This can serve as a useful source of information for guiding focused biological experiments.

Due to the recent advances in high throughput data acquisition technologies in biological sciences, there is a need for the development of sophisticated computational tools for characterization and prediction of protein families. In chapter 4, we solved two important protein families (antifreeze and bioluminescent proteins) which have great possibilities for medical and commercial advances. Our machine learning approach helps to annotate hypothetical proteins of antifreeze and bioluminescent protein families.

In chapter 5, we further discussed the bioinformatics and experimental approaches on protein translocation. We developed a novel prediction method that uses a support vector machine and pseudo amino acid composition to predict post-translocation pathway proteins. Further, we tested experimentally top ranked post-translocation pathway proteins (Bip, Rspo2 and Tmem9). Our analysis shows that Rspo2 and Tmem9 cannot translocate in vitro by the co and post-translational mechanism. We show that one large protein Bip translocate co-translationally (with the aid of signal recognition particle and ribosome) during its in vitro synthesis in the presence of dog microsomes.

Assigning putative functions to protein sequences remains one of the most challenging problems in functional genomics. Our methods will improve the annotation of newly sequenced genomes. It will give some useful insights into protein structure function relationships by exploring sequence regularities that are good predictors of function.

The results presented and discussed in chapters 3 to 5 reveal the fact that SVM and RF are promising data driven modeling techniques. The applications (theoretical as well as real life) uncovered the fact that the base algorithm can be easily modified to suit the necessities for the problem at hand.

7 References

Adams, J.M, Cory, S (1998) The Bcl-2 protein family: arbiters of cell survival. *Science*, 281, 1322-1326.

Aha, D, Kibler, D (1991) Instance-based learning algorithms. *Machine Learning*, 6, 37-66.

Altschul, S.F, Gish, W, Miller, W, Myers, E.W, Lipman, D.J (1990) Basic local alignment search tool. *J Mol Biol*, 215 (3), 403–410.

Altschul, S.F, Madden, T.L, Schaffer, A.A, Zhang, J, Zhang, Z, Miller, W, Lipman, D.J (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25 (17), 3389-3402.

Anand, A, Pugalenthi, G, Suganthan, P.N (2008) Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. *J. Theor. Biol*, 253 (2), 375-380.

Andreeva, A, Howorth, D, Chandonia, J.M, Brenner, S.E, Hubbard, T.J, Chothia, C, Murzin, A.G (2008) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res*, 36 (Database issue), D419-25.

Aszódi, A, Legate, K.R, Nakchbandi, I, Fässler, R (2006) What mouse mutants teach us about extracellular matrix function. *Annu Rev Cell Dev Biol*, 22, 591-621.

Attwood, T.K, Blythe, M.J, Flower, D.R, Gaulton, A, Mabey, J.E, Maudling, N, McGregor, L, Mitchell, A.L, Moulton, G, Paine, K, Scordis, P (2002) PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res*, 30, 239 -241.

Bairoch, A, Apweiler, R (2000) The SWISS-PROT Database and its Supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1), 45-48.

Baker, P.G, Goble, C.A, Bechhofer, S, Paton, N.W, Stevens, R, Brass, A (1999) An Ontology for Bioinformatics Application. *Bioinformatics*, 15(6), 510-520.

Barker, W.C, Pfeiffer, F, George, D.G (1996) Superfamily classification in PIR-International Protein Sequence Database. *Methods Enzymol*, 266, 59-71.

Bateman, J.F, Boot-Handford, R.P, Lamande, S.R (2009) Genetic diseases of connective tissues: cellular and extracellular effects of ECM mutations. *Nat Rev Genet*, 10, 173-183.

Benson, D, Karsch-Mizrachi, I, Lipman, D.J, Ostell, J, Rapp, B.A, Wheeler, D.L (2000) *GenBank. Nucleic Acids Research*, 28(1), 15-18.

Berman, H.M (2008) The Protein Data Bank: a historical perspective. *Acta Cryst*, A64, 88-95.

Berg, J, Tymoczko, J, Stryer L (2002) Biochemistry, W. H. Freeman and Company.

Bendtsen, J.D, Nielsen, H, Krogh, A, von Heijne, G, Brunak, S (2004a) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*, 340, 783-795.

Bendtsen, J.D, Jensen, L.J, Blom, N, Heijne, G.V, Brunak, S (2004b) Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel*, 17, 349–356.

Bishop, C.M (1995) Neural Networks for Pattern Recognition, Oxford, Oxford University Press.

Bishop, C.M (2006) Pattern Recognition and Machine Learning (Information Science and Statistics), Springer.

Blobel, G, Sabatini, D.D (1971) Ribosome-membrane interaction in eukaryotic cells. *Biomembranes*, 2, 193-5.

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

Bradley, A.P (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn*, 30, 1145-1159.

Breiman, L, Friedman, J.H, Olshen, R, Stone, C.J (1984) Classification and Regression

Tree, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.

Breiman, L (2001) Random forests. *Machine Learning*, 45, 5-32.

Breton, G, Danyluk, J, Ouellet, F, Sarhan, F (2000) Biotechnological applications of plant freezing associated proteins. *Biotechnol. Annu. Rev*, 6, 59-101.

Brown, M.P, Grundy, W.N, Lin, D, Cristianini, N, Sugnet, C.W, Furey, T.S, Ares, M. Jr, Haussler, D (1999) Knowledge-based analysis of microarray gene expression data using SVM. *Proceedings of the National Academy of Sciences*, 97(1), 262-267.

Bru, C, Courcelle, E, Carrère, S, Beausse, Y, Dalmar, S, Kahn, D (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res*, 33 (Database issue), D212- 5.

Bruckner-Tuderman, L, Bruckner, P (1998) Genetic diseases of the extracellular matrix: more than just connective tissue disorders. *Journal of molecular medicine (Berlin, Germany)*, 76, 226-237.

Burridge, K, Chrzanowska-Wodnicka, M (1996) Focal adhesions, contractility, and signaling. *Annu Rev Cell Dev Biol*, 12, 463-518.

Burbidge, R, Trotter, M, Buxton, B, Holden, S (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput Chem*, 26(1), 5-14.

Burges, C.J.C (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2).

Bylander, T (2002) Estimating generalization error on two class datasets using out-of-bag estimates. *Machine Learning*, 48, 287-297.

Cai, Y.D, Liu, X.J, Xu, X.B, Chou, K.C (2002a) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. *J. Cell. Biochem*, 84, 343-348.

Cai, Y.D, Liu, X.J, Xu, X.B, Chou, K.C (2002b) Prediction of protein structural classes by support vector machines. *Computers & Chemistry*, 26, 293-296.

Cai, Y.D, Liu, X.J, Xu, X.B, Chou, K.C (2002c) Support vector machines for predicting the specificity of GalNAc-transferase. *Peptides*, 23, 205-208.

Cai, Y.D, Liu, X.J, Xu, X.B, Chou, K.C (2002d) Support Vector Machines for predicting HIV protease cleavage sites in protein. *J Comput Chem*, 23, 267-274.

Cai, Y.D, Liu, X.J, Xu, X.B, Chou, K.C (2002e) Support vector machines for the classification and prediction of beta-turn types. *J Pept Sci*, 8, 297-301.

Cai, Y.D, Zhou, G.P, Chou, K.C (2003a) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical Journal*, 84, 3257-3263.

Cai, Y.D, Lin, S, Chou, K.C (2003b) Support vector machines for prediction of protein signal sequences and their cleavage sites. *Peptides*, 24, 159-161.

Cai, Y.D, Feng, K.Y, Li, Y.X, Chou, K.C (2003c) Support vector machine for predicting alpha-turn types. *Peptides*, 24, 629-630.

Cai, Y.D, Pong-Wong, R, Feng, K, Jen, J.C.H, Chou, K.C (2004a) Application of SVM to predict membrane protein types. *Journal of Theoretical Biology*, 226, 373-376.

Cai, Y.D, Zhou, G.P, Jen, C.H, Lin, S.L, Chou, K.C (2004b) Identify catalytic triads of serine hydrolases by support vector machines. *Journal of Theoretical Biology*, 228, 551-557.

Campbell, N.E, Kellenberger, L, Greenaway, J, Moorehead, R.A, Linnerth-Petrik, N.M, Petrik, J. (2010) Extracellular matrix proteins and tumor angiogenesis. *J Oncol*, 586905.

Champe, P.C, Harvey, R.A, Ferrier, D.R (2004) Lippincott's Illustrated Reviews: Biochemistry (3rd ed.) Hagerstwon, MD: Lippincott Williams & Wilkins.

Chang, C.C, Lin, C.J (2001) LIBSVM: A Library for Support Vector Machines. www.csie.ntu.edu.tw/~cjlin/libsvm.

Chalfie, M, Tu, Y, Euskirchen, G, Ward, W.W, Prasher, D.C (1994) Green fluorescent protein as a marker for gene expression. *Science*, 263, 802-805.

Cheng, C.H (1998) Evolution of the diverse antifreeze proteins. *Curr. Opin. Genet. Dev*, 8 (6), 715.

Chen, Y.L, Li, Q.Z (2004) Prediction of the subcellular location apoptosis proteins using the algorithm of measure of diversity. *Acta Sci. Natur. Univ. NeiMongol*, 25, 413-417.

Chen, X.W, Liu, M (2005) Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, 21, 4394-4400.

Chen, Y.L, Li, Q.Z (2007a) Prediction of the subcellular location of apoptosis proteins. *J. Theor. Biol*, 245, 775–783.

Chen, Y.L, Li, Q.Z (2007b) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. *J Theor. Biol*, 248, 377-381.

Chou, K.C (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. *J. Mol. Biol*, 223 (2), 509-517.

Chou, K.C (2001) Prediction of protein cellular attributes using pseudo-amino-acid composition. *Proteins: Structure, Function, and Genetics*, 43, 246-255.

Chou, K.C (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 21, 10 -19.

Chou, K. C (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics*, 6, 262-274.

Chou, K.C, Cai, Y.D (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem*, 277, 45765-45769.

Chou, K.C, Cai, Y.D (2005) Prediction of membrane protein types by incorporating amphipathic effects. *J. Chem. Inf. Modeling*, 45, 407-413.

Chou, K.C, Shen, H.B (2006a) Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochem.Biophys. Res. Commun*, 347, 150-157.

Chou, K.C, Shen, H.B (2006b) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Proteome Res*, 5, 1888-1897.

Chou, K.C, Shen, H.B (2006c) Large-scale plant protein subcellular location prediction. *J. Cell. Biochem*, 100, 665-678.

Chou, K.C, Shen, H.B (2006d) Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. *Biopolymers*, 85, 233-240.

Chou, K.C, Shen, H.B (2006e) Large-scale predictions of Gram-negative bacterial protein subcellular locations. *J. Proteome Res*, 5, 3420–3428.

Chou, K. C, Shen, H. B (2007a) Review: Recent progresses in protein subcellular location prediction. *Analytical Biochemistry*, 370, 1-16.

Chou, K.C, Shen, H.B (2007b) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *Journal of Proteome Research*, 6, 1728-1734.

Chou, K.C, Shen, H.B (2009) Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci*, 1, 63–92.

Chou, K. C, Shen, H. B (2008) Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols*, 3, 153-162.

Chou, K.C, Shen, H.B (2010a) Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Natural Science*, 2, 1090-1103.

Chou, K.C, Shen, H.B (2010b) 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.

Chou, K. C, Shen, H. B (2010c) Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. *PLoS ONE*, *5*, e11335.

Chou, K.C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J Theor Biol*, 273(1), 236-47.

Cleves, A.E (1997) Protein transports: the nonclassical ins and outs. *Curr. Biol*, 7, R318–R320.

Cortes, C, Vapnik, V (1995) Support vector networks. *Machine Learning*, 20, 273-297.

Contag, C.H, Bachmann, M.H (2002) Advances in in vivo bioluminescence imaging of gene expression. *Annu. Rev. Biomed. Eng*, 4, 235-260.

Cormier, M.J, Lee, J, Wampler, J.E (1975) Bioluminescence: Recent Advances. *Annual Review of Biochemistry*, 44, 255-272.

Crick, F.H.C (1958) On Protein Synthesis. *Symp. Soc. Exp. Biol. XII*, 139-163.

Cristianini, N, Campbell, C, Shawe-Taylor, J (2000) Dynamically adapting Kernels in support vector machines. In Advances in Neural Information Processing Systems, MIT Press.

Davies, P.L, Hew, C.L (1990) Biochemistry of fish antifreeze proteins. *FASEB J*, 4, 2460-2468.

Davies, P.L, Sykes, B.D (1997) Antifreeze proteins. *Curr. Opin. Struct. Biol*, 7 (6), 828-834.

Davies, P.L, Baardsnes, J, Kuiper, M.J, Walker, V.K (2002) Structure and function of antifreeze proteins. *Philos. Trans. R. Soc. London B Biol. Sci*, 357 (1423), 927-935.

Denny, P.W, Gokool, S, Russell, D.G, Field, M.C, Smith, D.F (2000) Acylation dependent protein export in Leishmania. *J. Biol. Chem*, 275, 11017- 11025.

Deshaies, R.J, Schekman, R (1987) A yeast mutant defective at an early stage in import of secretory protein precursors into the endoplasmic reticulum. *J. Cell Biol*, 105, 633-645.

Diaz-Uriarte, R, Alvarez de Andres, S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7, 3.

Dipilato, L.M, Cheng, X, Zhang, J (2004) Fluorescent indicators of cAMP and Epac activation reveal differential dynamics of cAMP signaling within discrete subcellular compartments. *Proc. Natl Acad. Sci. USA*, 101, 16513-16518.

Ding Y.S, Zhang T.L (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. *Pattern Recogn Lett*, 29, 1887-1892.

Dixon, A.S, Kakar, M, Schneide, K.M, Constance, J.E, Paullin, B.C, Lim, C.S (2009) Controlling subcellular localization to alter function: Sending oncogenic Bcr-Abl to the nucleus causes apoptosis. *J Control Release*, 140(3), 245-9.

Doolittle, R. F (1986) Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences, Series University Science Books, Mill Valley, CA.

Downward, J (2001) The ins and outs of signaling. *Nature*, 411,759.

Dudoit, S, Fridlyand, J, Speed, T (2002) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97, 77-87.

Eddy, S.R (1998) Profile hidden Markov models. *Bioinformatics*, 14 (9), 755-763.

Evan, G, Littlewood, T (1998) A matter of life and cell death. *Science*, 281, 1317–1322.

Ewart, K.V, Lin, Q, Hew, C.L (1999) Structure, function and evolution of antifreeze proteins. *Cell Mol. Life Sci*, 55 (2), 271-283.

Fetrow, J.S, Skolnick, J (1998) Method for prediction of protein function from sequence using the sequence-to-structure to-function paradigm with application to glutaredoxins/thioredoxins and T-1 ribonucleases. *J. Mol. Biol*, 281(5), 949-68.

Finn, R.D, Mistry, J, Tate, J, Coggill, P, Heger, A, Pollington, J.E, Gavin, O.L, Gunasekaran, P, Ceric, G, Forslund, K, Holm, L, Sonnhammer, E.L, Eddy, S.R, Bateman, A (2010) The Pfam protein families database. *Nucleic Acids Res*, 38 (Database issue), D211-22.

Frank, E, Witten, I.H (1998) Generating Accurate Rule Sets Without Global Optimization. *In: Fifteenth International Conference on Machine Learning*, 144-151.

Frank, E, Hall, M, Trigg, L, Holmes, G, Witten, I.H (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, 20, 2479-2481.

Freund, Y, Schapire, R.E (1996) Experiments with a new boosting algorithm. *In: Thirteenth International Conference on Machine Learning, San Francisco*, 148-156.

Galperin, M.Y, Cochrane, G.R (2011) The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res*, 39 (Database issue), D1- 6.

Garg, A, Raghava, G.P.S (2008) A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity search. *In Silico. Biol*, 8 1-12.

Gene Ontology Consortium (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, 38 (Database issue), D331-D335.

George, H. J, Langley, P (1995) Estimating continuous distributions in bayesian classifiers. *Eleventh Conf. Uncertainty Artif. Intell. San Mateo,* 338-345.

Glucksman, H (1966) On improvement of a linear separation by extending the adaptive process with a stricter criterion. *IEEE Transactions on Electronic Computers*, EC-15 (6), 941-944.

Gonzalez, J.E, Negulescu, P.A (1998) Intracellular detection assays for high-throughput screening. *Curr. Opin. Biotechnol*, 9, 624-631.

Goldberg, D. E (1989) Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley.

Gouridis, G, Karamanou, S, Gelis, I, Kalodimos, C.G, Economou, A (2009) Signal peptides are allosteric activators of the protein translocase. *Nature*, 462(7271), 363-7.

Graether, S.P, Kuiper, M.J, Gagne, S.M, Walker, V.K, Jia, Z, Sykes, B.D, Davies, P.L (2000) Beta-helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. *Nature*, 406, 325-328.

Greene, L.H, Lewis, T.E, Addou, S, Cuff, A, Dallman, T, Dibley, M, Redfern, O, Pearl, F, Nambudiry, R, Reid, A, Sillitoe, I, Yeats, C, Thornton, J.M, Orengo, C.A (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res*, 35 (Database issue), D291-7.

Green, K.A, Lund, L.R (2005) ECM degrading proteases and tissue remodelling in the mammary gland. *Bioessays*, 27, 894-903.

Griffith, M, Ewart, K.V (1995) Antifreeze proteins and their potential use in frozen foods. *Biotechnol. Adv*, 13 (3), 375-402.

Griffith, M, Antikainen, M, Hon, W.C, Pihakaski-Maunsbach, K, Yu, X.M, Chun, J.U, Yang (1997) Antifreeze proteins in winter rye. *Physiol. Plant*, 100, 327-332.

Grønborg, M, Kristiansen, T.Z, Iwahori, A, Chang, R, Reddy, R, Sato, N, Molina, H, Jensen, O.N, Hruban, RH, Goggins, M.G, Maitra, A, Pandey, A (2006) Biomarker discovery from pancreatic cancer secretome using a differential proteomic approach. *Mol. Cell Proteomics*, 5, 157-171.

Guda, C (2006) pTARGET: a web server for predicting protein subcellular localization. *Nucleic Acids Res*, 34, W210-213.

Gunn, S (1997) Support Vector Machines for Classification and Regression. *ISIS Technical Report.*

Haddock, S.H.D, Moline, M.A, Case, J.F (2010) Bioluminescence in the Sea. *Ann. Rev. Mar. Sci*, 2, 293-343.

Haindl, M, Somol, P, Ververidis, D, Kotropoulos, C (2006) Feature Selection Based on Mutual Correlation, *Progress in Pattern Recognition, Image Analysis and Applications*, 4225, 569 – 577.

Hann, B.C, Walter, P (1991) The signal recognition particle in Saccharomyces cerevisiae. *Cell*, 67, 131-144.

Hastings, J.W (1995) Bioluminescence. Academic Press, New York.

Hayes, M.J, Merrifield, C.J, Shao, D, Ayala-Sanmartin, J, Schorey, C.D, Levine, T.P, Proust, J, Curran, J, Bailly, M, Moss, S.E (2004) Annexin 2 Binding to Phosphatidylinositol 4, 5-

Bisphosphate on Endocytic Vesicles Is Regulated by the Stress Response Pathway. *J.Biol. Chem*, 279, 14157-14164.

Heijne, G (1990) The signal peptide. *J. Membr. Biol*, 115, 195-201.

Hinton, G.E, Sejnowski, T.J (1999) Unsupervised Learning and Map Formation: Foundations of Neural Computation, MIT Press.

Horton, P, Park, K.J, Obayashi, T, Fujita, N, Harada, H, Adams-Collier, C.J, Nakai, K (2007) WoLF PSORT: Protein localization predictor. *Nucleic Acids Res*, 35, W585-W587.

Holland, J. H (2001) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, 6 edition, MIT Press.

Hsu, C.W, Lin, C.J (2002) A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Networks*, 13, 415-425.

Huang, R.B, Du, Q.S, Wei, Y.T, Pang, Z.W, Wei, H, Chou, K.C (2009) Physics and chemistry-driven artificial neural network for predicting bioactivity of peptides and proteins and their design. *J. Theor. Biol*, 256, 428-435.

Huang, J, Shi, F (2005) Support vector machines for predicting apoptosis proteins types. *Acta Biotheor*, 53, 39-47.

Hughes, R.C (1999) Secretion of the galectin family of mammalian carbohydrate binding proteins. *Biochim. Biophys. Acta*, 1473, 172-185.

Hunter, S, Apweiler, R, Attwood, T.K, Bairoch, A, Bateman, A, Binns, D, Bork, P, Das, U, Daugherty, L, Duquenne, L, Finn, R.D, Gough, J, Haft, D, Hulo, N, Kahn, D, Kelly, E, Laugraud, A, Letunic, I, Lonsdale, D, Lopez, R, Madera, M, Maslen, J, McAnulla, C, McDowall, J, Mistry, J, Mitchell, A, Mulder, N, Natale, D, Orengo, C, Quinn, A.F, Selengut, J.D, Sigrist, C.J, Thimma, M, Thomas, P.D, Valentin, F, Wilson, D, Wu, C.H, Yeats, C (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*, 37 (Database Issue), D224-228.

Hulo, N, Bairoch, A, Bulliard, V, Cerutti, L, De Castro, E, Langendijk-Genevaux, P.S, Pagni, M, Sigrist, C.J (2006) The PROSITE database. *Nucleic Acids Res*, 34 (Database issue), D227-30.

Jacobson, M.D, Weil, M, Raff, M.C (1997) Programmed cell death in animal development. *Cell*, 88, 347- 354.

Jiang, P, Wu, H, Wang, W, Ma, W, Sun, X, Lu, Z (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features.

Nucleic Acids Res, 35, W339-344.

Jung, J, Ryu, T, Hwang, Y, Lee, E, Lee, D (2010) Prediction of extracellular matrix proteins based on distinctive sequence and domain characteristics. *J Comput Biol*, 17(1), 97-105.

Kain, S.R (1999) Green fluorescent protein (GFP): applications in cell-based assays for drug discovery. *Drug Discov Today*, 4(7), 304-312.

Kandaswamy, K.K, Pugalenthi, G, Hartmann, E, Kalies, K.U, Möller, S, Suganthan, P.N, Martinetz, T (2010) SPRED: A machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes. *Biochem Biophys Res Commun*, 391(3), 1306-11.

Kanehisa, M, Goto, S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28, 27-30.

Kawashima, S, Pokarowski, P, Pokarowska, M, Kolinski, A, Katayama, T, Kanehisa, M (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36, D202–205.

Kersey, P.J, Duarte, J, Williams, A, Karavidopoulou, Y, Birney, E, Apweiler, R (2004) The International Protein Index: An integrated database for proteomics experiments. *Proteomics,* 4 (7), 1985-1988.

Kerr, J.F, Wyllie, A.H, Currie, A.R (1972) Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br J Cancer*, 26, 239-257.

Kim, S.H, Turnbull, J.E, Guimond. S.E (2011) Extracellular matrix and cell signalling - the dynamic cooperation of integrin, proteoglycan and growth factor receptor. *J Endocrinol*, 209(2), 139-51.

Kohavi, R (1995) The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189.

Kohavi, R, John, G. H (1997) Wrappers for feature subset selection. *Artificial Intelligence*, 97, 273-324.

Koonin, E.V, Tatusov, R.L, Galperin, M.Y (1998) Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struc. Biol*, 8, 355.

Krane, D.E, Raymer, M.L (2006) Fundamental Concepts of Bioinformatics, Pearson Education.

Krogh, A, Larsson, B, von Heijne, G, Sonnhammer, E.L (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*, 305, 567-580.

Landriscina, M, Soldi, R, Bagala, C, Micucci, I, Bellum, S, Tarantini, F, Prudovsky, I, Maciag, T (2001) S100A13 participates in the release of fibroblast growth factor 1 in response to heat shock in vitro. *J. Biol. Chem*, 276, 22544 - 22552.

Lau, A.Y, Chasman, D.I (2004) Functional classification of proteins and protein variants. *Proc Natl Acad Sci U S A*, 101(17), 6576-81.

Lee, J.W, Lee, J.B, Park, M, Song, S.H (2005) An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48, 869-885.

Lewin, B, Cassimeris, L, Lingappa, V, Plopper, G, eds (2007) The extracellular matrix and cell adhesion, in Cells, Sudbury, MA, Jones and Bartlett.

Lewitt, J (1980) Responses of Plants to Environmental Stresses, Academic Press, New York.

Li, W, Jaroszewski, L, Godzik, A (2001) Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics*, 17, 282-283.

Liebermeister, W, Rapoport, T, Heinrich, R (2000) Ratcheting in posttranslational protein translocation - a mathematical model. *J. Mol. Biol*, 305(3), 643-56.

Lloyd, J.E (1978) Insect Bioluminescence, Academic Press, New York.

Lipman, D.J, Pearson, W.R (1985) Rapid and sensitive protein similarity searches. *Science*, 227 (4693), 1435-41.

Logsdon, J.M, Doolittle, W.F (1997) Origin of antifreeze protein genes: a cool tale in molecular evolution. *Proc. Natl. Acad. Sci. USA*, 94 (8), 3485-3487.

Marchler-Bauer, A, Lu, S, Anderson, J.B, Chitsaz, F, Derbyshire, M.K, DeWeese-Scott, C, Fong, J.H, Geer, L.Y, Geer, R.C, Gonzales, N.R, Gwadz, M, Hurwitz, D.I, Jackson, J.D, Ke, Z, Lanczycki, C.J, Lu, F, Marchler, G.H, Mullokandov, M, Omelchenko, M.V, Robertson, C.L, Song, J.S, Thanki, N, Yamashita, R.A, Zhang, D, Zhang, N, Zheng, C, Bryant, S.H (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res*, 39 (Database issue), D225-9.

Mason, N, Ciufo, L. F, Brown, J. D (2000) Elongation arrest is a physiologically important function of signal recognition particle. *EMBO J*, 19, 4164-4174.

McDowall, J, Hunter, S (2011) InterPro protein classification. *Methods Mol Biol*, 694, 37-47.

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

Mehul, B, Hughes, R.C (1997) Plasma membrane targetting, vesicular budding and release of galectin 3 from the cytoplasm of mammalian cells during secretion. *J. Cell. Sci*, 110, 1169-1178.

Mignatti, P, Rifkin, D.B (1991) Release of basic fibroblast growth factor, an angiogenic factor devoid of secretory signal sequence: a trivial phenomenon or a novel secretion mechanism?. *J Cell. Biochem*, 47, 201-207.

Mohamad, M.S, Omatu, S, Deris, S, Yoshioka, M (2009) An Iterative GASVM-Based Method: Gene Selection and Classification of Microarray Data, *Lecture Notes in Computer Science*, 5518,187-194.

Moriyama, M, Abe, J, Yoshida, M, Tsurumi, Y, Nakayama, S (1995) Seasonal changes in freezing tolerance, moisture content and dry weight of three temperate grasses. *Grassland Sci*, 41, 21-25.

Mothes, W, Jungnickel, B, Brunner, J, Rapoport, T (1997) Signal sequence recognition in cotranslational translocation by protein components of the endoplasmic reticulum membrane. *J. Cell Biol*, 142(2), 355-64.

Müsch, A, Hartmann, E, Rohde, K, Rubartelli, A, Sitia, R, Rapoport, T.A (1990) A novel pathway for secretory proteins?. *Trends Biochem Sci*, 15, 86-88.

Nelson, C.M, Bissell, M.J (2006) Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. *Annu Rev Cell Dev Biol*, 22, 287-309.

Nelson, D.L, Cox, M.M (2005) Lehninger's Principles of Biochemistry, 4th edition, W. H. Freeman and Company, New York.

Nickel, W (2003) The mystery of nonclassical protein secretion. *Eur. J. Biochem*, 2109-2119.

Palade, G (1975) Intracellular aspects of the process of protein synthesis. *Science,* 329, 347-358.

Panzner, S, Dreier, L, Hartmann, E, Kostka, S, Rapoport, T (1994) Posttranslational protein transport in yeast reconstituted with a purified complex of sec proteins and Kar2p. *Cell*, 81(4), 561-70.

Pearson, W. R (1997) Identifying distantly related protein sequences. *Comp. App. Biosci*, 13, 325-332.

Pearson,W.R, Lipman, D.J (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America,* 85 (8), 2444-8.

Peng, H.C, Long, F, Ding, C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226-1238.

Prados, J, Kalousis, A, Sanchez, J.C, Allard, L, Carrette, O, Hilario, M (2004) Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics*, 4(8), 2320-32.

Qi, Y, Klein-Seetharaman, J, Bar-Joseph, Z (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac Symp Biocomput*, 531-542.

Qiu, J.D, Luo, S.H, Huang, J.H, Liang, R.P (2009) Using support vector machines to distinguish enzymes: approached by incorporating wavelet transform. *J. Theor. Biol*, 256 (4), 625-631.

Quinlan, R (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.

Raff, M (1998) Cell suicide for beginners. *Nature*, 396, 119-122.

Rapoport, T. A (1990) Protein transport across the ER membrane. *Trends Biochem Sci*, 15, 355-358.

Rapoport, T. A (2007) Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*, 450(7170), 663-9.

Raymer, M. L, Punch, W. F, Goodman, E. D, Kuhn, L. A, Jain, A (2000) Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 4(2), 164-171.

Reed, J.C, Paternostro, G (1999) Post mitochondrial regulation of apoptosis during heart failure. *Proc. Natl. Acad. Sci. USA*, 96, 7614-7616.

Reinhardt, A, Hubbard, T (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*, 26, 2230-2236.

Roberts, R. J (2001) PubMed Central: The GenBank of the published literature. *Proceedings of the National Academy of Sciences*, 98 (2), 381-382.

Ross Quinlan, J (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco.

Rothman, J.E, Wieland, F.T (1996) Protein sorting by transport vesicles. *Science*, 272, 227-234.

Rubartelli, A, Bajetto, A, Allavena, G, Wollman, E, Sitia, R (1992) Secretion of thioredoxin by normal and neoplastic cells through a leaderless secretory pathway. *J. Biol. Chem*, 267, 24161-24164.

Saeys, Y, Inza, I, Larrañaga, P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.

Sakai, A, Larcher, W (1987) Frost Survival of Plants. Springer-Verlag, Heidelberg, Germany.

Sanger, F, Nicklen, S, Coulson, A.R (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, 74, 5463 -5467.

Schatz, G, Dobberstein, B (1996) Common principles of protein translocation across membranes. *Science*, 271, 1519-1526.

Schuster, S.C (2008) Next-generation sequencing transforms today's biology. *Nat Methods*, 5(1), 16-8.

Schwartz, M.A, Schaller, M.D, Ginsberg, M.H (1995) Integrins : emerging paradigms of signal transduction. *Annu Rev Cell Dev Biol*, 11, 549-99.

Scholander, P.F, Van Dam, L, Kanwisher, J.W, Hammel, H.T, Gordon, M.S (1957) Supercooling and osmoregulation in Arctic fish. *J. Cell. Comp. Physiol*, 49, 5–24.

Schulz, J.B, Weller, M, Moskowitz, M.A (1999) Caspases as treatment targets in stroke and neurodegenerative diseases. *Ann. Neurol*, 45, 421-429.

Scholkopf, B, Platt, J.C, Shawe-Taylor, J, Smola, A.J (1999) Estimating the support of a high dimensional distribution. Technical Report MSR-TR-99-87, Microsoft Research, Redmond, WA, USA. Online version: http://www.kernelmachines.org/papers/oneclass-tr.ps.gz.

Sformo, T, Kohl, F, McIntyre, J, Kerr, P, Duman, J.G, Barnes, B.M (2009) Simultaneous freeze tolerance and avoidance in individual fungus gnats, Exechia nugatoria. *J. Comp.Physiol.B*, 179(7), 897-902.

Shen, H.B, Chou, K.C (2008) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Analytical Biochemistry*, 373, 386-388.

Shen, H. B, Chou, K. C (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Analytical Biochemistry*, 394, 269-274.

Shi, J.Y, Zhang, S.W, Pan, Q, Cheng, Y.M, Xie, J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids*, 33(1), 69-74.

Shi, J.Y, Zhang, S.W, Pan, Q, Zhou, G.P (2008) Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution. *Amino Acids*, 35, 321-327.

Sigrist, C.J, Cerutti, L, de Castro, E, Langendijk-Genevaux, P.S, Bulliard, V, Bairoch, A, Hulo, N (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*, 38 (Database issue), D161-6.

Smith, C (2008) Subcellular targeting of proteins and drugs http://www.biocompare.com/Articles/TechnologySpotlight/976/Subcellular-Targeting-Of-Proteins-And-Drugs.html.

Sonnhammer, E.L, Eddy, S.R, Durbin, R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3), 405-420.

Sorokin, L (2010) The impact of the extracellular matrix on inflammation. *Nat Rev Immunol*, 10,712-723.

Statnikov, A, Wang, L, Aliferis, C.F (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9, 319.

Stoesser, G, Baker, W, van den Broek, A, Camon, E, Garcia-Pastor, M, Kanz, C, Kulikova, T, Leinonen, R, Lin, Q, Lombard, V, Lopez, R, Redaschi, N, Stoehr, P, Tuli, M.A, Tzouvara, K, Vaughan, R (2002) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*, 30(1), 21-6.

Strobl, C (2005) Statistical Sources of Variable Selection Bias in Classification Tree Algorithms Based on the Gini Index, *Discussion Paper 420, SFB Statistical Analysis of Discrete Structures, Munich, Germany.*

Suzuki, M, Youle, R.J, Tjandra, N (2000) Structure of Bax: coregulation of dimer formation and intracellular localization. *Cell*, 103, 645-654.

Sumner, M, Frank, E, Hall, M (2005) Speeding up Logistic Model Tree Induction. *In: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 675-683.

Svetnik, V, Liaw, A, Tong, C, Culberson, J.C, Sheridan, R.P Feuston, B.P (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci*, 43, 1947-1958.

Tarca, A.L, Carey, V.J, Chen, X.W, Romero, R, Drăghici, S (2007) Machine learning and its applications to biology, *PLoS Comput. Biol*, 3 (6), 954-963.

Trotman, L.C, Achermann, D.P, Keller, S, Straub, M, Greber, U.F (2003) Non-classical export of an adenovirus structural protein. *Traffic*, 4, 390-402.

Ubartelli, A, Bajetto, A, Allavena, G, Wollman, E, Sitia, R (1992) Secretion of thioredoxin by normal and neoplastic cells through a leaderless secretory pathway. *J. Biol. Chem*, 267 (34), 24161-24164.

Urrutia, M.E, Duman, J.G, Knight, C.A (1992) Plant thermal hysteresis proteins. *Biochim. Biophys. Acta*, 1121 (1-2), 199-206.

Vapnik, V (1995) The Nature of Statistical Learning Theory, Springer.

Vogelstein, B, Lane, D, Levine, A.J (2000) Surfing the p53 network. *Nature*, 408(6810), 307-10.

Von Heijne, G (1985) Signal sequences. The limits of variation. *J. Mol. Bio*, 184, 99-105

Ward, W.W, Swiatek, G.C, Gonzalez, D.G (2000) Green fluorescent protein in biotechnology education. *Methods Enzymol*, 305, 672-680.

Wary, K. K, Mainiero, F, Isakoff, S. J, Marcantonio, E. E, Giancotti, F. G (1996) The adaptor protein Shc couples a class of integrins to the control of cell cycle progression. *Cell*, 87, 733-43.

Walter, P, Gilmore, R, Blobel, G (1984) Protein translocation across the endoplasmic reticulum. *Cell*, 38, 5-8.

Walter, P, Blobel, G (1980) Translocation of proteins across the endoplasmic reticulum III. signal recognition protein (SRP) causes signal sequence-dependent and site-specific arrest of chain elongation that is released by microsomal membranes. *J. Cell Biol*, 91, 557-561.

Walter, P, Blobel, G (1982) Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*, 299, 691-698.

White, E.H, Rapaport, E, Seliger, H.H, Hopkins, T.A (1971) The chemi- and bioluminescence of firefly luciferin: an efficient chemical production of electronically excited states. *Bioorg. Chem*, 1, 92-122.

Wilson, T (1995) Comments on the mechanisms of chemi- and bioluminescence. *Photochem. Photobiol*, 62, 601-606.

Wilson, T, Hastings, J.W (1998) Bioluminescence. *Annual Review of Cell and Developmental Biology*, 14, 197-230.

Wu, B, Abbott, T, Fishman, D, McMurray, W, Mor, G, Stone, K, Ward, D, Williams, K, Zhao, H (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19, 1636-1643.

Wu, C.H, Huang, H, Yeh, L.S, Barker, W.C (2003) Protein family classification and functional annotation.*Comput Biol Chem*, 27(1), 37-47.

Wu, C.H, Huang, H, Nikolskaya, A, Hu, Z, Barker, W.C (2004) iProClass: an integrated database of protein family, function, and structure information. *Comput Biol Chem*, 28(1), 87-96.

Yoshida, M, Abe, J, Moriyama, M, Shimosakawa, S, Nakamura, Y (1997) Seasonal changes in the physical state of crown water associated with freezing tolerance in winter wheat. *Physiol. Plant*, 99, 363-370.

Yu, X.M, Griffith, M (2001) Winter rye antifreeze activity increases in response to cold and drought, but not abscisic acid. *Physiol. Plant*, 112, 78-86.

Zernov, V.V, Balakin, K.V, Ivaschenko, A.A, Savchuk, N.P, Pletnev, I.V (2003) Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci*, 43(6), 2048-56.

Zhang, Z.H, Wang, Z.H, Zhang, Z.R, Wang, Y.X (2006) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS Lett*, 580, 6169-6174.

Zhang, Y, Ding, C, Li, T (2008) Gene selection algorithm by combining reliefF and mRMR. *BMC Genomics*, 9, 2, S27.

Zhou, G.P, Doctor, K (2003) Subcellular location prediction of apoptosis proteins. *Proteins Struct Funct Genet*, 50, 44-48.

Zhou, X.B, Chen, C, Li, Z.C, Zou, X.Y (2007) Using Chou's amphiphilic pseudo- amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol*, 248, 546-551.

Krishna Kumar Kandaswamy
University of Luebeck,
D-23538 Luebeck, Germany,
Phone: +4917632276638,
E-mail: krishna@inb.uni-luebeck.de.

Personal Summary:

- Published **15 International publications in different disciplines - Sequence analysis, Biomarker discovery and Protein Engineering**
- Elaborate research experience on **protein translocation assays**
- Lead workflow development team for developing customer hub for **Inforsense** (http://www.inforsense.com) in the area of **statistical genetics, gene expression and data mining in bioinformatics and chemoinformatics**
- Developed **ADME models** for CRO companies

Education:

2008-Present Ph. D in Computational Biology,
University of Luebeck, Germany.
Thesis Title: **Sequence function classification by machine learning methods**
Supervisors: Prof. Thomas Martinetz and Prof. Enno Hartmann

2003- 2006 Engineering in Biotechnology (B. Tech),
Anna University, India.
Thesis Title: **Isolation and cloning of Goat mitochondrial cytochrome b**

2000-2003 Diploma in Chemical Technology (D.Ch.T),
CIT Sandwich Polytechnic College, India
Thesis Title: **Extraction of piperine from pepper by Downstream processing methods**

Experience:

12/2008 - Present Research staff, University of Luebeck, Germany,
12/2008 - Present External Consultant, Systems Biology Worldwide, India
06/2008 - 11/2008 Project Manager, Systems Biology Worldwide, India
08/2006 - 05/2008 Team Leader, Insilico Consulting, India
04/2006 - 07/2006 Project Assistant, National Chemical Laboratory (NCL), India

International Conferences/workshop:

- **Kandaswamy, K.K**, Pugalenthi, G, Hazrat, M.K, Kalies, K, Möller, E, Martinetz, T. BLProt: Prediction of Bioluminescent proteins based on Discrete Wavelet Transform and Support Vector Machine. The annual international conference on Intelligent Systems for Molecular Biology (ISMB), Vienna, Austria, 2011.

- **Kandaswamy, K.K**, Pugalenthi, G, Möller, S, Hartmann, E, Kalies, K, Suganthan, P, Martinetz, T, Prediction of subcellular localization of apoptotic protein with Genetic Algorithms and Support-Vector-Machines, Fourteenth International Conference on Research in Computational Molecular biology, Lisbon, Portual, 2010.

- Shameer K, Pugalenthi G, **Kandaswamy, K.K**, Suganthan PN, Archunan G, Sowdhamini R, Prediction of domain swapping events from protein sequence using support vector machine approach, The Eighth Asia Pacific Bioinformatics Conference, India, 2010

- **Kandaswamy, K.K**, Hartmann, E, Martinetz, T, Fladea, I, Moeller, S. Can we predict the translocation pathway from the sequences of signal peptides? 1st International ISoLA Workshop on Modeling, Analyzing, Discovering Complex Biological Structures, Berlin, Germany, 2009.

Service as a Reviewer:

Refereed manuscript(s) for the following international journals

Bioinformatics, BMC bioinformatics, Amino acids (Springer publishers), Pattern Recognition (Elsevier), Pattern Recognition letters (Elsevier), Protein & Peptide Letters (Bentham science publishers),

Awards and Achievements:

- Doctoral Fellowship awarded by the Graduate School for Computing in Medicine and Life Sciences, Luebeck, Germany.
- Received the 2010 Most Cited Paper Award from Pattern Recognition Letters journal.
- One of our paper chosen as Fast Breaking Paper in the field of Engineering selected by Thomson Reuters' Essential Science Indicators.
- Scope of Work = 154 citations (Data mining , Bioinformatics, Engineering).
- Received a certificate (Tierversuchskunde-Nachweis) to operate animal experiments with mice.
- Associate editor for International Journal of Advanced Bioinformatics applications and Research and International Journal of Biotechnology and Biosciences.

***Reference*:**

1. Prof. Thomas Martinetz
   Institute for Neuro- and Bioinformatics
   University of Luebeck
   D-23538 Lübeck, Germany
   Phone: +494515005500
   Fax: +494515005502
   Email: martinetz@informatik.uni-luebeck.de

2. Prof. Enno Hartmann
   Centre for Structural and Cell Biology in Medicine
   Institute of Biology
   University of Lübeck
   D-23538 Lübeck, Germany
   Phone: +494515004100
   Fax: +494515004815
   Email: ennohart@bio.uni-luebeck.de