# PROTEIN FUNCTION AND INHIBITOR PREDICTION BY STATISTICAL LEARNING APPROACH

*Founded 1905*

**HAN LIANYI**

*(M.Sc. ChongQing Univ.)*

**A THESIS SUBMITTED**
**FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**
**DEPARTMENT OF COMPUTATIONAL SCIENCE**
**NATIONAL UNIVERSITY OF SINGAPORE**

**2005**

# ACKNOWLEDGEMENTS

I would like to present my sincere thanks to my supervisor, Professor Chen YuZong, for his invaluable guidance and being a wonderful mentor and friend. I have benefited tremendously from his profound knowledge, expertise in research, as well as his enormous support. My appreciation for his mentorship goes beyond my words.

I would like to thank Ms. Har Jiayi for her collaboration and resourceful suggestions in my project for doing HIV PIs prediction. This project cannot be well fulfilled without her contributions.

I also gratefully acknowledge Prof Martti Tammi, Prof Low Boon Chuan and Prof Meena Sakharkar for their invaluable suggestions and helpful comments about this work.

Special thanks go to our BIDD Group members. In particulars, I would like to thank Dr. Cao Zhiwei, Dr. Ji Zhiliang, Dr. Chen Xin, Dr. Yap ChunWei, Ms Sun LiZhi, Mr Wang JiFeng, Ms. Zheng Chanjuan, Ms Yao LiXia, Mr. Lin Honghuang, Mr. Li Hu, Mr. Ung CY, Ms. Cui Juan, Ms.Tang Zhiqun, Ms. Zhang Hailei, Mr.Xie Bin etc. and our research staffs: Dr. Cai CongZhong, Dr. Li ZeRong, and Dr. Xue Ying. Without their help and group effort, this work cannot be properly finished.

I am profoundly grateful to my parents and my wife for your love, encourage and accompany.

A special appreciation goes to all my friends for love and support.

# TABLE OF CONTENTS

# SUMMARY

A fundamental understanding of how biological systems work requires knowledge of the proteins and interactions of biomolecules. The role of proteins as well as small molecules participating in interactions can be interpreted as their functions. This is becoming an increasingly important means for better understanding of biological process and for facilitating modern drug discoveries. This thesis presents the predicting of protein functional families and protein inhibitors by statistical machine learning approach.

Development of methods and computational tools for the prediction of functional families of protein is one of the main objectives of this study. Protein function classification systems were designed to assign functional families from proteins' primary sequence irrespective of sequence similarity. In this work, a number of protein classification problems such as enzyme families, transporter families and RNA-binding proteins were studied and the classification models were further evaluated by using independent evaluation sets. The independent evaluation results showed a prediction accuracy above 70% for 53 out of 72 protein functional families in this study.

In order to evaluate the capability of the prediction system for assigning functional class of proteins without any sequence similarity in protein sequence databases and proteins with similar sequence but different functions, novel proteins from bacterial, viral and plant species were selected and tested to examine to us what extent, their function can be predicted by using our prediction systems. It was shown that the accuracy for predicting their function is in an acceptable range of 67% ~ 85%, whereas other approaches solely based sequence similarity approach may not suitable for this task. These results suggest that an SVM-based prediction system is useful for facilitating the prediction of the function of novel proteins in the genomes of bacteria, virus, plants as well as other organisms and major functional groups, such as enzymes.

Another aim of this work is to predict protein inhibitors by statistical learning approach in order to cope with an increasing need of the discovery of inhibitors of therapeutically important proteins, particularly those with crystal 3D structures available. These inhibitors can be used as potential leads for drug development. Prediction of HIV-protease inhibitors (PIs) is used as an example, as it is of relevance of drug discovery and there are substantial structures and inhibitors to develop a statistical machine learning system. In the current use of HIV-1 protease inhibitors for anti-HIV therapies, the main concerns are the rapid emergence of drug resistance and many physiological side effects. Thus it is in high demand for speeding up drug discovery in the fight against with HIV infections by properly choosing HIV PIs candidates. In this study, a set of 4291 inhibitors and 10000 non-inhibitors were selected to develop a SVM classifier, which gave a prediction accuracy of 97.05% for a random selection of independent evaluation set composed of 3424 compounds. This result suggests that the classification model is self-consistent and has certain capability in the selection of probable HIV-1 PI candidates. Recursive feature selection has been employed to select significant molecular descriptors and it was shown that molecular connectivity and shape, flexibility, and hydrogen bond interactions are among the most distinguishing features for discriminating HIV-1 protease inhibitors. The results of this study indicate that the statistical learning approach is useful for PIs prediction, the methods implemented in this work can be extended to the other inhibitor/agonist/substrate prediction problems.

# LIST OF TABLES

# LIST OF FIGURES

# 1. Introduction

Knowledge of proteins is essential in the understanding of biological processes such as gene regulation and disease pathology[1, 2]. The demand and possibility for probing protein function and interactions with other biomolecules have been increasing along with the progress of genomics and proteomics. Resulting from large-scale genome sequencing projects, the gap between the large amounts of sequences information and their function characterization is continuously increasing[3, 4]. Thus, the understanding of protein function is important for facilitating drug target search, drug discovery and systematically study of biological events. The availability of the flood of biological information brings us both the chance and the challenge to probe the knowledge of the biomolecules interactions, proteins function and biological process, which not only helps us to understand and interpret the biological events in the molecular level but also enables us to study regions which are not accessible experimentally or which would imply very expensive experiments. Prediction of protein functions and protein inhibitors (normally protein inhibitors are referring to molecules that can inhibit the protein functions ) are two challenges in biology and drug discovery, that are investigated by a statistical learning method – Support Vector Machines in this thesis.

## 1.1. Introduction to protein function prediction

Increasing effort has been directed for predicting protein functions from their sequence. Various methods have been used for protein function prediction from their sequence, such as sequence similarity searching[5-7], evolutionary analysis[8, 9], structure-based approach[10], protein/gene fusion[11, 12], protein interaction[13, 14] and family classification by sequence clustering[15, 16].

Methods based on sequence similarity, such as *FASTA*[17], *BLAST*[18], Motifs[19] and Prosite[20], have frequently been used for protein function prediction. However, with decreasing in sequence similarities, the criteria for comparison of distantly-related proteins become increasingly difficult to formulate [16]. Moreover, not all homologous proteins have similar functions [8]. Even a shared domain within a group of proteins does not necessarily imply that these proteins have the same function[21]. These problems often hinder some of the sequence similarity based methods [15].

Unlike sequence similarity based approach, structure-based methods can determine protein function from the structure function relationship without solely relying on sequence similarities. Although the structure information may provide insights into protein function[22], a hypothetical function obtained by identifying the similar 3D folds in the absence of clear sequence identity does not reflect the real function with high confidence[23-26]. Structure-based approaches are not limited in finding clues between function and similar 3D folds. Several other approaches, such as structure descriptors[27], patterns in non-homologous tertiary structures[28] and geometric hashing[29], have been successfully implemented by using 3D templates known to be associated with functions to scan new structures against the profile library. However, the limited ability to locate 3D profiles automatically and the restriction of sequence variation of 3D templates methods[30] are the practical drawbacks of these methods.

Apart from the methods for determining specific protein function on the basis of similarities either in structure or in sequence, another approach to predict protein function is to classify proteins into their functional families on the basis of their sequences, which is expected to be particularly useful in the cases described above. To fulfill the task of protein functional families classification for facilitating protein function prediction, artificial intelligence statistical learning methods, such as support vector machine (SVM)[31-33] and neural network[34], have been reported. The strategy normally used is that samples of proteins in a functional family and those outside the family are used to train a system for protein classification. And the preliminary results[31-34] suggest that Support Vector Machine can be trained and used to recognize proteins with characteristics for a particular function if there are sufficient samples of proteins with specific function.

In summary, there are three principal strategies, sequence similarity based, structure based and statistical learning based methods relying on sequence or structures, to estimate function of a protein by using bioinformatics approaches.

### 1.1.1. Sequence similarity based approaches

As introduced in the previous section, various approaches have been implemented for facilitating the protein function assignment for the primary sequence, such as sequence alignment, clustering and pattern identification, remote homology searching, statistical methods and artificial intelligence. The most prominent and commonly used one among them is sequence alignment method. Based on sequence-structure-function relationship, proteins with high similarity in sequence are more likely to have the similarity in structure and function. This method normally starts by aligning the sequences of proteins with unknown function and proteins with known function together with a certain level of sequence similarities. By determining the level of sequence similarity, one can predict the potential functions.

As early in 1970, Needleman-Wunsch algorithm was proposed by Saul Needleman and Christian Wunsch[35] for solving the global pairwise sequence alignment problem where all the characters in both sequences participate in the alignment. Another famous algorithm, Smith-Waterman algorithm was first proposed by Temple Smith and Michael Waterman in 1981[36] for performing local sequence alignment to find related regions within sequences.

Pairwise sequence alignment methods are concerned with finding the best-matching piecewise local or global alignments of protein (DNA) sequences, however, it could be time consuming to perform a large sequence database scan in order to identify the sequences homologous.

In order to cope with the task of large-scale sequence database searching, *FASTA*[17] was proposed by David J. Lipman and William R. Pearson in 1985, which was latter superseded by *BLAST*[18] proposed by Stephen Altschul *etc* in 1990. *BLAST* became the most widely used bioinformatics programs because it addresse a fundamental problem and the algorithm emphasizes the balance between the speed and sensitivity. It is an important fact that biomolecules could share the similar structures and functions even if their sequences have low level of similarity or if they are dissimilar. In order to find distant relatives of a protein and identify weak but biologically relevant similarities, *PSI-BLAST*[37] has been introduced by Altschul and Koonin in 1998. It iteratively searches protein databases for sequences similar to one or more protein query sequences. *PSI-BLAST* is similar to *BLAST* except that it uses position-specific scoring matrices derived during the search. In addition to the usual PSI-blast criteria for matching, Pattern-Hit Initiated BLAST[38] (*PHI-BLAST*) is introduced to enforce the presence of a pattern in database searching for protein sequences that also contain the input pattern and have significant similarity to the query sequence near the pattern occurrences.

In many cases, a protein can perform certain functional activity if it contains a conserved sequence[20], thus motif based methods, such as Motifs[19] , Prosite[20] and

Sequence Clustering[15] that have been developed in recent years, also show certain capability of identifying proteins with weak similarities by using patterns, rules and profiles search.

However, identification of protein functions solely based on the sequence similarities is impractical for proteins without any homology in sequence[16]. In addition, proteins with similar sequences may not have similar functions[8]. Although the motif/pattern based methods could cluster proteins by identifying shared domains within a functional group, it does not necessarily imply that clustered proteins have the same function[21].

### 1.1.2. Structure based approaches

Unlike sequence-based approaches, structure–based approaches rely on the analysis of the protein 2D/3D structures. Based on assumption that proteins with similar structure have similar functions, one can predict the protein function or get clues on protein function from its structure.

Based on the knowledge of structure-function relationship, one can infer function from the corresponding protein structure[22]. Homology modeling approaches[27-29, 39] have been successfully implemented by using 3D templates known to be associated with functions to scan new structures against the profile library. However, the restriction of sequence variation in the templates[30] is the main limitation.

By studying the relationships between protein fold and functions, one is able to analyze the protein functions from the shared protein folds[40]. However, there are two concerns. Firstly, function identification that solely relies on the homologous fold identification without considering sequence similarity is of low confidence[23-26]. Secondly, the relationship between the 3D folds and protein function is usually very complex, and even ambiguous in many cases[41].

The gap between the amount of protein sequences and solved protein structures is increasing rapidly. Although a combination of techniques such as comparative protein modeling and experimental protein structure determination techniques[42] are widely used to determine protein structures, only about 15% of sequenced protein have 3D structures. The lack of solved structures limits the application of structure-based methods for predicting protein functions.

### 1.1.3. Statistical learning based approach

The sequence similarity based approaches and structure based approaches require certain similarities in their sequences or their structures. Thus it is necessary to look for alternative approaches to predict the protein function without considering similarities in either structures or sequences. Statistical learning based approach is one potential solution to address this problem.

Various statistical learning approaches have been developed to explore protein functions from its primary sequence by using statistical learning methods including discretized naïve Bayes, C4.5 decision trees, and instance-based leaning[33], neural networks[34] and support vector machines (SVM)[31-33, 43-46]. These methods rely on the model generated by training the protein examples from a specific functional class and negative examples outside the functional class. The features representing the protein sequence information have been obtained by several methods such as binary coding, amino acid composition, hydrophobicity, normalized Van der Waals volume, polarity, polarizability or their combinations[14, 31, 43, 47-49]. Some of these methods, use sequence derived features without considering sequence similarities, are capable of facilitating protein function prediction without considering sequence similarities.

The statistical learning approaches require certain number of representative examples for learning. Thus the effective data collection and negative examples selection are very important to obtain pre-classified functional protein examples and representative negative examples. However, the problem of effective examples remains unsolved.

## 1.2. Introduction to protein inhibitor prediction

Many drugs target on enzymatic proteins and act as competitive inhibitor of the enzymes, are commonly referred to as inhibitors[50]. Interactions between inhibitors and proteins such as enzymes and carrier proteins can be either reversible or irreversible. One of the common roles for inhibitors' activity is to hinder its target protein's normal reaction or to regulate the function of its target. For example, the cyclo-oxygenase inhibition by aspirin that irreversible acetylates a serine residue at the top of the main cytoclooxygenase site[51]; HIV-1 protease inhibition by indinavir, which block its peptide binding, site to prevent the binding of its peptide[51]. While not all inhibitors can be used as valid drugs due to the unwanted effects and poor pharmacokinetic properties, prediction of protein inhibitors is important for finding drug leads, probing protein inhibition mechanisms and designing better drugs and for protein enginering. Intensive efforts on designing inhibitors have lead to the advent of computer aided drug design[52-55], that aims to help the rapid and efficient discovery of drug leads.

Many existing computational approaches focused on the improvement of interaction between target proteins and their inhibitors. One approach studies the relationship between protein and its inhibitors to simulate the interactions and binding activities of protein-substrate system by finding if there is a stable energy minimum by protein-ligand docking approach[56], which requires 3D structures of both proteins and substrates. Other methods widely used to speed up the inhibitors identification in the early stage of drug discovery are statistical learning methods[57-60] and Quantitative Structure Activity Relationship(QSAR)[61-64] study. These approaches can be used to speed up the drug development circle by eliminating false drug leads in earlier stage. Various approaches have their requirements for achieving the study objective. Thus, it is necessary to have a close look on these approaches for facilitating protein inhibitor research.

### 1.2.1. Quantitative Structure Activity Relationship (QSAR)

It has been a century since Crum-Brown and Fraser proposed the idea that the physiological action of a substance is a function of its chemical composition and constitution[17] and about 40 years since the quantitative structure-activity relationship (QSAR) paradigm was practically used in chemistry and pharmacology[65]. Quantitative Structure Activity Relationship (QSAR) stands for the quantitative study of relationships between molecules' physical-chemical properties and their biological activities. In other words, QSAR is to study molecule behaviors in a biological event.

QSAR can be used to identify chemical structures that have good inhibitory effects on specific protein target. Optimal molecular properties are considered to develop the relationship between a list of compounds structure and their quantitative activities. And this relationship can be used to predict quantitative activities of new compounds from their structures. Unlike the docking and other molecular modeling approaches, the 3D structure of the protein target is not required.

QSAR process provids the usefully clues of which descriptors are important for the biological response. For example, the LogP is an important measure used in identifying "drug-likeness" according to Lipinski's Rule of Five[66], the LogP of 2.77-3.76 was found to be ideal for LOX inhibitors[67]; a logP value of 2.92 or higher, 18-atom-long or longer molecular length and a high Ehomo value etc are required for an effective p-glycoprotein inhibitor[68]; other important measures like chi (first-order Randic connectivity index) is for identification of carbonic anhydrase inhibitors[69]. The proposed important descriptor during the QSAR analysis can be used as a rule for virtual screening the new inhibitors that are likely to produce the desired activities.

Normally the development of QSAR model is based on a group of compounds with certain common structure, the diversity of the studied compounds is not enough for predicting novel inhibitors without the common structure. Thus, the use of QSAR for novel inhibitors design might not adequate as it requires a large number of compounds with experimental activity data to develop many QSAR models.

### 1.2.2. Molecular Docking Approach

Molecular docking is a widely used technique for screening and rapid testing of large amount of compounds to identify new binders of a selected protein target[56]. The identified new binders are candidates of new drug leads. It is an advance for docking brought by the development of empirical force fields. The automated docking techniques allow de novo drug design with the capacity of allowing assessment of relative binding strength and drug specificity[70].

This approach has been used widely in probing new inhibitor candidates. DesJarlais [71]suggested that the Targeted-DOCK can be used for the design of a novel non-peptide inhibitor of HIV-1 protease. Benzylamino acetylcholinesterase inhibitor-like compound screening is another successful application of docking approach by Yamamoto[72]. Other studies of protein inhibitors, such as human rhinovirus-14 inhibitors[73], glucoamylase inhibitors[74], thrombin inhibitors [75, 76] etc, especially the study of HIV protease inhibitors[70, 71, 77-79] which attracts a lot of interests, show that docking approach can be used for inhibitor screening.

However, the use of molecular docking approach requires 3D structure of the target proteins, which is essential for calculating the binding affinity from molecular mechanics/modeling. Because there are only limited number of proteins with 3D structures available, the molecular docking approach is not applicable in many other cases. Moreover, molecular docking normally prefers the conformation of the binding site of the protein target is rigid other than flexible, thus the flexibility of the protein structure can affect the screening accuracy.

### 1.2.3. Statistical learning approaches for protein inhibitor prediction

Statistical leaning methods have been applied in QSAR studies for facilitating inhibitors identification as the implementation of relationship analytical mothods[80-83]. On the other hand, the direct use of statistical learning methods for this purpose mainly focused on classification, such as distinguishing between inhibitors and non-inhibitors, or regression analysis between the molecular structure and the measurement of inhibition[57-60]. One of the advantages is that the direct use of statistical learning methods do not require the 3D structure of protein target, thus these methods are potentially applicable to the case that the target structure is unknown or very flexible. Another advantage of statistical learning methods for protein inhibitor prediction is the diversity in training samples, which allows us to predict diversified compounds.

Douali *et al* [80] approach the prediction of anti-HIV activity of HEPT by use of neural networks. Daszykowski *et al* [57] analysis of biological activity of Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs) by using tree based approach - Classification And Regression Trees. Mager[82] overview the work for using the neural approach to optimize the desired actions and to lower the side effects of non-nucleoside HIV-1 reverse transcriptase inhibitors.

However, a well-trained statistical learning model requires more inhibitor samples that QSAR approach to construct the decision function. Moreover, the proper selection of non-inhibitors is also very important because the decision function of statistical learning methods is usually determined by both positive and negative samples. Unfortunately, this problem remains unsolved because the compounds are enormous in numbers and they are very diverse. In work, we are going to approach this problem as well as other important issues such as data unbalance problem, predominant feature selections.

One of the well-known examples in the field of rational drug design is the discovery and development of drugs for the treatment of AIDS[84]. The major targets for the development of new chemotherapeutic agents are Protease, Intergrease, and Reverse Transcriptase. Protease inhibitors are known as effective antiviral agents in increasing the effectiveness of antiretroviral therapy and prolonging the survival of patients with HIV infection/AIDS. Thus, development of new HIV PIs is also in high demand for anti-HIV therapy. However, due to the poor pharmacokinetic properties and side effects, the discovery of novel PIs is a difficult task. In this study, the prediction of HIV PIs is taken as an example to illustrate our approach for protein inhibitors predictions.

## 1.3. Introduction to HIV protease inhibitors prediction

As of December 2004, an estimated 39.4 million ~ 37.2 million adults and 2.2 million children younger than 15 years – are infected with Human Immunodeficiency Virus (HIV) or living with AIDS. The rate of increase of the new infection is alarming. An estimation of 4.9 million new HIV infections occurred worldwide during 2004, amounting to about 14,000 infections each day[85]. In view of the huge worldwide impact of AIDS and the spreading speed of the AIDS pandemic, there have been intense global efforts towards understanding the biology and life cycle of HIV-1 and the host response to HIV-1 infection. These advances have led to the development of several new drugs that target the viral life cycle which are effective against HIV-1.

Currently, there are 20 approved antiretroviral agents for anti-HIV-1 clinical therapy[86], and each of those drugs could target one of the two viral enzymes protease or reverse transcriptase. Although the cocktail method[87] is introduced, the success of treatment is still limited due to the HIV-1 target drug resistant mutations[88, 89] which is the main cause of anti-HIV drug failure. Besides the drug resistant mutations that occurred in long term therapy, protease inhibitors are known as effective antiviral agents to increase the effectiveness in antiretroviral therapy and to prolong the survival of patients with HIV infection/AIDS. Efforts have been directed to development of new HIV protease inhibitors that could be potentially used for anti-HIV therapy. Development of new HIV PIs is also in high demand for anti-HIV therapy because the appearance of drug-resistant mutants and even multi-drug-resistance mutants is the main cause of the drug failure. Thus, it is time to have a clear look on HIV protease and its inhibitors.

### 1.3.1. HIV protease and protease inhibitors

The HIV-1 protease is responsible for the maturation of new infectious HIV particles. It cleaves the Gag protein to yield the functional core proteins, i.e. the capsid protein, matrix protein, and nucleocapsid protein. It also synthesizes the polymerase protein (Pol) of HIV-1 as a Gag-pol (Pr160$^{\text{Gag-Pol}}$) fusion polyprotein[90, 91].

HIV-1 PI inhibits the protease from properly cleaving Gag-pol polyprotein into its smaller functional units. The currently available HIV-1 protease inhibitors (PIs) can be classified into two broad classes[85, 86]: 1) Peptide-based inhibitors, which can be subdivided into peptides, peptidomimetics and symmetry-based inhibitors; and 2) non-peptide based inhibitors.

Peptides are short amino acid polymers in which the individual amino acid residues are linked by amide bonds (CO-NH). In this study, amino acids, amines and amides are categorized under peptides. Amines are compounds containing one or more substituents that are organic bonded to a nitrogen atom, i.e. RNH2, R2NH or R3N. Examples of amines among the positive samples are aminoglycosides, benzimidazole, indoles, pyrroles and decahydroisoquinolines. Amides are compounds containing –CONR2 functional groups, such as carboxyamides and sulfonamides[92].

Peptidomimetics are protease substrate analogues that have a non-hydrolysable amino acid at the scissile bond. They have been designed to mimic the tetrahedral transition-state intermediate formed during the HIV-1 PR catalysis event. The transition state of the aspartic proteinase-catalyzed reaction occurs with the addition of a water molecule, coordinated by the active site of aspartates, to the peptide bond. These substrate-based inhibitors have many chemical forms, but they assume similar conformations in the substrate-binding cleft of the protease[93]. Examples of peptidomimetic drugs approved by FDA, are Saquinavir (Ro 31-8959) and Indinavir

(L-735, 524).

C2 symmetry and pseudo-symmetry drugs are also peptide-based, they have less peptidic nature and they exploit protease-specific symmetry of the active site. Although symmetry is not thought to be an absolute requirement for the design of HIV PIs, these drugs were designed as an improvement of peptidic drugs with the expectation that the less peptidic nature of inhibitors might enhance stability. An example of symmetry-based drug is Ritonavir (ABT-538).

Non-peptidic inhibitors are inhibitors with moieties to displace water molecules in the active site cleft. Specifically, the binding features of the surrounded water are incorporated into the inhibitor. These classes of compounds have proved to be quite promising, and their discovery has provided a new starting point for designing of HIV-1 PR inhibitors. However, no inhibitor from this group is in clinical use yet.

The United States Food and Drug Administration (FDA) has approved nine protease inhibitors for marketing in the United States since the release of Saquinavir in 1995. As a part of the Highly Active Antiretroviral Therapy (HAART), all of the HIV-PIs are used in combination with other antiretroviral agents for the treatment of HIV-1 infection.

### 1.3.2. Current problems with the use of HIV-1 PIs

While existing HIV-1 PIs show promising results in antiretroviral therapy and prolonging the survival of patients with HIV infection/AIDS, most patients taking protease inhibitors alone show an increase in plasma viral RNA to near baseline levels by the end of the year of drug administration[94] and the occurrence of PI-resistance HIV. It has been discovered that there are two major problems related to the use of HIV-1 PIs, drug resistance and side effects due to drug toxicity.

Resistance mutations in the protease gene may result from amino acid substitutions at or near the active site. This interferes with inhibitor binding because of conformational perturbations and the properties change around the active binding site. Substitution of amino acids lying outside the active region compensates for the deleterious effects of primary mutations[95, 96].

Resistance to PIs can emerge rapidly when these inhibitors are administered at inadequate doses or as part of suboptimal regimens[97]. The interpretation of protease mutants is further complicated by the extensive polymorphisms found in the protease gene of HIV-1 isolates from untreated patients. In one study, variation was noted in nearly 48% of protease codons compared with the consensus (wild-type) sequence[98]. The significance of these polymorphisms in determining treatment outcome remains uncertain, since most studies have not found any correlation between the presence of these polymorphisms and virologic response, or the rate at which PI resistance emerges.

One other shortcoming of the present treatment involving protease inhibitors is the adverse effects, drug interactions, and other risks associated with their use. Generally, all protease inhibitors may cause hyperglycemia, diabetes mellitus and redistribution or accumulation of body fat and may increase the risk of bleeding in patients with hemophilia. They are also the causes of gastrointestinal adverse events such as nausea and diarrhea.

Other adverse reactions occur less commonly, and some are primarily associated with the use of a particular protease inhibitor. The widely used HIV-PI Saquinavir was found to be the most toxic in majority of cell types[99]. Atazanavir causes asymptomatic hyperbilirubinemia, which may be accompanied by jaundice in many patients, although it is reversible upon discontinuation of treatment. The use of Ritonavir and

Lopinavir/Ritonavir has been associated with large increases in total cholesterol and triglyceride concentrations, and in some cases, pancreatitis. Some patients treated with Amprenavir have experienced severe and life-threatening skin reactions, including Stevens-Johnson syndrome. Thus the development of new effective PIs for antiretroviral therapy with less toxicity and improved enzyme-inhibitor interaction is in high demand.

## 1.4. Introduction to Statistical learning methods

The key concepts of the learning methods are data and hypotheses[100]. As such, statistical learning methods are capable of learning from the evidence and predicting the new observations. The mathematical analysis of the learning process began when the first learning machine, Perceptron, was suggested by F.Rosenblatt in 1960s[101]. The Perceptron addressed the pattern reorganization problem by generalizing rules from given examples for recognizing their specific patterns. The Perceptron was soon widely known as it brought a general model of learning phenomenon. Over the past 50 years, a number of machine learning methods have been introduced for solving real-life problems, for examples, Decision Trees, Hidden Markov Model, Neural Networks and Support Vector Machines.

From the conceptual point of view, statistical learning methods are carried out in two flavors: supervised learning and unsupervised learning. During supervised learning, the observations are divided into two groups: explanatory part and one (or more) dependent part that was treated as the consequence of the explanatory part. The purpose of the learning process is to specify a relationship between the explanatory part and the dependent part. The application of supervised learning requires a sufficiently large number of data. Approaches under this category such as K-Nearest neighbor, Linear Learning Machines, Support Vector Machines, Probalistic Neural Networks, etc. were widely applied in the field of pattern reorganization. During unsupervised learning, all data under investigation are allowed to speak for themselves and they are treated evenly. They are forming naturally without any interference, i.e. the unsupervised learning methods do not happen to have advanced indication of correct or incorrect answers; instead, they adjust through direct confrontation with new experiences. This learning process is called self-organization. Many machine learning methods, such as Self Organization Map, clustering methods including both hierarchal clustering and partitional clustering, are implemented in the unsupervised manner.

Many statistical learning algorithms have been successfully applied in the pattern reorganization problems such as text reorganization and protein function classifications. In the following several sections, we will focus on some of the machines learning algorithms that have been employed in solving biological problems.

### 1.4.1. K- Nearest Neighbor

Learning from the observations is the centre of machine leaning system. KNN is an intuitional approach to demonstrate such learning process. An important feature of KNN is instance orientation. The decision procedure of KNN is very simple and intuitional by assuming that observations that are close together will share the same domain. The learned observations are pre-labeled while the new observation will be evaluated based on a similarity measure. The conclusions are based on the rule of “majority wins” voted by the K nearest neighbors closest to the new observation, whereas the remaining pre-labeled observations will not be considered for making decisions. The K, number of nearest neighbors, is a manageable variable optimized during the model training. Practically, K should be smaller with respect to the number of observations in order to make the data points close enough to produce an accurate estimate of the new observations. On the other hand, the K should be large enough to minimize the misclassification error due to biased examples involved in decision-making process.

Various forms of K-nearest neighbor methods have been applied widely in dealing with biological information. Because of its conceptual simplicity and good performance in particular problems, it has become a basic method for solving information centric problems such as pattern reorganization problems in bioinformatics. Moreover, it is usually selected as a benchmark tool for comparison.

The problem setup of KNN in the analysis of biology data is mostly for pattern recognitions, such as the detection of ventricular arrhythmia[102], the study of Quantitative Structure-Activity Relationship(QSAR)[62, 103-106], the classification of protein families based on certain characteristics such as protein function[107] and protein allergenicities[108].

The similarity measure used in KNN could be a drawback, because it treats all features equally based on computational similarities of distances. Since the nature of KNN is that only K nearest neighbors is considered for decision-making, this probably can lead to poor classification accuracy.

### 1.4.2. Clustering Methods

No matter how the learning problem is complicated, the information that the machine are learning could be enormous. Clustering method is one of the statistical learning approaches to reduce the amount of data by categorizing or grouping similar data items together.

Clustering methods[109-115] come in two basic types: hierarchical and partitional clustering. There exist a wealth of subtypes and different algorithms across a wide variety of communities for these two basic types of clustering methods.

Hierarchical clustering is implemented either by merging small clusters into larger ones, or by splitting large clusters into small ones. The clustering methods differ in the strategies on deciding which two small clusters should be merged together or which large cluster should be further divided. The end of the clustering procedure is a tree of clusters, which is also called a dendrogram. The obtained clusters are related together by sharing the root, which is like a tree with many branches and leafs. By cutting the dendrogram at a desired level, one can obtain a clustering of the data items into disjoint groups. Partitional clustering, on the other hand, attempts to decompose the data set into a set of disjoint clusters. The clustering algorithm tries to minimize the objective function by assigning clusters to the peaks in the probability density function. One of the partitional clustering algorithms is K-means clustering which is minimizing dissimilarity in the samples within each cluster and meanwhile maximizing the dissimilarity between clusters.

Many biological problems require the information categorizing to extract hints or clues for interpreting biological phenomenon. Such as the study of genotypic and phenotypic relationships[116, 117], Clustering receptors[118, 119], disease feature clustering[116, 120] etc. Although it is useful to abstract the flood of biological information for extracting easy understandable rules, it should be used with caution when the problem domain is very complex. The knowledge exploration of clustering approach requires little or no prior knowledge and start from the understanding of the whole data set, which makes the clusters very difficult to maintain. Grouped clusters based on the distance similarity can be easily affected by the input data with poor similarity measure.

### 1.4.3. Decision Trees

The Decision Tree is a popular machine learning algorithms in the application of data mining and pattern reorganization. Compared with many other machine learning approaches, such as neural networks, support vector machines and instance centric methods, Decision Tree is simple to construct efficient in decision making. It can produce human readable and interpretable rules. These rules provide an insight into the problem domains.

Decision Trees generate a series of rules from the input examples and then apply these rules to new examples for prediction. These rules are linked together and are shaped in a tree structure. The working flow starts from the topmost node and every decision of the node determines the direction of next node movement until the end of the tree branch node is reached. Therefore, the topmost node is the root of the decision tree, the variable playing this role is evaluated first as everything should start from the root of the tree. The variable on the root of the decision tree is one of the highest information gains. That is where the constriction of Decision tree starts form. Branches nodes of Decision trees can be calculated in the same way as a recursive procedure. Many elegant algorithms for building decision trees with the desirable quality have been introduced and applied in many real life problems, for example, C4.5[121](derived from ID3),CART[122], CHAID[123] are well known programs for decision trees construction.

Decision Tree has been demonstrated useful for common medical clinical problems where uncertainties are unlikely[124-128]. The logic flow of constructed Decision Trees can be an aid for the physician choosing a clinical strategy that offers the patient with the greatest expected value[124, 129]. Various application of Decision Trees in medical applications[126-128, 130, 131] are shown. The wise designed tree logic with wise administrative and flexibly understanding of the decision could benefit both economy and patients. Decision tree also has been applied in some biological information analysis problems, such as motif identification approach to explain T cell responses[132], leiomyomatous tumors characterizations[133], exons and introns identification in DNA sequences[134].

The construction of the decision trees usually requires large number of samples to produce a meaningful classifier in biological problems. Additionally, Decision trees may not perform well than other methods when the problem is complex. Because it is difficult or even impossible find enough samples to describe the problem, the rules generated by Decision trees may be biased or even misleading [125].

### 1.4.4. Neural Networks

It has been a long time we understand about how the human brain working differ from the traditional data analysis methods either in performance or in learning process. From the basic conceptual point of view, mathematical methods designed to mimic the way of information processing and the knowledge acquisition in human brain are neural networks. As its name indicated, neural networks consist of group of neurons that are linked together into a network. Increasing efforts were directed to the study of the learning problem by various neural networks since the so-called back propagation method was proposed to simultaneously compute the weight coefficient of neurons within the networks[135, 136] in 1986. The use of neural networks is still a hot research area in current machine learning research, such as pattern reorganization, association, and transformation to modeling in process control or expert system.

A neural network trains a hidden-layer-containing network and uses the output of this layer to recognize patterns from the input feature vectors [137, 138], where each vector representing the various data of an observation. A classifier for NN is $y = g\sum_{j} w_{0j}h_j$ , where $w_{0j}$ is the output weight of a hidden node $j$ to an output node; $g$ is the output function; $h_j$ is the value of a hidden layer node: $h_j = \delta(\sum_j w_{ji} x_j + w_j$ , $w_{ji}$ is the input weight from an input node i to a hidden node j, $\mathbf{w}_j$ is the threshold weight from an input node of value 1 to a hidden node j, and $\delta$ is a sigmoid function. The learning process is to optimize the weight vectors of all the neurons. The knowledge is gained from the learning and acquired by these weight vectors. Therefore, the optimized network that could act as a classifier can be used for determining whether or not a new input data of an observation response to a specific pattern.

The most widely used transfer active function in many neural network applications is the sigmoid function, $f(x) = \frac{1}{1+e^{-x}}$ . Other alternative activation functions like Gaussian have also been used widely in neural networks, e.g. probabilistic neural networks[139]. Although the underlying principle of every kind of neural networks start from the human neurons simulation, different approaches may have different performance for different problems. In the study of anesthesia, intensive care, and emergency medicine by neural network, it has been shown that "complex, non-linear, and time depending relationships can be modeled and forecasted"[140]. The encouraging results obtained in drug lead discovery and development also demonstrate it as a robust tool[141]. The successfully implementation of NN approaches bioinformatics problem have been demonstrated in protein structure prediction[142-147], protein function and protein-protein interaction prediciton[148-150].

Unfortunately, there are still several concerns[137, 138] for using neural networks to solve our problems. Firstly, it requires a great deal of computational effort to minimize overfitting. Secondly, the individual relations between the input variables and the output variables are not developed without analytical basis so that the model tends to be a black box. Thirdly, neural networks have a number of weight parameters that are consequently increasing the computation costs for model training.

### 1.4.5. Support Vector Machines

The basis of Support Vector Machine (SVM) learning theory had been brought forth by Vapnik[151] in 1979. It receives increasing attention since it was officially introduced by Vapnik[152] in 1995 and further explained by Dr. Burges[153] in 1998. Because of the successful fundamental construction of the theory and the prominent learning power, much more efforts have been directed into both the study of its theoretical aspects and the potential of its applications. SVM has been applied to a wide range of problems including text categorization[154-156], hand-written digit recognition[152], tone recognition[157], image classification and object detection[158-161]; flood stage forecasting[162]; cancer diagnosis[163-165], microarray gene expression data analysis[166], inhibitor classification[167], prediction of protein solvent accessibility[48], protein fold recognition[47], protein secondary structure prediction[49], prediction of protein-protein interaction[14] and protein functional class classification[31, 43, 45]. These studies have demonstrated that SVM is consistently superior to other supervised learning methods including classification methods[43, 166, 167]. Thus in this study, we selected SVM as the main statistical learning approach for predicting protein functions and inhibitors.

SVM is based on the structural risk minimization (SRM) principle from statistical learning theory[152]. In linearly separable cases, SVM constructs a hyperplane that separates two different classes of vectors with a maximum margin. Examples are tested by placing them onto this input space to recognize the classification label based on their relative positions to the hyperplane. As real world problems are most likely in non-linear forms, SVM can be extended by introducing kernel mappings that are able to project the samples from non-separable space onto a high-dimensional feature space in which the training examples can be linearly separated. The optimal separation hyperplane obtained in this high-dimensional feature space corresponds to the nonlinear decision boundary in the input space.

### *1.4.5.1. Theory and algorithm*

The beauty of SVM is not only in its successful applications in a wide range of real world classification problems, but also from where it starts.

Support vector machine aims to recognize patterns by learning process. A function mapping is described by training data set ($\mathbf{x}_i$, $y_i$) for pattern recognition:

$$f : R^N \rightarrow \{\pm 1\} \qquad (1)$$

where $\mathbf{x}_i$ are the n-dimensional feature vectors and $y_i$ are the corresponding class label.

Every data point is under the same probability distribution $P(\mathbf{x}, y)$,

$$(x_1, y_2), (x_2, y_2), ...(x_l, y_l) \in R^N X \{\pm 1\} \qquad (2)$$

The function $f$ is well generalized so that the training dataset ($\mathbf{x}_i$, $y_i$), i = 1, 2, …, $l$, satisfy $f$ ($\mathbf{x}_i$) = $y_i$. Through the learning, the function $f$ is usually able to correctly recognize new examples ($\underline{\mathbf{x}}_{\underline{j}}$, $\underline{y}_{\underline{j}}$), by satisfying $f$ ($\underline{\mathbf{x}}_{\underline{j}}$) = $\underline{y}_{\underline{j}}$. However, the fact is that the generalized function $f$ from the training dataset may have the poor performance on predicting new samples. That is, for any test dataset ($\underline{\mathbf{x}}_{\underline{j}}$, $\underline{\mathrm{y}}_{\underline{j}}$) ∈ $R^{\mathrm{N}}$ X {±1} and ∩ {$\underline{\mathbf{x}}_1$, $\underline{\mathbf{x}}_2$, …, $\underline{\mathbf{x}}_i$} = { }, there exists another function $f^*$ such that $f^*$ ($\mathbf{x}_i$) = $f$ ($\mathbf{x}_i$) for all $i$ and $f^*$ ($\mathbf{x}_j$) ≠ $f$ ($\mathbf{x}_j$) for all $j$.

Thus, there no way to decide which decision function is better than the other. In order to minimize the testing error, the statistical learning theory or the Vapnik-Chervonenkis

(VC) theory[101] is thus introduced to add the bounds on the test error. The minimization of these bounds, which depend on both the empirical risk (training error) and the capacity of the function class, leads to the principle of structural risk minimization[151]. The best-known capacity concept of VC theory is the VC dimension, defined as the largest number $h$ of points that can be separated in all possible ways using functions of given class. If the $h < l$ is the VC dimension of the class of functions that the machine learning can implement, then for all functions of that class, the bound with a probability of at least 1- $\eta$ will be

$$R(\alpha) \leq R_{emp}(\alpha) + \phi(\frac{h}{l}, \frac{\log(\eta)}{l}) \qquad (3)$$

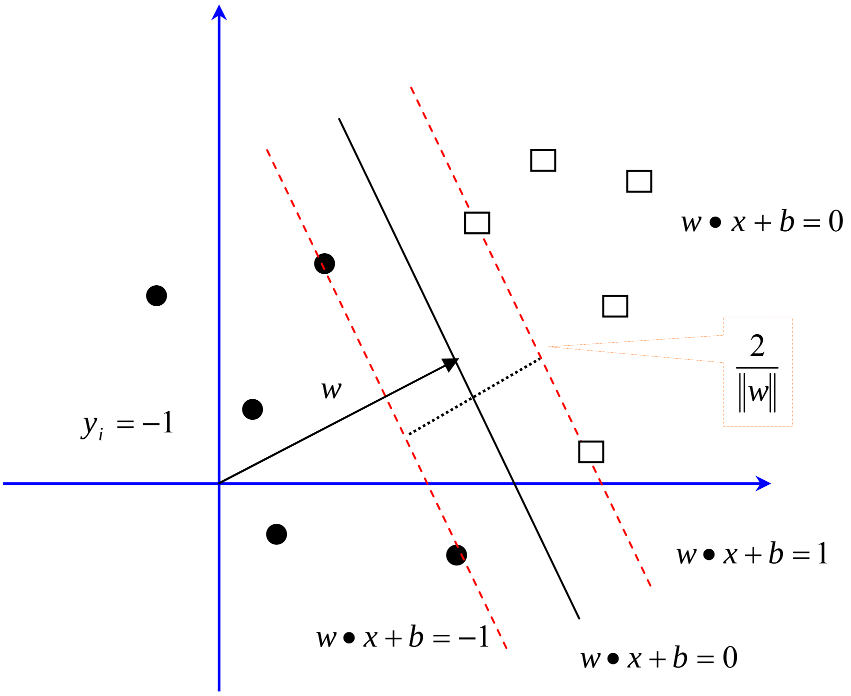
where the confidence term $\phi$ is defined as

$$\phi(\frac{h}{l}, \frac{\log(\eta)}{l}) = \sqrt{\frac{h(\log\frac{2l}{h} + 1) - \log(\frac{\eta}{4})}{l}} \qquad (4)$$

From the above function, in order to increase the capacity, a large VC dimension $h$ should be considered; the increase of $h$ is accompanied by the increase of the confidence term $\phi$.

The aim of SVM learning is to find the optimal separation hyperplane (OSH) that can separate the positive and negative samples by achieving maximum margins as shown in Figure 1-1.

Figure 1-1. The binary classification and the hyperplane. Hyperplanes $w \bullet x + b = \pm 1$ are boundaries of two classes of examples denoted by circles and squares. The OSH $w \bullet x + b = 0$ is decision hyperplane to separate the positive and negative samples.



Any hyperplane that can separate the input samples in the n-dimensions space can be described as follows:

$$(w \cdot x) + b = 0 \qquad (w \in R^N, b \in R) \qquad (5)$$

where ***w*** is the weight vector and the corresponding decision functions

$$f(x) = sign((w \cdot x) + b) \qquad (6)$$

It has been proved that the OSH is a unique one among the hyperplanes described in equation (5) which could yield the maximum margin of separation between the classes[152],

$$\frac{\max}{w,b} \quad \min\{\| x - x_i \| : x \in \mathrm{R}^{\mathrm{N}}, (w \cdot x) + b = 0, i = 1,2,...,l\} \qquad (7)$$

The construction of the Optimal Hyperplane is achieved by solving the following optimization problem:

minimize $\quad \tau(w) = \frac{1}{2} \| w \|^2 \qquad (8)$

subject to $\quad y_i \cdot ((w \cdot x_i) + b) \geq 1, i = 1,2,...,l \qquad (9)$

To solve the constrained optimization problem, the Langrangian and the Lagrange multiplier $\alpha_i$ is introduced,

$$L(w,b,\alpha) = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{l} \alpha_i (y_i \cdot ((x_i \cdot w) + b) - 1) \qquad (10)$$

Where $\alpha_i \geq 0$. The Lagrangian *L* has to be minimized with respect to the primal variables ***w*** and *b* and maximized with respect to the dual variables $\alpha_i$. ***w*** here has an expansion $w = \sum_i \alpha_i y_i x_i$ in terms of a subset of the training patters, called *Support*

*Vector* while $\alpha_i$ is non-zero. Solving the formula (10) subject to $\sum_{i=1}^{l} \alpha_i y_i = 0$ and $\alpha_i \geq 0$, the hyperplane decision function can thus be written as

$$f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i \cdot (x \cdot x_i) + b) \qquad (11)$$

where b is calculated by

$$\alpha_i \cdot [y_i((x_i \cdot w) + b) - 1] = 0, i = 1,2,...,l. \qquad (12)$$

### *1.4.5.2. Feature Spaces and Kernels*

When the examples is inseparable by linear SVM, the SVM OSH is developed by mapping data from input dimension space into higher dimension space where the problem can be solved by linear approach. The kernel function non-linearly maps samples into a higher dimensional space, so it can handle the case when the relation between class labels and attributes is nonlinear:

$$\phi : R^N \to F \qquad (13)$$

F is the hyperspace where the original problem becomes linear.

This requires the evaluation of dot products by a simple kernel function,

$$k(x, y) := (\phi(x) \cdot \phi(y)) \qquad (14)$$

If *F* is high-dimensional, then kernel function, polynomial kernel,

$$k(x, y) = (x \cdot y)^d \qquad (15)$$

can be shown to correspond to a map $\phi$ into the space spanned by all products of exactly d dimensions of $R^N$. For example, d = 2 and $\boldsymbol{x}, y \in R^2$, then

$$(x \cdot y) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = (\phi(x) \cdot \phi(y)) \qquad (16)$$

defining $\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ . For every kernel that gives rise to a positive matrix $(k(x_i, x_j))_{ij}$ , a map $\phi$ can be constructed.

A very useful kernel is Gaussian radial basis function (RBF):

$$K(x, y) = \exp(-\frac{\|x - y\|^2}{2\sigma^2}) \qquad (17)$$

The RBF function is chosen in this study because it has few numbers of parameters that influence the complexity of model selection. Furthermore, it reduces computation cost compared with polynomial kernels that kernel values may go to infinity or zero while the degree is large. In addition, RBF kernel has been commonly used in other SVM protein studies with consistently better performance than other kernels such as linear and polynomial[47, 168].

# 2. Scope and Research Objective

One of the main purposes of this study is to develop a classification system for predicting protein functions from their primary sequences. There are four focuses for this objective. Firstly, the features vectors are constructed from protein primary sequence. Our designed physico-chemical properties derived from sequence are independent to sequence similarities. Secondly, the strategy employed in this work is to classify proteins according to their functional families from their primary sequences by using Support Vector Machines (SVM). SVM is a relatively new and promising algorithm for binary classification by means of supervised learning. Although the studies of SVM used to solve various problems have demonstrated that SVM is consistently superior to other supervised learning methods including classification methods[166, 167], problems, such as data unbalance and over fitting are still critical when the optimal separation problem is addressed. Thirdly, the prediction system based on the well established SVM models are developed for solving the multiple-class classification problem. Various protein functional classes' classification problems are properly solved before their use for protein function prediction. Lastly, the potential of our designed protein function prediction system for predicting novel proteins' function are evaluated.

Because the problems of resistance development and physiological side effects remain in current HIV-1 protease inhibitors, methods for facilitating early elimination of potential HIV-1 protease inhibitors are useful for speeding up new drug discovery. Another main objective of this study is to predict HIV protease inhibitors by statistical learning approach. In order to fulfill this task, four important components are brought forward. Firstly, thousands of HIV-1 protease inhibitors are manually collected and checked to ensure the data quality. Secondly, the non-protease inhibitors used as negative control are representatively selected by distribution analysis. In order to diversify the negative control data set, compounds database containing large number of compound structures is constructed for diversity analysis. Thirdly, feature selection is considered to select distinguishing features for identification of HIV-1 protease inhibitors. Lastly, the prediction system is developed for protease inhibitor prediction and novel HIV-Protease inhibitor design.

# 3. Methods used in this study

## 3.1. Protein functional family classification and prediction

### 3.1.1. Feature vector construction

Construction of the feature vector for each protein is based on the formula used for the prediction of protein-protein interaction [14], protein fold recognition[47], and protein family classification[31, 43, 45, 46]. Each feature vector is constructed from the encoded representations of tabulated residue properties including amino acids composition, hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility[14, 43].

Amino acid composition can be computed directly. Some of the methods for computing each of the other properties can be found from the literature[14, 31, 43, 47, 49]. For calculating each group of properties, amino acids are divided into three groups such that those in a particular group are considered to have the same property. For instance, amino acids can be divided into hydrophobic (CVLIMFW[*]), neutral (GASTPHY), and polar (RKEDQN) groups. The groupings of amino acids for each of the properties are given in Table 3-1. Three descriptors, composition (C), transition (T), and distribution (D), are used to describe global composition of each of the properties. C is the number of amino acids of a particular property divided by the total number of amino acids in a protein sequence. T characterizes the percent frequency with which amino acids of a particular property is followed by amino acids of a different property. D measures the chain length within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular property is located respectively.

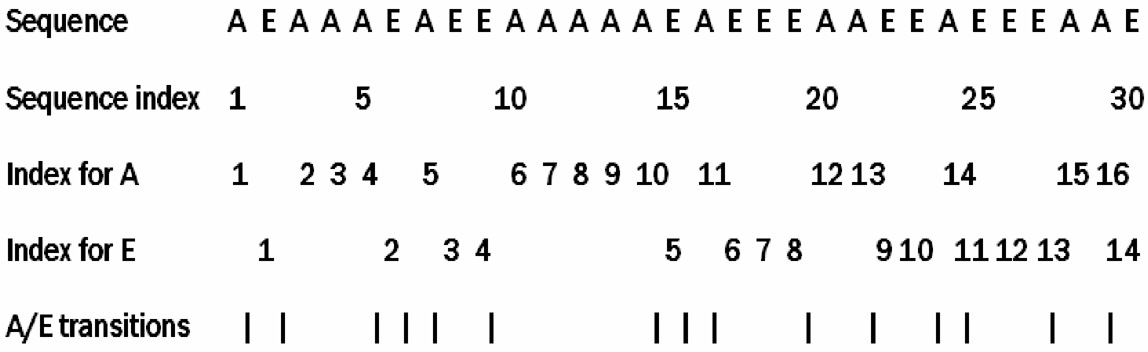
A hypothetical protein sequence AEAAAEAEEAAAAAEAEEEAAEEAEEEAAE, as

[*] List of amino acid in standard one letter amino acid codes

shown in Figure 3-1 , has 16 alanines (n1=16) and 14 glutamic acids (n2=14). The composition for these two amino acids are n1×100.00/(n1+n2)=53.33 and n2×100.00/(n1+n2)=46.67 respectively. There are 15 transitions from A to E or from E to A in this sequence and the percent frequency of these transitions is (15/29)×100.00=51.72. The first, 25%, 50%, 75% and 100% of alanines are located within the first 1, 5, 12, 20, and 29 residues respectively. The D descriptor for alanines is thus 1/30 ×100.00=3.33, 5/30×100.00=16.67, 12/30×100.00=40.0, 20/30×100.00=66.67, 29/30×100.00=96.67. Likewise, the D descriptor for glutamic is 6.67, 26.67, 60.0, 76.67, 100.0. Overall, the amino acid composition descriptors for this sequence are C=(53.33, 46.67), T=(51.72), and D=(3.33, 16.67, 40.0, 66.67, 96.67, 6.67, 26.67, 60.0, 76.67, 100.0) respectively. Descriptors for other properties can be computed by a similar procedure.

Overall, there are 21 elements representing these three descriptors: 3 for C, 3 for T and 15 for D. The feature vector of a protein is constructed by combining the 21 elements of all of these properties and the 20 elements of amino acid composition in sequential order. Table 3-2 gives the computed descriptors of the Purinergic receptor (Swiss-Prot AC O70397) with 474 amino acids. The feature vector of a protein is commutated by combining all of the descriptors in sequential order.

Figure 3-1 The sequence of a hypothetic protein and the illustration of feature vector derivation from its sequence. Sequence index indicates the position of an amino acid in the sequence. The index for each type of amino acids in the sequence (A or E) indicates the position of the first, second, third, … of that type of amino acid (The position of the first, second, third, …, A is at 1, 3, 4, …). A/E transition indicates the position of AE or EA pairs in the sequence.



```
Sequence         A  E  A  A  A  E  A  E  E  A  A  A  A  A  E  A  E  E  E  A  A  E  E  A  E  E  E  A  A  E
Sequence index   1           5              10             15             20             25             30
Index for A      1     2  3  4     5        6  7  8  9  10    11          12 13       14          15 16
Index for E         1           2     3  4                 5     6  7  8        9  10    11 12 13       14
A/E transitions    |  |        |  |  |     |              |  |  |        |     |     |  |        |     |
```

Table 3-1 Division of amino acids into 3 different groups for different physicochemical properties

| **Property** | | **Group 1** | **Group 2** | **Group 3** |
|---|---|---|---|---|
| **Hydrophobicity** | Type | *Polar* | *Neutral* | *Hydrophobic* |
| | Amino Acids in Group | RKEDQN | GASTPHY | CVLIMFW |
| **van der Waals volume** | Value | *0~2.78* | *2.95~4.0* | *4.43~8.08* |
| | Amino Acids in Group | GASCTPD | NVEQIL | MHKFRYW |
| **Polarity** | Value | *4.9~6.2* | *8.0~9.2* | *10.4~13.0* |
| | Amino Acids in Group | LIFWCMVY | PATGS | HQRKNED |
| **Polarizability** | Value | *0~0.108* | *0.128~0.186* | *0.219~0.409* |
| | Amino acids | GASDT | CPNVEQIL | KMHFRYW |

Table 3-2 Characteristic descriptors of Purinergic Receptor (Swiss-Prot AC O70397). The feature vector of this protein is constructed by combining all of the descriptors in sequential order.

| **Property** | **Elements of Descriptors of Purinergic Receptor (Swiss-Prot AC O70397)** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Amino acid composition** | 6.54 | 2.95 | 4.43 | 4.01 | 4.43 | 7.38 | 3.16 | 6.33 | 5.06 | 8.02 |
| | 1.05 | 3.16 | 7.81 | 4.22 | 5.06 | 7.17 | 6.54 | 6.96 | 1.89 | 3.80 |
| **Hydrophobicity** | 25.95 | 42.41 | 31.65 | 23.04 | 18.39 | 25.16 | 1.48 | 23.00 | 47.68 | 75.11 |
| | 99.79 | 0.42 | 27.00 | 53.16 | 79.96 | 99.58 | 0.21 | 22.57 | 48.10 | 71.52 |
| | 100.0 | | | | | | | | | |
| **Van der waals volume** | 42.83 | 32.70 | 24.47 | 28.12 | 20.51 | 14.80 | 0.42 | 28.48 | 53.80 | 80.38 |
| | 99.58 | 2.53 | 20.04 | 49.36 | 73.42 | 100.0 | 0.21 | 21.94 | 46.41 | 66.46 |
| | 98.95 | | | | | | | | | |
| **Polarity** | 35.44 | 35.44 | 29.11 | 23.47 | 23.04 | 21.35 | 0.21 | 22.36 | 47.26 | 70.68 |
| | 100.0 | 0.42 | 27.00 | 54.01 | 81.22 | 99.58 | 1.05 | 24.47 | 48.31 | 75.11 |
| | 99.79 | | | | | | | | | |
| **Polarizability** | 32.07 | 43.46 | 24.47 | 30.66 | 15.01 | 20.30 | 0.42 | 28.48 | 53.16 | 76.16 |
| | 99.58 | 1.27 | 23.42 | 50.63 | 76.79 | 100.0 | 0.21 | 21.94 | 46.41 | 66.46 |
| | 98.95 | | | | | | | | | |

### 3.1.2. Effective selection of examples

Statistical learning process of binary classification SVM requires both positive examples and negative samples, for example, proteins examples from a particular functional family and those outside of this family. The positive samples of a family include all of the known distinct proteins in that family. Ideally, the negative samples of the given family should include all of the proteins outside of the family. Because the proteins are enormous, it is impractical to include all of the proteins outside of the family as negative examples for statistical learning. Thus, the negative samples selected statistical learning should be restricted to a manageable level by using a minimum set of representative proteins.

Our approach to select effective proteins examples is to choose a few distinct proteins from each protein family. The negative samples of a family can be selected from seed proteins of the curated protein families in the Pfam database[169] excluding those families that have at least one member belongs to the Pfam family. The purpose of using Pfam families to generate negative protein examples is to ensure that the negative examples are more evenly distributed in the protein space than random selection. However, only the selection of negative examples is involved in using of Pfam families that are based on sequence similarity; the positive examples are collected without any consideration of sequence similarity. Thus, our approach for protein functional family classification is to some extent independent of sequence similarity.

### 3.1.3. Support Vector Machine classification

As the theory of SVM has been described in the previous section, only a brief description of our strategy of the implementation is given here.

In nonlinearly separable cases, SVM maps feature vectors into a high dimensional feature space using a kernel function $K(x_i, x_j)$. The kernel function employed in this work is the Gaussian kernel, which has been extensively used in a number of protein classification studies[14, 31, 44, 47-49, 153]:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma^2} \tag{1}$$

The linear SVM procedure is then applied to the feature vectors in this feature space and the decision function for their classification is given by:

$$f(\mathbf{x}) = sign(\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b) \qquad (2)$$

Where the coefficients $\alpha_i^0$ and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad (3)$$

Under conditions:

$$a_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (4)$$

A positive or negative value from Equation (2) determines whether the vector x belongs to the positive or negative group. In order to reduce the complexity of parameter selection, hard margin SVM with threshold instead of soft margin SVM is used in our program.

As in the case of all discriminative methods[170, 171], the performance of SVM classification can be measured by the quantity of true positive ***TP*** (correctly predicted members), false negative ***FN*** (incorrectly predicted as non-members), true negative ***TN*** (correctly predicted non-members), and false positive ***FP*** (non-members incorrectly predicted as members). In this work, protein functional family classification is a one-against-other multi-class prediction problem, thus the unique accuracy[47] specifically designed for evaluation of multi-class prediction is used. Due to the imbalanced number of positive and negative samples for each sub-class, two unique accuracies $Q_p$ and $Q_n$ are used to measure the accuracy of positive prediction (proteins that belong to a specific functional class) and negative prediction (proteins that do not belong to a given functional class) [47]:

$$Q_p = \frac{TP}{TP + FN} \qquad (5)$$

$$Q_n = \frac{TN}{TN + FP} \qquad (6)$$

Another quantity suitable for evaluating the classification accuracy of imbalanced positive and negative samples is the Matthews correlation coefficient C[172], which is given by:

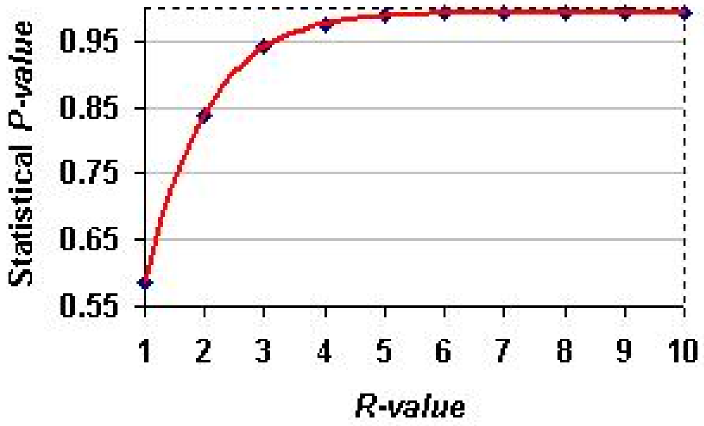
$$C = \frac{TP \cdot TN + FN \cdot FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \qquad (7)$$

A reliability index of SVM protein family prediction, R-value, has been introduced in this study to represent the level of the signal of decision function.

$$R - value = \begin{cases} 1 & if \quad d < 0.2 \\ d/0.2 + 1 & if \quad 0.2 \le d < 1.8 \\ 10 & if \quad d \ge 1.8 \end{cases} \qquad (8)$$

Where d is the distance between the position of the vector of a classified protein and the optimal separating hyperplane in the hyperspace. A statistical correlation between R-value and expected classification accuracy or probability of correct classification[49] as show in Figure 3-2. Another quantity, P-value, is introduced to indicate the expected classification accuracy. P-value is derived from the statistical relationship between the R-value and actual classification accuracy based on the analysis of 9,932 positive and 45,999 negative samples of proteins [43].

Figure 3-2 Expected classification accuracy P-value (probability of correct classification) versus R-value. It is derived from the statistical relationship between the R-value and actual classification accuracy based on the analysis of 9,932 positive and 45,999 negative samples of proteins.



### 3.1.4. Protein functional family classification systems-SVMProt

In this work, we have developed a protein functional family classification systems, SVMProt[153], based on support vector machines (SVM) for protein functional family prediction. The construction of SVMProt protein function prediction system is currently containing 97 protein functional classes as listed in Appendix A. The 97 protein functional classes include 46 enzyme families, 9 channel/transporter families, 21 transporter families, 4 RNA-binding protein families, DNA-binding proteins, 5 G-protein coupled receptors, nuclear receptors, tyrosine receptor kinases, cell adhesion proteins, coat proteins, envelope proteins, outer membrane proteins, structural proteins, and growth factors. Two broadly defined families of antigens and transmembrane proteins are also included.

Every protein function classification model in SVMProt has been trained and tested by using a large number of proteins. A training set contains positive examples those proteins belong to a functional family, and negative examples referring to those outside a family. The negative examples of a protein family are collected from representative proteins of the Pfam families without a member in that class A training set needs to be both diverse and kept as small as possible in order to ensure adequate representation and to reduce un-necessary noise generated from data redundancy.

The numbers of member and non-member protein sequences in the training sets are in the range of 14~3,892 and 513~7,299 respectively, and those of the independent evaluation sets are in the range of 7~4,841 and 986~7,291 respectively. For examples, 945 glycosyltransferases proteins and 1,896 non- glycosyltransferases are used for training glycosyltransferases (EC2.4) family, and there are 288 glycosyltransferases proteins and 4,926 non- glycosyltransferases are used for independent evaluation of glycosyltransferases (EC2.4) family. The number of sequences in all classes can be found in the Appendix A.

We develop the protein classification model in the following manner. First, every protein sequence is represented by specific feature vectors assembled from encoded representations of tabulated residue properties including amino acid composition, hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility for each residue in the sequence. The feature vectors of the positive and negative samples are used to train a SVMProt classifier. The trained SVMProt classifier is used to determine whether a protein belongs to this protein functional class or not.

The SVMProt training system for each family is optimized and tested using separate testing sets of both positive and negative samples. Those proteins outside of the training set for each functional family are positive examples of the testing set, and all the representative seed proteins in Pfam curated families not used for model training are negative examples of the testing set. The performance of SVMProt classification is further evaluated by using independent sets consisted of both positive and negative examples. There is no overlap in each training, testing or independent evaluation set.

Not all of the protein functional classes in SVMProt are at the same hierarchical level. These protein functional classes are mixtures of subfamilies, families and super-families. Some classes, such as antigen, are superfamily. While it is desirable to define all of the classes at the same level, this is not yet possible because of insufficiency of data for the sub-hierarchies of some families and super-families. Because of independency of SVMProt classifiers, different classification models can work simultaneously.

## 3.2. Methods for protein inhibitor prediction

### 3.2.1. Molecular descriptors

As shown in Table 3-3, a set of 159 molecular descriptors were used for quantitative description of structural and physiochemical properties of molecules in this study. There are 18 simple molecular properties, 28 molecular connectivity and shape descriptors, 84 descriptors computed from electro-topological state, 13 quantum chemical properties and 16 geometrical properties. All of these descriptors are calculated from the 3D structure of compound by using our previously published molecular descriptor program[173].

Table 3-3 Molecular Descriptors used in this work

| Descriptor class | Number of descriptors | Descriptors |
|---|---|---|
| Simple molecular properties | 18 | Molecular weight, number of ring structures, number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, element counts |
| Molecular connectivity and shape | 28 | Molecular connectivity indices, valence molecular connectivity indices, molecular shape, Kappa indices, Kappa alpha indices, flexibility index |
| Electro-topological state | 84 | Electrotopological state indices and atom type electrotopological state indices |
| Quantum chemical properties | 13 | Atomic charge on the most positively charged H atom, largest negative charge on an non-H atom, polarizability index, hydrogen bond acceptor basicity (covalent BAB),<br>hydrogen bond donor acidity (covalent HBDA), molecular dipole moment, absolute hardness, softness, ionization potential, electron affinity, chemical potential, electronegativity index, electrophilicity index |
| Geometrical properties | 16 | Molecular size vectors (distance of the longest separated atom pairs, combined distance of the longest separated three atoms, combined distance of the longest separated four atoms), molecular van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, polar molecular surface area, sum of solvent accessible surface areas of positively charged atoms, sum of solvent accessible surface areas of negatively charged atoms, sum of charge weighted solvent accessible surface areas of positively charged atoms, sum of charge weighted solvent accessible surface areas of negatively charged atoms, sum of van der Waals surface areas of positively charged atoms, sum of van der Waals surface areas of negatively charged atoms, sum of charge weighted van der Waals surface areas of positively charged atoms, sum of charge weighted van der Waals surface areas of negatively charged atoms. |

### 3.2.2. Selection of HIV-1 PI candidates

4291 HIV-1 PI candidates are selected as positive training samples from the HIV/OI Enzyme Inhibition Database[†] of the National Institute of Allergy and Infectious Diseases, National Institutes of Health.

Since the quality of input data have a direct effect in the training of the SVM model, the positive samples from the database were further examined by checking each of the PIs against PubMed Database[174] to ascertain that they have been described as HIV-1 protease inhibitors. Only those with reported IC50[‡] (inhibitory concentration 50%) in the literature were selected. Meanwhile, since PIs constitute a large chemical space (sulfonamides, benzopyrans, piperazines, benzimidazoles, urethanes, symmetry-based dihydroxy, epoxies etc.) with varying potencies for uses in different contexts, those with reported log (IC50) of log units -7.85 to 3.30 were selected.

### 3.2.3. Selection of HIV-1 non-PI candidates

There are numerous and diversified compounds. Thus, it is impractical to include all compounds outside of a specific family as negative examples. It is reliable to use experimentally determined negative compound examples, such as those compounds with non-inhibition activities to a specific protein target. However, only a small number of true negative examples have been report for some protein target. As such, it is inadequate to use those compounds to approximate the complete negative compounds space.

Our approach to generate comprehensive negative examples is to choose representative compounds from the compound space that is not covered by positive examples.

In order to analyze the distribution of positive examples within the compounds space, a

[†] HIV/OI Enzyme Inhibition Database: http://www.niaid.nih.gov/daids/

[‡] IC50 (or EC50 - effective concentration 50%) is the concentration required for 50% inhibition.

chemical database composed of 85,000 is constructed in this study. These compounds and their 3D structures are selected from MDDR (MDL Drug Data Report), ACD (MDL Available Chemicals Directory) and ChemIDPlus[§], ChemFinder[**] databases. Subsequently, we use the hierarchal clustering method to cluster these compounds into 8,000 subfamilies according to the Euclidean distance of their descriptors.

By this means, the compound space is condensed into 8,000 subfamilies instead of the original 85,000 compounds. The distribution of positive compounds is calculated based on the 8,000 subfamilies. Thus the distribution of negative examples could be obtained from the compound space that not occupied by positive examples.

The selection of Non-PIs is based on the distribution of PIs. In this study, 12,453 negative samples are selected to ensure data balance in the 2-class c-SVM model that is used. There are basically two requirements for the selection of negative example: 1) their structures are vary from each other, and 2) the distribution of the negative examples should be diversified enough to form an effective representation of negative compound space.

The crude 3-dimensional structures collected from the databases are converted into accurate, energy-based geometry optimized 3-dimensional structures by using commercial software, Concord™[††].

### 3.2.4. Recursive feature elimination within non-linear SVM

The purpose of feature or variable selection is to eliminate irrelevant variables to enhance the generalization performance of a given learning algorithm. The selected predominated variables show some insight about the concept to be learned [175].

[§] http://chem.sis.nlm.nih.gov/chemidplus/

[**] http://chemfinder.cambridgesoft.com/

[††] Tripos product Sheet, 2004

In this work, recursive feature elimination (RFE) using SVM-based Criteria has been employed to select important features for identifying HIV-protease inhibitors. Linear RFE-SVM algorithm has been introduced by Guyon [176] for gene selection on cancer classification. In this work, the algorithm is extended to solve non-linear cases.

As Kohavi and John[177] suggested, the ranking criterion for feature selection can be computed from feature's influence on the objective function. In this study, the objective function is represented by a cost function for the ith feature in the training set.

The basic idea of RFE is to find and remove the smallest change in cost function resulting from the features. In our case, the cost function to be minimized is:

$$J = \frac{1}{2} a^T H a - a^T L \qquad (9)$$

under the constraints $0 \le \alpha_i \le C$ and_$\sum \alpha_i y_i = 0$;

where *H* is the matrix with elements $y_i y_j K(x_i, x_j)$, *K* is a RBF kernel function that measures the similarity between $x_i$ and $x_j$, and L is an *l* dimensional vector of ones. One can compute change in cost function by assuming no change in the value of the $a$. Thus, one avoids having to retrain a classifier for every candidate feature to be eliminated. In order to compute the change in cost function caused by removing input component $\delta$ , we can leaves the $a$ unchanged and one re-computes matrix *H*. This corresponds to calculate

$$y_i y_j K(x_i(exclude\delta), x_j(exclude\delta)) \qquad (10)$$

yielding matrix $H(exclude\delta)$, where the notation $exclude\delta$ means that component $\delta$ has been removed. The resulting ranking coefficient is:

$$DJ(\delta) = 0.5 a^T H a - 0.5 a^Y H(exclude\delta) a \qquad (11)$$

The input corresponding to the smallest difference $DJ(\delta)$ is removed. This procedure is iterated until the final list of predominated feature is obtained.

# 4. Protein functional family classification based on primary sequence by Support Vector Machines

**The work in this chapter has been published in:**

**I)** Enzyme Family Classification by Support Vector Machines. C.Z. Cai, L.Y. Han, Z.L. Ji, Y.Z. Chen .Proteins. 55,66-76 (2004).

**II)** Prediction of RNA-Binding Proteins from Primary Sequence by Support Vector Machine Approach. L.Y. Han, C.Z. Cai, S. L. Lo, Maxey C. M. Chung,Y. Z. Chen. RNA. 10(3),355-368. (2004).

**III)** Prediction of Transporter Family by Support Vector Machine Approach H. H. Lin, L.Y. Han, C.Z. Cai, Z. L. Ji, and Y.Z. Chen. Proteins. 62 (1): 218-31 (2006)

Determination of protein function is essential for understanding biological processes. Our approach for predicting protein function is based on the protein functional family classification. The method used in our study for protein functional family classification starts from the analysis of physicochemical properties of a protein derived from its primary sequence. In the coming sections, our results of classification on some specific protein functional families, such as enzymes and transporters, are described.

## 4.1. Enzyme Family Classification (Paper I)

Enzymes represent the largest and most diverse group of all proteins, catalyzing chemical reactions an organism needs to survive. In addition, enzymes are well classified into functional families according to the recommendation by the classification of enzyme nomenclature committee of IUBMB[178]. Therefore, enzymes are ideal for comprehensive testing of SVM protein family classification systems. In our study for protein family classification, enzymes from protein sequence database have been classified into 46 enzyme families and classifier for each enzyme family has been further tested by independent evaluation. The optimized enzyme classifiers are also evaluated for their capability in the classification of distantly related enzymes and homologous enzymes with different function.

### 4.1.1. Methods

The definitions of enzyme families are obtained from BRENDA database[179]. As sufficient number of samples are required for developing a SVM classification system with statistical significance, only 46 enzyme families with more than 100 non-redundant protein entries for each family in Swiss-Prot Enzyme database[180] are selected in our study. Appendix A gives the list of enzyme families along with the number of enzymes for each family used for training, testing and evaluating SVM classification system.

All distinct members in each enzyme family found in Swiss-Prot database[180] are used to construct positive samples for training SVM. Based on the definition of enzyme families in BRENDA[179] and annotations in Pfam database [180], the negative samples for each enzyme family are selected from seed proteins of the curated protein families in the Pfam database[169]. Negative samples of one enzyme family include proteins from other enzyme families and non-enzyme proteins such as receptors, transporters, channels and matrix proteins. The redundancy in the selected datasets has been further removed by sequence comparisons.

Every enzyme sequence is represented by specific feature vectors assembled from encoded representations of tabulated residue properties including amino acid composition, hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility for each residue in the sequence[14, 31, 47-49, 181, 182]. There is some level of overlap in the descriptors for hydrophobicity, polarity, and surface tension. Our study of these descriptors by principle component analysis suggests that the use of the PCA-reduced descriptors only moderately improves the accuracy for only some of the families. As it is also noted that reasonably accurate results have been obtained in various protein classification studies using these overlapping descriptors[14, 31, 47-49, 181, 182]. Thus, we choose the whole set of descriptors for our SVM study.

The constructed feature vectors of both positive samples  and negative samples are then input into SVM classification system to train it to identify features that separate positive and negative samples. The trained SVM systems can thus be used to classify an enzyme into either the positive group or the negative group of each family. One enzyme is predicted to be a member of a family if it is classified into the positive group of that family. Likewise, it is predicted to not belong to a family if it is classified into the negative group of that family.

The theory of SVM has been earlier described in Chapter 2. Thus, only the method for performance measurement is given here. As in the case of all discriminative methods[171, 183], the performance of SVM classification can be measured by the quantity of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*). Enzyme family classification is a one-against-others multi-classes classification problem, thus the unique accuracy[47], sensitivity and specificity, for evaluation of multi-class prediction is used in this study. Because the number of positive and negative samples for each family is imbalanced in size, an additional measure, Matthews correlation coefficient $C$[172] (appears in Chapter 2),  is used to measure the randomness and the performance of SVM prediction.

### 4.1.2. Result and Discussion

#### *4.1.2.1. Assessment of overall accuracy of SVM enzyme family classification*

The results for the classification of the 46 enzyme families are given in *Appendix Table A*. All the computed *TP*, *TN*, *FP*, and *FN* for the testing sets and independent evaluation sets of these families are as shown in the Table. *Table A* also gives the unique classification accuracies $Q_p$ and $Q_n$ and Matthews correlation coefficient *C* for every family measured by using independent evaluation sets. The computed $Q_p$, $Q_n$ and *C* for the 46 enzyme families are in the range of 53.0% to 99.3%, 82.1.0% to 100%, and 54.1% to 96.1% respectively. These numbers on average are better improved from that obtained in other SVM studies of proteins[14, 31, 47-49, 181, 182]. One possible reason for this improvement is the use of representative proteins of Pfam curated families as negative samples for SVM classification, they provides a more comprehensive sampling of proteins not in an enzyme family.

Table 4-1 lists a number of randomly selected enzyme entries from Swiss-Prot database[180] which are not correctly classified into the corresponding family by our developed SVM classifiers . Amino acid sequence of each of these enzyme entries is examined to find out whether or not the classification error is caused by sequence-related problems such as fragment, incomplete chain, and mutations. As shown in Table 4-2, the composition of the negative samples for a specific enzyme family is diversified, thus these sequence-related problems do not appear to be a significant factor for the classification error. BLAST sequence alignment of each of these enzymes against other members of its family suggests these substantial portions (61.3%) of incorrectly classified enzymes are of low sequence similarity to other members in its family. Here, the threshold for sequence similarity score E value is 0.05.

The percentage of low sequence similarity proteins in a family is not expected to be very high. Therefore, our study seems to suggest that sequence similarity has certain level of influence on the accuracy of SVM classification.

Table 4-1.Randomly selected enzyme entries from Swiss-Prot database which are not correctly classified into their corresponding family in our study.

| EC Family number | Swiss Prot Accession number | Protein Name | Sequence feature* | Sequence similarity to other members of family* |
|---|---|---|---|---|
| EC 1.1 | Q8YH79 | Alcohol dehydrogenase | C | L |
| EC 1.14 | P79078 | Delta-9 fatty acid desaturase | C | S |
| EC 1.14 | Q8TE42 | Truncated steroid 21-hydroxylase | IC | L |
| EC 1.14 | P14791 | Heme oxygenase | C | L |
| EC 1.2 | O67724 | N-acetyl-γ-glutamyl-phosphate reductase | C | L |
| EC 1.2 | Q57658 | Aspartate-semialdehyde dehydrogenase | C | L |
| EC 2.1 | Q9ZE37 | tRNA (Guanine-N(1)-)-methyltransferase | C | S |
| EC 2.1 | Q9PJ28 | Methionyl-tRNA formyltransferase | C | S |
| EC 2.1 | Q9UX08 | Aspartate carbamoyltransferase | C | L |
| EC 2.1 | P96111 | PyrBI protein | C | L |
| EC 2.7 | Q9JR61 | Phosphatidylserine synthase | C | L |
| EC 2.7 | Q9ZE96 | Phosphatidylglycerophosphate synthase | C | L |
| EC 3.1 | Q62087 | Serum paraoxonase/arylesterase 3 | C | L |
| EC 3.1 | Q97VT7 | Aryldialkylphosphatase, putative | C | S |
| EC 3.2 | Q9EVP3 | Stx2fA protein subunit | C, subunit | L |
| EC 3.2 | Q9S9E4 | rRNA-glycosidase | C | L |
| EC 3.2 | Q41216 | Trichosanthin | C | L |

* C—Complete sequence; IC—Incomplete sequence; C,subunit—Complete sequence of subunit; C,chain—Complete sequence of chain; L—Low sequence similarity to other enzymes in a particular family; S—Significant sequence similarity to other enzymes in a particular family

| | | | | |
|---|---|---|---|---|
| EC 3.5 | P32320 | Cytidine deaminase | C, subunit | L |
| EC 3.5 | Q01432 | AMP deaminase 3 | C, subunit | L |
| EC 3.5 | Q49135 | Methenyltetrahydrofolate cyclohydrolase | C, subunit | S |
| EC 4.2 | P73715 | Endonuclease III | C | S |
| EC 4.2 | Q8RI68 | Cystathionine gamma-synthase | C | S |
| EC 4.3 | Q8XMJ8 | Argininosuccinate lyase | C | S |
| EC 5.1 | Q980W1 | UDP-glucose 4-epimerase | C | S |
| EC 5.1 | P21955 | Aldose 1-epimerase | C | L |
| EC 5.3 | P29954 | Mannose-6-phosphate isomerase | C | S |
| EC 5.4 | Q8Z8D7 | UDP-galactopyranose mutase | C | S |
| EC 6.1 | Q8YH72 | Alanyl-tRNA synthetase | C | L |
| EC 6.1 | Q9ZDF8 | Lysyl-tRNA synthetase | C | L |
| EC 6.1 | Q9HJM5 | Glutamyl-tRNA synthetase | C | L |
| EC 6.1 | Q55486 | Arginyl-tRNA synthetase | C | L |
| EC 6.3 | P57245 | Carbamoyl-phosphate synthase, small chain | C, chain | S |

Table 4-2 Composition of the negative samples for EC2.7 family. Here "other proteins" include proteins known to not belong to any of the families listed and those enzymes whose EC number is not specified at the time of our data Collection

| Family | No. of Entries | Family | No. of Entries |
|---|---|---|---|
| EC 1.1 | 10 | EC 3.3 | 2 |
| EC 1.2 | 3 | EC 3.4 | 12 |
| EC 1.3 | 17 | EC 3.5 | 9 |
| EC 1.4 | 6 | EC 3.6 | 33 |
| EC 1.5 | 2 | EC 4.1 | 28 |
| EC 1.6 | 7 | EC 4.2 | 18 |
| EC 1.7 | 2 | EC 4.4 | 7 |
| EC 1.8 | 1 | EC 4.6 | 5 |
| EC 1.9 | 24 | EC 5.1 | 7 |
| EC 1.10 | 8 | EC 5.4 | 3 |
| EC 1.11 | 4 | EC 5.5 | 1 |
| EC 1.13 | 4 | EC 5.99 | 9 |
| EC 1.14 | 1 | EC 6.1 | 1 |
| EC 1.15 | 3 | EC 6.2 | 1 |
| EC 1.18 | 2 | EC 6.3 | 20 |
| EC 2.1 | 11 | EC 6.4 | 6 |
| EC 2.3 | 20 | EC 6.5 | 9 |
| EC 2.4 | 20 | Receptors | 17 |
| EC 2.5 | 4 | Transporters | 53 |
| EC 3.1 | 30 | Channels | 11 |
| EC 3.2 | 33 | Other proteins | 1455 |

The quality of each of SVM classifiers trained for classification of a particular enzyme family can be further assessed by conducting direct two-way tests. For such a purpose, a set of 3000 enzymes in a randomly selected enzyme family EC1.6 is used for testing the accuracy of positive classification for that family. It is found that 76.8% of these enzymes are correctly classified into the EC1.6 family by our SVM system. A set of 2850 randomly selected non-enzyme proteins is used for assessing the accuracy of negative classification for that enzyme family. It is found that 98.5% of these non-enzyme proteins are correctly classified as not belong to the EC1.6 family. This result is comparable to the independent evaluation of the EC1.6 family in out study, where the sensitivity and specificity are 94.5 and 98.2 respectively.

#### *4.1.2.2. Independent evaluation and 10-fold cross validation*

In this work, independent evaluation sets were used to determine the accuracy of enzyme family classification. To examine whether it can provide sufficiently accurate assessment of prediction accuracy, we have conducted 10 fold cross validation on three randomly selected families to compare with our results from independent evaluation.

Table 4-3 show the results of the 10-fold cross validation study for the EC1.9, EC4.4 and EC5.2 family respectively. For comparison, the results from our study are also included in the respective Table. It is found that the computed $Q_p$, $Q_n$, and $C$ for each of these families using our method is roughly similar to those obtained by using 10-fold cross validation study. This suggests that our method may be used to assess the quality of SVM enzyme family classification, with a comparable accuracy as that of n-fold cross validation study.

Table 4-3 Ten-fold Cross Validation Results of EC1.9, EC4.4 and EC5.2 family. The true positive ***TP*** means number of correctly predicted members, false negative ***FN*** is the number of incorrectly predicted as non-members, true negative ***TN*** is the number of correctly predicted non-members, and false positive ***FP*** is the number of non-members incorrectly predicted as members. Sensitivity $Q_p$ and specificity $Q_n$ are defined as Qp=TP/(TP+FN), Qn=TN/(TN+FP), Matthews correlation coefficient C[172], which is given by equation (7) in Chapter 1.

| **EC family** | | **Performance measures** | | |
|---|---|---|---|---|
| | | **Qn(%)** | **Qp(%)** | **C** |
| EC1.9 | 10 CV | 94.2 | 99.3 | 0.947 |
| | Independent Evaluations | 95.7 | 99.5 | 0.961 |
| EC4.4 | 10 CV | 65.7 | 99.9 | 0.791 |
| | Independent Evaluations | 50.0 | 99.9 | 0.679 |
| EC5.2 | 10 CV | 66.7 | 99.9 | 0.800 |
| | Independent Evaluations | 65.3 | 99.8 | 0.776 |

### 4.1.3. Conclusion remark

Our study suggests the potential usefulness of SVM in classification of enzymes into functional families. The developed SVM models by using sequence derived physico-chemical properties are able to discriminate enzymes into their functional families with comparable accuracies and even better than other protein function prediction methods[14, 31, 47-49, 181, 182]. Moreover, it shows the capability for classification of enzymes with very low sequence similarities. The enzyme classification SVM models are very useful for classifying an unknown protein. As it is revealed in our study, the quality and diversity of enzyme protein samples and proteins as negative samples is very important for developing a SVM model with both good sensitivity and specificity. Our results also suggest that the developed SVM classification models could be a useful tool for facilitating protein function prediction.

## 4.2. Classification of RNA-Binding Proteins (Paper II)

Knowledge about how proteins interact with each other and with other molecules is essential in the understanding of cellular processes[184-187]. With the accumulation of sequence information, attention has been paid to the development of methods for predicting protein function[188] and protein-protein interactions[14, 189, 190] from sequence. Several computational methods have been developed for the prediction of protein-protein interactions using support vector machines[14] and for the prediction of protein-protein interaction maps by Rosetta/gene fusion[12, 191], phylogenetic profile[192], gene neighbor[189, 190], and interacting domain profile pair[193] methods.

While progress has been made in the development of predictive methods for protein-protein interactions, there is no effort has been made for predicting protein-RNA interactions by using machine-learning approach. Most cellular RNAs work in concert with protein partners and protein-RNA interactions are critically important in regulation of different steps of gene expression[186]. Moreover, binding of proteins to some catalytic RNA molecules are known to activate or enhance the activity of these molecules [194]. Therefore, prediction of protein-RNA interactions is very important for understanding how cellular processes and biological network works.

In this work, the use of SVM for the prediction of RNA-binding proteins from protein primary sequence was explored. SVM is used for the prediction of individual classes of rRNA-, mRNA-, tRNA-binding proteins as well as all RNA-binding proteins. There are other groups of RNA-binding proteins, such as snRNA-binding and snoRNA-binding proteins, with small number of proteins and fewer available sequences[195, 196]. A search of protein family and sequence databases finds a total of 60 sequences of snRNA-binding proteins and 21 sequences of snoRNA-binding proteins, which is fewer than the number of 80~100 sequences typically needed to properly train an SVM protein classification system. Non-the-less, to evaluate its performance on classification of a small protein class, SVM is used for the prediction of snRNA-binding proteins. Proteins of small RNA-binding classes as well as other RNA-binding proteins are included in training and testing SVM classification of all RNA-binding proteins.

### 4.2.1. Selection of RNA-binding proteins and non- RNA- binding proteins

All RNA-binding proteins used in this study are from a comprehensive search of Swiss-Prot database [180]. A total number of 4458 RNA-binding protein sequences are obtained, which include 2054 rRNA-, 570 mRNA-, 259 tRNA-, 60 snRNA-, and 21 snoRNA-binding proteins. The distribution of RNA-binding proteins in different kingdoms and in top 10 host species is given in Appendix Table B and that of each class of RNA-binding proteins is given in Table 4-4. As shown in the table, these RNA-binding proteins are from diverse range of species and all species appear to be adequately represented.

Table 4-4 Distribution of rRNA-, mRNA-, tRNA- and snRNA-binding proteins in different kingdoms and in top 10 host species. Not all protein sequences studied in this work are included because the host species information of some protein sequences is not yet available in the protein sequence database.

| | **rRNA-binding** | | **mRNA-binding** | | **tRNA-binding** | | snRNA-binding | |
|---|---|---|---|---|---|---|---|---|
| | Kingdom or species | No. of proteins | Kingdom or species | No. of proteins | Kingdom or species | No. of proteins | Kingdom or species | No. of proteins |
| **Protein distribution in kingdom** | Eucaryote | 493 | Eucaryote | 310 | Eucaryote | 19 | Eucaryote | 50 |
| | Eubacteria | 1330 | Eubacteria | 235 | Eubacteria | 230 | Eubacteria | - |
| | Archaea | 181 | Archaea | - | Archaea | 10 | Archaea | - |
| **Protein distribution in top 10 species** | Thermus thermophilus | 32 | Homo sapiens | 77 | Thermus thermophilus | 6 | Homo sapiens | 18 |
| | Aquifex aeolicus | 29 | Candida albicans | 41 | Homo sapiens | 5 | Candida albicans | 15 |
| | Mycobacterium leprae | 28 | Mus musculus | 36 | Bacillus subtilis | 5 | Mus musculus | 5 |
| | Chlamydia pneumoniae | 28 | Schizosaccharomyces pombe | 21 | Escherichia coli | 5 | Xenopus laevis | 3 |
| | Helicobacter pylori | 28 | Escherichia coli | 21 | Pasteurella multocida | 4 | Drosophila melanogaster | 3 |
| | Rickettsia prowazekii | 28 | Arabidopsis thaliana | 19 | Mycoplasma genitalium | 4 | Schizosaccharomyces pombe | 3 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Thermotoga maritima | 28 | Caenorhabditis elegans | 18 | Deinococcus radiodurans | 4 | Caenorhabditis elegans | 2 |
| | Chlamydia trachomatis | 28 | Drosophila melanogaster | 15 | Neisseria meningitidis (serogroup A) | 4 | Rattus norvegicus | 2 |
| | Borrelia burgdorferi | 28 | Rattus norvegicus | 14 | Helicobacter pylori | 4 | Arabidopsis thaliana | 2 |
| | Buchnera aphidicola | 28 | Nicotiana tabacum | 11 | Campylobacter jejuni | 4 | Macropus eugenii | 1 |

Not all of the protein sequences in each of the RNA-binding classes are specified as such in the protein sequence database. We have manually checked all the selected RNA-binding protein sequences to ensure the data quality. The number of known snRNA- and snoRNA-binding proteins is much lower than those in the other groups[195, 196] and it is substantially below the number of 80~100 sequences needed to properly train a SVM protein classification system. In order to evaluate the performance of SVM on classification of a small protein class, the classification of snRNA binding proteins was also studied in this work.

All distinct members in each group are used to construct positive samples for training, testing and independent evaluation of SVM classification system. The negative samples for training and testing are selected from seed proteins of the curated protein families in the Pfam database[169] excluding those that belong to the group of RNA-binding proteins under study. For each group of non-rRNA-, non-mRNA-, non-tRNA-, non-snRNA-binding proteins, distinct members in the other three groups are added to the negative samples of each of the training, testing and independent evaluation set. It is expected that the number of negative samples in each of these three groups may be higher than that in the group of negative samples for all RNA-binding proteins.

### 4.2.2. Results and discussion

The number of positive and negative samples for each of the training, testing and independent evaluation set for each group of RNA-binding proteins is given in Table 4-5. The training set is composed of 708 rRNA-binding and 972 non-rRNA-binding proteins, 277 mRNA-binding and 2106 non-mRNA-binding proteins, 94 tRNA-binding and 792 non-tRNA-binding proteins, 33 snRNA-binding proteins and 1988 non-snRNA-binding proteins, and 2161 RNA-binding proteins and 2965 non-RNA-binding proteins. The testing set is comprised of 1245 rRNA-binding and 9044 non-rRNA-binding proteins, 129 mRNA-binding and 10164 non-mRNA-binding proteins, 114 tRNA-binding and 9297 non-tRNA-binding proteins, and 1850 RNA-binding proteins and 6816 non-RNA-binding proteins. The independent evaluation set is made of 101 rRNA-binding and 4997 non-rRNA-binding proteins, 164 mRNA-binding and 6046 non-mRNA-binding proteins, 51 tRNA-binding and 5033 non-tRNA-binding proteins, 20 snRNA-binding and 6151 non-snRNA-binding proteins, and 447 RNA-binding proteins and 4881 non-RNA-binding proteins.

Table 4-5 Prediction accuracies and number of positive and negative samples in the training, testing, and independent evaluation set of rRNA-, mRNA-, tRNA-, and snRNA-binding proteins and of all RNA-binding proteins respectively. Predicted results are given in TP (true positive), FN (false negative), TN (true negative), FP (false positive), sensitivity SE=TP/(TP+FN), specificity SP=TN/(TN+FP), and Q (overall accuracy, Q=(TN+TP)/(TP+FN+TN+FP)). Number of positive or negative samples in the testing and independent evaluation sets is TP+FN or TN+FP respectively.

| **Protein family** | **Training set** | | **Testing set** | | | | **Independent evaluation set** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | positive | negative | positive | | negative | | positive | | | negative | | | *Q*(%) |
| | | | *TP* | *FN* | *TN* | *FP* | *TP* | *FN* | *SE* (%) | *TN* | *FP* | *SP* (%) | |
| RNA<br>-binding | 2161 | 2965 | 1844 | 6 | 6802 | 14 | 437 | 10 | 97.8 | 4685 | 196 | 96.0 | 96.1 |
| rRNA<br>-binding | 708 | 972 | 1243 | 2 | 9031 | 13 | 95 | 6 | 94.1 | 4931 | 66 | 98.7 | 98.6 |
| mRNA<br>-binding | 277 | 2106 | 129 | 0 | 10164 | 0 | 130 | 34 | 79.3 | 5833 | 213 | 96.5 | 96.0 |
| tRNA<br>-binding | 94 | 792 | 114 | 0 | 9295 | 2 | 48 | 3 | 94.1 | 5028 | 5 | 99.9 | 99.8 |
| snRNA<br>-binding | 33 | 1988 | 7 | 0 | 10373 | 1 | 9 | 11 | 41.0 | 6133 | 18 | 99.7 | 99.5 |

### 4.2.2.1. Overall prediction accuracy

The numbers and prediction results of specific class of RNA-binding proteins and non-class-members are given in Table 4-5. In this table, TP stands for true positive (correctly predicted RNA-binding proteins of specific class), FN stands for false negative (specific class of RNA-binding proteins incorrectly predicted as non-class-members), TN stands for true negative (correctly predicted non-class-members), and FP stands for false positive (non-class-members incorrectly predicted as specific class of RNA-binding proteins). The predicted sensitivity SE for rRNA-, mRNA-, tRNA-, snRNA-binding proteins and all RNA-binding proteins, which measures the overall prediction accuracy for each class of RNA-binding proteins, is 94.1%, 79.3%, 94.1%, 41.0% and 97.8% respectively. The predicted specificity SP for non-rRNA-, non-mRNA-, non-tRNA-, non-snRNA-binding proteins and all non-RNA-binding proteins, which measures prediction accuracy for each group of non-RNA-binding proteins, is 98.7%, 96.5%, 99.9% 99.7% and 96.0% respectively.

A direct comparison with results from previous protein studies is inappropriate because of the differences in the specific aspects of proteins classified, dataset, descriptors and classification methods. Nevertheless, a tentative comparison may provide some crude estimate regarding the level of accuracy of our method with respect to those achieved by other studies of proteins. With the exception of snRNA-binding proteins, the range of accuracy for the prediction of each class of RNA-binding proteins from our study is from 79.3% to 97.8%, which is comparable to or better than the level of accuracy obtained from other SVM studies of proteins[32, 43, 44, 197-201] as summarized in the Table 4-6.

.

Table 4-6. Performance of Support Vector Machines for predicting protein functional classes as reported in the literature. All of the data and results were collected from the original papers. N+, N- and N are the number of class members, non-members and all proteins (members + non-members) respectively, SE and SP are prediction accuracy for class members and non-members respectively, Q is the overall accuracy.

| **Protein Functional Class** | **Protein Sub-Classes** | **Protein Descriptors** | **Number of Proteins in Training Set N (N+/N-)** | **Validation Method** | **Reported Prediction Accuracy** | | | **Ref** |
|---|---|---|---|---|---|---|---|---|
| | | | | | **SE (%)** | **SP (%)** | **Q(%)** | |
| G-protein coupled receptors | All GPCRs | Physicochemical properties | 2247 (927/1320) | Independent evaluation | 95.6 | 98.1 | 97.4 | 43 |
| | | Dipeptide composition | 3302(778/2524) | 5-fold CV | 98.6 | 99.8 | 99.5 | 197 |
| | Gi/o binding type | Structural characteristics (extra cellular loops, intracellular loops etc) | 132(61/71) | 4-fold CV | 77.0 | 78.3 | | 198 |
| | Gq/11 binding type | | 132(47/85) | 4-fold CV | 68.1 | 72.7 | | |
| | Gs binding type | | 132(24/108) | 4-fold CV | 83.3 | 95.2 | | |
| Nuclear receptors | | Amino acid composition | 282 | 5-fold CV | | | 82.6 | 32 |
| | | Dipeptide composition | 282 | 5-fold CV | | | 97.5 | |
| | | Physicochemical properties | 872(334/538) | Independent evaluation | 89.5 | 97.6 | | 43 |
| DNA-binding proteins | | Amino acid composition, limited range correlation of hydrophobicity, solvent accessible surface area | 12507 (7739/4768) | 10-fold CV | 92.8 | 77.1 | 86.8 | 44 |
| | | Surface and overall composition, overall charge and positive potential patches on the protein | 359 (121/238) | 5-fold CV | 89.1 | 82.1 | 93.9 | 199 |
| | | | | Jackknife | 90.5 | 81.8 | 94.9 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | surface | | leave 1-pair holdout | 86.3 | 80.6 | 87.5 | |
| Transmembrane proteins | Functional Domain Composition | 2059 | jackknife test | | | 86.3 | 200 |
| | | | independent test | | | 67.5 | |
| | | | self-consistency | | | 93.9 | |
| | Pseudo-amino acid composition | 2059 | jackknife test | | | 82.4 | 201 |
| | | | independent test | | | 90.3 | |
| | | | self-consistency | | | 99.9 | |
| | Physicochemical properties | 4668(2105/2563) | Independent evaluation | 90.1 | 86.7 | 86.7 | 43 |

As a statistical learning method, sufficient number of samples is needed in order to properly train and test a SVM classification system. The total number of available snRNA-binding protein sequences is only 60, from which a very small training set of 33 sequences is used in this work. As less positive examples tends to be less adequate or not enough in representing all types of proteins in a class, it is thus not surprising to find that the prediction accuracy for this RNA-binding class is at a very low level of 40%, in contrast to the level of 79.3% to 97.8% for other RNA-binding classes.

The prediction accuracy for each group of non-RNA-binding proteins appears to be better than that for the corresponding group of RNA-binding proteins. The higher prediction accuracy for non-RNA-binding proteins likely results from the availability of sufficiently diverse set of non-RNA-binding proteins than that of RNA-binding proteins, which enables SVM to perform a better statistical learning for recognition of non-RNA-binding proteins. Based on the statistics provided on the webpage of Pfam database, there are more than 7,000 families of proteins, from which one can generate a diverse set of non-RNA-binding proteins.

Inspection of individual misclassified protein sequences of different RNA-binding and non-RNA-binding classes, including those false negatives and false positives in the independent evaluation data sets, shows that a significant portion of these wrongly predicted protein sequences are either protein fragment or described as hypothetical, probable, or putative. Sequence incompleteness likely contributes to some of the prediction errors in this work. Many of the hypothetical, probable, and putative proteins are described primarily based on some form of distant sequence similarity relationship with existing proteins of known functions.

The accuracy measures of the SVM prediction suggested that the prediction on

RNA-binding proteins is less accurate than that of non-RNA-binding proteins. One possible reason is that SVM based on an unbalanced datasets tends to produce feature vectors that push the hyperplane towards the side with smaller number of data [202], that can lead to a reduced accuracy for the set either with a smaller number of samples or of less diversity. It is however inappropriate to simply reduce the size of non-RNA-binding proteins to artificially match that of RNA-binding proteins, since this compromises the diversity needed to fully represent all non-RNA-binding proteins. Computational methods for re-adjusting biased shift of hyperplane have been introduced[166]. Application of these methods may help improving SVM prediction accuracy in this and other cases involving unbalanced data.

#### *4.2.2.2. Classification of proteins with specific characteristics*

A number of RNA-binding proteins have a molecular structure and contain RNA-binding domains of 70-150 amino acids that mediate RNA recognition [203, 204]. Three classes of RNA-binding domains have been documented to bind RNA in a sequence independent manner, and these domains are RNA-recognition motif (RRM), double-stranded RNA-binding motif (dsRM), and K-homology (KH) domain[204]. A fourth class of RNA-binding domain, S1 RNA-binding domain, has also been found in a number of RNA-associated proteins[205]. These domains have distinguished structural features responsible for RNA recognition and binding. Thus the performance of SVM classification of RNA-binding proteins can be evaluated by examining whether or not proteins containing one of these domains can be correctly classified as RNA-binding proteins.

A search of protein family and sequence databases shows that there are a total of 260, 74, 190, and 41 RNA-binding protein sequences known to contain RRM, dsRM, KH and S1 RNA-binding domain respectively. The majority of these sequences are included in the training and testing set of all RNA-binding proteins. In the corresponding independent evaluation set, there are 35, 16, 93, and 10 sequences containing RRM, dsRM, KH, and S1 RNA-binding domain respectively. The prediction status and examples of these protein sequences are given in Table 4-7. All but one protein sequence are correctly classified as RNA-binding by SVM, which shows the capability of our trained SVM classification system. The only incorrectly predicted protein sequence is HnRNP-E2 protein fragment in the group that contains KH domain. The incompleteness of this sequence might partially contribute to its incorrect prediction by SVM. Thus, it is suggested that one must be aware of the pitfalls in statistical analysis of prediction accuracies of testing data.

Some RNA-binding proteins are in a primarily sequence-specific manner. Typical examples are ribosomal proteins[187] and a U8 snoRNA-specific binding protein[196]. Majority of the ribosomal protein entries are correctly predicted as rRNA-binding proteins. Inspection of the ribosomal protein entries that are incorrectly predicted as a non-rRNA-binding protein shows that some of these entries are protein fragment and some are described as hypothetical, probable, or putative. The prediction error for some of these sequences may be partly due to sequence incompleteness or low sequence similarity to those of other protein sequences in each class. Some ribosomal proteins are known to bind to mRNA and tRNA as well as rRNA, examples of these proteins are 30S ribosomal protein S1, S3, S4. The multiple binding natures of these proteins likely makes it more difficult for a statistical classification system such as SVM to unambiguously distinguish the features between rRNA-binding, mRNA-binding and tRNA-binding, which is another possible reason for the inaccurate classification of these sequences.

Some proteins, such as dihydrofolate reductase and thymidylate synthase, are known to bind to their own mRNA[206]. Not all of these proteins are listed as RNA binding proteins in protein sequence databases. As a result, these mRNA-binding proteins may not be included in the right protein group, which likely affects prediction accuracy on these proteins. Hence, additional work is needed to search for these proteins and include them in the group of mRNA-binding proteins.

Table 4-7 Prediction statistics, examples and host species of RNA-binding protein sequences known to contain one of the RNA-recognition motif (RRM), double-stranded RNA-binding motif (dsRM), K-homology (KH), and S1 RNA-binding domain. Only those RNA-binding proteins in the independent evaluation sets are included. Host species of some protein sequences are not provided because the relevant information is not yet available in the protein sequence database. The only incorrectly predicted protein sequence with KH domain is HnRNP-E2 protein fragment.

| RNA-Binding Domain | RNA-Binding Proteins Known to Contain Domain | | | |
|---|---|---|---|---|
| | Number of RNA-binding proteins with domain | Number of Proteins Correctly Predicted as RNA-biding | Example of correctly predicted protein (host species) | Prediction Accuracy (%) |
| **RRM** | 35 | 35 | CUG triplet repeat RNA-binding protein 1 (Homo sapiens)<br>ELAV-like protein 2 (Mus musculus)<br>ELAV-like protein 4 (Homo sapiens, Rattus norvegicus)<br>Heterogeneous nuclear ribonucleoprotein A1 (Mus musculus)<br>Heterogeneous nuclear ribonucleoprotein A3 (Homo sapiens, Xenopus laevis)<br>Heterogeneous nuclear ribonucleoprotein H (Homo sapiens)<br>Matrin 3 (Rattus norvegicus)<br>Nuclear polyadenylated RNA-binding protein NAB4 (Candida albicans)<br>Polypyrimidine tract-binding protein 1 (Rattus norvegicus)<br>RNA-binding protein FUS (Mus musculus)<br>RNA-binding region containing protein 2 (Mus musculus)<br>Splicing factor, arginine/serine-rich 4 (Mus musculus)<br>Splicing factor, arginine/serine-rich 5 (Homo sapiens)<br>Splicing factor U2AF 65 kDa subunit (Mus musculus, Caenorhabditis elegans) | 100% |
| **dsRM** | 16 | 16 | ATP-dependent RNA helicase A (Bos taurus)<br>Interleukin enhancer-binding factor 3 (Mus musculus, Rattus norvegicus)<br>Ribonuclease III (Escherichia coli, Ralstonia solanacearum, Brucella melitensis, Salmonella typhi, Yersinia pestis, Rhizobium meliloti, Staphylococcus aureus (strain N315), Neisseria meningitidis (serogroup A), Neisseria meningitidis (serogroup B), Chlamydia muridarum, Helicobacter pylori J99) | 100% |

| | | | | |
|---|---|---|---|---|
| | | | SON protein (Mus musculus) | |
| **KH** | 94 | 93 | 30S ribosomal protein S3 (Mycobacterium bovis, Escherichia coli, Mycoplasma pneumoniae, Buchnera aphidicola (subsp. Acyrthosiphon kondoi), Acholeplasma florum, Buchnera aphidicola (subsp. Acyrthosiphon pisum), Synechocystis sp. (strain PCC 6803), Thermus thermophilus, Phytoplasma sp. (strain STRAWB2), Mycoplasma capricolum, Acholeplasma sp. (strain ATCC J233), Fusobacterium nucleatum (subsp. nucleatum), etc.)<br>A kinase anchor protein 1 (Homo sapiens, Mus musculus)<br>GTP-binding protein era homolog (Streptococcus pyogenes (serotype M3), Streptococcus pneumoniae, Fusobacterium nucleatum (subsp. nucleatum), Clostridium perfringens, Anabaena sp. (strain PCC 7120), Mycoplasma pulmonis, Staphylococcus aureus (strain Mu50 / ATCC 700699), Neisseria meningitidis (serogroup A), Neisseria meningitidis (serogroup B), Bacillus halodurans, Lactococcus lactis (subsp. lactis), Helicobacter pylori J99)<br>Hypothetical UPF0109 protein TC0030 (Chlamydia muridarum)<br>N utilization substance protein A homolog (Bacillus halodurans, Rickettsia conorii)<br>Poly(rC)-binding protein 1 (Oryctolagus cuniculus)<br>Poly(rC)-binding protein 2 (Homo sapiens)<br>Poly(rC)-binding protein 3 (Mus musculus)<br>Poly(rC)-binding protein 4 (Mus musculus)<br>Polyribonucleotide nucleotidyltransferase (Bacillus subtilis, Buchnera aphidicola (subsp. Schizaphis graminum))<br>Probable exosome complex RNA-binding protein 1 (Methanosarcina mazei, Thermoplasma acidophilum, Pyrococcus abyssi)<br>Heterogeneous nuclear ribonucleoprotein K (Oryctolagus cuniculus)<br>Vigilin (Gallus gallus)<br>Zipcode-binding protein 2 (Gallus gallus) | 98.9% |
| **S1 RNA binding domain** | 10 | 10 | 30S ribosomal protein S1 (Chlamydia trachomatis, Chlamydia pneumoniae)<br>Eukaryotic translation initiation factor 2 (Rattus norvegicus)<br>N utilization substance protein A homolog (Buchnera aphidicola (subsp. Schizaphis graminum))<br>Probable translation initiation factor 2 alpha subunit (Methanopyrus kandleri, Pyrococcus furiosus, Sulfolobus tokodaii, Pyrococcus abyssi)<br>Ribonuclease E (Buchnera aphidicola (subsp. Schizaphis graminum)) | 100% |

.

### 4.2.2.3. Contribution of feature properties to the classification of RNA-binding proteins

We choose a total of nine feature properties for describing physicochemical characteristics of each protein, which have been routinely used in previous studies of proteins [14, 43, 47, 168, 182]. However, not all feature vectors contribute equally to the classification of proteins, some have been found to play relatively more prominent role than others in specific aspects of proteins [47]. It is therefore of interest to examine which feature properties play more prominent role in classification of RNA-binding proteins.

The contribution of individual feature property to protein classification is investigated by separately conducting classification using each feature property. Our analysis on the classification of all RNA-binding proteins suggests that, in order of prominence, the amino acid composition, charge, polarity, hydrophobicity play more prominent role than other feature properties. Amino acid composition and hydrophobicity are important factors for the interaction of a protein with other biomolecules as well as for structural folding. On the other hand, charge and polarity is important for electrostatic interactions and hydrogen-bonding to RNA. As the backbone of RNA is charged, charge and polarity are expected to be particularly important feature properties for the binding of a protein with its RNA-substrate. A recent study of the dynamics of protein-RNA interfaces showed that actions condensed around RNA affect the binding of protein to RNA [207], which is indicative of the strong effect of charges and polarity.

## 4.3. Classification of Transporters (Paper III)

Transporters play key roles in transporting cellular molecules across cell and cellular compartment boundaries, mediating the absorption and removal of various molecules, and regulating the concentration of metabolites and ionic species [208-210]. Specific transporters have been explored as therapeutic targets [211-213] and a variety of transporters are responsible for the absorption, distribution and excretion of drugs [214, 215]. Functional assignment of transporters is important for facilitating functional study of genomes, for probing molecular mechanism of cellular processes and diseases, and for searching new therapeutic targets and pharmacologically relevant transporters.

There are active and passive transporters. Active transporters couple solute transport to the input of energy and these can be divided into two classes: ion-coupled and ATP-dependent transporters. Ion-coupled transporters link uphill solute transport to downhill electrochemical ion gradients. ATP-dependent transporters are directly energized by the hydrolysis of ATP and they transport a heterogeneous set of substrates. Passive transporters include facilitated transporters and channels, which allow the diffusion of solutes across membranes. These transporters evolve from common themes into families of different architectures [208, 216, 217].

Functional families of transporters are described by the transporter classification (TC) system (http://www-biology.ucsd.edu/msaier/transport/) based on their mode of transport, energy coupling mechanism, molecular phylogeny and substrate specificity [217]. In particular, transport mode and the energy coupling mechanism have been used as the primary basis for transporter family classification due to their relatively stable characteristics [217]. Therefore, transporters in a TC family share common transport modes and mechanisms. In cases that the precise function of a transporter is unknown, prediction of its TC family provides useful hint about its broad transportation role, mode of action and substrate classes.

TC families are classified at four levels (TC class, TC sub-class, TC family, and TC sub-family) as indicated by a specific TC number TC I.X.J.K.L. Here I=1, …, 9 represents each of the 9 TC classes, X=A, B, C, D, E, … represents each of the TC sub-classes that belong to a TC class, J=1, … represents each of the TC families that belong to a TC sub-class, K=1, … represents each of the TC sub-families that belong to a TC family, and L=1, … represents individual transporters under a sub-family.

So far, sequence alignment and clustering are used widely for predicting the TC family as well as the function of transporters [218, 219]. Some transporters are known to have no or low homology to other proteins of known function [220-223]. Substantial portions of transporters in different TC families have very low sequence identity to other family members. For instance, a member of the multidrug transporter family, bmr3, has only 7% sequence identity and 17% similarity to another family member blt [223]. A K+ channel, TASK-2, has 18-22% sequence identity to other members of the two pore domain K+ channel family such as TWIK-1, TREK-1, TASK-1, and TRAAK [224]. Two members of the major facilitator family, GlpT and LacY, are 21% identical to each other [225]. Thus the function of some of these transporters may be difficult to assign based solely on homology [16, 226], and methods that predict protein function without the use of sequence similarity are needed.

Several methods have been developed for predicting protein function without sequence alignment and clustering. Some of these explore structural features [10], interaction profiles [13, 14], and protein/gene fusion data [11, 12]. Others conduct functional family assignment by using statistical learning methods including neural networks and SVM [31, 34, 43, 45, 46]. These methods have been tested by using a variety of proteins including enzymes, receptors and transmembrane proteins. While these methods have not been specifically tested for transporters, some of these methods are expected to be applicable to transporters.

One approach for protein functional classification has shown useful capability for functional family assignment of distantly related proteins as well as homologous proteins at high accuracy rates [43, 227, 228]. Some SVM systems have been developed to classify proteins into functional families defined from activities and physicochemical properties rather than sequence similarity [14, 31, 43, 45, 46]. In training a SVM classification system, proteins represented by their sequence-derived physicochemical properties are projected onto a hyperspace where proteins in a family are separated from those outside the family by a hyperplane. By projecting a new sequence onto this hyperspace, the SVM system can determine whether the corresponding protein belongs to the family based on its location with respect to the hyperplane. To some extent, no sequence similarity is required in this process. The overall accuracy of functional family prediction is 87%, based on the test of 34,582 proteins. The accuracy for the correct assignment of non-family-members is 97%, based on the test of 310,000 proteins [43, 45, 46]. Thus SVM appears to be a useful alternative approach for predicting the TC family of transporters irrespective of sequence similarity.

So far, SVM and other statistical learning methods have not been explored for predicting transporter families, due in large part to the limited information about transporters. The relevant data in the transporter databases [229, 230] has now reached to a level useful for using SVM to predict transporter families. A survey of the transporter databases [229, 230] finds that the number of known transporters in each of the 13 TC sub-classes and 8 TC families is no less than 80-100, which is typically needed for properly training a SVM protein classification system [43]. Thus, in this work, transporter family classification is conducted at the sub-class level for the 13 TC sub-classes and at the family level for the 8 TC families.

### 4.3.1. Selection of transports and non-members of TC sub-classes and TC families

The seed transporters for each of the 20 known TC sub-class are from the TCDB database [230]. A BLAST search is conducted to scan the Swissprot database [231] for finding additional transporters in each sub-class. There is no seed transporter for the TC1.D, TC2.B and TC9.C sub-classes, and the number of collected transporters in the TC3.B, TC3.C, TC5.A and TC5.B sub-classes are substantially less than the number of 80-100 typically needed for properly training a SVM protein classification system [43]. Moreover, there are 8 TC families found to have more than 80 transporters. Thus 13 TC sub-classes with a combined number of 14,987 transporters, and 8 TC families with a combined number of 2684 transporters are studied in this work. All distinct members in each sub-class are used to construct positive samples for training, testing and independent evaluation of the SVM classification system.

The negative samples of each TC sub-class/family for training and testing our SVM classification systems refer to those proteins outside this sub-class/family which include both non-transporter proteins and transporters of other sub-classes and families. These negative samples are selected from seed proteins of the 7,316 curated protein families in the Pfam database [232] that have no protein as a member of that sub-class. Each negative set contains at least one randomly selected seed protein from each of the 7,316 Pfam families. For the group of negative samples of a sub-class, distinct members in the other sub-classes are added to the group of each of the training, testing and independent evaluation set. It is expected that the number of negative samples in each of these groups may be higher than that of non-transporter proteins.

The performance of SVM classification is further evaluated by using an independent evaluation set, which is composed of all of the proteins in each sub-class/family and those outside the sub-class/family that have not been used in the training and optimization the SVM system. No duplicate protein entry is used in the training, testing and independent evaluation set for each group. The number of positive and negative samples for each of the training, testing and independent evaluation set for each of the 13 transporter sub-classes is given in Appendix A, as indicated in “Protein family” as transporters.

### 4.3.2. Results and Discussion

Statistics of the datasets and prediction results of each of the 13 TC sub-classes of transporters and those of the 8 TC families are given in Appendix Table A(as indicated in “Protein family” as transporters). The computed TP, TN, FP, FN, $Q_p$ and $Q_n$ and C for each TC sub-class and family by using the respective testing and independent evaluation sets are also given respectively. The computed $Q_{p,}$, $Q_n$ and C for the 13 TC sub-classes is in the range of 70.7% to 96.1%, 97.6% to 99.9% and 69.7% to 96.5% respectively, and those for the 8 TC families is in the range of 60.6%~97.1% and 91.5%~99.4% respectively. The overall accuracies for the assignment of 4,351 and 770 transporters into their respective TC sub-class and TC family are 83.4% and 88.0% respectively, and those for the correct assignment of 83,151 and 57, 951 non-members of TC sub-classes and families are 99.3% and 96.6% respectively. These accuracies are comparable to the overall accuracy of 86% for the SVM assignment of the enzymes classification previously in section 4.1.

In order to evaluate the capability of SVM classification systems for distinguishing between transporters of a particular TC sub-class and transmembrane proteins outside that sub-class, all of the transmembrane proteins known to not belong to each of the 13 investigated TC sub-classes are collected and used to test the corresponding SVM classifier. A total of 26,139 such transmembrane proteins are found from the SwissProt database[231]. The number of transmembrane proteins outside each of the 13 TC sub-classes and the SVM prediction results are given in Table 4-8. It is shown that 90.4% to 99.6% of the transmembrane proteins outside each TC sub-class are correctly predicted to be non-members of that sub-class, suggesting that our SVM classification systems have certain level of capability for separating transporter members and transmembrane non-members of transporter families.

Table 4-8 Transmembrane proteins outside each of the TC families and SVM prediction results for these proteins.

| **Transporter sub-class** | **Transmembrane proteins outside the sub-class** | **Prediction results** | | |
|---|---|---|---|---|
| | | **Predicted as non-member** | **Predicted as member** | **Prediction accuracy** |
| TC 1.A α-Type channels | 25456 | 24599 | 857 | 96.6% |
| TC 1.B β-Barrel porins | 26011 | 25816 | 195 | 99.3% |
| TC 1.C Pore-forming toxins | 26061 | 23565 | 2496 | 90.4% |
| TC 1.E Holins | 26101 | 26001 | 100 | 99.6% |
| TC 2.A Porters (uniporters, symporters, and antiporters) | 25439 | 24321 | 1118 | 95.6% |
| TC 2.C Ion gradient-driven energizers | 26100 | 26049 | 51 | 99.8% |
| TC 3.A Diphosphate bond hydrolysis-driven transporters | 25559 | 23244 | 2315 | 90.9% |
| TC 3.D Oxidoreduction-driven transporters | 24266 | 23498 | 768 | 96.8% |
| TC 3.E Light absorption-driven transporters | 24929 | 24684 | 245 | 99.0% |
| TC 4.A Phosphotransfer-driven group translocators | 26062 | 25753 | 309 | 98.8% |
| TC 8.A Auxiliary transport proteins | 26053 | 25915 | 138 | 99.5% |
| TC 9.A Transporters of unknown biochemical mechanism | 26085 | 25647 | 438 | 98.3% |
| TC 9.B Putative but uncharacterized transport proteins | 25815 | 24246 | 1569 | 93.9% |

The prediction accuracy for the non-members of each TC sub-class/family appears to be higher than that for the transporters in the sub-class/family. It is likely resulting from the availability of a significantly more diverse set of non-transporter proteins than that of transporters, which enables the training of a system with higher capability for recognizing non-members of a TC sub-class or family. There are over 7,316 families of proteins Pfam database [232], from which a diverse set of non-members for each TC sub-class or family can be generated.

Examples of the predicted true positive, false negative, true negative and false positive protein sequences of each of the 13 sub-classes are given in Table 4-9. Inspection of the false negative transporters and the false positive non-members of each sub-class show that a substantial percentage of these incorrectly predicted proteins are actually sequence fragment entries of the corresponding protein, which likely contributes to some of the prediction errors in this work.

Table 4-9 Examples of the predicted true positive (TP), true negative (TN), false positive (FP), false negative (FN) protein entries of different TC sub-classes. Only proteins in the independent evaluation sets are included in this Table. Host species of some protein sequences are not provided because the relevant information is not yet available in the protein sequence database.

| Protein class | Prediction category | Example of predicted proteins |
|---|---|---|
| TC 1.A α-Type channels | TP | Cyclic-nucleotide-gated cation channel<br>Calcium transporter CaT1<br>Transient receptor potential cation channel protein<br>P2X purinoceptor 1<br>Glycine receptor alpha-1 chain precursor<br>Glutamate-gated chloride channel<br>Outwardly rectifying chloride channel<br>CLC-Nt2 protein<br>Urea transporter; Structural polyprotein P130<br>Magnesium and cobalt transport protein corA<br>VPU protein |
| | TN | V1A arginine vasopressin receptor<br>ATP-binding protein of ABC transporter<br>Chorionic gonadotropin beta subunit (Fragment)<br>16 kDa heat shock protein A<br>Methyl-accepting chemotaxis protein<br>Ribulose-1,5-bisphosphate carboxylase<br>alsyntenin-1 precursor<br>DsRNA-binding protein<br>NADH gehydrogenase 8 subunit (Fragment) |
| | FP | Probable G-protein-coupled receptor Mth-like 10 precursor<br>CG18678 protein<br>Envelope glycoprotein (Fragment)<br>Sulfonylurea receptor-1 (Fragment)<br>Hfq protein; Short transient receptor potential channel 2<br>Lantibiotic epidermin precursor<br>Neuromedin U-25<br>P0492F05.25 protein<br>Phosphatidylserine synthase-2<br>Dentatorubro-pallidoluysian atrophy protein (Fragment)<br>RNA replicase beta chain (Fragment)<br>C14orf1-like protein |

| | | |
|---|---|---|
| | FN | PBCV-1 K+ ion channel protein<br>Melastatin 1<br>Channel protein (Hypothetical protein)<br>Calcium-activated chloride channel protein 2<br>Non-selective cation channel<br>NADPH thyroid oxidase 2<br>BspA protein precursor<br>YKUT protein. |
| TC 1.B β-Barrel porins | TP | Outer membrane protein C precursor<br>Sucrose porin precursor<br>Voltage-dependent anion-selective channel protein 3<br>Long-chain fatty acid transport protein precursor<br>Hemoglobin receptor<br>Lactoferrin-binding protein B precursor<br>Cation efflux system protein cusC precursor<br>Porin B precursor |
| | TN | GRH receptor-2<br>Putative pheromone receptor<br>ATPase alpha subunit (Fragment)<br>Translation elongation factor 1-alpha (Fragment)<br>Chaperone protein dnaK<br>Decoy TNF receptor<br>ADP-ribosylation factor-like protein<br>Cytochrome b (Fragment)<br>Cytochrome b. |
| | FP | R09B5.5 protein.<br>Hypothetical 21.7 kDa protein.<br>Photosystem II reaction center X protein.<br>PAR-1a protein<br>Putative FKBP-type peptidyl-prolyl cis-trans isomerase<br>Replication-related protein.<br>HrpF.<br>Putative nitrate-induced protein.<br>Homeodomain protein vaamana.<br>Probable soluble cytochrome b562 2 precursor |
| | FN | GnRH receptor-2.<br>Putative pheromone receptor.<br>ATPase alpha subunit (Fragment).<br>Ribulose 1,5-bisphosphate carboxylase large subunit (fagment).<br>2010109I03Rik protein.<br>Hypothetical protein All2748.<br>Blastomere-cadherin precursor<br>Cytochrome c.<br>Ribonuclease III |
| TC 1.C Pore-forming toxins | TP | Alpha-toxin.<br>Bifunctional hemolysin-adenylate cyclase precursor<br>Plantaricin S beta protein precursor.<br>Mastoparan B.<br>Crabrolin.<br>Myeloid cathelicidin 1 precursor.<br>Defensin precursor.<br>Cytotoxin L |

| | | |
|---|---|---|
| | TN | Re6 receptor long splice variant.<br>Cell division protein FTSE<br>ATP synthase beta subunit (Fragment).<br>Translation elongation factor 1 alpha (Fragment).<br>Methyl-accepting chemotaxis protein.<br>Tumor necrosis factor receptor superfamily member 11B precursor<br>Matrix metalloproteinase 9 precursor<br>Hemoglobin alpha chain. |
| | FP | NAD-glycohydrolase.<br>Apolipophorin-III precursor<br>Photosystem I reaction centre subunit XII precursor.<br>Steroid receptor coactivator 1a<br>Bll2849 protein.<br>Probable spore cortex-lytic enzyme.<br>Hypothetical protein NMA0089. |
| | FN | Enterocin P precursor.<br>Myeloid secondary granule protein.<br>Countin.<br>Lactococcin 972 precursor.<br>Hemolysin BL lytic component L2.<br>Beta2-toxin. |
| TC 1.E Holins | TP | Lysis protein S.<br>Extracellular secretory protein.<br>Holin.<br>LrgA family protein. |
| | TN | Long-wavelength opsin (Fragment).<br>G-protein-coupled receptor Mth2 precursor<br>Hypothetical protein CBU1189.<br>ABC transporter ATP-binding protein-oligopeptide transport.<br>ATP synthase beta subunit (Fragment).<br>Glycoprotein hormone beta 5 precursor<br>CG14207-PB.<br>NADP-dependent malate dehydrogenase (Fragment).<br>GRAAL2 protein precursor.<br>LDL receptor-related protein 6. |
| | FP | NADH dehydrogenase.<br>YVLD.<br>Cytochrome c oxidase, cbb3-type, CcoQ subunit.<br>Probable transmembrane protein. |
| | FN | Long-wavelength opsin (Fragment).<br>G-protein-coupled receptor Mth2 precursor<br>Hypothetical protein CBU1189.<br>ABC transporter ATP-binding protein-oligopeptide transport.<br>ATP synthase beta subunit (Fragment).<br>Glycoprotein hormone beta 5 precursor<br>CG14207-PB. |

| | | |
|---|---|---|
| TC 2.A Porters (uniporters, symporters, and antiporters) | TP | Hexose transporter 1.<br>Metabolite transport protein.<br>Integral membrane protein.<br>Inorganic phosphate transporter 1.<br>Organic cation transporter.<br>Feline leukemia virus subgroup C receptor FLVCR.<br>Aromatic amino acid and leucine permease. |
| | TN | Orphan G protein-coupled receptor Ren 1.<br>Glucagon receptor.<br>Thyrotropin beta subunit precursor.<br>Elongation factor 1a (Fragment).<br>U-plasminogen activator receptor form 2-human (Fragment).<br>Actin I.<br>At2g14250 protein. |
| | FP | Glutamate receptor 3.1 precursor<br>NADH-ubiquinone oxidoreductase chain 1<br>Manganese transport system membrane protein mntC.<br>Iron ABC transporter, permease protein.<br>NADH dehydrogenase subunit 4.<br>Pollen coat oleosin.<br>G-protein coupled receptor GPR110. |
| | FN | Carboxypeptidase II (Fragment).<br>Glucose uptake protein.<br>Lysine and histidine specific transporter.<br>Sodium proton exchanger NHX1 (Fragment).<br>KtrB protein.<br>Purine nucleoside permease.<br>MNHG NA+/H+ antiporter subunit<br>Bilitranslocase.<br>Threonine export carrier. |
| TC 2.C Ion gradient-driven energizers | TP | TonB protein<br>Biopolymer transport exbB protein<br>TolQ protein<br>TolR protein |
| | TN | Olfactory receptor-like protein 42-2 (Fragment).<br>Metabotropic glutamate receptor 7 variant 3.<br>Tat-binding homolog 7, AAA ATPase family protein.<br>Transcription termination factor Rho.<br>Thyroptin beta chain (Fragment).<br>C901 protein.<br>Phospholipase A2-3 (Fragment).<br>Coenzyme A disulfide reductase.<br>Ras-related protein Rab-12 (Fragment). |
| | FP | TAU-1a (Fragment).<br>S164 (Fragment).<br>F22F1.3 protein.<br>Outer-membrane lipoproteins carrier protein precursor.<br>CG13097 protein (SD02943p).<br>46-kDa surface lipoprotein (Fragment).<br>Hypothetical protein XAC3753.<br>Complexin 2 (Synaphin 1) (921-L).<br>Conserved hypothetical protein. |

| | | |
|---|---|---|
| | FN | Peptidoglycan-associated lipoprotein [Precursor]<br>TolA protein<br>TolB protein precursor. |
| TC 3.A Diphosphate bond hydrolysis-driven transporters | TP | ABC transporter ATP-binding subunit.<br>Bacitracin export permease protein bceB.<br>P-type ATPase.<br>Copper-transporting ATPase, P-type<br>Plasma membrane calcium-transporting ATPase 1 |
| | TN | Olfactory receptor (Fragment).<br>Bovine growth hormone-releasing hormone receptor (Fragment).<br>Seven transmembrane helix receptor.<br>Gonadotropin beta-II chain precursor<br>Atonal-like protein 3.<br>Inhibin alpha subunit (Fragment).<br>UL144 protein.<br>Ras-like small monomeric GTP-binding protein. |
| | FP | Chemotaxis sensory transducer protein.<br>Putative signal peptidase IB.<br>Hydrogenase expression/formation protein HypE (Fragment).<br>Major outer membrane protein OmpA.<br>PugilistDominant (Fragment).<br>Ribosomal small subunit pseudouridine synthase A.<br>Inward rectifier potassium channel 4<br>NADH dehydrogenase subunit 4 (Fragment). |
| | FN | DNA translocase ftsK<br>Tra protein<br>Gene I protein<br>ComG operon protein 1<br>Putative mitochondrial F0-ATPase, mammalian subunit b |
| TC 3. D Oxidoreduction-driven transporters | TP | Cytochrome c oxidase polypeptide IVB<br>Ubiquinol oxidase polypeptide I<br>Cytochrome O ubiquinol oxidase subunit III<br>Protoheme IX farnesyltransferase |
| | TN | SH3P13S.<br>Growth differentiation factor 9B (Fragment).<br>D13L protein.<br>0610005K03Rik protein.<br>Actin (Fragment).<br>Similar to ankyrin-like protein<br>CG7802 protein<br>Aspartic protease Bla g 2 precursor<br>Cyclic nucleotide-gated channel 2b. |
| | FP | Flagella-related protein G.<br>Hypothetical protein VCA0629.<br>Sarcolipin.<br>Hypothetical protein.<br>Protein pufQ.<br>Putative membrane protein MMPS2.<br>Chromosome IV reading frame ORF YDL072C. |

| | | |
|---|---|---|
| | FN | NADH-quinone oxidoreductase chain 2<br>Mbh12 membrane bound hydrogenase alpha<br>Hypothetical protein PF1431.<br>Protoheme IX farnesyltransferase<br>Ubiquinol oxidase polypeptide II precursor |
| TC 3.E Light absorption-driven transporters | TP | Cytochrome b.<br>Photosystem Q(B) protein.<br>Cytochrome B6-F complex iron-sulfur subunit<br>Photosystem II 44 kDa reaction center protein |
| | TN | G protein-coupled receptor 119<br>Glucagon receptor.<br>Metabotropic glutamate receptor 2 precursor<br>ClpB protein.<br>ATP synthase beta subunit (Fragment).<br>Gonadotropin beta-II chain precursor<br>Sugar transport related protein.<br>Coagulation factor XIII, beta subunit. |
| | FP | Cytochrome oxidase 1 (Cox1 protein) (Fragment).<br>Beta polypeptide.<br>NAS-20 protein (Fragment).<br>Glycine betaine transporter (Fragment).<br>Lysosomal-associated transmembrane protein 4A<br>Conserved hypothetical protein. |
| | FN | Photosystem II 44 kDa reaction center protein<br>Photosystem Q(B) protein<br>Cytochrome b6-f complex subunit 4 |
| TC 4.A Phosphotransfer-driven group translocators | TP | PTS system, glucose-specific IIBC component<br>PTS system, fructose-specific IIBC component<br>PTS system, mannitol-specific IIABC component<br>PTS system, lactose-specific IIA component<br>PTS system, N,N'-diacetylchitobiose-specific IIC component<br>AgaC. |
| | TN | Opsin Rh6<br>Bride of sevenless protein precursor.<br>ATP-dependent Clp protease subunit.<br>Hemin transport system ATP-binding protein hmuV.<br>FSH beta-subunit.<br>Latent transforming growth factor beta binding protein 3(Fragment). |
| | FP | PsbA protein.<br>Hypothetical protein SAV2534<br>Ammonia permease.<br>DNA-directed RNA polymerase subunit K (EC 2.7.7.6).<br>Transporter.<br>Integral membrane protein.<br>Carbon starvation protein. |

| | | |
|---|---|---|
| | FN | PtsC1 protein.<br>PTS system, lichenan-specific IIB component<br>Putative phosphotransferase D-arabitol specific component IIC (Fragment).<br>AgaW.<br>Putative PTS system, glucitol/sorbitol-specific enzyme II |
| TC 8.A Auxiliary transport proteins | TP | Tyrosine-protein kinase etk<br>Voltage-gated potassium channel beta-1 subunit<br>Large conductance calcium-dependent potassium ion channel beta 4subunit.<br>Sodium channel beta-1 subunit precursor.<br>Phosphocarrier protein HPr |
| | TN | Olfactory receptor MOR256-1<br>Latrophilin 2 splice variant bbabe.<br>Calcium-sensing receptor related protein 4 (Fragment).<br>ATP-dependent Clp protease ATP-binding subunit clpX.<br>ATP synthase beta subunit (Fragment).<br>Luteinizing hormone beta subunit.<br>Ovomucoid (Fragment).<br>Hepatocyte growth factor precursor<br>Alpha 3B chain of laminin-5 (Fragment). |
| | FP | Ebh protein.<br>HapP1 protein precursor.<br>Similarity to late embryogenesis abundant protein.<br>Prospero-related homeobox 1 (Fragment).<br>DNA helicase-primase complex component.<br>Protein F14.<br>Conserved hypothetical protein. |
| | FN | Chromosome XII COSMID 9449.<br>Potassium voltage-gated channel subfamily E member 1<br>Cardiac phospholamban<br>Transport accessory protein.<br>Pediocin PA-1 biosynthesis protein pedC |
| TC 9.A Transporters of unknown biochemical mechanism | TP | MerC.<br>Peroxisomal targeting signal 1 receptor.<br>IRON(II) transport protein<br>High-affinity iron permease CaFTR2.<br>Lysosome-associated membrane glycoprotein 2 precursor<br>MgtE.<br>MG2+ transporter. |
| | TN | Muscarinic acetylcholine receptor M1.<br>Latrophilin 3 splice variant bbbh.<br>Nuclear valosin-containing protein-like<br>ABC transporter (Fragment).<br>ATP synthase beta chain<br>Thyrotropin beta subunit.<br>Fibrillin 1 precursor |

| | | |
|---|---|---|
| | FP | NADH dehydrogenase subunit I<br>Phosphoenolpyruvate carboxylase, isoform 1 (Fragment).<br>MHBs protein<br>DNA-repair protein complementing XP-G cells homolog<br>Calsequestrin 1.<br>Lysosomal amino acid transporter 1. |
| | FN | Ubiquitin-conjugating enzyme E2-21 kDa<br>PbrT protein.<br>ComC.<br>Putative mercuric ion binding protein.<br>Probable tryptophan transport protein. |
| TC 9.B Putative but uncharacterized transport proteins | TP | Galectin-9.<br>Cytochrome c-type biogenesis protein ccl1.<br>Long-chain-fatty-acid--CoA ligase.<br>Putative chloroquine resistance transporter.<br>Bax inhibitor-1<br>Magnesium and cobalt efflux protein corC. |
| | TN | Olfactory receptor MOR114-5<br>Bovine growth hormone-releasing hormone receptor (Fragment).<br>Gamma-aminobutyric acid type B receptor, subunit 1 precursor<br>Endopeptidase Clp ATP-binding chain B.<br>ATP synthase alpha chain, sodium ion specific<br>Thyroptin beta chain (Fragment).<br>Fibrillin-1 (Fragment). |
| | FP | Rainbow trout DNA for mature peptide, exon2 (Fragment).<br>Annexin max4.<br>Protein export.<br>Vng1454c.<br>Similar to C-14 sterol reductase.<br>19kD alpha zein B5 (Fragment). |
| | FN | Very-long-chain acyl-CoA synthetase<br>Probable crotonobetaine/carnitine-CoA ligase<br>Retrograde regulation protein 3<br>Beta-(1-3)-glucosyl transferase.<br>Conserved hypothetical protein.<br>Hemolysin-related protein, containing CBS domain. |

Because the number of transporters is significantly less than that of non-members, there is an unbalance between the positive and negative training dataset for each sub-class and family. SVM based on an imbalanced dataset tends to generate a hyper-plane closer to the side with smaller number of samples [202], which can lead to a lower prediction accuracy for these samples compared to those on the other side of hyper-plane. This partly explains why the accuracy for assigning the sub-class of transporters is lower than that for the non-members. It is however inappropriate to simply reduce the size of non-members of each sub-class to artificially match that of transporters in the same sub-class, since this compromises the diversity needed to fully represent all non-members. Computational methods for re-adjusting biased shift of the hyper-plane are being developed and evaluated [233]. These methods, when sufficiently developed, may help improving SVM prediction accuracy in this and other cases involving unbalanced data.

# 5. Prediction of the functional class of novel proteins - Specific Case Studies

**The work in this chapter has been published in:**

**IV)** Predicting Functional Family of Novel Enzymes Irrespective of Sequence Similarity: A Statistical Learning Approach. L.Y.Han, C.Z.Cai, Z.L.Ji, Z.W.Cao,J.Cui, Y.Z.Chen Nucleic Acids Res.32(21): 6437-6444(2004).

**V)** Prediction of Functional Class of Novel Viral Proteins by a Statistical Learning Method Irrespective of Sequence Similarity. L.Y.Han, C.Z Cai, Z. L. Ji, Y.Z. Chen. Virology 331(1):136-143 (2005).

**VI)** Prediction of Functional Class of Novel Plant Proteins by a Statistical Learning Method. L. Y. Han, C. J. Zheng, H. H. Lin, J. Cui, H. Li, H. L. Zhang, Z. Q. Tang, and Y. Z. Chen, New Phytologist. 168:109-121(2005)

**VII)** Prediction of Functional Class of Novel Bacterial Proteins without the Use of Sequence Similarity by a Statistical Learning Method.J. Cui, L. Y. Han, C. Z. Cai, C.J.Zheng, Z. L. Ji, and Y. Z. Chen.J. Mol. Microbiol. Biotech. 9 (2): 86-100 (2005)

A fundamental understanding of how biological systems work requires knowledge of protein functions as well as protein interactions. Finding clues of functions is becoming an increasingly important means for better understanding in biological process. For example, exploring functions of certain novel sequences of some bacterial species could help elucidate their pathogenesis potential and reveal novel pathways for drug intervention; large DNA viruses such as poxviruses encode for a variety of proteins that can specifically manipulate the function of host immune factors/messengers, e.g. interferon, interleukins and chemokines. Hints of these novel viral proteins functions are very important for interpreting how the viruses use to interact with their hosts and for searching molecular targets of antiviral therapeutics.

The gap between the large amounts of sequences information resulting from large-scale genome sequencing projects and their function characterization is continuously increasing. In the completely sequenced genome of *Arabidopsis*, the function of 30% of the putative protein-coding open reading frames (ORFs) is unknown [234, 235]. Similar percentage of unknown ORFs is expected in other plant genomes. The function of a substantial percentage (17-20%) of the putative protein-coding open reading frames (ORFs) in many bacterial genomes is unknown [236, 237]. There is also a substantial percentage of the unknown ORFs in the recently determined genomes of Fer-de-lance virus[238], Grapevine fleck virus[239], Indian citrus ringspot virus[240], and SARS coronavirus [241] etc. The same problem arisen when we shift the visual angle from the aspect of taxonomy to the aspect of the protein functional group. There is a large amount of proteins that could perform a specific function or could play a certain biological role haven't been discovered. Thus, increasing efforts have been directed to development of methods for probing protein functions. However, as the sequence of these ORFs mentioned above has no significant similarity to those of known proteins, their functions are difficult to formulate by using sequence alignment and clustering methods. In addition, approaches building upon direct sequence comparisons were lacking of sensitivity and were even unable to identify those novel proteins with remote homologous.

Since our approach for protein function prediction is based on the statistical learning from the physico-chemical properties derived from the primary sequence instead of the sequence comparison, it is possible to predict protein functional class irrespective of sequence similarities. In order to extensively evaluate the potential and usefulness of the protein functional class prediction system 'SVMProt' developed in this study, novel proteins that are distinctly related to other known functional proteins, diversely covered novel enzymes, novel viral ORFs, novel bacterial ORFs, as well as novel protein in plants are selected to examine the usefulness of our prediction system.

## 5.1. Prediction of Functional Family of Novel Enzymes (Paper IV)

Enzymes are proteins that act as catalysts that could affect the rate of chemical reactions. As Enzymes play a central role in every aspect of life involves in chemical reactions, as well as they provided a means for regulating the reactions in the metabolic pathways of the body[242], the knowledge on enzymes is highly in demand for facilitating the understanding of biological processes.

Enzymes have been systematically classified by the International Commission on Enzymes [178] into six major groups: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. These six major groups are further subdivided according to the more specific type of reactions that these enzymes involve. By classifying a protein into a specific enzyme family, one can get the hints of the enzyme function as well as the type of reactions that the enzyme may catalyze.

Large amount of proteins that perform an enzymatic function have not been discovered. Thus, it is essential to interpret biological process at a deep level, especially when we consider discovering or even developing new therapeutic strategies. As mentioned previously, the function of an enzyme that has low sequence similarities of known function is difficult to assign based on their sequence similarity. The same problem may arise for homologous enzymes with different functions. In order to evaluate the capability of our developed prediction system for assignment of distantly related enzymes and homologous enzymes with different functions, two different groups of enzymes were tested for their functional class assignment.

### 5.1.1. Methods

In this work, two groups of enzymes, obtained from protein databases and literatures and subsequently verified by PSI-BLAST[38], are used to assess the capability of SVM for predicting the functional family of novel enzymes.

One group includes enzymes that are without a homolog in protein database based on PSI-BLAST search of these databases. A similarity threshold E-value of 0.05 is used for protein sequence similarity searching. Those novel enzymes are firstly searched from the Swiss-Prot database [180] by using the key word “novel”, “distinct”, or “unrelated” combined with “enzyme”. The next step is to eliminate those with at least one homolog of known function (except for hypothetical proteins) by conducting a PSI-BLAST[38] search against the NR databases that include all non-redundant GenBank, CDS translations, PDB, SwissProt, PIR, and PRF databases. This ensures that only those truly having no homolog in protein databases are selected. While the selected enzymes from this process are without a homolog, their function has been determined experimentally and these were reported in the literature and subsequently described in the Swiss-Prot database. The last step is to remove the redundancy.

The second group contains pairs of homologous enzymes of different families. A low similarity E-value threshold of $10^{-6}$ is used for selecting these enzyme pairs to ensure the high sequence similarities. In a hypothetical situation that one enzyme in a pair of homologous enzymes of different families is newly discovered and the other is the only known protein of similar sequence, the function of the first enzyme can be incorrectly assigned to that of the second enzyme by using sequence similarity methods. Thus, it is of interest to examine to what extent SVM can be used as an alternative approach for facilitating functional assignment for these enzymes. These enzymes are further checked to remove the redundancy.

### 5.1.2. Results and Discussion

As shown in Table 5-1, 12 enzymes without a homolog in the NR databases (group NR) and additional 38 enzymes without a homolog in the SwissProt database (group SP) are selected from the process introduced in 5.1.1 methods section. None of them is in the SVM training sets. SVMProt correctly assigns 8 out of 12 (67%) enzymes in the group "NR" and 28 out of 38 (73.7%) enzymes in the group "SP" to the respective family. The overall accuracy is 72%, which is comparable to the average sensitivity for the enzyme families, and it is consistent with the sequence-similarity-independent nature of SVM functional assignment.

These 8 pairs of homologous enzymes of different families from previous publications [45, 243] that satisfy the low E-value criterion, which together with SVMProt predicted top family for each enzyme are given in Table 5-2**.** It is found that 5 or 62% of these enzyme pairs are correctly assigned by SVMProt, such accuracy is comparable to the average sensitivity for the enzyme families and indicative of the sequence-similarity-independent nature of SVM functional assignment.

These results suggest that our prediction system has the capability for functional family assignment of novel enzymes without any sequence similarities in protein database, and for distinguishing homologous enzymes of different functions. The overall accuracy of SVM prediction system is not yet at the same level of that of sequence alignment for homologous proteins. One reason is the imbalance between the number of positive and negative samples. The total number of distinct enzymes in some families is less than 200, which is significantly smaller than that of a few thousand representative proteins used as the negative samples of the respective family. Such a large data imbalance is known to affect the accuracy of a SVM classification system and methods for solving these problems are being developed [233]. It is likely that not all possible types of proteins, particularly those of distantly related members, are adequately represented in some families. This can be improved along with the availability of more protein data. Not all distantly related proteins from one functional families have similar structural and chemical features due to the flexibility at the active site [26]. These improvements will enable the development of SVM into a useful tool for facilitating functional study of novel proteins.

Table 5-1 List of enzymes without a homolog in the NR and SwissProt databases and the results of SVM functional family assignment. The symbol +, *, and – represent the cases that the predicted family with highest ranking, one of the predicted families, and none of the predicted families matches the enzyme function respectively.

| Enzyme (EC number)[SwissProt Accession number] | Database Containing No Homolog | SVM assigned functional family(probability of correct prediction) | Assignment Status |
|---|---|---|---|
| Thiocyanate hydrolase beta subunit (EC 3.5.5.8) [O66186]. | NR | EC 3.5 Hydrolase of non-Peptide Carbon-Nitrogen Bonds (98.9%)<br>EC 2.6 Transferases of Nitrogenous Groups (62.2%) | + |
| Potential cysteine protease avirulence protein avrPpiC2 (EC 3.4.22.-) [Q9F3T4]. | NR | EC 4.2 Carbon-Oxygen Lyase (93.6%)<br>EC 2.3 Acyltransferase (83.9%)<br>EC 4.1 Carbon-Carbon Lyase (71.3%)<br>Outer membrane (58.6%) | - |
| Extracellular phospholipase (EC 3.1.1.5) [P82476] | NR | EC 3.1 Hydrolase of Ester Bonds (98.7%) | + |
| Cytochrome c oxidase polypeptide IV, mitochondrial precursor (EC 1.9.3.1) [P30815]. | NR | EC 1.9 Oxidoreductase of a heme group of donors (99.0%) | + |
| Cytochrome c oxidase polypeptide VI (EC 1.9.3.1) [P26310]. | NR | EC 1.9 Oxidoreductase of a heme group of donors (98.4%)<br>Transmembrane (98.3%)<br>EC 3.1 Hydrolase of Ester Bonds (62.2%) | + |
| Alginate lyase precursor (EC4.2.2.3) [P39049]. | NR | Transmembrane (65.4%)<br>Outer membrane (58.6%)<br>EC 2.1 Transferase of One-Carbon Groups (58.6%) | - |
| DNA alpha-glucosyltransferas e (EC 2.4.1.26) [P04519] | NR | EC 2.4 Glycosyltransferase (80.4%);<br>EC 2.7 Transferase of Phosphorus-Containing Groups (68.5%) | + |
| Endonuclease CviAII (EC 3.1.21.4 [P31117] | NR | EC 3.1 Hydrolase of Ester Bonds (99.0%) | + |
| Type II restriction enzyme CviJI (EC 3.1.21.4) [P52283] | NR | EC 3.1 Hydrolase of Ester Bonds (99.0%);<br>rRNA-binding Proteins (98.8%) ;<br>EC 3.4 Peptidase (68.5%) | + |
| DNA-directed RNA polymerase, subunit 10 homolog (EC 2.7.7.6) [P42488] | NR | EC 2.7 Transferase of Phosphorus-Containing Groups (99.0%)<br>7 transmembrane receptor metabotropic glutamate family (58.6%) | + |
| Endonuclease IV (EC 3.1.21.-) [P39250] | NR | No function predicted | - |
| Beta-agarase precursor (EC3.2.1.81) [P13734]. | NR | EC 4.1 Carbon-Carbon Lyase (96.7%)<br>EC 2.4 Glycosyltransferase (71.3%) | - |
| Phenylacetaldoxime dehydratase (EC 4.2.1.-) [P82604]. | SwissProt | Transmembrane (98.2%)<br>EC 3.4 Peptidase (96.4%)<br>EC 3.3 Hydrolase of Ether Bonds (80.4%)<br>EC 2.7 Transferase of Phosphorus-Containing Groups (73.8%) | - |
| ATP synthase H chain, mitochondrial precursor (EC3.6.3.14) [ Q12349]. | SwissProt | EC 3.6 Hydrolase of Acid Anhydrides (99.0%)<br>RNA-binding Protein (58.6%) | + |

| | | | |
|---|---|---|---|
| Peptide-N(4)-(N-acetyl-beta-D-glucosaminyl)asparagine amidase F precursor (EC 3.5.1.52) [P21163] | SwissProt | EC 3.5 Hydrolase of non-Peptide Carbon-Nitrogen Bonds (99.0%)<br>Beta-Barrel porin (58.6%) | + |
| S-adenosyl-L-methionine hydrolase (EC 3.3.1.2) [P07693] | SwissProt | EC 3.3 Hydrolase of Ether Bonds (99.0%)<br>EC 2.7 Transferase of Phosphorus-Containing Groups (71.3%)<br>DNA-binding Protein (65.4%) | + |
| Hypothetical 52.8 kDa protein in VPS15-YMC2 intergenic region .(EC 3.1.22.-) [P38257] | SwissProt | DNA-binding Protein (89.3%)<br>Outer membrane (58.6%) | - |
| Hypothetical protein BBB03 (EC3.1.22.-) [O50979]. | SwissProt | EC 2.7 Transferase of Phosphorus-Containing Groups (88.1%)<br>EC 3.4 Peptidase (86.8%)<br>EC 2.3 Acyltransferase (71.3%)<br>EC 4.1 Carbon-Carbon Lyase (65.4%) | - |
| Telomere elongation protein (EC2.7.7.-) [P17214]. | SwissProt | EC 2.7 Transferase of Phosphorus-Containing Groups (99.1%)<br>DNA-binding Protein (78.4%) | + |
| Fucose-1-phosphate guanylyltransferase (EC 2.7.7.30) [O14772] | SwissProt | EC 2.7 Transferase of Phosphorus-Containing Groups (99.1%)<br>7 transmembrane receptor metabotropic glutamate family (58.6%) | + |
| DNA-directed RNA polymerase I 14 kDa polypeptide (EC 2.7.7.6) [P50106]. | SwissProt | EC 2.7 Transferase of Phosphorus-Containing Groups (99%)<br>DNA-binding Protein (62.2%)<br>Beta-Barrel porin (58.6%)<br>EC 3.4 Peptidase (58.6%) | + |
| DNA polymerase III, theta subunit (EC 2.7.7.7) [P28689]. | SwissProt | EC 2.7 Transferase of Phosphorus-Containing Groups (99.0%)<br>EC 4.2 Carbon-Oxygen Lyase (58.6%) | + |
| Cytochrome c oxidase polypeptide IV (EC 1.9.3.1) [P77921] | SwissProt | EC 1.9 Oxidoreductase of a heme group of donors (97.0%)<br>Envelope protein (58.6%)<br>Transmembrane (58.6%) | + |
| Cytochrome c oxidase polypeptide VII (EC 1.9.3.1) [P10174]. | SwissProt | EC 1.9 Oxidoreductase of a heme group of donors (98.3%)<br>Transmembrane (58.6%) | + |
| Cytochrome c oxidase polypeptide VIII, mitochondrial precursor (EC 1.9.3.1) [P04039]. | SwissProt | EC 1.9 Oxidoreductase of a heme group of donors (99.0%)<br>Transmembrane (58.6%)<br>RNA-binding Protein (58.6%) | + |
| Cytochrome c oxidase polypeptide VIIA precursor (EC1.9.3.1) [P07255]. | SwissProt | EC 1.9 Oxidoreductase of a heme group of donors (97.8%)<br>Transmembrane (93.8%)<br>EC 1.10 Oxidoreductase of diphenols and related substances as donors (58.6%)<br>Alpha-Type channel (58.6%) | + |
| Heme-copper oxidase subunit IV (EC 1.9.3.-) [Q9YDX4]. | SwissProt | EC 1.9 Oxidoreductase of a heme group of donors (99.0%)<br>Transmembrane (99.0%) | + |
| Aminoglycoside 2'-N-acetyltransferase (EC 2.3.1.-) [P95219] | SwissProt | EC 2.7 Transferase of Phosphorus-Containing Groups (78.4%)<br>EC 4.2 Carbon-Oxygen Lyase (58.6%) | - |
| Glycosyl transferase alg8 (EC2.4.1.-) [Q887P9]. | SwissProt | Transmembrane (99.0%)<br>EC 2.4 Glycosyltransferase (98.6%) | * |

| | | | |
|---|---|---|---|
| Beta-agarase B (EC 3.2.1.81) [P48840]. | SwissProt | Outer membrane (58.6%)<br>Beta-Barrel porin (58.6%) | - |
| CM (EC 5.4.99.5) [P19080] | SwissProt | EC 5.4. Intramolecular Transferase (99.0%)<br>EC 4.2. Carbon-Oxygen Lyase (58.6%)<br>Outer membrane (58.6%) | + |
| DNA beta-glucosyltransferase (EC 2.4.1.27) [P04547] | SwissProt | EC 2.4 Glycosyltransferases (95.7%);<br>EC 2.5 Transferase of Alkyl or Aryl Groups, Other than Methyl Groups (80.4%) | + |
| dNMPkinase (EC 2.7.4.13) [P04531] | SwissProt | EC 2.7 Transferase of Phosphorus-Containing Groups (99.0%);<br>EC 2.4 Glycosyltransferase (96.4%);<br>EC 1.1 Oxidoreductase of the CH-OH group of donors (71.3%) | + |
| Endonuclease II (EC 3.1.21.1) [P07059] | SwissProt | EC 3.1 Hydrolase of Ester Bonds (99.0%) | + |
| Endonuclease V (EC 3.1.25.1) [P04418] | SwissProt | EC 3.1 Hydrolase of Ester Bonds (99.0%) | + |
| Exonuclease (EC 3.1.11.3) [P03697] | SwissProt | EC 3.1 Hydrolase of Ester Bonds (99.0%);<br>EC 4.1 Carbon-Carbon Lyases (88.1%);<br>EC 2.7 Transferase of Phosphorus-Containing Groups (68.5%);<br>EC 1.1 Oxidoreductase of the CH-OH group of donors (58.6%) | + |
| Ribonuclease (EC 3.1.-.-)[P13312] | SwissProt | EC 3.1 Hydrolase of Ester Bonds (99.0%) | + |
| Intron-associated endonuclease 1 (EC 3.1.-.-) [P13299] | SwissProt | EC 3.1 Hydrolase of Ester Bonds (99.0%);<br>DNA-binding Protein (83.9%) | + |
| Intron-associated endonuclease 2 (EC 3.1.-.-) [P07072] | SwissProt | EC 3.1 Hydrolase of Ester Bonds (99.0%) | + |
| Putative adenine-specific methylase (EC 2.1.1.72) [P51715] | SwissProt | EC 2.1 Transferase of One-Carbon Groups (99.0%)<br>Outer membrane (58.6%)<br>mRNA-binding Protein (58.6%) | + |
| Protein kinase (EC 2.7.1.37) [P00513] | SwissProt | EC 2.7 Transferase of Phosphorus-Containing Groups (99.0 %) | + |
| Slt35 (EC 3.2.1.-) [P41052] | SwissProt | Outer membrane (99.0%)<br>EC 1.1. Oxidoreductase acting on the CH-OH group of donors (89.3%)<br>EC 4.1. Carbon-Carbon Lyase (62.2%) | – |
| Ammonia monooxygenase (EC 1.13.12.- )[ Q04508] | SwissProt | EC 1.13. oxygenase (99.0%)<br>Transmembrane (99.0%)<br>EC 2.4. Glycosyltransferases (83.9%) | + |
| 2-aminomuconate deaminase (EC 3.5.99.5) [P81593] | SwissProt | EC 3.5. Hydrolase acting on Carbon-Nitrogen Bonds, other than Peptide bonds (99.0%)<br>EC 3.4. Peptidase (58.6%) | + |
| ADP-ribosyltransferase (EC2.4.2.37) [P14299] | SwissProt | Transmembrane (92.9%)<br>EC 2.4. Glycosyltransferase (90.3%)<br>Outer membrane (58.6%) | * |
| Alpha-N-AFase II (EC 3.2.1.55 ) [P82594] | SwissProt | EC 3.4. Peptidase (91.3%) | – |

| | | | |
|---|---|---|---|
| Aminopeptidase G (EC 3.4.11.-) [Q54340] | SwissProt | EC 3.4. Peptidase (99.0%)<br>TC 1.C. Pore-forming toxins (proteins and peptides) (58.6%) | + |
| Alginate lyase (EC 4.2.2.3 ) [Q59478] | SwissProt | Transmembrane (96.4%)<br>EC 3.1.-.-: Hydrolases - Acting on Ester Bonds (78.4%)<br>Outer membrane (58.6%) | – |
| ATPE_YEAST (EC 3.6.3.14) [P21306] | SwissProt | RNA-binding Proteins (58.6%) | – |
| AhdA2cA1c (EC1.14.-.- ) [BAC65427.1] | SwissProt | EC 3.1. Hydrolase acting on Ester Bonds (82.2%)<br>DNA-binding Protein (80.4%)<br>Transmembrane (58.6%) | – |

Table 5-2 List of pairs of homologous enzymes of different families and the results of SVM functional family assignment. E1→ F1 or E2 → F2 indicates that enzyme E1 or E2 is assigned into family F1 and F2 respectively. E1→ W or E2 → W indicates that enzyme E1 or E2 is assigned into a wrong family respectively. The symbol + or - represents the cases that SVM is able or unable to distinguish the two enzymes and exclusively assign them into the respective family.

| Enzyme E1 (SwissProt Accession number) | EC Class (F1) | Enzyme E2 (SwissProt Accession number) | EC Class (F2) | Sequence Similarity (BLAST E-Value) | SVM Functional Family Assignment | Assignment Status |
|---|---|---|---|---|---|---|
| Glycolateoxidase (P05414) | EC 1.1 | IPP isomerase (Q8PW37) | EC 5.3 | 3.00E-07 | E1→F1;E2→F2 | + |
| Creatine amidinohydrolase (P38488) | EC 3.5 | Prolinedipeptidase (O58885) | EC 3.4 | 3.00E-15 | E1→F1;E2→F2 | + |
| Cystathionine gamma-synthase (P38675) | EC 2.5 | Methionine gamma-lyase (P13254) | EC 4.4 | 2.00E-15 | E1→W;E2→F2 | - |
| Exocellobiohydrolase 1 (P38676) | EC 3.2 | Cystathionine gamma-lyase (Q8VCN5) | EC 4.4 | 1.00E-12 | E1→W;E2→F2 | - |
| Maleylacetoacetate isomerase (P57109) | EC 5.2 | Glutathione S-transferase zeta class (P57108) | EC 2.5 | 1.00E-51 | E1→F1;E2→F2 | + |
| Tyrosine-protein kinase FRK (P42685) | EC 2.7 | Intestinalguanylate cyclase (P70106) | EC 4.6 | 2.60E-12 | E1→F1;E2→F1 | - |
| Glutamate-1-semialdehyde aminotransferase (Q06774) | EC 5.4 | 4-aminobutyrate aminotransferase (P22256) | EC 2.6 | 5.70E-32 | E1→F1;E2→F2 | + |
| Exodeoxyribonuclease (P37454) | EC 3.1 | DNA- (apurinic or apyrimidinic site) lyase (P43138) | EC 4.2 | 1.60E-96 | E1→F1;E2→F2 | + |

## 5.2. Prediction of Functional Class of Novel Viral Proteins (Paper V)

### 5.2.1. Introduction of exploring knowledge of novel viral proteins

The need to explore functions of novel viral proteins is required for better understanding of how viruses interact with their host. For example, large DNA viruses such as poxviruses encode for a unique variety of proteins that can specifically manipulate the function of important host immune factors/messengers, e.g. interferon, interleukins and chemokines. The complete genomes of 1,536 viruses have been sequenced (Viral genomes at NCBI[244] as of September 2004). Knowledge of these genomes is very important for mechanistic study of viral infections and identification of molecular targets of antiviral therapeutics [245-247]. However, the function of over 15% of the putative protein-coding open reading frames (ORFs) in these viral genomes is unknown [245, 247, 248]. Determination of the function of these unknown ORFs is important for a more comprehensive understanding of the molecular mechanism of specific virus and for searching novel targets for antiviral drug development.

The sequence of many of these unknown ORFs has no significant similarity to proteins of known functions, and their functions are difficult to predict based on sequence similarity. For instance, 50%, 100%, 20% and 67% of the unknown ORFs in the recently determined genomes of Fer-de-lance virus[238], Grapevine fleck virus[239], Indian citrus ringspot virus[240], and SARS coronavirus[241] are without a homolog in Swissprot database[180] based on BLAST search against Swiss-Prot database[180] as of September 2004. This suggests that a large number of new viral proteins are likely to have no known sequence homolog.

In the absence of clear sequence or structural similarities, the criteria for comparison of distantly-related proteins become increasingly difficult to formulate [16]. Moreover, not all homologous proteins have analogous functions [8]. The presence of shared domain within a group of proteins does not necessarily imply that these proteins perform the same function[21]. Therefore, careful evaluation is required to determine whether method is useful for facilitating functional study of novel viral proteins with no homology to proteins of known function.

This work evaluates the usefulness of SVMProt for predicting the functional class of viral ORFs of unknown function. It is assessed by using novel viral proteins that (1) have no homology in the Swissprot database[180] based on sequence similarity search; (2) have a clear function indication described in the literature and (3) were not in the training set of SVMProt. These proteins are collected from an unbiased search of Medline [244] and Swiss-Prot database [180]. The SVMProt predicted functional classes of these proteins are compared with the function described in the literature and databases to evaluate to what extent SVMProt are useful for functional class assignment of novel viral proteins. The prediction accuracy for assignment of these novel proteins is compared with the overall accuracy of the SVMProt assignment of a large number of proteins to examine the level of sequence similarity independence of SVMProt classification.

### 5.2.2. Methods

The key words, “novel protein virus” or “novel viral protein”, are used to search the Medline [244] and the Swissprot database [180] for finding viral proteins that are both described as novel and with their precise function provided. As the search of the Medline is confined to the abstracts, those proteins whose function is not explicitly mentioned in an abstract are excluded. Thus, the selected proteins likely account for a portion of the known novel viral proteins with available functional information. PSI_BLAST[38] sequence analysis is subsequently conducted on each of these novel viral proteins against all SwissProt entries in the SwissProt protein database [180] so that those with at least one sequence homolog of known function (including that of the same protein in different species) are removed. The commonly-used criterion[38] for homologs, the similarity score e-value < the inclusion threshold value of 0.005, is used in this work. Finally, those proteins that are in the training sets of SVMProt are removed. 25 novel viral proteins are identified in this process. These protein and their protein accession number, literature-described functional indications and related references are given in Table 5-3.

Table 5-3 Novel viral proteins, literature-described functional indications as suggested from experiment and/or sequence analysis, and SVMProt predicted functions. The SVMProt predicted functions are categorized in one of the four classes: The first class is M (matched), in which all of the literature-described functional indications are predicted. The second is PM (partially matched), in which some of the literature-described functional indications are predicted. The third is WC (weakly consistent), in which some of the predicted functions can be considered to be consistent with literature-described functional indications on an inconclusive basis. The fourth is NM (not matched), in which No function predicted of the literature-described functions matched or consistent with a predicted function.

| **Protein (SwissProt or NCBI accession number)** | **Virus** | **Literature Described Function (reference)** | **Function characterized by SVMProt (probability of correct characterization P-value)** | **Prediction status** |
|---|---|---|---|---|
| ADOMetase (P07693) | Bacteriophage T3 | Adenosylmethionine hydrolase (EC 3.3.1.2) [249] | EC 3.3: Hydrolase of Ether Bonds (99.0%);<br>EC 2.7: Transferase of Phosphorus-Containing Groups (71.3%);<br>DNA-binding Proteins ( 65.4%); | M |
| AGT (P04519) | Enterobacteria phage T4 | DNA alpha-glucosyltransferase (EC 2.4.1.26) [246] | EC 2.4 Glycosyltransferase (80.4%);<br>EC 2.7 Transferase of Phosphorus-Containing Groups (68.5%) | M |
| BGT (P04547) | Enterobacteria phage T4 | DNA beta-glucosyltransferase (EC 2.4.1.27) [246, 250] | EC 2.4 Glycosyltransferases (95.7%);<br>EC 2.5 Transferase of Alkyl or Aryl Groups, Other than Methyl Groups (80.4 %) | M |
| DNA-directed RNA polymerase (P42488) | African swine fever virus (strain BA71V) | DNA-directed RNA polymerase, subunit 10 homolog (EC 2.7.7.6) [251]. | EC 2.7 Transferase of Phosphorus-Containing Groups (99.0%); | M |
| DNK (P04531) | Enterobacteria phage T4 | dNMPkinase (EC 2.7.4.13) [252] | EC 2.7 Transferase of Phosphorus-Containing Groups (99.0%);<br>EC 2.4 Glycosyltransferase (96.4%);<br>EC 1.1 Oxidoreductase of the CH-OH group of donors (71.3%) | M |
| Endonuclease II (P07059) | Enterobacteria phage T4 | Endonuclease II (EC 3.1.21.1) [253]. | EC 3.1 Hydrolase of Ester Bonds (99.0%) | M |
| Endonuclease IV (P39250) | Enterobacteria phage T4 | Endonuclease IV (EC 3.1.21.-) [246] | No function predicted | NM |
| Endonuclease V (P04418) | Enterobacteria phage T4 | Endonuclease V (EC 3.1.25.1) [254] | EC 3.1 Hydrolase of Ester Bonds(99.0%) | M |

| | | | | |
|---|---|---|---|---|
| Exonuclease (P03697) | Bacteriophage lambda | Exonuclease (EC 3.1.11.3) [255]. | EC 3.1 Hydrolase of Ester Bonds(99.0%);<br>EC 4.1 Carbon-Carbon Lyases (88.1%);<br>EC 2.7 Transferase of Phosphorus-Containing Groups(68.5%);<br>EC 1.1 Oxidoreductase of the CH-OH group of donors (58.6%) | M |
| FALPE (Q65010) | Amsacta moorei Entomopoxvirus | Associated with unique cytoplasmic structures, filament-associated protein [256] | No function predicted | NM |
| Gp61.9 (P13312) | Enterobacteria phage T4 | Ribonuclease (EC 3.1.-.-) [257] | EC 3.1 Hydrolase of Ester Bonds(99.0%) | M |
| IRF protein (P13299) | Enterobacteria phage T4 | Intron-associated endonuclease 1 (EC 3.1.-.-) [258] | EC 3.1 Hydrolase of Ester Bonds(99.0 %);<br>DNA-binding Protein (83.9%) | M |
| I-TevII (P07072) | Enterobacteria phage T4 | Intron-associated endonuclease 2 (EC 3.1.-.-) [259] | EC 3.1 Hydrolase of Ester Bonds(99.0%) | M |
| MotA protein (P22915) | bacteriophage T4 | DNA-binding, transcription regulation [260] | DNA-binding Proteins (99.0 %);<br>EC 3.1: Hydrolase - Acting on Ester Bonds (68.5%) | M |
| ORF13 (P51715) | Haemophilus phage HP1 | Putative adenine-specific methylase (EC 2.1.1.72) [261] | EC 2.1 Transferase of One-Carbon Groups (99.0%);<br>Outer membrane (58.6%);<br>mRNA-binding Protein (58.6%) | M |
| Outer capsid protein VP4 (P35746) | Bovine rotavirus (serotype 10 / strain B223) | surface outer capsid protein [262] | Coat protein (99.0%) | M |
| possible CC chemokine (NP_042976) | Human herpesvirus 6 | chemokine like [263] | No function predicted | NM |
| Protein kinase (P00513) | Enterobacteria phage T7 | Protein kinase (EC 2.7.1.37) [264] | EC 2.7 Transferase of Phosphorus-Containing Groups ( 99.0 %) | M |
| Putative BARF0 protein (Q8AZJ4) | Epstein-Barr virus | Membrane associated and encodes three arginin-rich motifs of RNA-binding properties [265] | EC 4.1.-.-: Carbon-Carbon Lyase (58.6%) | NM |
| R.CviAII (P31117) | Paramecium bursaria Chlorella virus 1 | Endonuclease CviAII (EC 3.1.21.4) [266] | EC 3.1 Hydrolase of Ester Bonds (99.0%) | M |

| | | | | |
|---|---|---|---|---|
| R.CviJI (P52283) | Chlorella virus IL3A | Type II restriction enzyme CviJI (EC 3.1.21.4)[267] | EC 3.1 Hydrolase of Ester Bonds (99.0%);<br>rRNA-binding Proteins(98.8%) ;<br>EC 3.4 Peptidase (68.5%) | M |
| SeMNPV ORF18 (AAF33548) | Spodoptera exigua nucleopolyhedrovirus | Transferase[268] | No function predicted | NM |
| SPLT137 (NP_258405) | SpLtMNPV virus | A noval envelope protein[269] | No function predicted | NM |
| TRL10 (AAL27474) | Human cytomegalovirus (HCMV) | Structural envelop glycoprotein[270] | Transmembrane (98.2%) | NM |

### 5.2.3. Results and Discussion

SVMProt predicted functional classes for each of the 25 novel viral proteins together with their literature-described function are given in Table 5-3. SVMProt may characterize more than one class for each protein and the probability of correct prediction for each class is given in the table. There are 18 proteins with the top hit of the SVMProt assigned functional class matching their functions described in the literature, representing 72% of the novel viral proteins studied in this work. These proteins are MotA protein of bacteriophage T4 [260], outer capsid protein VP4 of bovine rotavirus (serotype 10 / strain B223) [262], ADOMetase of bacteriophage T3 [249], R.CviJI of chlorella virus IL3A [267], exonuclease of bacteriophage lambda [255], R.CviAII of paramecium bursaria chlorella virus 1 [266], ORF13 of haemophilus phage HP1 [261], Protein kinase of enterobacteria phage T7 [264], DNA-directed RNA polymerase of African swine fever virus (strain BA71V)[251], AGT [246], BGT [248, 250], DNK [252], Endonuclease II [253], Endonuclease V [254], Gp61.9 [257], IRF protein [258], and I-TevII [259] of enterobacteria phage T4.

MotA protein of bacteriophage T4 has been found to be a transcription activator that binds to DNA [260] and the far-C-terminal region of the sigma70 subunit of Escherichia coli RNA polymerase [271]. The top hit of SVMProt predicted functional class for this protein is the DNA-binding, which matches with literature-described functions. Bovine rotavirus is a double-stranded RNA virus which is naked. Thus the outer capsid protein VP4 of bovine rotavirus (serotype 10 / strain B223) is located at the viral surface acting as part of the viral coat [262]. This protein is predicted by SVMProt as a coat protein, which is consistent with literature-described function. The other 14 proteins are enzymes and SVMProt correctly assigns all these to the respective enzyme EC class.

Because these proteins have no homolog of known function in the SwissProt entries of Swissprot database based on PSI-BLAST search, our study suggests that SVMProt has certain level of capability for providing useful hint about the functional class of novel proteins with no or low homology to known proteins, and this capability is not based on sequence similarity or clustering. The overall accuracy of 72% for the assignment of the novel viral proteins is smaller than that of 87% for SVMProt functional class assignment of 34,582 proteins. This indicates certain level of the sequence-similarity-independent nature of SVM protein classification.

Several factors may affect the accuracy of SVMProt for functional characterization of novel plant proteins. One is the diversity of protein samples used for training SVMProt. It is likely that not all possible types of proteins, particularly those of distantly related members, are adequately represented in some protein classes. This can be improved along with the availability of more protein data. Not all distantly related proteins of the same function have similar structural and chemical features. There are cases in which different functional groups, un-conserved with respect to position in the primary sequence, mediate the same mechanistic role, due to the flexibility at the active site [272]. This plasticity is unlikely to be sufficiently described by the physicochemical descriptors currently used in SVMProt. Therefore, SVMProt in the present form is not expected to be capable of classification of these types of distantly related enzymes.

Some of the SVMProt functional classes are at the level of families and superfamilies that may include a broad spectrum of proteins. It has been shown that, SVM works not as well as HMM for distinguishing proteins in a superfamily, but may be more accurate with subfamily discrimination [31]. Thus, the use of some large families and superfamilies as the basis for classification may affect the prediction accuracy of SVMProt to some extent.

# 5.3. Prediction of functional class of novel plant proteins (Paper VI)

## 5.3.1. Introduction of probing function of unknown ORFs in plant

Plants have the well known advantages for the production of clinically-useful, therapeutic proteins, such as low-cost, large-scale production of safe and biologically active mammalian proteins[273]. In the completely sequenced genome of *Arabidopsis*, the function of 30% of the putative protein-coding open reading frames (ORFs) remain uncovered [234, 235]. Similar percentage of unknown ORFs is expected in other plant genomes. The sequence of these ORFs has no significant similarity to those of known proteins, and their functions are difficult to probe by using sequence alignment and clustering methods. It is thus desirable to explore complementary methods or combination of methods for providing useful hint about the function of unknown ORFs.

Various methods for probing protein function have been developed. These include evolutionary analysis [8, 9], hidden Markov models [274], structural consideration [10, 27], protein/gene fusion [11, 12], protein-protein interactions [14], motifs [19], family classification by sequence clustering [12], and functional family prediction by statistical learning methods [31, 34, 43, 45, 46]. In the absence of clear sequence or structural similarities, the criteria for comparison of distantly-related proteins become increasingly difficult to formulate [16]. Moreover, not all homologous proteins have analogous functions [8]. The presence of shared domain within a group of proteins does not necessarily imply that these proteins perform the same function [21]. Therefore careful evaluation is needed to determine which method or combination of methods is useful for facilitating functional study of novel proteins with no homology to proteins of known function.

In this work, SVMProt is assessed for its capability in prediction of the functional class of a number of literature-described novel plant proteins that have no homolog in the SwissProt entries of the SwissProt database based on PSI-BLAST search and with their functional indications provided in the literature. There are 49 plant proteins selected from a comprehensive search of Medline abstracts and SwissProt databases in 1999-2004 to test SVMProt. These proteins are selected based on 1) no sequence similar proteins in Swissprot protein database, 3) not in our dataset for training SVMProt and 3) with precise functional indications provided by the literature. These proteins represent unique proteins whose functions cannot be confidently predicted by sequence alignment and clustering methods at present. The predicted functional class of 31 proteins is consistent, and that of 4 other proteins is weakly consistent with literature-described functions. Overall, the functional class of 71.4% of these proteins is consistent or weakly consistent with literature described functional indications. SVMProt shows certain level of capability for providing useful hint about the function of novel plant proteins un-similar to known proteins.

### 5.3.2. Methods of novel plant proteins selection

The key words "novel plant protein" is used to search two sources for finding plant proteins that were both described as novel and with their precise functional indications provided. One is the abstracts of Medline [244] published during 1999-2004. The sequences of these proteins are obtained by querying the protein database. As the search is confined to the abstracts, those proteins whose functional indication is not apparently hinted in an abstract are excluded. Thus, the selected proteins likely account for a portion of the known novel plant proteins with available functional indications. The second source is the SwissProt database[180]. The key words "novel plant" is used to search the description field of the plant protein entries to find those with precise functional indications provided. There are 413 proteins selected from these two search procedures.

Some of these selected proteins may become less novel than originally described because of the subsequent findings of additional proteins. Thus PSI_BLAST [38] search is conducted for each of these proteins against all SwissProt entries in the SwissProt protein database [180] to determine whether it has a sequence homolog (including that of the same protein of different species). The commonly-used criterion for homologs, the similarity score e-value < the inclusion threshold value of 0.005 [38], is used in this work. Based on PSI-BLAST analysis, 49 of these proteins have no sequence homolog in the SwissProt entries of SwissProt database and they are not in the training sets of SVMProt.

These 49 proteins, along with their NCBI protein accession number, or Swiss-Prot accession number, literature-described functional indications and related references, are given in Table 5-4. Only a few proteins published before 2001 are selected primarily because more proteins published in earlier years tend to have their homologs available than those published more recently. Because of the lack of a sequence homolog, sequence alignment and clustering tools would not confidently predict the function of these proteins. They are thus ideal for testing the feasibility of using SVMProt for facilitating functional characterization of novel plant proteins.

### 5.3.3. Prediction results and discussions

Table 5-4 gives SVMProt ascribed functional classes for each of the 49 novel plant proteins together with their literature described functional indications. SVMProt may characterize more than one functional class and the probabilities of correct prediction for each class are given in Table 5-4. There are 31 proteins with SVMProt predicted class to be consistent with literature-described functional indications, 20 of which are enzymes with their enzyme classification (EC) number assigned in the literature. The predicted functional class of these enzymes can thus be confirmed based on the comparison with their respective EC number. These enzymes are SPP of *Aegilops speltoides* [275], CPDase[276] and GddR of *Arabidopsis thaliana*, Cucumisin of *Cucumis melo var reticulates*[277], AOC of *Hordeum vulgare*, Spp of *Hordeum vulgare var distichum* [275], AOC[278] and RdRP[279] of *Lycopersicon esculentum*, Beta-1,2-xylosyltransferase and AOC of *Oryza sativa*, GrG of Phaseolus angularis [280], PAT1 and rfs of *Pisum sativum*, Sucrose-phosphatase of *Secale cereale* [275], CR6 of *Solanum tuberosum* [281], CPDase, SPP1,SPP2,SPP3 and fut12 of *Triticum aestivum.* Some of these enzymes do not yet have a reference because they have been submitted to Swissprot database prior to their publications [180].

Four proteins are predicted as transmembrane and another one as a DNA-binding protein by SVMProt, which can be directly compared with their respective literature described functional indications. PSI-O of *Arabidopsis thaliana* is known to have two transmembrane helices [282]. PM19 of *Hordeum vulgare* has been described as a putative plasma membrane protein [283]. OsBLE2 of Oryza sativa has been suggested to contain nine possible transmembrane regions [284]. NEC1 of *Petunia x hybrida* has been found to be reminiscent of a transmembrane protein with possible role in sugar metabolism and nectar secretion [285]. MYB-related transcription factor EPR1 of *Arabidopsis thaliana* is part of a regulatory feedback loop that suppresses its own expression, and it is known to specifically recognizes the DNA sequence 5'-YAAC[GT]G-3' [286]. The SVMProt predicted transmembrane or DNA-binding property for each of these proteins appears to be consistent with literature descriptions.

The predicted functional class of the other four proteins also appears to be consistent with literature described functional indications based on our analysis. NCP1 of *Lycopersicon esculentum* has been described as a nuclear matrix protein and a candidate for a plant-specific structural protein with a function both in the nucleus and cytoplasm [287]. The top hit of SVMProt predicted functional classes for this protein is the structural protein class that includes matrix proteins, core proteins, viral occlusion body, and keratins. This prediction is consistent with literature-described function. Antimicrobial peptide 2, 3 and 4 of *Pinus sylvestris* are known to interfere with cell wall synthesis [288]. The top hit of SVMProt predicted class for each of these proteins is EC3.4 peptidase enzyme family. It is known that members of peptidase family such as penicillin-binding protein 5 (EC 3.4.16.4) polymerize and modify peptidoglycan, the stress-bearing component of the bacterial cell wall, thereby helping to create the morphology of the peptidoglycan exoskeleton together with cytoskeleton proteins that regulate septum formation and cell shape [289]. While other mechanisms cannot be ruled out yet, EC3.4 peptidase enzymatic activity is certainly an interesting possibility for the observed interference of each of these proteins with cell wall synthesis.

There are 4 proteins whose SVMProt predicted function may possibly explain literature described functional indications non-conclusively. The predicted functional class of each of these proteins is thus considered to be weakly consistent with literature descriptions pending further studies. PLATZ1 of *Pisum sativum* has been found to be responsible for A/T-rich sequence-mediated transcriptional repression [290]. The top ranked SVMProt predicted class for this protein is the nuclear receptor class. Nuclear receptors such as thyroid hormone T3 receptor have been known to be involved in transcriptional repression [291]. Thus, there is a possibility that PLATZ1 is a nuclear receptor. SPA15 of *Ipomoea batatas*, has been found to be specifically associated with the cell wall and involved in oligogalacturonides signaling during leaf senescence [292]. SVMProt predicts this protein as an outer membrane protein, which is possible to possess both properties.

SVMProt predicts three of these proteins as DNA-binding protein. OsGRF1 of *Arabidopsis thaliana* has been described as a putative transcription factor possibly playing a regulatory role in stem elongation [293]. bnKCP1 of *Brassica napus* contains a putative kinase-inducible domain and it may function as a transcription factor [294]. Transcription factors primarily exert their function through DNA-binding[295], thus these two proteins are likely DNA-binding proteins. HvS40 of *Hordeum vulgare subsp. Vulgare* has been described as a novel nucleus-targeted protein [296]. The nuclear HvS40 protein belongs to the group of nuclear proteins that possess two putative NLSs, one belonging to the SV40 class, the other to the class of bipartite NLSs. In the case of the maize transcription factor opaque 2, the bipartite NLS has an additional function in DNA binding [296]. Although there is no other evidence, it is possible that HvS40 of *Hordeum vulgare subsp. Vulgare* is a DNA-binding protein like the other of bipartite

NLS containing proteins such as the maize transcription factor opaque 2.

Another protein, HvCaBP1 of *Hordeum vulgare*, has been described as a putative calcium binding protein [297]. One of the SVMProt predicted classes for this protein is outer membrane class. It is known that some outer membrane proteins, such as the 40 kDa outer membrane protein, form spheroplast at a high rate in an isotonic medium in the presence of calcium and the calcium-protein complex helps maintaining the structural integrity of the cell wall [298]. Thus, there is some possibility that HvCaBP1 is a calcium-binding outer membrane protein.

Overall, SVMProt characterized functions of 71.4% of the 49 novel plant proteins studied in this work are found to be consistent or weakly consistent with the functional indications described in the literature. Because all of these proteins have no homolog in the SwissProt entries of Swissprot database based on PSI-BLAST search, our study suggests that SVMProt has certain level of capability for probing the functional class of novel plant proteins with no or low homology to known proteins, and this capability is not based on sequence similarity or clustering.

Table 5-4 Novel plant proteins, literature-described functional indications as suggested by the literature and SVMProt predicted functional classes. The SVMProt predicted functional classes are categorized in one of the four classes: The first class is C (consistent with literature-described functional indications), the second is WC (weakly consistent with literature-described functional indications, i.e., the predicted functional class can be considered to be consistent to the literature-described functions on an inconclusive basis.), the third is NC (not consistent with literature-described functional indications), and the fourth is represented by a question mark "?" (Currently available information is insufficient to determine prediction status).

| *Host Plant* | Protein (NCBI or SwissProt Accession number) | Literature-described function (Reference) | SVMProt predicted functional class (probability of correct prediction) | Prediction Status |
|---|---|---|---|---|
| ***Aegilops speltoides*** | SPP(AAO33156) | Sucrose-phosphatase (EC 3.1.3.24) [275] | EC 3.1 Hydrolases - Acting on Ester Bonds(94.7%);<br>EC 2.7 Transferases - Transferring Phosphorus-Containing Groups (76.2%);<br>TC 1.C Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%) | C |
| ***Arabidopsis thaliana*** | MYB-related transcription factor EPR1 (BAC98462) | DNA-binding protein, specifically recognizes the sequence 5'-YAAC[GT]G-3' [286] | DNA-binding Protein (98.8%) | C |
| | OsGRF1 (AAM52876) | putative transcription factor playing a regulatory role in stem elongation[293] | DNA-binding Protein (97.0%) | WC |
| | PSI-O (CAD37939) | contains two transmembrane helices [282] | Transmembrane (68.5%);<br>EC 5.3 Intramolecular Oxidoreductase (58.6%) | C |
| | ERN1 (CAA75349) | a novel ethylene-regulated nuclear protein, putative transcription factor [299] | EC 4.2 Carbon-Oxygen Lyase (58.6%);<br>7 transmembrane receptor metabotropic glutamate family (58.6%) | NC |
| | CPDase (O04147) | Cyclic phosphodiesterase (EC 3.1.4.-) [276] | DNA-binding Proteins(71.3%);<br>EC 3.1 Hydrolases - Acting on Ester Bonds(58.6%) | C |
| | GddR precursor (Q9FPU3) | Glutathione dependent dehydroascorbate reductase (EC 1.8.5.1) * | EC 1.8 Oxidoreductases - Acting on a sulfur group of donors(99.0%);<br>Transmembrane(58.6%);<br>TC 1.C Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%) | C |
| ***Brassica napus*** | bnKCP1 (AAO53442) | contains a putative kinase-inducible domain, may function as a transcription factor [294] | DNA-binding Protein (68.5%) | WC |

| | | | | |
|---|---|---|---|---|
| ***Cucumis melo var reticulatus*** | Cucumisin (Q940D5) | serine protease(EC 3.4.21.25)[277] | EC 3.4 Hydrolases - Acting on peptide bonds (Peptidases) (99.0%);<br>EC 3.1 Hydrolases - Acting on Ester Bonds(78.4%) | C |
| ***Glycine max*** | CPP1 (CAA09028) | DNA-binding protein interacting with the promoter of the soybean leghemoglobin gene Gmlbc3 [300] | No function predicted | NC |
| | GmN6L (AAL86737) | both as a soluble protein and as a peripheral membrane protein bound to the peribacteroid membrane, a late nodulin [301] | EC 1.1 Oxidoreductase acting on CH-OH group of donors (73.8%);<br>EC 3.6 Hydrolase Acting on Acid Anhydrides (71.3%); | ? |
| ***Hordeum vulgare*** | Lem1 (AAK58425) | possibly associated with membranes, may play a role in organ development [302] | EC 3.4 Peptidase (58.6%);<br>Lectin (58.6%) | NC |
| | HvCaBP1 (AAK92225) | putative calcium binding protein [297] | EC 1.3 Oxidoreductase acting on CH-CH group of donors (85.4%);<br>EC 4.1 Carbon-Carbon Lyase (62.2%);<br>Outer membrane (58.6%) | WC |
| | PM19 (AAF29532) | putative plasma membrane protein [283] | Transmembrane (68.5%) | C |
| | AOC (Q711R0) | Allene oxide cyclase precursor (EC 5.3.99.6)* | EC 5.3 Isomerases - Intramolecular Oxidoreductases (95.7%);<br>EC 1.10 Oxidoreductases - Acting on diphenols and related substances as donors (65.4%) | C |
| ***Hordeum vulgare subsp. vulgare*** | HvS40 (CAC36956) | a novel nucleus-targeted protein with connection to the degeneration of chloroplasts [296] | DNA-binding Protein (78.4%);<br>Nuclear Receptor (65.4%);<br>EC 2.1 Transferase of One-Carbon Groups (58.6%);<br>RNA-binding Protein (58.6%) | WC |
| | SnIP1 (CAB97356) | interacts with SNF1-related protein kinase [303] | EC 3.4 Peptidase (71.3%);<br>EC 5.3 Intramolecular Oxidoreductase (68.5%);<br>EC 1.3 Oxidoreductase acting on CH-CH group of donors (65.4%);<br>EC 3.5 Hydrolase acting on Carbon-Nitrogen Bonds other than Peptide Bonds (62.2%);<br>7 transmembrane receptor secretin family (58.6%) | NC |
| ***Hordeum vulgare var distichum*** | Spp (Q84ZX7) | Sucrose-phosphatase (EC 3.1.3.24) [275] | EC 3.1 Hydrolases - Acting on Ester Bonds (97.7%);<br>EC 2.4 Transferases - Glycosyltransferases(91.3%) | C |

| | | | | |
|---|---|---|---|---|
| ***Ipomoea batatas*** | SPA15 (AAK08655) | specifically associated with the cell wall [292] | Outer membrane (58.6%) | C |
| ***Lilium longiflorum*** | LlSCL (BAC77269) | strong activity of transcriptional activation [304] | No function predicted | NC |
| ***Lycopersicon esculentum*** | NCP1 (AAK83083) | Nuclear Matrix Protein, structural protein with a function both in the nucleus and cytoplasm [287] | Structural protein (99.0%)<br>EC 5.4 Intramolecular Transferase (85.4%)<br>DNA-binding Proteins (65.4%) | C |
| | AOC (Q9LEG5) | Allene oxide cyclase precursor (EC 5.3.99.6) [278] | EC 5.3 Isomerases - Intramolecular Oxidoreductases (99.0%);<br>EC 4.1 Lyases - Carbon-Carbon Lyases(58.6%) | C |
| | RdRP (Q9ZR58) | RNA-directed RNA polymerase(EC 2.7.7.48) [279] | EC 2.7 Transferases - Transferring Phosphorus-Containing Groups (99.1%) | C |
| | LeMan3 (Q9FUQ6) | Endo-beta-mannanase precursor(EC 3.2.1.78) * | EC 2.4 Transferases - Glycosyltransferases(95.2%) ;<br>EC 2.3 Transferases - Acyltransferases (58.6%) | NC |
| | MAN5 (Q6YM50) | Mannan endo-1,4-beta-mannanase precursor (EC 3.2.1.78)[305] | EC 2.3 Transferases - Acyltransferases (68.5%) | NC |
| ***Oenothera bertiana*** | A6L (P07513) | ATP synthase protein 8(EC 3.6.3.14) [180] | EC 3.1 Hydrolases - Acting on Ester Bonds(58.6%);<br>Transmembrane(58.6%);<br>mRNA-binding Proteins(58.6%) | NC |
| ***Oryza sativa*** | OsBLE2 (BAB88327) | contains nine possible transmembrane regions, involved in BL-regulated growth and development processes [306] | Transmembrane (99.1%) Alpha-Type channel (58.6%) | C |
| | OsMYBS2 (AAN63153) | trans-activates a promoter containing the TATCCA element, interacts with other protein factors [307] | Transmembrane (71.3%);<br>7 transmembrane receptor secretin family (58.6%) | NC |
| | Beta-1,2-xylosyltransferase (Q703H1) | Beta-1,2-xylosyltransferase (EC 2.4.2.38)* | EC 2.4 Transferases - Glycosyltransferases (98.8%) ;<br>EC 4.2 Lyases - Carbon-Oxygen Lyases (58.6%);<br>Outer membrane (58.6%) | C |
| | AOC(Q8L6H4) | Allene oxide cyclase (EC 5.3.99.6) * | EC 5.3 Isomerases - Intramolecular Oxidoreductases(99.0%);<br>EC 1.10 Oxidoreductases - Acting on diphenols and related substances as donors (58.6%) | C |

| | | | | |
|---|---|---|---|---|
| | Aspartate aminotransferase (Q42991) | Aspartate aminotransferase (EC 2.6.1.1) * | TC 1.C Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%);<br>RNA-binding Proteins (58.6%) | NC |
| ***Petunia x hybrida*** | NEC1 (AAG34696) | reminiscent of a transmembrane protein, possible role in sugar metabolism and nectar secretion [285] | Transmembrane (97.3%) | C |
| ***Phaseolus angularis*** | GrG (Q9SBZ0) | Galactinol-raffinose galactosyltransferase (EC 2.4.1.67) [280] | EC 2.4 Transferases - Glycosyltransferases (96.4%); EC 4.2 Lyases - Carbon-Oxygen Lyases (78.4%) | C |
| ***Pinus sylvestris*** | antimicrobial peptide 1 (AAL05052) | Interferes with cell wall synthesis [288] | Transmembrane (58.6%) | NC |
| | antimicrobial peptide 2 (AAL05053) | interferes with cell wall synthesis [288] | EC 3.4 Peptidase (58.6%);<br>EC 4.1 Carbon-Carbon Lyase (58.6%) | C |
| | antimicrobial peptide 3 (AAL05054) | interferes with cell wall synthesis [288] | EC 3.4 Peptidase (58.6%) | C |
| | antimicrobial peptide 4 (AAL05055) | interferes with cell wall synthesis [288] | EC 3.4 Peptidase (58.6%);<br>EC 4.1 Carbon-Carbon Lyase (58.6%) | C |
| ***Pisum sativum*** | PLATZ1 (BAB69816) | zinc-dependent DNA-binding protein responsible for A/T-rich sequence-mediated transcriptional repression [290] | Nuclear Receptor (68.5%);<br>EC 3.1 Hydrolase Acting on Ester Bonds (62.2%);<br>EC 4.1 Carbon-Carbon Lyase (58.6%) | C |
| | PAT1 (Q43085) | Phosphoribosylanthranilate transferase (EC 2.4.2.18) * | EC 2.4 Transferases - Glycosyltransferases (99.1%); Transmembrane (96.1%);<br>EC 2.7 Transferases - Transferring Phosphorus-Containing Groups (76.2%) ;<br>7 transmembrane receptor (Secretin family) (58.6%) ;<br>7 transmembrane receptor (metabotropic glutamate family) (58.6%) | C |
| | rfs (Q8VWN6) | Raffinose synthase (EC 2.4.1.82) * | EC 2.4 Transferases - Glycosyltransferases (98.6%); Aptamer-binding protein (98.0%);<br>EC 4.2 Lyases - Carbon-Oxygen Lyases (78.4%) | C |
| ***Secale cereale*** | Sucrose-phosphatase (Q84ZX9) | Sucrose-phosphatase (EC 3.1.3.24) [275] | EC 3.1 Hydrolases - Acting on Ester Bonds (86.8%);<br>EC 2.7 Transferases - Transferring Phosphorus-Containing Groups (62.2%) | C |

| | | | | |
|---|---|---|---|---|
| ***Solanum tuberosum*** | CR6(P48505) | Ubiquinol-cytochrome C reductase complex 6.7 kDa protein (EC 1.10.2.2) [281] | EC 1.10 Oxidoreductases - Acting on diphenols and related substances as donors (99.0%);<br>EC 3.4 Hydrolases - Acting on peptide bonds (Peptidases) (58.6%) | C |
| ***Triticum aestivum*** | CPDase (P62809) | Cyclic phosphodiesterase (EC 3.1.4.-) [308] | EC 1.9 Oxidoreductases - Acting on a heme group of donors (58.6%);<br>EC 3.1 Hydrolases - Acting on Ester Bonds (58.6%);<br>EC 3.4 Hydrolases - Acting on peptide bonds (Peptidases) (58.6%);<br>Transmembrane (58.6%) ;<br>Aptamer-binding protein (58.6%) | C |
| | SPP3 (Q9ARG8) | Sucrose-6F-phosphate phosphohydrolase SPP3 (EC 3.1.3.24) [*] | EC 3.1 Hydrolases - Acting on Ester Bonds (96.4%);<br>EC 2.7 Transferases - Transferring Phosphorus-Containing Groups (92.1%);<br>TC 1.C. Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%) | C |
| | SPP2 (Q9AXK5) | Sucrose-6F-phosphate phosphohydrolase SPP2 (EC 3.1.3.24) * | EC 3.1 Hydrolases - Acting on Ester Bonds (93.6%);<br>EC 2.7 Transferases - Transferring Phosphorus-Containing Groups (68.5%);<br>TC 1.C. Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%) | C |
| | SPP1 (Q9AXK6) | Sucrose-6F-phosphate phosphohydrolase SPP1 (EC 3.1.3.24) * | EC 3.1 Hydrolases - Acting on Ester Bonds (96.4%);<br>EC 2.1 Transferases - Transferring One-Carbon Groups (58.6%);<br>TC 1.C.Channels/Pores - Pore-forming toxins (proteins and peptides) (58.6%) | C |

[*] NOTE: Some of these enzymes do not yet have a reference because they have been submitted to Swissprot database prior to their publications

| | | | | |
|---|---|---|---|---|
| | fut12 (Q7XAG0) | GDP-fucose protein-O-fucosyltransferase 1 (EC 2.4.1.221) * | EC 2.4 Transferases - Glycosyltransferases (98.4%);<br>EC 2.7 Transferases - Transferring Phosphorus-Containing Groups (96.7%) | C |
| ***Vigna unguiculata*** | FGARAT (Q8W160) | Formylglycinamide ribonucleotide amidotransferase (EC 6.3.5.3) * | EC 2.4 Transferases - Glycosyltransferases (95.2%);<br>DNA-binding Proteins (73.8%);<br>EC 2.7 Transferases - Transferring Phosphorus-Containing Groups (62.2%) | NC |
| ***Zea mays*** | ATPase (Q6V916) | Putative AAA-type ATPase (EC 3.6.4.8) * | EC 2.4 Transferases - Glycosyltransferases (71.3%) | NC |

## 5.4. Prediction of the functional class of novel bacterial proteins (Paper VII)

### 5.4.1. Overview of function prediction of novel bacterial ORFs

The complete genomes of a growing number of bacteria have been sequenced. The total number of distinct complete genomes in the bacterial genome database at NCBI (http://www.ncbi.nlm.nih.gov/genomes/static/eub_g.html) has reached 202. Knowledge of these genomes has facilitated the mechanistic study of bacterial growth and infections [309, 310], and the search of antibacterial targets [311-314]. The function of a substantial percentage (17-20%) of the putative protein-coding open reading frames (ORFs) in these genomes is unknown [236, 237]. Determination of the function of these ORFs is important for a more comprehensive understanding of the molecular mechanism of bacterial growth and infections and for searching novel antibacterial targets.

The sequence of these ORFs has no significant similarity to proteins of known functions. As a result, their functions are difficult to probe based on sequence similarity alone. The system developed in this work has shown some potential for assignment of a functional class of distantly related proteins and homologous proteins of different functions as well as homologous proteins of similar functions[43, 45, 46, 227]. It classifies proteins into functional classes defined based on activities or physico-chemical properties rather than sequence similarity [14, 31, 43, 45, 228]. This work is intended to further evaluate the capability of SVMProt for predicting the functional class of bacterial proteins of unknown function. It is assessed by using novel bacterial proteins that are without a single homolog in the Swiss-Prot database [180] and not included in the training sets of SVMProt. The precise functions of these proteins are described in the literature. These proteins are collected from an unbiased search of Medline [244] and

Swiss-Prot database [180]. The SVMProt predicted functional classes of these proteins are compared with the reported function to evaluate to what extent SVMProt is useful for the functional class assignment of these proteins. The prediction accuracy for the assignment of these novel proteins is compared with the overall accuracy of the SVMProt assignment of a large number of proteins to examine to which extent that sequence similarity affects the prediction accuracy of SVMProt.

### 5.4.2. Selection of novel bacterial proteins

The key words, "novel protein bacterium" or "novel bacterial protein", are used to search the Medline [244] and the Swiss-Prot database [180] for finding bacterial proteins that are both described as novel and with their precise function provided. As the search of the Medline is confined to the abstracts, those proteins whose function is not explicitly hinted in an abstract are not selected. Thus the selected proteins likely account for a portion of the known novel viral proteins with available functional information. PSI-BLAST[38] sequence analysis is subsequently conducted on each of these novel viral proteins against all Swiss-Prot entries in the Swiss-Prot protein database [180] so that those with at least one sequence homolog of known function (including that of the same protein in different species) are removed. The commonly-used criterion for homologs, the similarity score e-value less than the inclusion threshold value of 0.005 [38], is used in this work. Finally, those proteins that are in the training sets of SVMProt are removed. 46 novel bacterial proteins are identified in this process, which together with their protein accession number and literature-described functional indications and related references are given in Table 5-5.

### 5.4.3. Results and discussion of functional class prediction of novel bacterial proteins

Table 5-5 gives SVMProt ascribed functional classes for each of the 46 novel bacterial proteins together with their literature-described function. As shown in Table 5-5, there are 26 proteins with the top hit and 5 proteins with one of the hits of the SVMProt assigned functional classes matching the reported function, representing 67.4% of the novel bacterial proteins studied in this work. The 26 top-hit-matching proteins are Nhe of *Bacillus cereus* [315], AAC(6') of *Enterobacter aerogenes* [316], alpha-clostripain of *Clostridium histolyticum* [317], AMDASE of *Bordetella bronchiseptica* [318], aminopeptidase G of *Streptomyces lividans* [319], 2-aminomuconate deaminase of *Pseudomonas pseudoalcaligenes* [320], ammonia monooxygenase of *Nitrosomonas europaea* [321] , AmpE protein of *Escherichia coli* [322], esterase precursor of *Streptomyces scabies* [323], CM of *Bacillus subtilis* [324], cytochrome c oxidase polypeptide IV of *Paracoccus denitrificans* [325], 2-dehydro-3-deoxygalactonokinase of *Escherichia coli* [326], DNA polymerase III theta subunit of *Escherichia coli* [327], Extracellular lipase of *Aeromonas hydrophila* [328], Extracellular serine protease of *Bacteroides nodosus* [329], flp-1 of *Actinobacillus actinomycetemcomitans* [330], Histidine protein kinase of *Lactobacillus plantarum* [331], Monofunctional chorismate mutase precurs of *Erwinia herbicola* [332], PNGase F Glycopeptide N-glycosidase N-glycanase of *Flavobacterium meningosepticum* [333], Precorrin-6A reductase of *Pseudomonas denitrificans* [334], Putative cytochrome P450 128 of *Mycobacterium tuberculosis* [335], Thiocyanate hydrolase beta subunit of *Thiobacillus thioparus* [336], Thiaminase I [Precursor] of *Paenibacillus thiaminolyticus* [337], DNA alpha-glucosyltransferase of *Bacteriophage T4* [246], Type II restriction enzyme ScaI of *Streptomyces caespitosus* [338], and ATP synthase C chain of *Rhodospirillum rubrum, Paenibacillus thiaminolyticus* [339].

Flp-1 protein of *Actinobacillus actinomycetemcomitans* has been found to be associated with the bacterial cell surface and smaller structures, involved in fibril formation and cell adherence [330]. The top hit of SVMProt predicted functional class for this protein is the cell adhesion class, which matches with the literature-described functions. AmpE protein of *Escherichia coli* has been reported to be an integral membrane protein with a likely ATP-binding site between the second and third putative transmembrane region [340]. This protein is predicted as a transmembrane protein by SVMProt, which is consistent with the reported function. The other 28 correctly assigned proteins are enzymes and their SVMProt predicted EC class matches with the corresponding EC number.

Because these proteins have no homolog of known function in Swiss-Prot database based on PSI-BLAST search, our study suggests that SVMProt has a certain capability for providing useful hint about the functional class of novel proteins with no or low homology to known proteins, and this capability is not based on sequence similarity or clustering. The overall accuracy of 67.4% for the assignment of the novel bacterial proteins is smaller than that of 87% for the SVMProt functional class assignment of 34,582 proteins that have at least one homolog of known function. This indicates the sequence-similarity-independent nature of SVM protein classification.

Several factors may affect the accuracy of SVMProt for functional characterization of novel bacterial proteins. One is the diversity of protein samples used for training SVMProt. It is likely that not all possible types of proteins, particularly those of distantly related members, are adequately represented in some protein classes. This can be improved along with the availability of more protein data. Not all distantly related proteins of the same function have similar structural and chemical features. There are cases in which different functional groups, un-conserved with respect to position in the primary sequence, mediate the same mechanistic role, due to the flexibility at the active site [272]. This plasticity is unlikely to be sufficiently described by the physicochemical descriptors currently used in SVMProt. Therefore, SVMProt in the present form is not expected to be capable of classification of these types of distantly related proteins.

Some of the SVMProt functional classes are at the level of families and superfamilies that may include a broad spectrum of proteins. It has been shown that performance of SVM may not better than HMM for distinguishing proteins in a superfamily, but may be more accurate with subfamily discrimination [31]. Thus, the use of some large families and superfamilies as the basis for classification may affect the prediction accuracy of SVMProt to some extent.

In this evaluation work, SVMProt shows a certain level of capability for predicting the functional class of a number of novel bacterial proteins. This suggests that SVMProt is potentially useful to a certain extent for providing useful hints about the function of distantly related proteins in the genomes of bacteria as well as in other organisms.

Table 5-5 Novel bacterial proteins, literature-described functional indications as suggested from experiment and/or sequence analysis, and SVMProt predicted functions. The SVMProt predicted functions are categorized in one of the three classes: The first class is M (matched), in which all of the literature-described functional indications are predicted. The second is PM (partially matched), in which some of the literature-described functional indications are predicted. The third is NM (not matched), in which No function predicted of the literature-described functions matched or were consistent with a predicted function.

| **Protein [Swiss-Prot or NCBI accession number]** | **Bacterium** | **Literature Described Function (reference)** | **Function characterized by SVMProt (probability of correct characterization P-value)** | **Prediction status** |
|---|---|---|---|---|
| AAC(2')-IC [P95219] | *Mycobacterium tuberculosis; Mycobacterium bovis* | Aminoglycoside 2'-N-acetyltransferase (EC 2.3.1.-) [341] | EC 2.7 Transferase of Phosphorus-Containing Groups (78.4%)<br>EC 4.2 Carbon-Oxygen Lyase (58.6%) | NM |
| Nhe [P81242] | *Bacillus cereus.* | Non-hemolytic enterotoxin 105 kDa component (EC 3.4.24.-) [315] | EC 3.4. Hydrolases - Acting on peptide bonds (Peptidases) (99.0%)<br>EC 3.1. Hydrolases - Acting on Ester Bonds (65.4%) | M |
| AAC(6') [P50858] | *Enterobacter aerogenes (Aerobacter aerogenes)* | Aminoglycoside N(6')-acetyltransferase type 1 (EC 2.3.1.82) [316]. | EC 2.3 Acyltransferases (99.0%)<br>EC 3.1.Hydrolases - Acting on Ester Bonds(86.8% )<br>EC2.7.Transferases-Transferring Phosphorus-Containing Groups (68.5%)<br>EC 4.2. Carbon-Oxygen Lyases(62.2%)<br>EC 4.1. Carbon-Carbon Lyases(58.6%)<br>Outer membrane (58.6%) | M |
| ADP-ribosyltransferase [P14299] | *Rhodospirillum rubrum* | ADP-ribosyltransferase(EC 2.4.2.37) [342] | Transmembrane (92.9%)<br>EC 2.4. Glycosyltransferase (90.3%)<br>Outer membrane (58.6%) | M |
| Limonene-1,2-epoxide hydrolase [Q9ZAG3] | *Rhodococcus erythropolis..* | Limonene-1,2-epoxide hydrolase (EC 3.3.2.8)[343] | EC 3.3 Hydrolases - Acting on Ether Bonds (99.0%)<br>EC 4.2. Carbon-Oxygen Lyases (71.3%)<br>Transmembrane (62.2%)<br>Outer membrane (58.6%) | M |

| | | | | |
|---|---|---|---|---|
| AhdA2cA1c [BAC65427.1] | *Sphingobium sp. strain P2* | a salicylate 1-hydroxylase (EC 1.14.-.-)[344] | EC 3.1. Hydrolase acting on Ester Bonds (82.2%)<br>DNA-binding Protein (80.4%)<br>Transmembrane (58.6%) | NM |
| Alginate lyase [Q59478] | *Klebsiella pneumoniae* | Alginate lyase(EC 4.2.2.3)[345] | Transmembrane (96.4 %)<br>EC 3.1. Hydrolases - Acting on Ester Bonds (78.4%)<br>Outer membrane (58.6%) | NM |
| Alpha-clostripain [P09870] | *Clostridium histolyticum* | Clostridiopeptidase B (EC3.4.22.8 )[317] | EC 3.4. Peptidases (99.0%)<br>TC 1.B. Beta-Barrel porin (58.6%) | M |
| Alpha-N-AFase II [P82594] | *Streptomyces chartreusis* | Arabinosidase II (EC 3.2.1.55) [346] | EC 3.4. Peptidase (91.3%) | NM |
| AMDASE [Q05115] | *Bordetella bronchiseptica (Alcaligenes bronchisepticus)* | Arylmalonate decarboxylase (EC 4.1.1.76) [318] | EC 4.1. Carbon-Carbon Lyases (99.0%)<br>Transmembrane (93.6%)<br>EC 1.1. Oxidoreductases - Acting on the CH-OH group of donors (68.5%)<br>EC 4.2. Carbon-Oxygen Lyases (62.2%) | M |
| Aminopeptidase G [Q54340] | *Streptomyces lividans* | Aminopeptidase G(EC 3.4.11.-)[319] | EC 3.4. Hydrolases - Acting on peptide bonds (Peptidases) (99.0%)<br>TC 1.C. Pore-forming toxins (proteins and peptides) (58.6%) | M |
| Aminopeptidase [AAK69184.1] | *Sphingomonas capsulata* | A novel aminopeptidase with unique substrate specificity, no significant homology to any known aminopeptidases (EC3.4.-.-)[347] | EC 3.5. Hydrolase acting on Carbon-Nitrogen Bonds, other than Peptide Bonds (78.4%)<br>EC 1.1. Oxidoreductase acting on the CH-OH group of donors (76.2%)<br>Outer membrane (58.6%)<br>TC 1.B. Beta-Barrel porins (58.6%) | NM |

| | | | | |
|---|---|---|---|---|
| 2-aminomuconate deaminase [P81593] | *Pseudomonas pseudoalcaligenes* | 2-aminomuconate deaminase (EC 3.5.99.5) [320] | EC 3.5. Hydrolase acting on Carbon-Nitrogen Bonds, other than Peptide bonds (99.0%)<br>EC 3.4. Peptidase (58.6%) | M |
| Ammonia monooxygenase [Q04508] | *Nitrosomonas europaea* | Ammonia monooxygenase(EC 1.13.12.-)[321] | EC 1.13. Oxidoreductases - Acting on single donors with incorporation of molecular oxygen (oxygenases) (99.0%)<br>Transmembrane(99.0%)<br>EC 2.4. Transferases - Glycosyltransferases (83.9%) | M |
| protein C5 [Q9RSH3] | *Deinococcus radiodurans.* | Ribonuclease P protein component(EC 3.1.26.5 ) [348] | EC 3.1.Hydrolases - Acting on Ester Bonds (99.0%)<br>RNA-binding Proteins (99.0%)<br>rRNA-binding Proteins (78.4%)<br>DNA-binding Proteins (62.2%)<br>mRNA-binding Proteins (58.6%)<br>TC 1.A. Alpha-Type channels (58.6%) | M |
| AmpE protein [P13017] | *Escherichia coli ; Shigella flexneri* | an integral membrane protein with a likely ATP-binding site between the second and third putative transmembrane region [340] | Transmembrane (99.0%)<br>7 transmembrane receptor (Odorant receptor) (58.6%) | M |
| Esterase precursor [P22266] | *Streptomyces scabies.* | Esterase precursor (EC 3.1.1.-)[323] | EC 3.1. Hydrolases - Acting on Ester Bonds (99.0%)<br>Transmembrane (86.8%)<br>EC 3.4. Hydrolases - Acting on peptide bonds (Peptidases) (62.2%) | M |
| Beta-agarase B [P48840] | *Vibrio sp. (strain JT0107)* | Beta-agarase B (EC 3.2.1.81)[349] | Outer membrane (58.6%)<br>TC 1.B. Beta-Barrel porin (58.6%) | NM |

| | | | | |
|---|---|---|---|---|
| cehA [BAB85626.1] | *Rhizobium sp. strain AC100* | a novel carbaryl hydrolase (EC 3.5.1.-)[350] | EC 1.4. Oxidoreductases - Acting on the CH-NH2 group of donors (85.4%)<br>Transmembrane (82.2%)<br>EC 2.7. Transferases - Transferring Phosphorus-Containing Groups (73.8%)<br>EC 6.4. Ligases - Forming Carbon-Carbon Bonds (65.4%) | NM |
| CM [P19080] | *Bacillus subtilis* | CM(EC 5.4.99.5)[324] | EC 5.4. Intramolecular Transferase (99.0%)<br>EC 4.2. Carbon-Oxygen Lyase (58.6%)<br>Outer membrane (58.6%) | M |
| Curlin genes transcriptional activatory protein [P24251] | *Escherichia coli* | Curlin genes transcriptional activatory protein[351] | EC 2.7. Transferases - Transferring Phosphorus-Containing Groups (78.4%) | NM |
| Cytochrome c oxidase polypeptide IV [P30815] | *Paracoccus denitrificans* | Cytochrome c oxidase polypeptide IV (EC 1.9.3.1)[325] | EC 1.9 Oxidoreductase of a heme group of donors (97.0%)<br>Envelope protein (58.6%)<br>Transmembrane (58.6%) | M |
| 2-dehydro-3-deoxygalactonokinase [P31459] | *Escherichia coli* | 2-dehydro-3-deoxygalactonokinase (EC 2.7.1.58)[327] | EC 2.7. Transferases - Transferring Phosphorus-Containing Groups (99.1%)<br>EC 2.3. Transferases - Acyltransferases (76.2%)<br>EC 4.1. Carbon-Carbon Lyases(65.4% )<br>EC 4.2. Carbon-Oxygen Lyases (58.6%) | M |
| DNA polymerase III, theta subunit [P28689] | *Escherichia coli; Shigella flexneri* | DNA polymerase III, theta subunit (EC 2.7.7.7)[326] | EC 2.7 Transferase of Phosphorus-Containing Groups (99.0%)<br>EC 4.2 Carbon-Oxygen Lyase (58.6%) | M |
| Extracellular lipase[P40600] | *Aeromonas hydrophila* | Triacylglycerol lipase (EC 3.1.1.3)[328] | EC 3.1. Hydrolase acting on Ester Bonds (99.0%)<br>EC 1.3. Oxidoreductase acting on the CH-CH group of donors (65.4%)<br>Outer membrane (58.6%)<br>TC 1.B. Beta-Barrel porin (58.6%) | M |

| | | | | |
|---|---|---|---|---|
| Extracellular serine protease [P19577] | *Bacteroides nodosus (Dichelobacter nodosus)* | Extracellular serine protease (EC 3.4.21.-)[329] | EC 3.4. Peptidase (99.0%)<br>TC 1.C. Pore-forming toxins (62.2%) | M |
| flp-1 [Q9ANW5] | *Actinobacillus actinomycetemcomitans* | Associated with the bacterial cell surface and smaller structures, involved in fibril formation and cell adherence[330] | Cell adhesion (58.6%)<br>Outer membrane (58.6%)<br>Seven transmembrane receptor secretin family(58.6%) | PM |
| Glycosyl transferase alg8 [Q887P9] | *Pseudomonas syringae (pv. tomato)* | Glycosyl transferase alg8 (EC 2.4.1.-)[352] | Transmembrane (99.0%)<br>EC 2.4 Glycosyltransferase (98.6%) | M |
| HbpA [AAK68926.1] | *Treponema denticola* | iron-regulated 44-kDa outer membrane protein (HbpA) with hemin binding ability[353] | EC 3.4. Peptidase (86.8%)<br>EC 3.1. Hydrolase acting on Ester Bonds (62.2%)<br>EC 1.7. Oxidoreductase acting on other nitrogenous compounds as donors (62.2%) | NM |
| Histidine protein kinase [Q88S61] | *Lactobacillus plantarum* | Histidine protein kinase (EC 2.7.3.-)[331] | EC 2.7. Transferases - Transferring Phosphorus-Containing Groups (91.3%)<br>TC 2.C.Electrochemical Potential-driven transporters - Ion-gradient-driven energizers (73.8%)<br>TC 3.A.Primary Active Transporters - P-P-bond-hydrolysis-driven transporters (73.8%) | M |
| Hypothetical protein BBB03 [O50979] | *Borrelia burgdorferi (Lyme disease spirochete)* | Hypothetical protein BBB03 (EC 3.1.22.-)[354] | EC 2.7 Transferase of Phosphorus-Containing Groups (88.1%)<br>EC 3.4 Peptidase (86.8%)<br>EC 2.3 Acyltransferase (71.3%)<br>EC 4.1 Carbon-Carbon Lyase (65.4%) | NM |
| Monofunctional chorismate mutase precurs [P42517] | *Erwinia herbicola* | Monofunctional chorismate mutase precursor (EC 5.4.99.5)[332] | EC 5.4. Isomerases - Intramolecular Transferases (99.0%) | M |
| omp28 [AAD51843.1] | *Porphyromonas gingivalis* | outer membrane protein[355] | No function predicted | NM |
| opcA [AAL67945.1] | *Neisseria polysaccharea* | outer membrane protein[356] | EC 3.1 Hydrolase acting on Ester Bonds (82.2%)<br>EC 4.2 Carbon-Oxygen Lyase (62.2%)<br>Outer membrane (58.6%)<br>TC1.B Beta-Barrel porin (58.6%) | M |

| | | | | |
|---|---|---|---|---|
| Phenol hydroxylase P4 protein [P19733] | *Pseudomonas sp. (strain CF600)* | Phenol 2-monooxygenase P4 component (EC 1.14.13.7 )[357] | EC 1.9. Oxidoreductase acting on a heme group of donors (78.4%)<br>EC 3.4. Peptidase (58.6%)<br>EC 4.1.Carbon-Carbon Lyase (58.6%) | NM |
| Phenylacetaldoxime dehydratase [P82604] | *Bacillus sp. (strain OxB-1)* | Phenylacetaldoxime dehydratase (EC 4.2.1.-)[358] | Transmembrane (98.2%)<br>EC 3.4 Peptidase (96.4%)<br>EC 3.3 Hydrolase of Ether Bonds (80.4%)<br>EC 2.7 Transferase of Phosphorus-Containing Groups (73.8%) | NM |
| PNGase F Glycopeptide N-glycosidase N-glycanase [P21163] | *Flavobacterium meningosepticum (Chryseobacterium meningosepticum)* | Peptide-N(4)-(N-acetyl-beta-D-glucosaminyl)asparagine amidase F precursor (EC 3.5.1.52)[333] | EC 3.5 Hydrolase of non-Peptide Carbon-Nitrogen Bonds (99.0%)<br>Beta-Barrel porin (58.6%) | M |
| Precorrin-6A reductase [P21920] | *Pseudomonas denitrificans* | Precorrin-6A reductase (EC 1.3.1.54)[334] | EC 1.3. Oxidoreductases - Acting on the CH-CH group of donors (99.0%)<br>EC 3.5. Hydrolases - Acting on Carbon-Nitrogen Bonds, other than Peptide Bonds (58.6%)<br>Outer membrane (58.6%) | M |
| Putative cytochrome P450 128 [Q59572] | *Mycobacterium tuberculosis* | Putative cytochrome P450 128 (EC 1.14.-.-)[335] | EC 1.14. Oxidoreductases - Acting on paired donors with incorporation or reduction of molecular oxygen (99.0%)<br>EC 2.3. Transferases – Acyltransferases(86.8%)<br>EC 4.1. Carbon-Carbon Lyases(85.4% )<br>EC 4.2. Carbon-Oxygen Lyases (83.9%) | M |
| Slt35 [P41052] | *Escherichia coli* | Membrane-bound lytic murein transglycosylase B (EC 3.2.1.-) [359] | Outer membrane (99.0%)<br>EC 1.1. Oxidoreductase acting on the CH-OH group of donors (89.3%)<br>EC 4.1. Carbon-Carbon Lyase (62.2%) | NM |
| Thiocyanate hydrolase beta subunit [O66186] | *Thiobacillus thioparus* | Thiocyanate hydrolase beta subunit (EC 3.5.5.8)[336] | EC 3.5 Hydrolase of non-Peptide Carbon-Nitrogen Bonds (98.9%)<br>EC 2.6 Transferases of Nitrogenous Groups (62.2%) | M |

| | | | | |
|---|---|---|---|---|
| Thiaminase I [Precursor] [P45741] | *Paenibacillus thiaminolyticus (Bacillus thiaminolyticus).* | Thiaminase I precursor (Thiamine pyridinylase)( EC 2.5.1.2) [337] | EC 2.5. Transferases - Transferring Alkyl or Aryl Groups, Other than Methyl Groups (99.0%)<br>EC 2.7. Transferases of Phosphorus-Containing Groups (94.7%)<br>Transmembrane (90.3%) | M |
| DNA AGT [P04519] | *Bacteriophage T4.* | DNA alpha-glucosyltransferase (EC 2.4.1.26) [246] | EC 2.4. Glycosyltransferases (80.4%)<br>EC 2.7. Transferases - Transferring Phosphorus-Containing Groups (68.5%) | M |
| Hydroxyneurosporene dehydrogenase [Q9F723] | *Chlorobium tepidum.* | Hydroxyneurosporene dehydrogenase (EC 1.-.-.-) [360] | EC 4.1.Carbon-Carbon Lyases (65.4%) | NM |
| Type II restriction enzyme ScaI [O52691] | *(R.ScaI).Streptomyces caespitosus.* | Type II restriction enzyme ScaI (Endonuclease ScaI) (EC 3.1.21.4) [338] | EC 3.1. Hydrolases acting on Ester Bonds (99.0%)<br>TC 1.C Pore-forming toxins(proteins and peptides) (58.6%) | M |
| ATP synthase C chain [P15014] | *Rhodospirillum rubrum.* | ATP synthase C chain (Lipid-binding protein) (EC 3.6.3.14) [339] | EC 3.6. Hydrolases acting on Acid Anhydrides (99.0%)<br>Transmembrane (58.6%) | M |

# 6. Prediction of Protein Inhibitors by Statistical Learning Approach, HIV-1 Protease as a case study

The in depth understanding of the drug-target interaction mechanism and rapid advances in biochemistry and organic chemistry lead to the advent of computer aided drug design[52-55, 361-364], which aims to help the rapid and efficient discovery of drug leads. Existing computational investigations mostly focused on how to improve the interaction between protease and inhibitor. One approach is to simulate HIV-1 protease with a substrate by finding if there is a stable energy minimum by molecular dynamics[365-367] or docking[368]. Another method to speed up the PI development process is the identification of PIs in the early stage of drug discovery using statistical learning methods. As such, drug candidates that are not involved in protease inhibition can be eliminated earlier and the cost effectiveness of the drug discovery process can be improved. In a study by Patankar and Jurs[58], radial basis function neural networks were used for classify HIV-PI. The model was trained with a limited set of only 123 compounds and tested using 12 compounds. Although the predictive ability was in the high 80% range for the external prediction set, the model is not robust due to the small representation of compounds and statistically insignificant prediction set.

In this study, support vector machine is implemented for HIV-1 protease inhibitors exploration by using new strategy and more comprehensive data set.

## 6.1. Methods

### 6.1.1. HIV-1 Protease Inhibitors

An accurate SVM classification model requires large number of examples for both protease inhibitors and non-inhibitors. In this study, HIV-1 PIs are selected from the

HIV/OI Enzyme Inhibition Database of the National Institute of Allergy and Infectious Diseases (National Institutes of Health). The diversity analysis is based on the chemical family information that is obtained from the same data source. In our collected HIV protease inhibitors, 76.57% of them are peptide-based inhibitors and only 23.43% are non-peptide-based inhibitors. Among these peptide-based inhibitors, about 66%of them are peptidomimetics that are made up of a wide variety of compounds, and only about 5% are symmetry-based inhibitors.

Peptidomimetics can be described as compounds derived from peptides and proteins and are obtained by structural modification using unnatural amino acids, conformational restraints, isosteric replacement, cyclisation etc. The peptidomimetics bridge the gap between simple peptides and the nonpeptide synthetic structures and may be useful in delineating pharmacophores and in helping to translate peptides into small non-peptide compounds. Peptidomimetic is sometimes used in a broad sense to designate organic molecules mimicking some properties of peptide ligands[369].They are the most common starting point for HIV-1 inhibitor drug development and have been designed to mimic the tetrahedral transition-state intermediate formed during the HIV-1 protease catalysis event. In this study, 57.18% of the total positive samples found were peptidomimetics.

Although there is currently no non-peptide-based inhibitors reaching clinical trials, but there has been considerable interest in using non-peptide based compounds in HIV drug development. Thus, we also consider these non-peptidic inhibitors in this study.

### 6.1.2. HIV-1 Protease non-Inhibitors

The supervised statistical learning requires both substantial positive and negative examples to develop a prediction model with a certain generalization potential. Thus, the selection of effective negative example should be considered seriously in terms of the distribution and effectiveness.

In this work, Hierarchal Clustering is employed to analyze the compound distribution according to their descriptors. The comprehensive negative examples are chosen from the following conditions based on the distribution of positive examples: (1) They allocate in the chemical space not occupied by positive examples, (2) their structures must be sufficiently distinct from the positive samples, and (3) the distribution of the selected negative examples should be diverse enough to form an effective representation of negative examples within the chemical space. To simulate the entire chemical space and figure out the distribution of positive examples within this space, a compound database[*] composed of 85,000 entries is constructed. In this work, a total number of 12453 negative samples were selected to ensure data balance in the binary classification SVM model.

### 6.1.3. Positive and negative samples quantity

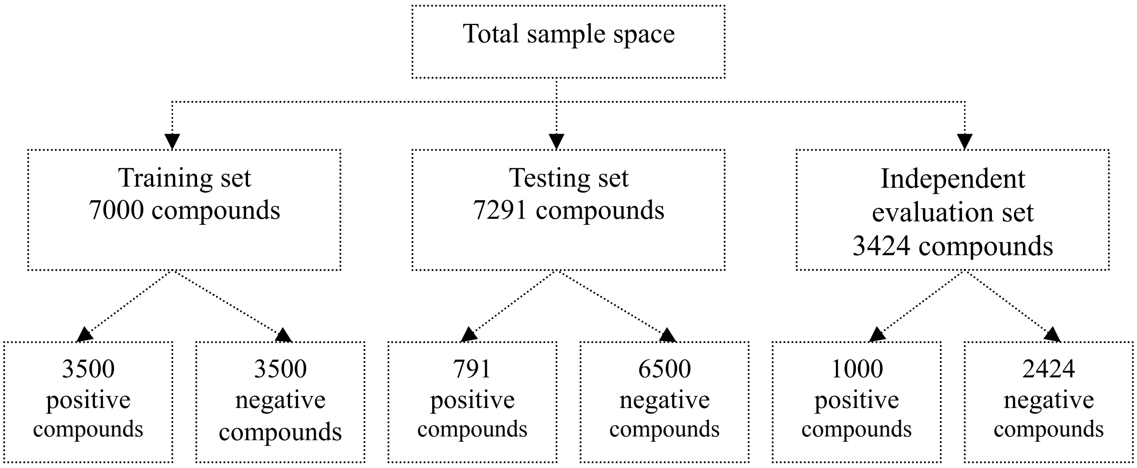
The use of comprehensive dataset for model training is required for developing a robust and reliable prediction system, in turn, a small sample size and range is inadequate in representing all chemical families of HIV-1 PIs and non-PIs, leading to biased learning and eventually poor accuracies.

In terms of robustness, reliability and statistical significance, the SVM model developed in this work is a significant improvement from the previously reported work for the prediction of HIV protease inhibitor using radial basis function neural

[*] These compounds are selected from MDDR, ACD and ChemIDPlus, ChemFinder databases with available 3D structures

networks[58] due to the notably larger, and more diversified dataset size for both positive and negative samples. The number and distribution of data used in each set are shown in Figure 6-1.

Figure 6-1 The distribution and number of samples in each set



## 6.2. Results and Discussion

### 6.2.1. Self- consistence testing accuracy

As shown in Table 6-1, The prediction sensitivity, specify and overall accuracy of the testing set are 80.7%, 94.1% and 92.7%respectively, which suggest the self consistency of the model.

It is noticed that the prediction accuracy of HIV-1 PIs (*sensitivity*) is lower than the prediction accuracy of non-HIV-1 PIs (*specificity*) as shown in Table 6-1. This may be explained by the smaller size of the positive sample dataset compared to that of the negative dataset. It has been known that SVM model based on unbalanced data set tends to produce feature vectors that push the hyperplane towards the side with a smaller number of data[370]. This can lead to reduced accuracy for the set either with a smaller number of samples or of less diversity. The higher prediction accuracy for non-PIs is likely the result of the availability of a more numerous and diverse set of samples as compared to the HIV-1 PIs due to the selection of one or few chemicals from each super family from the database. This enables SVM to perform a better statistical learning for recognition.

Table 6-1 The prediction accuracy of the testing set. Predicted results are given in TP (true positive), FN (false negative), TN (true negative), FP (false positive), HIV-PIs prediction accuracy (TP/(TP+FN)), and Non-HIV-PIs prediction accuracy (TN/(TN+FP)). Number of positive or negative samples in the testing sets is TP+FN or TN+FP respectively.

| **No.HIV-PIs** | | **No.Non-HIV-PIs examples** | | **Accuracies** | | |
|---|---|---|---|---|---|---|
| **TP** | **FN** | **TN** | **FP** | **HIV-PIs** | **Non-HIV-PIs** | **Overall Accuracy** |
| 638 | 153 | 6118 | 382 | 80.66% | 94.12% | 92.66% |

A direct comparison with results from an earlier study is impractical because of differences in the quantity and quality of data, molecular descriptors and classification methods and algorithms used. Nonetheless, a rough comparison with Patankar and Jurs's work[58] on HIV PI prediction by using neural networks shows that our approach improved the testing accuracy from the 80% to 92%.

### 6.2.2. Independent evaluation

The optimal separating hyperplane was constructed after the training process and it was subjected to further evaluation using an independent dataset that does not overlap with that which was used for model training and testing. The independent evaluation is aimed to show the potential of the model's generalization abilities. The independent evaluation results are as summarized in Table 6-2.

Table 6-2 The results of independent evaluation. Predicted results are given in TP (true positive), FN (false negative), TN (true negative), FP (false positive), HIV-PIs prediction accuracy (TP/(TP+FN)), and Non-HIV-PIs prediction accuracy (TN/(TN+FP)). Number of positive or negative samples in the testing sets is TP+FN or TN+FP respectively.

| **No.HIV-PIs** | | **No.Non-HIV-PIs examples** | | **Accuracies** | | |
|---|---|---|---|---|---|---|
| **TP** | **FN** | **TN** | **FP** | **HIV-PIs** | **Non-HIV-PIs** | **Overall Accuracy** |
| 989 | 11 | 2334 | 90 | 98.90% | 96.29% | 97.05% |

The prediction accuracies for positive samples, negative samples and overall sample set are in the range from 96.29% to 98.90%. This suggests that the classification model is quite reliable in terms of the prediction accuracies.

The positive sample prediction accuracy (sensitivity) of the independent evaluation set was 98.90%, where only 11 samples were incorrectly predicted out of 1000. One of the possible reasons is that not all individual compound sub-groups have the same accuracies, and further knowledge of this might be of help to provide a way to improve the overall accuracy of the model. The prediction accuracy for each chemical family is shown in Table 6-3. It was found that there are indeed a few groups that have good accuracies, such as amines, peptides, peptidomimetics and inhibitors without any specified class have precision of 100.00%. The prediction accuracy of non-peptides is also at the high-value rage between 90% and 97.30%. The groups with the lowest sensitivities are amides and symmetry-based inhibitors, with only 80.00% and 93.48% respectively. It was noted that the sample size for amides is small, and this sensitivity may not be representative of the true prediction power of the model when more samples are given. The sensitivity of the amides group in the testing set is 62.96% (18/27). The amides make up only 3.25% of the total positive sample size used. The relative lack of samples in this group may be inadequate to represent all compound subtypes in the family, leading to biased learning and poor accuracies eventually. The future inclusion of newly found amide protease inhibitors in SVM model training is likely to increase the accuracy. The same reason applies to the symmetry-based inhibitors, which only make up 3.84% of the training sample.

Table 6-3 The sensitivity of individual groups of compounds in the independent evaluation set

| Chemical class | Sensitivity<br>(Number of true positive/total number of compound) |
|---|---|
| Amides | 80.00% (4/5) |
| Amines | 100.00% (10/10) |
| Non-peptides | 97.30% (252/259) |
| Peptides | 100.00% (15/15) |
| Peptidomimetics | 100.00% (658/658) |
| Symmetry-based inhibitors | 93.48% (43/46) |
| Unspecified | 100.00% (7/7) |
| Overall average | 98.90% (989/1000) |

### 6.2.3. Recursive Feature Elimination

As introduced previously in Chapter 1.4, the feature selection provides the insight for discriminating the positive and negative examples. In this study, the non-linear recursive feature elimination (RFE) method was used to select the 20 predominant features for discriminating HIV PIs and non-HIV PIs.

#### *6.2.3.1. Selected significant features by RFE*

Table 6-4 gives the list of RFE-selected descriptors for HIV-1 PI classification in the order of importance, with the most significant feature on top.

As shown in Table 6-4, half of the important features selected by RFE methods are simple molecular connectivity chi indices, such as topological descriptors, which represent how constituent atoms are interconnected in the molecule. Three other molecular shape kappa indices quantitative the molecular structure from its shape. Besides, simple molecular properties related to electrostatic interaction such as the number of hydrogen atoms and the number of H-bond donors and acceptors are also show certain importance. The remaining dominant features are molecular weight, Kier molecular flexibility index, number of rotatable bonds and the number of rings – which dictates the rigidity of the inhibitors. This is consistent with a previous comparative quantitative structure activity relationship (QSAR) study[371] of inhibitory activity of HIV-1 protease inhibitor model. They suggested that topological, molecular connectivity, and kappa shape indices were important for binding. These features were interpretable as hydrogen bond donating ability, non-polar groups, skeletal branching, and molecular globularity.

Table 6-4 Molecular descriptors selected by the RFE method for the classification of HIV-1 PIs

| Descriptor selected | Description | Class |
|---|---|---|
| ${}^{3}\chi_{CH}$ | Simple molecular connectivity chi indices for cycles of 3 atoms | Connectivity and shape |
| ${}^{1}\kappa$ | Molecular shape kappa indices for one bond fragment | Connectivity and shape |
| ${}^{0}\chi$ | Simple molecular connectivity chi indices for path order 0 | Connectivity and shape |
| ${}^{3}\chi_{C}$ | Simple molecular connectivity chi indices for cluster | Connectivity and shape |
| nrot | Number of rotatable bonds | Simple molecular properties |
| ${}^{1}\chi$ | Simple molecular connectivity chi indices for path order 1 | Connectivity and shape |
| ${}^{5}\chi_{CH}$ | Simple molecular connectivity chi indices for cycles of 5 atoms | Connectivity and shape |
| ${}^{2}\chi$ | Simple molecular connectivity chi indices for path order 2 | Connectivity and shape |
| ndonr | Number of H-bond donors | Simple molecular properties |
| ${}^{4}\chi_{CH}$ | Simple molecular connectivity chi indices for cycles of 4 atoms | Connectivity and shape |
| ${}^{3}\kappa$ | Molecular shape kappa indices for 3 bond fragments | Connectivity and shape |
| ${}^{4}\chi_{PC}$ | Simple molecular connectivity chi indices for path/cluster | Connectivity and shape |
| ${}^{6}\chi_{CH}$ | Simple molecular connectivity chi indices for cycles | Simple molecular properties |

| | | |
|---|---|---|
| | of 6 atoms | |
| nhyd | Count of hydrogen atoms | Simple molecular properties |
| phi | Kier molecular flexibility index | Connectivity and shape |
| naccr | Number of H-bond acceptors | Simple molecular properties |
| $W_{mol}$ | Molecular weight | Simple molecular properties |
| $^{3}\chi_{P}$ | Simple molecular connectivity chi indices for path order 3 | Connectivity and shape |
| $^{2}\kappa$ | Molecular shape kappa indices for 2 bond fragments | Connectivity and shape |
| nring | Numbers of rings | Simple molecular properties |

From the RFE study, the absence of electrophilicity descriptors in the dominant feature list indicated that the importance of hydrophobicity is superseded by that of simple molecular properties, molecular connectivity, and kappa shape interactions. Our results suggest that the count of hydrogen atoms, the number of H-bond donors and acceptors are important to distinguish molecular descriptors of HIV-1 PIs. These molecular properties contribute directly to the properties of quantum chemical descriptors such as electrophilicity, polarizability and molecular dipole moment. This is consistent with the finding that hydrogen bonding is extremely crucial in the enzyme-inhibitor interaction[372].

In this study, all topological descriptors found to be significant features are simple molecular chi indices. The key concept in chi indices is the decomposition of the molecular graph into fragments of different size and complexity[373]. As half of the significant features obtained by REF are molecular connectivity descriptors, simple molecular connectivity chi indices for cycles of 3 atoms, path order 0, cluster, path order 1, cycles of 5 atoms, path order 2, cycles of 4 atoms, path/cluster, cycles of 6 atoms, and path order 3 are shown to be important to discriminateHIV-1 PIs and non-PIs. This is as understandable because the peptide-based inhibitors form a major group compounds which are typically heavy, long chain, complex molecules with many branch points and ring structures. Besides, the molecular topology for non-peptidic inhibitors is also intricate, such as carbohydrates, nucleoside conjugates, natural product and symmetry-based inhibitors have complicated structures with numerous rings and branch points. The difference in the complexity of molecules might cause the significant difference in their distinctive indices.

Apart from the molecular connectivity descriptors, molecular shape kappa indices for one, two, and three bonded fragments are found in the RFE-selected feature list. The kappa shape indices are the basis of a method of molecular structure quantization. In this study, the importance of Kier molecular shape indices suggests that HIV-1 protease is highly specific for their substrate in terms of their shape. The inhibitors of an enzyme should be of similar shape and chemical nature as the substrate in order to align properly with the active site and bind tightly to it. This approach has been widely used to design inhibitors for diverse enzymatic targets, including HIV-1 protease[374].

Our results also revealed that the number of rotatable bonds, number of rings, Kier molecular flexibility index, and molecular weight have important contributions to discriminated PIs and non-PIs. The Kier molecular flexibility index is a descriptor based on structural properties that restrict a molecule from being "infinitely flexible", the model for which is an endless chain of C ($sp^3$) atoms. The structural features considered to prevent a molecule from attaining infinite flexibility are: (a) fewer atoms, (b) the presence of rings, (c) branching, and (d) the presence of atoms with covalent radii smaller than those of C ($sp^3$)[375].

#### 6.2.3.2. Prediction accuracy by using selected significant features

The elimination of irrelevant molecular descriptors greatly reduced the computation costs. More importantly, the removal of noise-generating features could improve the accuracy of SVM models in some cases[376]. In this work, the testing sensitivity is improved from 80.66% to 84.70% by using the selected 20 features. The specificity and overall accuracy are comparable to that utilizing all of the features, with a slight decrease from 94.12% to 93.40% and 92.66% to 92.40% respectively.

## 6.3. Conclusion remark

This work has lead to a robust and intelligent classification system for predicting HIV-PIs with accuracies in the range of 90%s. As the basis of a statistical learning method, the significant number and diversity of the positive and negative datasets confer statistical significance to the results. Recursive feature elimination coupled Support Vector Machines was successfully employed in the automated selection of relevant molecular descriptors and noise reduction.

# 7. Conclusion

## 7.1. Protein functional class prediction

As the gap between the large amounts of sequences information and their function characterization is continuously increasing[3, 4], efforts has been directed in development of methods for probing protein functions. It is difficult to predict protein functions solely based on the sequence similarity if the protein sequence is dissimilar to the sequence. Moreover, the sequence similarity may not able to distinguish the protein functions for homologous proteins with different functions. Thus, it is desirable to explore methods that are not based on sequence similarity.

One of the main purposes of this study is to develop a prediction system that is able to classify proteins into functional classes based on primary sequence by statistical learning approach – Support Vector Machines. The classification system is designed to be able to assign functional families from proteins' primary sequence irrespective sequence similarity. Protein classification problems such as enzymes classification, transporters classification and RNA-binding proteins classification are studied and the classification models are further evaluated by using independent evaluation sets.

The SVMProt protein functional class prediction system was build on the basis of above described optimized classifiers. SVMProt has increased to 97 protein functional classes as listed in Appendix Table A. The functional classes of SVMProt include 46 enzyme families, 9 channel/transporter families, 21 transporter families, 4 RNA-binding protein families, DNA-binding proteins, 5 G-protein coupled receptors, nuclear receptors, Tyrosine receptor kinases, cell adhesion proteins, coat proteins, envelope proteins, outer membrane proteins, structural proteins, and growth factors.

The independent evaluation of the functional classes in SVMProt showed that the prediction accuracy for proteins belonging to a functional family was greater than 70% for 53 families, and 53%~70% for the remaining 19 families. The accuracy for proteins outside the given functional family was 82%~100% for all families. These accuracies are comparable to that from other SVM studies of proteins in terms of accuracies obtained in evaluation, such as G-protein coupled receptor classification[31] and Protein fold prediction [47]. The results of this work either revealed that substantial portions of misclassified proteins are with low similarity to most members in its family, or described as hypothetical, probable and putative. These findings indicate that sequence distance and novelty have some influence on prediction accuracy. Besides, limited diversity of proteins for some families may also affect the prediction accuracies.

Novel proteins such as novel enzymes, novel bacterial proteins, viral proteins and novel plant proteins are selected and evaluated by our developed protein function prediction system. The evaluation accuracy is in the range of 67~ 85%, it is suggested that the prediction system is useful for protein functional family assignment of distantly related proteins in the genomes of bacteria, virus as well as in other organisms and major functional groups such as enzymes.

The approach employed in this work is to classify proteins into functional classes. Some of functional classes are at the level of families and super families that may include a broad spectrum of proteins. Although the results of this study showed that the SVM prediction system may not work as well as HMM for distinguishing proteins in a super family, it may be more accurate with subfamily discrimination [31]. Thus, the use of some large families and super families as the basis for classification may affect the prediction accuracy of SVMProt to some extent. The prediction accuracy and prediction confidence level could be further improved by incorporating the domain based or multiple sequence alignment based approach with the SVM approach outlined in this study. Because the advantages of SVM in predicting function of novel proteins and advantages of sequence similarity based methods in probing functions from homology proteins could complement each other, the new combined prediction system may be an improvement in terms of prediction accuracy and prediction confidence.

Another issue is that the development of classifier for each protein functional family needs a certain number of representative examples. Only those protein functional classes with enough positive examples could be implemented in the classification system developed in this work. Although one can still build the SVM classifier based on the limited positive examples, the problem of data imbalance may be critical as it would affect the accuracy of a SVM classification system[233]. Thus, approaches to solve the data imbalance problem properly are necessary for further improvements in protein functional family coverage. In addition, improvements in SVM algorithm, reliable protein sample collection, distribution analysis on both positive and negative examples, and a more comprehensive and refined set of protein descriptors may enable the development of this prediction system into a practical and mature tool for facilitating functional study of unknown function proteins.

## 7.2. Prediction of protein inhibitors

As the problem of rapid resistance development and physiological side effects remain in current use of HIV-1 protease inhibitors for anti-HIV therapies, methods for facilitating early elimination of potential HIV-1 protease inhibitors are useful for speeding up new drug discovery in the battle with HIV infections. In this study, addition to the applications in functional families' calcification and prediction, support vector machine also show the potential in the application of HIV-1 protease inhibitors exploration. A set of 4291 inhibitors and 10000 non-inhibitors are selected to train and test an SVM classification system. This gave a prediction accuracy of 97.05% for an independent evaluation set composed of 1000 PIs and 2424 non-PIs. This shows that the classification model developed during the training, testing and independent evaluation process is self-consistent and has certain capable in the selection of probable HIV-1 PI candidates for further experimentation and assay. Recursive feature selection method is employed to select significant descriptors and it was shown that molecular connectivity and shape, flexibility and hydrogen bond interactions are among the most distinguishing features for discriminating HIV-1 protease inhibitors.

In order to get the insight from the significant features for discriminating the PIs and non-PIs, Recursive feature selection method is employed. The results of feature elimination show that half of the important features sleeted by RFE methods are molecular topological descriptors, molecular properties related to electronic interaction such as the number of hydrogen atoms and the number of H-bond donors and acceptors, and some other global properties such as molecular weight, rigidity (Kier molecular flexibility, number of rotatable bonds and the number of rings etc). which is consistent with a previous comparative quantitative structure activity relationship (QSAR) study of inhibitory activity of HIV-1 protease inhibitor model [371].

In conclusion, The results of this study indicated that the statistical learning approach was useful for protein inhibitors prediction, the methods implemented in this work could be extended to other fields in drug discovery and effort could be brought to the development new HIV PIs as well as new drug leads.

Future work entails further refinement of the SVM model through the inclusion of the most recently discovered HIV-PIs and the improvement of SVM algorithms, as well as compliment with other approaches for aiding in inhibitors prediction. As suggested in this study, SVM coupled with RFE is potentially useful as a classifier for facilitating the prediction of HIV-1 PIs. The introduction of weighted function into SVM-RFE is expected to improve the accuracy of the model. Moreover, a comprehensive collection of available compound is important for the compound diversity analysis. Qualified collection of positive examples and representative negative examples are the keys for developing model with generalization power.

# BIBLIOGRAPHY

1. Eisenberg, D., Marcotte, E.M., Xenarios, I. & Yeates, T.O. Protein function in the post-genomic era. *Nature* **405**, 823-826 (2000).
2. Bork, P. et al. Predicting function: from genes to genomes and back. *J Mol Biol* **283**, 707-725 (1998).
3. Hunkapiller, T., Kaiser, R.J., Koop, B.F. & Hood, L. Large-scale and automated DNA sequence determination. *Science* **254**, 59-67 (1991).
4. Roberts, L. Large-scale sequencing trials begin. *Science* **250**, 1336-1338 (1990).
5. Bork, P. & Koonin, E.V. Predicting functions from protein sequences--where are the bottlenecks? *Nat Genet* **18**, 313-318 (1998).
6. Baxevanis, A.D. Practical aspects of multiple sequence alignment. *Methods Biochem Anal* **39**, 172-188 (1998).
7. Schuler, G.D. Sequence alignment and database searching. *Methods Biochem Anal* **39**, 145-171 (1998).
8. Benner, S.A., Chamberlin, S.G., Liberles, D.A., Govindarajan, S. & Knecht, L. Functional inferences from reconstructed evolutionary biology involving rectified databases--an evolutionarily grounded approach to functional genomics. *Res Microbiol* **151**, 97-106 (2000).
9. Eisen, J.A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* **8**, 163-167 (1998).
10. Teichmann, S.A., Murzin, A.G. & Chothia, C. Determination of protein function, evolution and interactions by structural genomics. *Curr Opin Struct Biol* **11**, 354-363 (2001).
11. Marcotte, E.M. et al. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-753 (1999).
12. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90 (1999).
13. Aravind, L. Guilt by association: contextual information in genome analysis. *Genome Res* **10**, 1074-1077 (2000).
14. Bock, J.R. & Gough, D.A. Predicting protein--protein interactions from primary structure. *Bioinformatics* **17**, 455-460 (2001).
15. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).
16. Enright, A.J. & Ouzounis, C.A. GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* **16**, 451-457 (2000).
17. Lipman, D.J. & Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435-1441 (1985).
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
19. Hodges, H.C. & Tsai, J.W. 3D-Motifs: An informatics approach to protein function prediction. *FASB J* **16**, A543-A543 (2002).
20. Gattiker, A., Gasteiger, E. & Bairoch, A. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinformatics* **1**, 107-108 (2002).
21. Henikoff, S. et al. Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609-614 (1997).
22. Thornton, J.M., Todd, A.E., Milburn, D., Borkakoti, N. & Orengo, C.A. From structure to function: approaches and limitations. *Nat Struct Biol* **7 Suppl**, 991-994 (2000).
23. Bartlett, G.J., Borkakoti, N. & Thornton, J.M. Catalysing new reactions during evolution: economy of residues and mechanism. *J Mol Biol* **331**, 829-860 (2003).
24. Orengo, C.A., Todd, A.E. & Thornton, J.M. From protein structure to function. *Curr Opin Struct Biol* **9**, 374-382 (1999).
25. Shakhnovich, B.E., Dokholyan, N.V., DeLisi, C. & Shakhnovich, E.I. Functional fingerprints of folds: evidence for correlated structure-function evolution. *J Mol Biol* **326**, 1-9 (2003).
26. Todd, A.E., Orengo, C.A. & Thornton, J.M. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307**, 1113-1143 (2001).
27. Di Gennaro, J.A. et al. Enhanced functional annotation of protein sequences via the use of structural descriptors. *J Struct Biol* **134**, 232-245 (2001).
28. Stark, A. & Russell, R.B. Annotation in three dimensions. PINTS: Patterns in

Non-homologous Tertiary Structures. *Nucleic Acids Res* **31**, 3341-3344 (2003).

29. Wallace, A.C., Borkakoti, N. & Thornton, J.M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci* **6**, 2308-2323 (1997).
30. Pazos, F. & Sternberg, M.J. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A* **101**, 14754-14759 (2004).
31. Karchin, R., Karplus, K. & Haussler, D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **18**, 147-159 (2002).
32. Bhasin, M. & Raghava, G.P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J Biol Chem* **279**, 23262-23266 (2004).
33. des Jardins, M., Karp, P.D., Krummenacker, M., Lee, T.J. & Ouzounis, C.A. Prediction of enzyme classification from protein sequence without the use of sequence similarity. *Proc Int Conf Intell Syst Mol Biol* **5**, 92-99 (1997).
34. Jensen, L.J. et al. Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol* **319**, 1257-1265 (2002).
35. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-453 (1970).
36. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).
37. Altschul, S.F. & Koonin, E.V. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* **23**, 444-447 (1998).
38. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
39. Ivanciuc, O. et al. Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Curr Med Chem* **11**, 583-593 (2004).
40. Andreeva, A. et al. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**, D226-229 (2004).
41. Jones, S. & Thornton, J.M. Searching for functional sites in protein structures. *Curr Opin Chem Biol* **8**, 3-7 (2004).
42. Contreras-Moreira, B., Fitzjohn, P.W. & Bates, P.A. Comparative modelling: an essential methodology for protein structure prediction in the post-genomic era. *Appl Bioinformatics* **1**, 177-190 (2002).
43. Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X. & Chen, Y.Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res* **31**, 3692-3697 (2003).
44. Cai, Y.D. & Lin, S.L. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim Biophys Acta* **1648**, 127-133 (2003).
45. Cai, C.Z., Han, L.Y., Ji, Z.L. & Chen, Y.Z. Enzyme family classification by support vector machines. *Proteins* **55**, 66-76 (2004).
46. Han, L.Y., Cai, C.Z., Lo, S.L., Chung, M.C. & Chen, Y.Z. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *Rna* **10**, 355-368 (2004).
47. Ding, C.H. & Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17**, 349-358 (2001).
48. Yuan, Z., Burrage, K. & Mattick, J.S. Prediction of protein solvent accessibility using support vector machines. *Proteins* **48**, 566-570 (2002).
49. Hua, S. & Sun, Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol* **308**, 397-407 (2001).
50. H.P. Rang, M.M. Dale & Ritter, J.M. Pharmacology, Edn. 4th. (Churchill Livingstone, London; 1999).
51. Rang, H.P., Dale, M.M. & Ritter, J.M. Pharmacology, Edn. 4th. (Churchill Livingstone, London; 1999).
52. Struthers, R.S., Rivier, J. & Hagler, A.T. Molecular dynamics and minimum energy conformations of GnRH and analogs. A methodology for computer-aided drug design. *Ann N Y Acad Sci* **439**, 81-96 (1985).
53. Marshall, G.R. Computer-aided drug design. *Annu Rev Pharmacol Toxicol* **27**, 193-213 (1987).
54. Vedani, A. [Computer-Aided Drug Design: An Alternative to Animal Testing in the Pharmacological Screening]. *Altex* **8**, 39-60 (1991).
55. Veselovsky, A.V. & Ivanov, A.S. Strategy of computer-aided drug design. *Curr Drug Targets*

*Infect Disord* **3**, 33-40 (2003).

56. Lybrand, T.P. Ligand-protein docking and rational drug design. *Curr Opin Struct Biol* **5**, 224-228 (1995).

57. Daszykowski, M. et al. Classification and regression trees--studies of HIV reverse transcriptase inhibitors. *J Chem Inf Comput Sci* **44**, 716-726 (2004).

58. Patankar, S.J. & Jurs, P.C. Classification of HIV protease inhibitors on the basis of their antiviral potency using radial basis function neural networks. *J Comput Aided Mol Des* **17**, 155-171 (2003).

59. Wang, Y.H., Li, Y., Yang, S.L. & Yang, L. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J Chem Inf Model* **45**, 750-757 (2005).

60. Xia, X., Maliski, E.G., Gallant, P. & Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J Med Chem* **47**, 4463-4470 (2004).

61. Cos, P. et al. Structure-activity relationship and classification of flavonoids as inhibitors of xanthine oxidase and superoxide scavengers. *J Nat Prod* **61**, 71-76 (1998).

62. Kauffman, G.W. & Jurs, P.C. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *J Chem Inf Comput Sci* **41**, 1553-1560 (2001).

63. Mattioni, B.E. & Jurs, P.C. Development of quantitative structure-activity relationship and classification models for a set of carbonic anhydrase inhibitors. *J Chem Inf Comput Sci* **42**, 94-102 (2002).

64. Pirard, B. & Pickett, S.D. Classification of kinase inhibitors using BCUT descriptors. *J Chem Inf Comput Sci* **40**, 1431-1440 (2000).

65. C. Hansch & A. Leo Substituent Constants for Correlation Analysis in Chemistry and Biology. (John Wiley & Sons, New York; 1979).

66. Lipinski, C.A.L., F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Del. Rev.*, 3-25 (1997).

67. Eleni, P. & Dimitra, H.L. Review in quantitative structure activity relationships on lipoxygenase inhibitors. *Mini Rev Med Chem* **3**, 487-499 (2003).

68. Wang, R.B., Kuo, C.L., Lien, L.L. & Lien, E.J. Structure-activity relationship: analyses of p-glycoprotein substrates and inhibitors. *J Clin Pharm Ther* **28**, 203-228 (2003).

69. Agrawal, V.K. et al. QSAR study on carbonic anhydrase inhibitors: water-soluble sulfonamides incorporating beta-alanyl moieties, possessing long lasting-intra ocular pressure lowering properties-a molecular connectivity approach. *Eur J Med Chem* (2005).

70. Olson, A.J. & Goodsell, D.S. Automated docking and the search for HIV protease inhibitors. *SAR QSAR Environ Res* **8**, 273-285 (1998).

71. DesJarlais, R.L. & Dixon, J.S. A shape- and chemistry-based docking method and its use in the design of HIV-1 protease inhibitors. *J Comput Aided Mol Des* **8**, 231-242 (1994).

72. Yamamoto, Y., Ishihara, Y. & Kuntz, I.D. Docking analysis of a series of benzylamino acetylcholinesterase inhibitors with a phthalimide, benzoyl, or indanone moiety. *J Med Chem* **37**, 3141-3153 (1994).

73. Artico, M. et al. Investigation on QSAR and binding mode of a new class of human rhinovirus-14 inhibitors by CoMFA and docking experiments. *Bioorg Med Chem* **4**, 1715-1724 (1996).

74. Coutinho, P.M., Dowd, M.K. & Reilly, P.J. Automated docking of glucoamylase substrates and inhibitors. *Ann N Y Acad Sci* **799**, 164-171 (1996).

75. Knegtel, R.M. et al. Comparison of two implementations of the incremental construction algorithm in flexible docking of thrombin inhibitors. *J Comput Aided Mol Des* **13**, 167-183 (1999).

76. Bohm, H.J., Banner, D.W. & Weber, L. Combinatorial docking and combinatorial chemistry: design of potent non-peptide thrombin inhibitors. *J Comput Aided Mol Des* **13**, 51-56 (1999).

77. Jenwitheesuk, E. & Samudrala, R. Virtual screening of HIV-1 protease inhibitors against human cytomegalovirus protease using docking and molecular dynamics. *Aids* **19**, 529-531 (2005).

78. King, B.L., Vajda, S. & DeLisi, C. Empirical free energy as a target function in docking and design: application to HIV-1 protease inhibitors. *FEBS Lett* **384**, 87-91 (1996).

79. Vieth, M. & Cummins, D.J. DoMCoSAR: a novel approach for establishing the docking mode that is consistent with the structure-activity relationship. Application to HIV-1 protease inhibitors and VEGF receptor tyrosine kinase inhibitors. *J Med Chem* **43**, 3020-3032 (2000).

80. Douali, L., Villemin, D. & Cherqaoui, D. Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives. *Curr Pharm Des* **9**, 1817-1826 (2003).

81. Li, M., Du, L., Wu, B. & Xia, L. Self-organizing molecular field analysis on alpha(1a)-adrenoceptor dihydropyridine antagonists. *Bioorg Med Chem* **11**, 3945-3951 (2003).

82. Mager, P.P. Hybrid canonical-correlation neural-network approach applied to nonnucleoside HIV-1 reverse transcriptase inhibitors (HEPT derivatives). *Curr Med Chem* **10**, 1643-1659 (2003).

83. Robinson, D.D., Winn, P.J., Lyne, P.D. & Richards, W.G. Self-organizing molecular field analysis: a tool for structure-activity studies. *J Med Chem* **42**, 573-583 (1999).

84. Arnold, E. et al. Targeting HIV reverse transcriptase for anti-AIDS drug design: structural and biological considerations for chemotherapeutic strategies. *Drug Des Discov* **13**, 29-47 (1996).

85. UNAIDS Report on the Global AIDS Epidemic. http://www.unaids.org/bangkok2004/report.html. *Joint United Nations Programme on HIV/AIDS* (2004).

86. Imamichi, T. Action of anti-HIV drugs and resistance: reverse transcriptase inhibitors and protease inhibitors. *Curr Pharm Des* **10**, 4039-4053 (2004).

87. Henkel, J. Attacking AIDS with a 'cocktail' therapy? *FDA Consum* **33**, 12-17 (1999).

88. Emini, E.A., Byrnes, V.W., Condra, J.H., Schleif, W.A. & Sardana, V.V. The genetic and functional basis of HIV-1 resistance to nonnucleoside reverse transcriptase inhibitors. *Arch Virol Suppl* **9**, 11-17 (1994).

89. Re, M.C. et al. Drug failure during HIV-1 treatment. New perspectives in monitoring drug resistance. *New Microbiol* **26**, 405-413 (2003).

90. Jacks, T. et al. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331**, 280-283 (1988).

91. Wilson, W. et al. HIV expression strategies: ribosomal frameshifting is directed by a short sequence in both mammalian and yeast systems. *Cell* **55**, 1159-1169 (1988).

92. McMurry, J. Organic chemistry, Edn. 5th. (Brooks/Cole, Pacific Grove, CA; 2000).

93. Coffin, J.M., Hughes, S.H. & Varmus, H.E. Retrovirus, Vol. Chapter 15. (Cold Spring Harbor Laboratory Press, 1997).

94. Kaplan, A.H. Constraints on the sequence diversity of the protease of human immunodeficiency virus type 1: a guide for drug design. *AIDS Res Hum Retroviruses* **12**, 849-853 (1996).

95. Hertogs, K. et al. A rapid method for simultaneous detection of phenotypic resistance to inhibitors of protease and reverse transcriptase in recombinant human immunodeficiency virus type 1 isolates from patients treated with antiretroviral drugs. *Antimicrob Agents Chemother* **42**, 269-276 (1998).

96. Chen, Z. et al. Three-dimensional structure of a mutant HIV-1 protease displaying cross-resistance to all protease inhibitors in clinical trials. *J Biol Chem* **270**, 21433-21436 (1995).

97. Molla, A. et al. Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nat Med* **2**, 760-766 (1996).

98. Kozal, M.J. et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* **2**, 753-759 (1996).

99. Germinario, R.J. & Colby-Germinario, S.P. The effect of human immunodeficiency virus-1 protease inhibitors on the toxicity of a variety of cells. *In Vitro Cell Dev Biol Anim* **39**, 275-279 (2003).

100. Russell, S.J. & Norvig, P. Artificial intelligence: a modern approach. (Prentice Hall, Upper Saddle River, N.J.; [Great Britain]; 2003).

101. Vapnik, V.N. The nature of statistical learning theory. (Springer, New York; 1995).

102. Cabello, D., Barro, S., Salceda, J.M., Ruiz, R. & Mira, J. Fuzzy K-nearest neighbor classifiers for ventricular arrhythmia detection. *Int J Biomed Comput* **27**, 77-93 (1991).

103. Hoffman, B. et al. Quantitative structure-activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J Med Chem* **42**, 3217-3226 (1999).

104. Shen, M. et al. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J Med Chem* **45**, 2811-2823 (2002).

105. Xiao, Z., Varma, S., Xiao, Y.D. & Tropsha, A. Modeling of p38 mitogen-activated protein kinase inhibitors using the Catalyst HypoGen and k-nearest neighbor QSAR methods. *J Mol Graph Model* **23**, 129-138 (2004).

106. Asikainen, A.H., Ruuskanen, J. & Tuppurainen, K.A. Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR QSAR Environ Res* **15**, 19-32 (2004).

107. Shen, H. & Chou, K.C. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. *Biochem Biophys Res Commun* **334**, 288-292 (2005).

108. Zorzet, A., Gustafsson, M. & Hammerling, U. Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol* **2**, 525-534 (2002).

109. Anderberg, M.R. Cluster analysis for applications. (Academic Press, New York; London; 1973).

110. Hartigan, J.A. Clustering algorithms. (Wiley, New York; London (etc.); 1975).

111. Jain, A.K. & Dubes, R.C. Algorithms for clustering data. (Prentice Hall, Englewood Cliffs, N.J.; London; 1988).

112. Jardine, N. & Sibson, R. Mathematical taxonomy. (Wiley, London; New York; 1971).

113. Sneath, P.H.A. & Sokal, R.R. Numerical taxonomy. (W.H. Freeman, San Francisco; 1973).

114. Tryon, R.C. & Bailey, D.E. Cluster analysis. (McGraw-Hill, New York; Maidenhead; 1970).

115. Xu, R. & Wunsch, D., 2nd Survey of clustering algorithms. *IEEE Trans Neural Netw* **16**, 645-678 (2005).

116. Satko, S.G. & Freedman, B.I. The familial clustering of renal disease and related phenotypes. *Med Clin North Am* **89**, 447-456 (2005).

117. Shy, M.E. et al. Phenotypic clustering in MPZ mutations. *Brain* **127**, 371-384 (2004).

118. Griffith, L.C. Receptor clustering: nothing succeeds like success. *Curr Biol* **14**, R413-415 (2004).

119. Sourjik, V. Receptor clustering and signal processing in E. coli chemotaxis. *Trends Microbiol* **12**, 569-576 (2004).

120. Criswell, L.A. Familial clustering of disease features: implications for the etiology and investigation of systemic autoimmune disease. *Arthritis Rheum* **50**, 1707-1708 (2004).

121. Quinlan, J.R. C4.5: programs for machine learning. (Morgan Kaufmann Publishers, San Mateo, Calif.; 1993).

122. Breiman, L. Classification and regression trees. (Wadsworth International Group c1984, Belmont, Calif.; 1984).

123. Kass, G.V. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* **29**, 119-127 (1980).

124. Aspinall, P. & Hill, A.R. Clinical inferences and decisions--II. Decision trees, receiver operator curves and subjective probability. *Ophthalmic Physiol Opt* **4**, 31-38 (1984).

125. Kokol, P., Zorman, M., Stiglic, M.M. & Maleiae, I. The limitations of decision trees and automatic learning in real world medical decision making. *Medinfo* **9 Pt 1**, 529-533 (1998).

126. Sprogar, M. et al. Supporting medical decisions with vector decision trees. *Medinfo* **10**, 552-556 (2001).

127. Kuo, W.J., Chang, R.F., Chen, D.R. & Lee, C.C. Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. *Breast Cancer Res Treat* **66**, 51-57 (2001).

128. Wellman, M.P. et al. Automated critiquing of medical decision trees. *Med Decis Making* **9**, 272-284 (1989).

129. Fineberg, H.V. Decision trees: construction, uses, and limits. *Bull Cancer* **67**, 395-404 (1980).

130. Hazen, G.B. Factored stochastic trees: a tool for solving complex temporal medical decision models. *Med Decis Making* **13**, 227-236 (1993).

131. Hazen, G.B. Stochastic trees and the StoTree modeling environment: models and software for medical decision analysis. *J Med Syst* **26**, 399-413 (2002).

132. Savoie, C.J., Kamikawaji, N., Sasazuki, T. & Kuhara, S. Use of BONSAI decision trees for the identification of potential MHC class I peptide epitope motifs. *Pac Symp Biocomput*, 182-189 (1999).

133. Decaestecker, C. et al. Methodological aspects of using decision trees to characterise leiomyomatous tumors. *Cytometry* **24**, 83-92 (1996).

134. Salzberg, S., Chen, X., Henderson, J. & Fasman, K. Finding genes in DNA using decision trees and dynamic programming. *Proc Int Conf Intell Syst Mol Biol* **4**, 201-210 (1996).

135. Rumelhart, D.E., Hinton, D.E. & Williams, R.J. in parallel distributed processing:Explorations in macrostructure of cognition, Vol. I (Cambridge, MA.; 1986).

136. LeCun, Y. in Disordered systems and biological organiozations 233-240 (Springer, Les Houches; 1986).

137. Wang, D. & Larder, B. Enhanced prediction of lopinavir resistance from genotype by use of

artificial neural networks. *J Infect Dis* **188**, 653-660 (2003).

138. Draghici, S. & Potter, R.B. Predicting HIV drug resistance with neural networks. *Bioinformatics* **19**, 98-107 (2003).

139. Specht, D., F. Probabilistic Neural Networks. *Neural Networks* **3**, 109-118 (1990).

140. Traeger, M. et al. [Artificial neural networks. Theory and applications in anesthesia, intensive care and emergency medicine]. *Anaesthesist* **52**, 1055-1061 (2003).

141. Winkler, D.A. Neural networks as robust tools in drug lead discovery and development. *Mol Biotechnol* **27**, 139-168 (2004).

142. Cheng, J. & Baldi, P. Three-stage prediction of protein {beta}-sheets by neural networks, alignments and graph algorithms. *Bioinformatics* **21 Suppl 1**, i75-i84 (2005).

143. Lin, K., Simossis, V.A., Taylor, W.R. & Heringa, J. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* **21**, 152-159 (2005).

144. Riis, S.K. & Krogh, A. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J Comput Biol* **3**, 163-183 (1996).

145. Barlow, T.W. Feed-forward neural networks for secondary structure prediction. *J Mol Graph* **13**, 175-183 (1995).

146. Vanhala, J. & Kaski, K. Protein structure prediction system based on artificial neural networks. *Proc Int Conf Intell Syst Mol Biol* **1**, 402-410 (1993).

147. Holley, L.H. & Karplus, M. Neural networks for protein structure prediction. *Methods Enzymol* **202**, 204-224 (1991).

148. Yang, Z.R. Prediction of caspase cleavage sites using Bayesian bio-basis function neural networks. *Bioinformatics* **21**, 1831-1837 (2005).

149. Yang, Z.R. & Berry, E.A. Reduced bio-basis function neural networks for protease cleavage site prediction. *J Bioinform Comput Biol* **2**, 511-531 (2004).

150. Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. Prediction of protein--protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem* **269**, 1356-1361 (2002).

151. Vapnik, V. Estimation of Dependences Based on Empirical Data [in Russian]. [English translation: Springer Verlag, New York, 1982]. (1979).

152. Vapnik, V. The Nature of Statistical Learning Theory. (Springer, New York; 1995).

153. Burges, C. A tutorial on Support Vector Machine for pattern recognition. *Data Min. Knowl. Disc.* **2**, 121-167 (1998).

154. Kim, K.I., Jung, K., Park, S.H. & Kim, H.J. Support vector machine-based text detection in digital video. *Pattern Recognition* **34**, 527-529 (2001).

155. Drucker, H., Wu, D.H. & Vapnik, V.N. Support Vector Machines for Spam Categorization. *IEEE Transactions on Neural Networks* **10**, 1048-1054 (1999).

156. de Vel, O., Anderson, A., Corney, M. & Mohay, G. Mining e-mail content for author identification forensics. *Sigmod Record* **30**, 55-64 (2001).

157. Thubthong, N. & Kijsirikul, B. Support vector machines for Thai phoneme recognition. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems* **9**, 803-813 (2001).

158. Ben-Yacoub, S., Abdeljaoued, Y. & Mayoraz, E. Fusion Face and Speech Data for Person Identity Verification. *IEEE Transactions on Neural Networks* **10**, 1065-1074 (1999).

159. Karlsen, R.E., Gorsich, D.J. & Gerhart, G.R. Target classification via support vector machines. *Optical Engineering* **39**, 704-711 (2000).

160. Papageorgiou, C. & Poggio, T. A trainable system for object detection. *International Journal of Computer Vision* **38**, 15-33 (2000).

161. Huang, C., Davis, L.S. & Townshend, J.R.G. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing* **23**, 725-749 (2002).

162. Liong, S.Y. & Sivapragasam, C. Flood stage forecasting with support vector machines. *Journal of the American Water Resources Association* **38**, 173-186 (2002).

163. Rasmussen, M. & Bjorck, L. Unique regulation of SclB - a novel collagen-like surface protein of Streptococcus pyogenes. *Mol Microbiol* **40**, 1427-1438 (2001).

164. Furey, T.S. et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906-914 (2000).

165. Fritsche, H.A. Tumor Markers and Pattern Recognition Analysis: A New Diagnostic Tool for Cancer. *J. Clin. Ligand Assay* **25**, 11-15 (2002).

166. Brown, M.P. et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* **97**, 262-267 (2000).

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

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

169. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res* **32 Database issue**, D138-141 (2004).

170. Roulston, J.E. Screening with tumor markers: critical issues. *Mol Biotechnol* **20**, 153-162 (2002).

171. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412-424 (2000).

172. Rost, B. & Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232**, 584-599 (1993).

173. Xue, Y. et al. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J Chem Inf Comput Sci* **44**, 1630-1638 (2004).

174. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* **32 Database issue**, D35-40 (2004).

175. Alain, R. Variable Selection Using SVM-based Criteria. *Journal of Machine Learning Research* **3**, 1357-1370 (2003).

176. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46**, 389-422 (2002).

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

178. Enzyme-Nomenclature Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB). (Academic Press, New York; 1992).

179. Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* **30**, 47-49 (2002).

180. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-370 (2003).

181. Cai, Y.D., Wang, W.L. & Chen, Y.Z. Support Vector Machine Classification of Physical and Biological Datasets. *Inter.J.Mod.Phys.C* **14**, 575 - 585 (2003).

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

183. Burges, C.J.C. A tutorial on Support Vector Machine for pattern recognition. *Data Min Knowl Disc* **2**, 121-167 (1998).

184. Downward, J. The ins and outs of signalling. *Nature* **411**, 759-762 (2001).

185. Lengeler, J.W. Metabolic networks: a signal-oriented approach to cellular models. *Biol Chem* **381**, 911-920 (2000).

186. Siomi, H. & Dreyfuss, G. RNA-binding proteins as regulators of gene expression. *Curr Opin Genet Dev* **7**, 345-353 (1997).

187. Draper, D.E. & Reynaldo, L.P. RNA binding strategies of ribosomal proteins. *Nucleic Acids Res* **27**, 381-388 (1999).

188. Fetrow, J.S. & Skolnick, J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol* **281**, 949-968 (1998).

189. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**, 324-328 (1998).

190. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-2901 (1999).

191. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86 (1999).

192. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-4288 (1999).

193. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868 (1998).

194. Frank, D.N. & Pace, N.R. Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu Rev Biochem* **67**, 153-180 (1998).

195. Singh, R. RNA-protein interactions that regulate pre-mRNA splicing. *Gene Expr* **10**, 79-92 (2002).

196. Tomasevic, N. & Peculis, B. Identification of a U8 snoRNA-specific binding protein. *J Biol Chem* **274**, 35914-35920 (1999).

197. Bhasin, M. & Raghava, G.P. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res* **32**, W383-389 (2004).

198. Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H. & Suwa, M. GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. *Nucleic Acids Res* **33**, W148-153 (2005).

199. Bhardwaj, N., Langlois, R.E., Zhao, G. & Lu, H. Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res* **33**, 6486-6493 (2005).

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

201. Wang, M., Yang, J., Liu, G.P., Xu, Z.J. & Chou, K.C. Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. *Protein Eng Des Sel* **17**, 509-516 (2004).

202. Veropoulos, K., Campbell, C. and Cristianini, N. in Proceedings of the International Joint Conference on Artificial Intelligence (UCAI99). (ed. T. Dean) 55-60 (Morgan Kaufmann, Sweden; 1999).

203. Mattaj, I.W. RNA recognition: a family matter? *Cell* **73**, 837-840 (1993).

204. Perez-Canadillas, J.M. & Varani, G. Recent advances in RNA-protein recognition. *Curr Opin Struct Biol* **11**, 53-58 (2001).

205. Bycroft, M., Hubbard, T.J., Proctor, M., Freund, S.M. & Murzin, A.G. The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. *Cell* **88**, 235-242 (1997).

206. Zhang, K. & Rathod, P.K. Divergent regulation of dihydrofolate reductase between malaria parasite and human host. *Science* **296**, 545-547 (2002).

207. Hermann, T. & Westhof, E. Simulations of the dynamics at an RNA-protein interface. *Nat Struct Biol* **6**, 540-544 (1999).

208. Hediger, M.A. Structure, function and evolution of solute transporters in prokaryotes and eukaryotes. *J Exp Biol* **196**, 15-49 (1994).

209. Borst, P. & Elferink, R.O. Mammalian ABC transporters in health and disease. *Annu Rev Biochem* **71**, 537-592 (2002).

210. Seal, R.P. & Amara, S.G. Excitatory amino acid transporters: a family in flux. *Annu Rev Pharmacol Toxicol* **39**, 431-456 (1999).

211. Joet, T. et al. Why is the Plasmodium falciparum hexose transporter a promising new drug target? *Expert Opin Ther Targets* **7**, 593-602 (2003).

212. Birch, P.J., Dekker, L.V., James, I.F., Southan, A. & Cronk, D. Strategies to identify ion channel modulators: current and novel approaches to target neuropathic pain. *Drug Discov Today* **9**, 410-418 (2004).

213. Dutta, A.K., Zhang, S., Kolhatkar, R. & Reith, M.E. Dopamine transporter as target for drug development of cocaine dependence medications. *Eur J Pharmacol* **479**, 93-106 (2003).

214. Lee, W. & Kim, R.B. Transporters and renal drug elimination. *Annu Rev Pharmacol Toxicol* **44**, 137-166 (2004).

215. Kunta, J.R. & Sinko, P.J. Intestinal drug transporters: in vivo function and clinical importance. *Curr Drug Metab* **5**, 109-124 (2004).

216. Driessen, A.J., Rosen, B.P. & Konings, W.N. Diversity of transport mechanisms: common structural principles. *Trends Biochem Sci* **25**, 397-401 (2000).

217. Saier, M.H., Jr. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev* **64**, 354-411 (2000).

218. Zhou, X., Hvorup, R.N. & Saier, M.H., Jr. An automated program to screen databases for members of protein families. *J Mol Microbiol Biotechnol* **5**, 7-10 (2003).

219. Campbell, R.S. et al. Development and validation of a robust specific enzyme mediated assay for phenylalanine in serum. *Clin Chim Acta* **210**, 197-210 (1992).

220. Howard, E.M., Zhang, H. & Roepe, P.D. A novel transporter, Pfcrt, confers antimalarial drug resistance. *J Membr Biol* **190**, 1-8 (2002).

221. Sano, Y. et al. A novel two-pore domain K+ channel, TRESK, is localized in the spinal cord. *J Biol Chem* **278**, 27406-27412 (2003).

222. Zhang, Y., Jock, S. & Geider, K. Genes of Erwinia amylovora involved in yellow color formation and release of a low-molecular-weight compound during growth in the presence of copper ions. *Mol Gen Genet* **264**, 233-240 (2000).

223. Ohki, R. & Murata, M. bmr3, a third multidrug transporter gene of Bacillus subtilis. *J Bacteriol* **179**, 1423-1427 (1997).

224. Reyes, R. et al. Cloning and expression of a novel pH-sensitive two pore domain K+ channel

from human kidney. *J Biol Chem* **273**, 30863-30869 (1998).

225. Vardy, E., Arkin, I.T., Gottschalk, K.E., Kaback, H.R. & Schuldiner, S. Structural conservation in the major facilitator superfamily as revealed by comparative modeling. *Protein Sci* **13**, 1832-1840 (2004).

226. Whisstock, J.C. & Lesk, A.M. Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **36**, 307-340 (2003).

227. Han, L.Y., Cai, C.Z., Ji, Z.L. & Chen, Y.Z. Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity. *Virology* **331**, 136-143 (2005).

228. Han, L.Y. et al. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res* **32**, 6437-6444 (2004).

229. Ren, Q., Kang, K.H. & Paulsen, I.T. TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res* **32 Database issue**, D284-288 (2004).

230. Busch, W. & Saier, M.H., Jr. The transporter classification (TC) system, 2002. *Crit Rev Biochem Mol Biol* **37**, 287-337 (2002).

231. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45-48 (2000).

232. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res* **30**, 276-280 (2002).

233. Kim, H. & Park, H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins* **54**, 557-562 (2004).

234. Arabidopsis_Genome_Initiative Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).

235. Cho, Y. & Walbot, V. Computational methods for gene annotation: the Arabidopsis genome. *Curr Opin Biotechnol* **12**, 126-130 (2001).

236. Kalinowski, J. et al. The complete Corynebacterium glutamicum ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J Biotechnol* **104**, 5-25 (2003).

237. Henne, A. et al. The genome sequence of the extreme thermophile Thermus thermophilus. *Nat Biotechnol* **22**, 547-553 (2004).

238. Makeyev, E.V. & Bamford, D.H. Evolutionary potential of an RNA virus. *J Virol* **78**, 2114-2120 (2004).

239. Sabanadzovic, S., Ghanem-Sabanadzovic, N.A., Saldarelli, P. & Martelli, G.P. Complete nucleotide sequence and genome organization of Grapevine fleck virus. *J Gen Virol* **82**, 2009-2015 (2001).

240. Rustici, G., Milne, R.G. & Accotto, G.P. Nucleotide sequence, genome organisation and phylogenetic analysis of Indian citrus ringspot virus. Brief report. *Arch Virol* **147**, 2215-2224 (2002).

241. He, R. et al. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem Biophys Res Commun* **316**, 476-483 (2004).

242. Marks, D.B., Marks, A.D. & Smith, C.M. Basic medical biochemistry: a clinical approach. (Williams & Wilkins, Baltimore; 1996).

243. Rost, B. Enzyme function less conserved than anticipated. *J Mol Biol* **318**, 595-608 (2002).

244. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31**, 28-33 (2003).

245. Herniou, E.A., Olszewski, J.A., Cory, J.S. & O'Reilly, D.R. The genome sequence and evolution of baculoviruses. *Annu Rev Entomol* **48**, 211-234 (2003).

246. Miller, E.S. et al. Bacteriophage T4 genome. *Microbiol Mol Biol Rev* **67**, 86-156, table of contents (2003).

247. Marra, M.A. et al. The Genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399-1404 (2003).

248. Miller, E.S. et al. Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J Bacteriol* **185**, 5220-5233 (2003).

249. Hughes, J.A., Brown, L.R. & Ferro, A.J. Nucleotide sequence and analysis of the coliphage T3 S-adenosylmethionine hydrolase gene and its surrounding ribonuclease III processing sites. *Nucleic Acids Res* **15**, 717-729 (1987).

250. Tomaschewski, J., Gram, H., Crabb, J.W. & Ruger, W. T4-induced alpha- and beta-glucosyltransferase: cloning of the genes and a comparison of their products based on sequencing data. *Nucleic Acids Res* **13**, 7551-7568 (1985).

251. Yanez, R.J. et al. Analysis of the complete nucleotide sequence of African swine fever virus. *Virology* **208**, 249-278 (1995).

252. Broida, J. & Abelson, J. Sequence organization and control of transcription in the bacteriophage T4 tRNA region. *J Mol Biol* **185**, 545-563 (1985).

253. Sjoberg, B.M. et al. The bacteriophage T4 gene for the small subunit of ribonucleotide reductase contains an intron. *Embo J* **5**, 2031-2036 (1986).

254. Valerie, K., Henderson, E.E. & deRiel, J.K. Identification, physical map location and sequence of the denV gene from bacteriophage T4. *Nucleic Acids Res* **12**, 8085-8096 (1984).

255. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. & Petersen, G.B. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol* **162**, 729-773 (1982).

256. Alaoui-Ismaili, M.H. & Richardson, C.D. Identification and characterization of a filament-associated protein encoded by Amsacta moorei entomopoxvirus. *J Virol* **70**, 2697-2705 (1996).

257. Valerie, K., Stevens, J., Lynch, M., Henderson, E.E. & de Riel, J.K. Nucleotide sequence and analysis of the 58.3 to 65.5-kb early region of bacteriophage T4. *Nucleic Acids Res* **14**, 8637-8654 (1986).

258. Chu, F.K., Maley, G.F., West, D.K., Belfort, M. & Maley, F. Characterization of the intron in the phage T4 thymidylate synthase gene and evidence for its self-excision from the primary transcript. *Cell* **45**, 157-166 (1986).

259. Tomaschewski, J. & Ruger, W. Nucleotide sequence and primary structures of gene products coded for by the T4 genome between map positions 48.266 kb and 39.166 kb. *Nucleic Acids Res* **15**, 3632-3633 (1987).

260. Gerber, J.S. & Hinton, D.M. An N-terminal mutation in the bacteriophage T4 motA gene yields a protein that binds DNA but is defective for activation of transcription. *J Bacteriol* **178**, 6133-6139 (1996).

261. Esposito, D. et al. The complete nucleotide sequence of bacteriophage HP1 DNA. *Nucleic Acids Res* **24**, 2360-2368 (1996).

262. Hardy, M.E., Gorziglia, M. & Woode, G.N. Amino acid sequence analysis of bovine rotavirus B223 reveals a unique outer capsid protein VP4 and confirms a third bovine VP4 type. *Virology* **191**, 291-300 (1992).

263. Luttichau, H.R. et al. A highly selective CCR2 chemokine agonist encoded by human herpesvirus 6. *J Biol Chem* **278**, 10928-10933 (2003).

264. Dunn, J.J. & Studier, F.W. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J Mol Biol* **166**, 477-535 (1983).

265. Fries, K.L. et al. Identification of a novel protein encoded by the BamHI A region of the Epstein-Barr virus. *J Virol* **71**, 2765-2771 (1997).

266. Zhang, Y., Nelson, M., Nietfeldt, J.W., Burbank, D.E. & Van Etten, J.L. Characterization of Chlorella virus PBCV-1 CviAII restriction and modification system. *Nucleic Acids Res* **20**, 5351-5356 (1992).

267. Skowron, P.M. et al. Cloning and applications of the two/three-base restriction endonuclease R.CviJI from IL-3A virus-infected Chlorella. *Gene* **157**, 37-41 (1995).

268. WF, I.J., Roode, E.C., Goldbach, R.W., Vlak, J.M. & Zuidema, D. Characterization of Spodoptera exigua multicapsid nucleopolyhedrovirus ORF17/18, a homologue of Xestia c-nigrum granulovirus ORF129. *J Gen Virol* **83**, 2857-2867 (2002).

269. Yin, C. et al. Identification of a novel protein associated with envelope of occlusion-derived virus in Spodoptera litura multicapsid nucleopolyhedrovirus. *Virus Genes* **26**, 5-13 (2003).

270. Spaderna, S., Blessing, H., Bogner, E., Britt, W. & Mach, M. Identification of glycoprotein gpTRL10 as a structural component of human cytomegalovirus. *J Virol* **76**, 1450-1460 (2002).

271. Pande, S. et al. The bacteriophage T4 transcription activator MotA interacts with the far-C-terminal region of the sigma70 subunit of Escherichia coli RNA polymerase. *J Bacteriol* **184**, 3957-3964 (2002).

272. Todd, A.E., Orengo, C.A. & Thornton, J.M. Plasticity of enzyme active sites. *Trends Biochem Sci* **27**, 419-426 (2002).

273. Faye, L., Boulaflous, A., Benchabane, M., Gomord, V. & Michaud, D. Protein modifications in the plant secretory pathway: current status and practical implications in molecular pharming. *Vaccine* **23**, 1770-1778 (2005).

274. Fujiwara, Y. & Asogawa, M. Protein function prediction using hidden Markov models and neural networks. *NEC Res Dev* **43**, 238-241 (2002).

275. Lunn, J.E. Sucrose-phosphatase gene families in plants. *Gene* **303**, 187-196 (2003).

276. Yamada, K. et al. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **302**, 842-846 (2003).

277. Yamagata, H., Yonesu, K., Hirata, A. & Aizono, Y. TGTCACA motif is a novel cis-regulatory

enhancer element involved in fruit-specific expression of the cucumisin gene. *J Biol Chem* **277**, 11582-11590 (2002).

278. Ziegler, J. et al. Molecular cloning of allene oxide cyclase. The enzyme establishing the stereochemistry of octadecanoids and jasmonates. *J Biol Chem* **275**, 19132-19138 (2000).

279. Schiebel, W. et al. Isolation of an RNA-directed RNA polymerase-specific cDNA clone from tomato. *Plant Cell* **10**, 2087-2101 (1998).

280. Peterbauer, T. et al. Stachyose synthesis in seeds of adzuki bean (Vigna angularis): molecular cloning and functional expression of stachyose synthase. *Plant J* **20**, 509-518 (1999).

281. Jansch, L., Kruft, V., Schmitz, U.K. & Braun, H.P. Cytochrome c reductase from potato does not comprise three core proteins but contains an additional low-molecular-mass subunit. *Eur J Biochem* **228**, 878-885 (1995).

282. Knoetzel, J., Mant, A., Haldrup, A., Jensen, P.E. & Scheller, H.V. PSI-O, a new 10-kDa subunit of eukaryotic photosystem I. *FEBS Lett* **510**, 145-148 (2002).

283. Ranford, J.C., Bryce, J.H. & Morris, P.C. PM19, a barley (Hordeum vulgare L.) gene encoding a putative plasma membrane protein, is expressed during embryo development and dormancy. *J Exp Bot* **53**, 147-148 (2002).

284. Yang, A.S. Structure-dependent sequence alignment for remotely related proteins. *Bioinformatics* **18**, 1658-1665 (2002).

285. Ge, Y.X. et al. NEC1, a novel gene, highly expressed in nectary tissue of Petunia hybrida. *Plant J* **24**, 725-734 (2000).

286. Kuno, N. et al. The novel MYB protein EARLY-PHYTOCHROME-RESPONSIVE1 is a component of a slave circadian oscillator in Arabidopsis. *Plant Cell* **15**, 2476-2488 (2003).

287. Rose, A., Gindullis, F. & Meier, I. A novel alpha-helical protein, specific to and highly conserved in plants, is associated with the nuclear matrix fraction. *J Exp Bot* **54**, 1133-1141 (2003).

288. Asiegbu, F.O., Choi, W., Li, G., Nahalkova, J. & Dean, R.A. Isolation of a novel antimicrobial peptide gene (Sp-AMP) homologue from Pinus sylvestris (Scots pine) following infection with the root rot fungus Heterobasidion annosum. *FEMS Microbiol Lett* **228**, 27-31 (2003).

289. Popham, D.L. & Young, K.D. Role of penicillin-binding proteins in bacterial cell morphogenesis. *Curr Opin Microbiol* **6**, 594-599 (2003).

290. Nagano, Y., Furuhashi, H., Inaba, T. & Sasaki, Y. A novel class of plant-specific zinc-dependent DNA-binding protein that binds to A/T-rich DNA sequences. *Nucleic Acids Res* **29**, 4097-4105 (2001).

291. Tomita, A., Buchholz, D.R. & Shi, Y.B. Recruitment of N-CoR/SMRT-TBLR1 corepressor complex by unliganded thyroid hormone receptor for gene repression during frog development. *Mol Cell Biol* **24**, 3337-3346 (2004).

292. Yap, M.N., Lee, R.H., Huang, Y.J., Liao, C.J. & Chen, S.C. Molecular characterization of a novel senescence-associated gene SPA15 induced during leaf senescence in sweet potato. *Plant Mol Biol* **51**, 471-481 (2003).

293. Kim, J.H., Choi, D. & Kende, H. The AtGRF family of putative transcription factors is involved in leaf and cotyledon growth in Arabidopsis. *Plant J* **36**, 94-104 (2003).

294. Gao, M.J. et al. A novel protein from Brassica napus has a putative KID domain and responds to low temperature. *Plant J* **33**, 1073-1086 (2003).

295. Ulker, B. & Somssich, I.E. WRKY transcription factors: from DNA binding towards biological function. *Curr Opin Plant Biol* **7**, 491-498 (2004).

296. Krupinska, K. et al. A novel nucleus-targeted protein is expressed in barley leaves during senescence and pathogen infection. *Plant Physiol* **130**, 1172-1180 (2002).

297. Jang, C.S., Lee, M.S., Kim, J.Y., Kim, D.S. & Seo, Y.W. Molecular characterization of a cDNA encoding putative calcium binding protein, HvCaBP1, induced during kernel development in barley (Hordeum vulgare L.). *Plant Cell Rep* **22**, 64-70 (2003).

298. Tada, Y. & Yamaguchi, J. A function of 40 kDa outer membrane protein in Serratia marcescens. *Microbiol Immunol* **38**, 171-175 (1994).

299. Trentmann, S.M. ERN1, a novel ethylene-regulated nuclear protein of Arabidopsis. *Plant Mol Biol* **44**, 11-25 (2000).

300. Cvitanich, C. et al. CPP1, a DNA-binding protein involved in the expression of a soybean leghemoglobin c3 gene. *Proc Natl Acad Sci U S A* **97**, 8163-8168 (2000).

301. Trevaskis, B., Wandrey, M., Colebatch, G. & Udvardi, M.K. The soybean GmN6L gene encodes a late nodulin expressed in the infected zone of nitrogen-fixing nodules. *Mol Plant Microbe Interact* **15**, 630-636 (2002).

302. Skadsen, R.W. et al. Cloning of the promoter for a novel barley gene, Lem1, and its

organ-specific promotion of Gfp expression in lemma and palea. *Plant Mol Biol* **49**, 545-555 (2002).

303. Slocombe, S.P. et al. Identification of SnIP1, a novel protein that interacts with SNF1-related protein kinase (SnRK1). *Plant Mol Biol* **49**, 31-44 (2002).

304. Morohashi, K., Minami, M., Takase, H., Hotta, Y. & Hiratsuka, K. Isolation and characterization of a novel GRAS gene that regulates meiosis-associated gene expression. *J Biol Chem* **278**, 20865-20873 (2003).

305. Filichkin, S.A., Leonard, J.M., Monteros, A., Liu, P.P. & Nonogaki, H. A novel endo-beta-mannanase gene in tomato LeMAN5 is associated with anther and pollen development. *Plant Physiol* **134**, 1080-1087 (2004).

306. Yang, G., Matsuoka, M., Iwasaki, Y. & Komatsu, S. A novel brassinolide-enhanced gene identified by cDNA microarray is involved in the growth of rice. *Plant Mol Biol* **52**, 843-854 (2003).

307. Lu, C.A., Ho, T.H., Ho, S.L. & Yu, S.M. Three novel MYB proteins with one DNA binding repeat mediate sugar and hormone regulation of alpha-amylase gene expression. *Plant Cell* **14**, 1963-1980 (2002).

308. Genschik, P., Hall, J. & Filipowicz, W. Cloning and characterization of the Arabidopsis cyclic phosphodiesterase which hydrolyzes ADP-ribose 1",2"-cyclic phosphate and nucleoside 2',3'-cyclic phosphates. *J Biol Chem* **272**, 13211-13219 (1997).

309. Gruss, A. & Michel, B. The replication-recombination connection: insights from genomics. *Curr Opin Microbiol* **4**, 595-601 (2001).

310. Ge, Z. & Taylor, D.E. Contributions of genome sequencing to understanding the biology of Helicobacter pylori. *Annu Rev Microbiol* **53**, 353-387 (1999).

311. Donadio, S. et al. Targets and assays for discovering novel antibacterial agents. *J Biotechnol* **99**, 175-185 (2002).

312. Isaacson, R.E. Genomics and the prospects for the discovery of new targets for antibacterial and antifungal agents. *Curr Pharm Des* **8**, 1091-1098 (2002).

313. Buysse, J.M. The role of genomics in antibacterial target discovery. *Curr Med Chem* **8**, 1713-1726 (2001).

314. Meinke, A., Henics, T. & Nagy, E. Bacterial genomes pave the way to novel vaccines. *Curr Opin Microbiol* **7**, 314-320 (2004).

315. Lund, T. & Granum, P.E. The 105-kDa protein component of Bacillus cereus non-haemolytic enterotoxin (Nhe) is a metalloprotease with gelatinolytic and collagenolytic activity. *FEMS Microbiol Lett* **178**, 355-361 (1999).

316. Bunny, K.L., Hall, R.M. & Stokes, H.W. New mobile gene cassettes containing an aminoglycoside resistance gene, aacA7, and a chloramphenicol resistance gene, catB3, in an integron in pBWH301. *Antimicrob Agents Chemother* **39**, 686-693 (1995).

317. Dargatz, H., Diefenthal, T., Witte, V., Reipen, G. & von Wettstein, D. The heterodimeric protease clostripain from Clostridium histolyticum is encoded by a single gene. *Mol Gen Genet* **240**, 140-145 (1993).

318. Miyamoto, K. & Ohta, H. Cloning and heterologous expression of a novel arylmalonate decarboxylase gene from Alcaligenes bronchisepticus KU 1201. *Appl Microbiol Biotechnol* **38**, 234-238 (1992).

319. Butler, M.J. et al. Intracellular aminopeptidases in Streptomyces lividans 66. *J Ind Microbiol* **13**, 24-29 (1994).

320. He, Z. & Spain, J.C. A novel 2-aminomuconate deaminase in the nitrobenzene degradation pathway of Pseudomonas pseudoalcaligenes JS45. *J Bacteriol* **180**, 2502-2506 (1998).

321. Chain, P. et al. Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph Nitrosomonas europaea. *J Bacteriol* **185**, 2759-2773 (2003).

322. McTavish, H., Fuchs, J.A. & Hooper, A.B. Sequence of the gene coding for ammonia monooxygenase in Nitrosomonas europaea. *J Bacteriol* **175**, 2436-2444 (1993).

323. Wei, Y. et al. A novel variant of the catalytic triad in the Streptomyces scabies esterase. *Nat Struct Biol* **2**, 218-223 (1995).

324. Gray, J.V., Golinelli-Pimpaneau, B. & Knowles, J.R. Monofunctional chorismate mutase from Bacillus subtilis: purification of the protein, molecular cloning of the gene, and overexpression of the gene product in Escherichia coli. *Biochemistry* **29**, 376-383 (1990).

325. Witt, H. & Ludwig, B. Isolation, analysis, and deletion of the gene coding for subunit IV of cytochrome c oxidase in Paracoccus denitrificans. *J Biol Chem* **272**, 5514-5517 (1997).

326. Studwell-Vaughan, P.S. & O'Donnell, M. DNA polymerase III accessory proteins. V. Theta encoded by holE. *J Biol Chem* **268**, 11785-11791 (1993).

327. Burland, V., Plunkett, G., 3rd, Daniels, D.L. & Blattner, F.R. DNA sequence and analysis of 136 kilobases of the Escherichia coli genome: organizational symmetry around the origin of replication. *Genomics* **16**, 551-561 (1993).

328. Anguita, J., Rodriguez Aparicio, L.B. & Naharro, G. Purification, gene cloning, amino acid sequence analysis, and expression of an extracellular lipase from an Aeromonas hydrophila human isolate. *Appl Environ Microbiol* **59**, 2411-2417 (1993).

329. Moses, E.K., Rood, J.I., Yong, W.K. & Riffkin, G.G. Molecular analysis of one of multiple protease-encoding genes from the prototype virulent strain of Bacteroides nodosus. *Gene* **77**, 219-228 (1989).

330. Kachlany, S.C. et al. flp-1, the first representative of a new pilin gene subfamily, is required for non-specific adherence of Actinobacillus actinomycetemcomitans. *Mol Microbiol* **40**, 542-554 (2001).

331. Kleerebezem, M. et al. Complete genome sequence of Lactobacillus plantarum WCFS1. *Proc Natl Acad Sci U S A* **100**, 1990-1995 (2003).

332. Xia, T., Song, J., Zhao, G., Aldrich, H. & Jensen, R.A. The aroQ-encoded monofunctional chorismate mutase (CM-F) protein is a periplasmic enzyme in Erwinia herbicola. *J Bacteriol* **175**, 4729-4737 (1993).

333. Tarentino, A.L. et al. Molecular cloning and amino acid sequence of peptide-N4-(N-acetyl-beta-D-glucosaminyl)asparagine amidase from flavobacterium meningosepticum. *J Biol Chem* **265**, 6961-6966 (1990).

334. Crouzet, J. et al. Genetic and sequence analysis of an 8.7-kilobase Pseudomonas denitrificans fragment carrying eight genes involved in transformation of precorrin-2 to cobyrinic acid. *J Bacteriol* **172**, 5980-5990 (1990).

335. Cole, S.T. et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537-544 (1998).

336. Katayama, Y. et al. Cloning of genes coding for the three subunits of thiocyanate hydrolase of Thiobacillus thioparus THI 115 and their evolutionary relationships to nitrile hydratase. *J Bacteriol* **180**, 2583-2589 (1998).

337. Campobasso, N., Costello, C.A., Kinsland, C., Begley, T.P. & Ealick, S.E. Crystal structure of thiaminase-I from Bacillus thiaminolyticus at 2.0 A resolution. *Biochemistry* **37**, 15981-15989 (1998).

338. Xu, S.Y. et al. Cloning and expression of the ApaLI, NspI, NspHI, SacI, ScaI, and SapI restriction-modification systems in Escherichia coli. *Mol Gen Genet* **260**, 226-231 (1998).

339. Falk, G. & Walker, J.E. DNA sequence of a gene cluster coding for subunits of the F0 membrane sector of ATP synthase in Rhodospirillum rubrum. Support for modular evolution of the F1 and F0 sectors. *Biochem J* **254**, 109-122 (1988).

340. Lindquist, S., Galleni, M., Lindberg, F. & Normark, S. Signalling proteins in enterobacterial AmpC beta-lactamase regulation. *Mol Microbiol* **3**, 1091-1102 (1989).

341. Ainsa, J.A. et al. Aminoglycoside 2'-N-acetyltransferase genes are universally present in mycobacteria: characterization of the aac(2')-Ic gene from Mycobacterium tuberculosis and the aac(2')-Id gene from Mycobacterium smegmatis. *Mol Microbiol* **24**, 431-441 (1997).

342. Fitzmaurice, W.P., Saari, L.L., Lowery, R.G., Ludden, P.W. & Roberts, G.P. Genes coding for the reversible ADP-ribosylation system of dinitrogenase reductase from Rhodospirillum rubrum. *Mol Gen Genet* **218**, 340-347 (1989).

343. Arand, M. et al. Structure of Rhodococcus erythropolis limonene-1,2-epoxide hydrolase reveals a novel active site. *Embo J* **22**, 2583-2592 (2003).

344. Pinyakong, O., Habe, H., Yoshida, T., Nojiri, H. & Omori, T. Identification of three novel salicylate 1-hydroxylases involved in the phenanthrene degradation of Sphingobium sp. strain P2. *Biochem Biophys Res Commun* **301**, 350-357 (2003).

345. Baron, A.J. et al. Alginate lyase from Klebsiella pneumoniae, subsp. aerogenes: gene cloning, sequence analysis and high-level production in Escherichia coli. *Gene* **143**, 61-66 (1994).

346. Matsuo, N., Kaneko, S., Kuno, A., Kobayashi, H. & Kusakabe, I. Purification, characterization and gene cloning of two alpha-L-arabinofuranosidases from streptomyces chartreusis GS901. *Biochem J* **346 Pt 1**, 9-15 (2000).

347. Byun, T. et al. Aminopeptidase from Sphingomonas capsulata. *J Biol Chem* **276**, 17902-17907 (2001).

348. White, O. et al. Genome sequence of the radioresistant bacterium Deinococcus radiodurans R1. *Science* **286**, 1571-1577 (1999).

349. Sugano, Y., Matsumoto, T. & Noma, M. Sequence analysis of the agaB gene encoding a new beta-agarase from Vibrio sp. strain JT0107. *Biochim Biophys Acta* **1218**, 105-108 (1994).

350. Hashimoto, M., Fukui, M., Hayano, K. & Hayatsu, M. Nucleotide sequence and genetic structure of a novel carbaryl hydrolase gene (cehA) from Rhizobium sp. strain AC100. *Appl Environ Microbiol* **68**, 1220-1227 (2002).

351. Overbeeke, N., Bergmans, H., van Mansfeld, F. & Lugtenberg, B. Complete nucleotide sequence of phoE, the structural gene for the phosphate limitation inducible outer membrane pore protein of Escherichia coli K12. *J Mol Biol* **163**, 513-532 (1983).

352. Buell, C.R. et al. The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. *Proc Natl Acad Sci U S A* **100**, 10181-10186 (2003).

353. Xu, X., Holt, S.C. & Kolodrubetz, D. Cloning and expression of two novel hemin binding protein genes from Treponema denticola. *Infect Immun* **69**, 4465-4472 (2001).

354. Fraser, C.M. et al. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. *Nature* **390**, 580-586 (1997).

355. Slakeski, N. et al. Characterization and expression of a novel Porphyromonas gingivalis outer membrane protein, Omp28. *Oral Microbiol Immunol* **17**, 150-156 (2002).

356. Zhu, P. et al. Identification of opcA gene in Neisseria polysaccharea: interspecies diversity of Opc protein family. *Gene* **307**, 31-40 (2003).

357. Nordlund, I., Powlowski, J. & Shingler, V. Complete nucleotide sequence and polypeptide analysis of multicomponent phenol hydroxylase from Pseudomonas sp. strain CF600. *J Bacteriol* **172**, 6826-6833 (1990).

358. Kato, Y., Nakamura, K., Sakiyama, H., Mayhew, S.G. & Asano, Y. Novel heme-containing lyase, phenylacetaldoxime dehydratase from Bacillus sp. strain OxB-1: purification, characterization, and molecular cloning of the gene. *Biochemistry* **39**, 800-809 (2000).

359. Ehlert, K., Holtje, J.V. & Templin, M.F. Cloning and expression of a murein hydrolase lipoprotein from Escherichia coli. *Mol Microbiol* **16**, 761-768 (1995).

360. Eisen, J.A. et al. The complete genome sequence of Chlorobium tepidum TLS, a photosynthetic, anaerobic, green-sulfur bacterium. *Proc Natl Acad Sci U S A* **99**, 9509-9514 (2002).

361. Hibert, M.F., Gittos, M.W., Middlemiss, D.N., Mir, A.K. & Fozard, J.R. Graphics computer-aided receptor mapping as a predictive tool for drug design: development of potent, selective, and stereospecific ligands for the 5-HT1A receptor. *J Med Chem* **31**, 1087-1093 (1988).

362. Loew, G.H., Villar, H.O., Jung, W. & Davies, M.F. Computer-aided drug design for the benzodiazepine receptor site. *NIDA Res Monogr* **112**, 43-61 (1991).

363. Ooms, F. Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry. *Curr Med Chem* **7**, 141-158 (2000).

364. Reddy, M.R. & Erion, M.D. Computer-aided drug design strategies used in the discovery of fructose 1, 6-bisphosphatase inhibitors. *Curr Pharm Des* **11**, 283-294 (2005).

365. Harrison, R.W. & Weber, I.T. Molecular dynamics simulations of HIV-1 protease with peptide substrate. *Protein Eng* **7**, 1353-1363 (1994).

366. Jenwitheesuk, E. & Samudrala, R. Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations. *BMC Struct Biol* **3**, 2 (2003).

367. Trylska, J., Bala, P., Geller, M. & Grochowski, P. Molecular dynamics simulations of the first steps of the reaction catalyzed by HIV-1 protease. *Biophys J* **83**, 794-807 (2002).

368. Jenwitheesuk, E. & Samudrala, R. Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antivir Ther* **10**, 157-166 (2005).

369. Tollenaere, J.P. & Moret, E.E. Hyper-glossary of Terminology, Guidebook on Molecular Modelling in Drug Design. (Academic Press, N.C.; 1996).

370. Veropoulos, K., Campbell, C. & Cristianini, N. in Proceedings of the International Joint Conference on Artificial Intelligence (UCAI99). (ed. T. Dean) 55-60 (Morgan Kaufmann, Sweden; 1999).

371. Maw, H.H. & Hall, L.H. E-state modeling of HIV-1 protease inhibitor binding independent of 3D information. *J Chem Inf Comput Sci* **42**, 290-298 (2002).

372. Wlodawer, A. & Erickson, J.W. Structure-based inhibitors of HIV-1 protease. *Annu Rev Biochem* **62**, 543-585 (1993).

373. Bianucci, A.M. Application of Cascade Correlation Networks for Structures to Chemistry. *Applied Intelligence* **12**, 117-146 (2000).

374. Stoffler, D., Sanner, M.F., Morris, G.M., Olson, A.J. & Goodsell, D.S. Evolutionary analysis of HIV-1 protease inhibitors: Methods for design of inhibitors that evade resistance. *Proteins* **48**, 63-74 (2002).

375. Hall, L.H. & Kier, L.B. in Reviews in Computational Chemistry II. (eds. K.B. Lipkowitz & D.B. Boyd) 367-422 (VCH Publishers, New York; 1991).

376. Yu, H., Yang, J., Wang, W. & Han, J. Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines. *Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB)*, 220-228 (2003).

# APPENDICES

## Appendix A：List of protein families currently covered by SVMProt, statistics of datasets and prediction results[***].

| Protein family | Training set | | Testing set | | | | Independent evaluation set | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | positive | negative | positive | | negative | | positive | | | negative | | |
| | | | TP | FN | TN | FP | TP | FN | Sensitivity | TN | FP | Specificity |
| EC1.1 Oxidoreductases acting on the CH-OH group of donors | 1164 | 2324 | 1795 | 10 | 7594 | 14 | 494 | 105 | 82.5% | 4760 | 192 | 96.1% |
| EC1.2 Oxidoreductases acting on the aldehyde or oxo group of donors | 665 | 1960 | 705 | 14 | 8051 | 25 | 259 | 69 | 79.0% | 4908 | 77 | 98.5% |
| EC1.3 Oxidoreductases acting on the CH-CH group of donors | 491 | 1917 | 131 | 3 | 8090 | 17 | 73 | 37 | 66.4% | 4941 | 57 | 98.9% |
| EC1.4 Oxidoreductases acting on the CH-NH2 group of donors | 307 | 1869 | 92 | 2 | 8179 | 8 | 50 | 26 | 65.8% | 4990 | 26 | 99.5% |
| EC1.5 Oxidoreductases acting on the CH-NH group of donors | 276 | 1755 | 56 | 3 | 8278 | 5 | 41 | 29 | 58.6% | 4985 | 21 | 99.6% |
| EC1.6 Oxidoreductases acting on NADH or NADPH | 1333 | 2132 | 2189 | 21 | 7857 | 19 | 1118 | 65 | 94.5% | 4901 | 88 | 98.2% |
| EC1.7 Oxidoreductases acting on other nitrogenous compounds as donors | 170 | 1356 | 86 | 0 | 8703 | 2 | 29 | 15 | 65.9% | 5005 | 13 | 99.7% |
| EC1.8 Oxidoreductases acting on a sulfur group of donors | 299 | 1531 | 114 | 2 | 8500 | 13 | 40 | 28 | 58.8% | 4989 | 20 | 99.6% |
| EC1.9 Oxidoreductases acting on a heme group of donors | 561 | 807 | 9493 | 22 | 9246 | 24 | 4805 | 36 | 99.3% | 4978 | 48 | 99.0% |
| EC1.10 Oxidoreductases acting on diphenols and related substances as donors | 219 | 1348 | 88 | 0 | 8728 | 4 | 65 | 20 | 76.5% | 4996 | 30 | 99.4% |
| EC1.11 Oxidoreductases acting on a peroxide as acceptor | 344 | 1416 | 343 | 2 | 8664 | 5 | 146 | 22 | 86.9% | 5009 | 22 | 99.6% |

[***] Predicted results are given in TP (true positive), FN (false negative), TN (true negative), FP (false positive), Sensitivity (TP/(TP+FN)), and Specificity(TN/(TN+FP)). Number of positive or negative samples in testing and independent evaluation sets is TP+FN or TN+FP respectively

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EC1.13 Oxidoreductases acting on single donors with incorporation of molecular oxygen (oxygenases) | 152 | 1232 | 90 | 7 | 8832 | 4 | 29 | 23 | 55.8% | 5009 | 13 | 99.7% |
| EC1.14 Oxidoreductases acting on paired donors with incorporation reduction of molecular oxygen | 566 | 1896 | 786 | 0 | 8120 | 8 | 93 | 38 | 71.0% | 4941 | 57 | 98.9% |
| EC1.15 Oxidoreductases acting on superoxide as acceptor | 259 | 881 | 416 | 2 | 9214 | 7 | 222 | 18 | 92.5% | 5019 | 20 | 99.6% |
| EC1.17 Oxidoreductases acting on CH2 groups | 100 | 1308 | 109 | 4 | 8779 | 8 | 43 | 12 | 78.2% | 5026 | 9 | 99.8% |
| EC1.18 Oxidoreductases acting on iron-sulfur proteins as donors | 244 | 1229 | 232 | 1 | 8842 | 8 | 78 | 7 | 91.8% | 5005 | 22 | 99.6% |
| EC2.1 Transferases transferring one-carbon groups | 1509 | 2991 | 800 | 0 | 6903 | 2 | 190 | 89 | 68.1% | 4194 | 740 | 85.0% |
| EC2.2 Transferases transferring aldehyde or ketone residues | 35 | 1197 | 30 | 2 | 1121 | 0 | 26 | 5 | 83.9% | 1005 | 3 | 99.7% |
| EC2.3 Acyltransferases | 302 | 1001 | 246 | 0 | 1284 | 4 | 196 | 44 | 81.7% | 966 | 27 | 97.3% |
| EC2.4 Glycosyltransferases | 945 | 1896 | 1211 | 25 | 7940 | 41 | 203 | 85 | 70.5% | 4640 | 286 | 94.2% |
| EC2.5 Transferases transferring alkyl or aryl groups, other than methyl groups | 764 | 2174 | 519 | 24 | 7832 | 33 | 137 | 58 | 70.3% | 4915 | 93 | 98.1% |
| EC2.6 Transferases transferring nitrogenous groups | 343 | 1684 | 301 | 5 | 8395 | 6 | 75 | 32 | 70.1% | 4982 | 49 | 99.0% |
| EC2.7 Transferases transferring phosphorus-containing groups | 3892 | 5324 | 3761 | 4 | 6140 | 6 | 2463 | 553 | 81.7% | 5082 | 625 | 89.0% |
| EC2.8 Transferases transferring sulfur-containing groups | 203 | 1549 | 43 | 0 | 8531 | 7 | 20 | 10 | 66.7% | 5021 | 11 | 99.8% |
| EC3.1 Hydrolases acting on ester bonds | 2482 | 3859 | 1504 | 53 | 5677 | 100 | 379 | 154 | 71.1% | 4355 | 452 | 90.6% |
| EC3.2 Glycosylases | 337 | 867 | 379 | 2 | 1397 | 13 | 268 | 49 | 84.5% | 939 | 51 | 94.8% |
| EC3.3 Hydrolases acting on ether bonds | 97 | 1999 | 44 | 22 | 8053 | 49 | 32 | 22 | 59.3% | 5007 | 32 | 99.4% |
| EC3.4 Hydrolases acting on peptide bonds (Peptidases) | 2011 | 3402 | 1522 | 35 | 6207 | 29 | 264 | 90 | 74.6% | 4528 | 279 | 94.2% |
| EC3.5 Hydrolases acting on carbon-nitrogen bonds, other than peptide bonds | 1020 | 2498 | 440 | 2 | 7447 | 3 | 130 | 85 | 60.5% | 4849 | 110 | 97.8% |
| EC3.6 Hydrolases acting on acid anhydrides | 2195 | 2504 | 1449 | 1 | 7435 | 4 | 687 | 63 | 91.6% | 4742 | 220 | 95.6% |
| EC4.1 Carbon-carbon lyases | 546 | 1145 | 776 | 5 | 1113 | 17 | 547 | 62 | 89.8% | 881 | 105 | 89.4% |
| EC4.2 Carbon-oxygen lyases | 505 | 1231 | 382 | 1 | 1047 | 2 | 324 | 79 | 80.4% | 915 | 77 | 92.2% |
| EC4.3 Carbon-nitrogen lyases | 218 | 1068 | 194 | 2 | 9009 | 8 | 29 | 10 | 74.4% | 4994 | 37 | 99.3% |
| EC4.4 Carbon-sulfur lyases | 182 | 1999 | 53 | 23 | 8072 | 14 | 35 | 23 | 60.3% | 5024 | 7 | 99.9% |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EC4.6 Phosphorus-oxygen lyases | 200 | 1789 | 63 | 14 | 8250 | 7 | 55 | 27 | 67.1% | 4112 | 899 | 82.1% |
| EC5.1 Racemases and Epimerases | 379 | 1796 | 91 | 4 | 8249 | 19 | 35 | 31 | 53.0% | 4990 | 30 | 99.4% |
| EC5.2 Cis-trans-Isomerases | 35 | 1404 | 113 | 2 | 8671 | 11 | 72 | 36 | 66.7% | 5008 | 24 | 99.5% |
| EC5.3 Intramolecular oxidoreductases | 461 | 1122 | 92 | 3 | 1062 | 0 | 135 | 43 | 75.8% | 4910 | 99 | 98.0% |
| EC5.4 Intramolecular transferases | 329 | 1714 | 143 | 4 | 8337 | 16 | 42 | 35 | 54.5% | 4991 | 31 | 99.4% |
| EC5.5 Intramolecular lyases | 47 | 909 | 24 | 0 | 9196 | 0 | 75 | 32 | 70.1% | 4982 | 49 | 99.0% |
| EC5.99 Other Isomerases | 163 | 1038 | 393 | 8 | 9036 | 9 | 153 | 13 | 92.2% | 5007 | 22 | 99.6% |
| EC6.1 Ligases forming carbon-oxygen bonds | 281 | 1115 | 381 | 1 | 1185 | 13 | 286 | 29 | 90.8% | 980 | 27 | 97.3% |
| EC6.2 Ligases forming carbon-sulfur bonds | 149 | 1233 | 154 | 4 | 8858 | 4 | 51 | 13 | 79.7% | 5203 | 13 | 99.8% |
| EC6.3 Ligases forming carbon-nitrogen bonds | 381 | 1133 | 358 | 2 | 1148 | 3 | 294 | 57 | 83.8% | 946 | 45 | 95.5% |
| EC6.4 Ligases forming carbon-carbon bonds | 99 | 1543 | 45 | 0 | 8548 | 8 | 28 | 16 | 63.6% | 5033 | 4 | 99.9% |
| EC6.5 Ligases forming phosphoric ester bonds | 94 | 1679 | 36 | 2 | 8408 | 3 | 22 | 9 | 71.0% | 5027 | 6 | 99.9% |
| TC1.A alpha-type channels | 381 | 1786 | 272 | 8 | 10425 | 7 | 164 | 25 | 86.8% | 6037 | 44 | 99.3% |
| TC1.B beta-barrel porins | 221 | 2008 | 58 | 0 | 12452 | 2 | 65 | 27 | 70.7% | 7178 | 29 | 99.6% |
| TC1.C Pore-forming toxins (proteins and peptides) | 357 | 2007 | 33 | 14 | 12371 | 0 | 100 | 27 | 78.7% | 6452 | 15 | 99.8% |
| TC1.E Holins | 100 | 513 | 55 | 5 | 11837 | 14 | 55 | 15 | 78.6% | 6151 | 4 | 99.9% |
| TC2.A porters (symporters, uniporters, antiporters) | 629 | 1175 | 781 | 4 | 10938 | 13 | 370 | 54 | 87.3% | 5945 | 90 | 98.5% |
| TC2.C Ion-gradient-driven energizers | 166 | 1014 | 86 | 2 | 11325 | 10 | 91 | 28 | 76.5% | 6140 | 13 | 99.8% |
| TC3.A P-P-bond-hydrolysis-driven transporters | 1220 | 2549 | 1301 | 20 | 9568 | 15 | 897 | 243 | 78.7% | 5895 | 143 | 97.6% |
| TC3.D Oxidoreduction-driven transporters | 435 | 1529 | 981 | 1 | 12980 | 5 | 617 | 60 | 91.1% | 7197 | 36 | 99.5% |
| TC3.E Light absorption-driven transporters | 139 | 954 | 696 | 5 | 13648 | 2 | 395 | 16 | 96.1% | 7267 | 11 | 99.8% |
| TC4.A Phosphotransfer-driven group translocators | 197 | 887 | 212 | 8 | 11429 | 5 | 153 | 32 | 82.7% | 6120 | 21 | 99.7% |
| TC8.A Auxiliary transport proteins | 223 | 1388 | 169 | 5 | 10925 | 13 | 124 | 43 | 74.3% | 6120 | 15 | 99.8% |
| TC9.A Recognized transporters of unknown biochemical mechanism | 203 | 1034 | 188 | 1 | 11247 | 29 | 130 | 35 | 78.8% | 6085 | 43 | 99.3% |
| TC9.B Putative uncharacterized transport proteins | 869 | 2079 | 581 | 5 | 10153 | 5 | 469 | 116 | 80.2% | 6002 | 98 | 98.4% |
| G protein coupled receptors | 927 | 1320 | 4993 | 5 | 13212 | 4 | 2421 | 111 | 95.6% | 7104 | 140 | 98.1% |
| 7 transmembrane receptor (rhodopsin family & chemoreceptor ) | 729 | 1061 | 4604 | 7 | 13535 | 3 | 2223 | 71 | 96.9% | 7214 | 61 | 99.2% |
| 7 transmembrane receptor (secretin family) | 218 | 2007 | 71 | 0 | 12580 | 1 | 117 | 12 | 90.7% | 6900 | 370 | 94.9% |
| 7 transmembrane receptor (metabotropic glutamate family) | 116 | 2001 | 40 | 0 | 12613 | 0 | 62 | 7 | 89.9% | 6975 | 308 | 95.8% |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 transmembrane receptor (odorant receptor) | 130 | 1999 | 11 | 0 | 12631 | 0 | 38 | 1 | 97.4% | 7113 | 178 | 97.6% |
| DNA-binding proteins | 3260 | 4251 | 4146 | 115 | 4914 | 73 | 2469 | 1114 | 68.9% | 4065 | 464 | 89.8% |
| RNA-binding proteins | 2161 | 2965 | 1844 | 6 | 6802 | 14 | 437 | 10 | 97.8% | 4685 | 196 | 96.0% |
| mRNA-binding proteins | 277 | 2106 | 129 | 0 | 10164 | 0 | 130 | 34 | 79.3% | 5833 | 213 | 96.5% |
| rRNA-binding proteins | 708 | 972 | 1243 | 2 | 9031 | 13 | 95 | 6 | 94.1% | 4931 | 66 | 98.7% |
| tRNA-binding proteins | 94 | 792 | 114 | 0 | 9295 | 2 | 48 | 3 | 94.1% | 5028 | 5 | 99.9% |
| Structural proteins (Matrix protein,Core protein,Viral occlusion body,Keratin) | 858 | 1353 | 4977 | 4 | 8512 | 12 | 2615 | 41 | 98.5% | 4884 | 40 | 99.2% |
| Transmembrane | 2105 | 2563 | 11135 | 1722 | 8237 | 1368 | 3054 | 335 | 90.1% | 5254 | 809 | 86.7% |
| Outer membrane | 602 | 1539 | 547 | 0 | 8384 | 1 | 318 | 25 | 92.7% | 4276 | 672 | 86.4% |
| Cell adhesion | 513 | 1678 | 322 | 1 | 8208 | 15 | 232 | 38 | 85.9% | 4897 | 44 | 99.1% |
| Coat proteins | 346 | 1474 | 297 | 8 | 8344 | 26 | 167 | 30 | 84.8% | 4885 | 29 | 99.4% |
| Envelope proteins | 177 | 1999 | 112 | 11 | 7904 | 28 | 135 | 15 | 90.0% | 4927 | 25 | 99.5% |
| Nuclear receptors | 334 | 538 | 601 | 7 | 1755 | 6 | 221 | 26 | 89.5% | 962 | 24 | 97.6% |
| Tyrosine kinase receptors | 14 | 1197 | 3 | 0 | 1121 | 0 | 5 | 2 | 71.4% | 1006 | 2 | 99.8% |
| Growth factor | 329 | 1320 | 205 | 5 | 8695 | 4 | 142 | 21 | 87.1% | 4970 | 28 | 99.4% |
| Antigen | 836 | 1867 | 1200 | 2 | 7786 | 8 | 720 | 29 | 96.1% | 4747 | 74 | 98.5% |
| Chlorophyll | 189 | 603 | 945 | 3 | 14630 | 10 | 515 | 14 | 97.4% | 6965 | 11 | 99.8% |
| Chlorophyll biosynthesis | 309 | 1742 | 109 | 0 | 13424 | 0 | 153 | 24 | 86.4% | 6158 | 777 | 88.8% |
| Herbicide resistance | 227 | 1999 | 205 | 7 | 13196 | 2 | 199 | 10 | 95.2% | 6948 | 10 | 99.9% |
| Photoreceptor | 354 | 1537 | 893 | 3 | 13611 | 11 | 548 | 42 | 92.9% | 6896 | 26 | 99.6% |
| Photorespiration | 368 | 1672 | 8197 | 4 | 13504 | 76 | 4257 | 13 | 99.7% | 6955 | 24 | 99.7% |
| Photosynthesis | 1054 | 1914 | 544 | 0 | 12950 | 47 | 613 | 44 | 93.3% | 6664 | 132 | 98.1% |
| Photosystem I | 264 | 1491 | 392 | 70 | 13726 | 1 | 326 | 8 | 97.6% | 5900 | 1061 | 84.8% |
| Photosystem II | 506 | 986 | 2018 | 4 | 14120 | 46 | 1192 | 31 | 97.5% | 6890 | 36 | 99.5% |
| Plant defense | 559 | 1830 | 456 | 5 | 13302 | 14 | 289 | 37 | 88.7% | 6857 | 60 | 99.1% |

**Appendix B: Distribution of RNA-binding proteins in different kingdoms and in top 10 host species of each kingdom. Not all protein sequences studied in this work are included because the host species information of some protein sequences is not yet available in the protein sequence database.**

| **Kingdom** | **Eucaryote** | **Eubacteria** | **Archaea** |
|---|---|---|---|
| **Number of proteins in kingdom** | 986 | 1854 | 294 |
| **List of top 10 species and number of proteins in each species** | Homo sapiens (168) | Escherichia coli (75) | Methanococcus jannaschii (22) |
| | Mus musculus (78) | Bacillus subtilis (64) | Methanobacterium thermoautotrophicum (21) |
| | Candida albicans (77) | Haemophilus influenzae (60) | Archaeoglobus fulgidus (20) |
| | Schizosaccharomyces pombe (52) | Buchnera aphidicola (subsp. Acyrthosiphon pisum) (50) | Halobacterium sp (19) |
| | Drosophila melanogaster (45) | Helicobacter pylori (49) | Pyrococcus horikoshii (19) |
| | Arabidopsis thaliana (42) | Buchnera aphidicola (subsp. Schizaphis graminum) (47) | Pyrococcus abyssi (18) |
| | Xenopus laevis (30) | Aquifex aeolicus (45) | Sulfolobus solfataricus (18) |
| | Rattus norvegicus (28) | Mycobacterium tuberculosis (45) | Aeropyrum pernix (18) |
| | Caenorhabditis elegans (26) | Rickettsia prowazekii (44) | Methanopyrus kandleri (15) |
| | Porphyra purpurea (19) | Mycoplasma pneumoniae (43) | Thermoplasma volcanium (14) |

**__END__**